ISSN 2073-7394

Національний університет
“Полтавська політехніка імені Юрія Кондратюка”

National University
“Yuri Kondratyuk Poltava Polytechnic”

Системи управління, навігації та зв'язку

Control, navigation and communication systems

Випуск 2 (84)

Issue 2 (84)

Щоквартальне видання

Засноване у 2007 році

У журналі відображені результати наукових досліджень з розробки та удосконалення систем управління, навігації та зв’язку у різних проблемних галузях.

Засновник і видавець:
Національний університет
“Полтавська політехніка імені Юрія Кондратюка”

Телефон:
+38 (050) 302-20-71

E-mail редколегії:
kuchuk56@ukr.net

Інформаційний сайт:
http://journals.nupp.edu.ua/sunz

Quarterly

Founded in 2007

Journal represent the research results on the development and improvement of control, navigation and communication systems in various areas

Founder and publisher:
National University
"Yuri Kondratyuk Poltava Polytechnic"

Phone:
+38 (050) 302-20-71

E-mail of the editorial board:
kuchuk56@ukr.net

Information site:
http://journals.nupp.edu.ua/sunz

За достовірність викладених фактів, цитат та інших відомостей відповідальність несе автор

*Журнал індексується міжнародними наукометричними базами: Index Copernicus (ICV = **85.62**), General Impact Factor, Google Scholar, Academic Resource Index, Scientific Indexed Service*

*Затверджений до друку Вченою Радою Національного університету “*Полтавська політехніка імені Юрія Кондратюка*” (протокол від 4 травня 2026 року № 6).*

Ідентифікатор медіа R30-04135 згідно з рішенням Національної ради України з питань телебачення і радіомовлення від 25.04.2024 № 1416

Включений до “Переліку наукових фахових видань України, в яких можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора наук, кандидата наук та ступеня доктора філософії ” до категорії Б – наказами МОН України від 17.03.2020 № 409 та від 09.02.2021 № 157

Полтава • 2026

З М І С Т

НАВІГАЦІЯ ТА ГЕОІНФОРМАЦІЙНІ СИСТЕМИ

АВТОМОБІЛЬНИЙ, РІЧКОВИЙ, МОРСЬКИЙ ТА АВІАЦІЙНИЙ ТРАНСПОРТ

УПРАВЛІННЯ В СКЛАДНИХ СИСТЕМАХ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

ЦИВІЛЬНА БЕЗПЕКА

ЗВ'ЯЗОК, ТЕЛЕКОМУНІКАЦІЇ ТА РАДІОТЕХНІКА

Authors affiliation

Azerbaijan Technical University, Baku, Azerbaijan
National Defense University, Baku, Azerbaijan
Zhongke Shuguang, Tianjin, China
CNOOC Financial Shared Service Center PRD Branch, Shenzhen, China
University of the National Education Commission, Krakow, Poland
Geological Bureau BEG SA, Switzerland
State University of Trade and Economics, Kyiv, Ukraine
Ya. Usikov Institute for Radiophysics and Electronics of the NAS of Ukraine, Kharkiv, Ukraine
Kyiv National University of Construction and Architecture, Ukraine
Kremenchuk Mykhailo Ostrohradskyi National University, Kremenchuk, Ukraine
Interregional Academy of Personnel Management, Kyiv, Ukraine
Radio Scientific and Production Enterprise, Kropyvnytskyi, Ukraine
Research Institute of Military Intelligence, Ukraine
Hetman Petro Sahaidachnyi National Army Academy, Lviv, Ukraine
National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
National Technical University "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
Lviv Polytechnic National University, Ukraine
National Defence University of Ukraine, Kyiv
Odesa Polytechnic National University, Ukraine
National University "Yuri Kondratyuk Poltava Polytechnic", Poltava, Ukraine
Uzhhorod National University, Uzhhorod, Ukraine
Kharkiv National Automobile and Highway University, Kharkiv, Ukraine
Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine
Kharkiv National University of Internal Affairs, Ukraine
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine
Ukrainian State University of Railway Transport, Ukraine
Cherkasy State Technological University, Ukraine
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine

Навігація та геоінформаційні системи

УДК 528.88 + 515.127.1 **doi: 10.26906/SUNZ.2026.2.005**

Р. Е. Пащенко

Інститут радіофізики та електронники імені О. Я. Усикова НАН України, Харків, Україна

ЛОКАЛІЗАЦІЯ АНОМАЛІЙ НА КОСМІЧНИХ ЗНІМКАХ СІЛЬСЬКОГОСПОДАРСЬКИХ ЗЕМЕЛЬ ЗА ДОПОМОГОЮ ПОБУДОВИ “ПІРАМІДИ” ФРАКТАЛЬНИХ РОЗМІРНОСТЕЙ

Анотація. З використанням інформації, що отримується з космічних засобів дистанційного зондування Землі, можна оцінювати негативний стан сільськогосподарських земель які проявляються у вигляді різноманітних аномалій. **Предметом** дослідження є оцінка можливості локалізації аномалій на космічних знімках з використанням “піраміди” фрактальних розмірностей. **Об’єктом** дослідження є космічні знімки сільськогосподарських земель, які отримані з супутника Sentinel-2 з аномаліями і без аномалій. **Метою** є оцінка можливості локалізації аномалій на космічних знімках сільськогосподарських земель з використанням фрактального аналізу. **Отримані наступні результати**. Розглянуто можливість локалізації аномалій на космічних знімках сільськогосподарських земель з використанням “пірамідального” фрактального аналізу. Створення “піраміди” космічних знімків здійснюється за рахунок розподілу вихідного космічного знімку на знімки менших розмірів, які у подальшому індексуються. Під час побудови “піраміди” фрактальних розмірностей для кожного знімка “піраміди” космічних знімків розраховується середня (мінімальна) фрактальна розмірність з використанням “ковзаючого вікна”, тобто “вікна”, що рухається по космічному знімку з кроком переміщення, який дорівнює одиниці. **Висновки**. Застосування “піраміди” середніх (мінімальних) фрактальних розмірностей дозволяє локалізувати аномалії на космічних знімках, якщо величини середніх (мінімальних) фрактальних розмірностей космічних знімків на кожному рівні “піраміди” менше D = 2,9, то на них є аномалії. За допомогою фрактального аналізу космічних знімків без аномалій показано, що, якщо середні (мінімальні) фрактальні розмірності на всіх рівнях “піраміди” більше фрактальної розмірності D = 2,9, то на космічному знімку аномалій немає. На деякі особливості на космічному знімку вказують менші середні (мінімальні) фрактальні розмірності на нижчих рівнях “піраміди” порівняно з вищими рівнями “піраміди”.

Ключові слова: моніторинг стану сільськогосподарських земель; космічні знімки; аномалія; фрактальна розмірність.

Вступ

Агроекологічний моніторинг дозволяє проводити дослідження впливу природних, техногенних і антропогенних чинників на стан сільськогосподарських земель [1]. Дія всіх цих чинників може привести до негативних наслідків стану землі, її деградації та погіршення врожайності. Оцінити негативний стан і структуру сільськогосподарських земель також можна з використанням інформації, що отримується з космічних засобів дистанційного зондування Землі (ДЗЗ) [2]. На космічних знімках ділянки землі, що мають ознаки деградації, проявляються у вигляді різноманітних аномалій.

Сучасні супутники ДЗЗ можуть отримувати інформацію про стан земної поверхні у різних діапазонах хвиль, тобто отримувати багатоспектральні космічні знімки. Просторова роздільна здатність таких космічних знімків може дорівнювати 10-30 м, а періодичність їх отримання – один знімок у 5-8 діб на задану територію. Найбільш розповсюдженими космічними знімками, які використовуються для розв’язання різноманітних задач ДЗЗ і, які є у вільному доступі у мережі Інтернет, є космічні знімки супутників Sentinel-2 [3] і Landsat-8 [4].

В останні десятиріччя все частіше для оцінки стану і структуру земельних ділянок та посівних площ, а також для аналізу аномальних ділянок на космічних знімках застосовуються методи фрактального аналізу зображень [5, 6]. Розрахунок фрактальних розмірностей дозволяє за їх величиною розрізняти складність структури земних поверхонь на космічних знімках. При цьому можна використовувати космічні знімки супутників ДЗЗ в одному діапазоні хвиль. У роботах [6, 7] розглянуто можливість аналізу стану посівів кукурудзи та інших сільськогосподарських культур на різних фазах вегетації з використанням фрактального аналізу космічних знімків. У цих роботах досліджено як змінюються величини середніх фрактальних розмірностей космічних знімків полів засіяних кукурудзою, соняшником, гречкою, пшеницею і ячменем, але у цих роботах не розглядається можливість моніторингу змін стану сільськогосподарських земель.

У роботі [8] показано можливість моніторингу змін стану сільськогосподарських земель під впливом різних чинників за даними фрактального аналізу і наведено як візуалізація поля фрактальних розмірностей дозволяє наочно показати зміну стану сільськогосподарських земель та межі аномалій на знімках. Але у цій роботі не розглядається можливість локалізації і визначення розмірів аномальних ділянок на космічних знімках.

Таким чином, перспективним напрямком досліджень є оцінка можливості застосування фрактального аналізу для локалізації аномальних ділянок

на космічних знімках земель сільськогосподарського призначення.

Мета статті: розглянути можливість локалізації аномалій на космічних знімках сільськогосподарських земель з використанням фрактального аналізу.

Побудова “піраміди” космічних знімків сільськогосподарських земель

Розглянемо можливість локалізації аномальних ділянок сільськогосподарських земель на космічних знімках супутника Sentinel-2. При цьому вихідні космічні знімки були отримані з мережі Інтернет і на них були візуально визначені об’єкти дослідження – поля з аномаліями і без аномалій. Далі були визначені їх розміри і після цього здійснено вирізання зображень розміром 56х56 пікселів. На рис. 1 наведено вихідний космічний знімок каналу b8 супутника Sentinel-2 поля з пшеницею з аномаліями.

На рис. 1 видно, що на вихідному знімку поля з пшеницею візуально спостерігаються дві аномалії, перша знаходиться зверху праворуч на знімку, а друга – ліворуч у середині знімка.

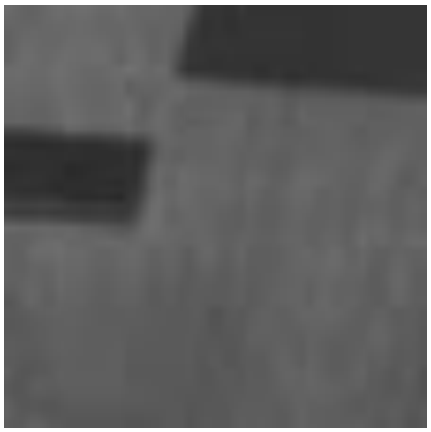


Рис. 1. Космічний знімок поля з пшеницею з аномаліями розмірами 56х56 пікселів

Далі була створена “піраміда” космічних знімків, коротко розглянемо порядок її побудови. На початку вихідний космічний знімок розміром 56х56 пікселів розбивається на чотири квадранти. Таким чином будується перший рівень “піраміди” – чотири знімки розміром 28х28 пікселів. На другому рівні “піраміди” отримуємо шістнадцять знімків розміром 14х14 пікселів, а на третьому рівні – шістдесят чотири знімка розміром 7х7 пікселів. Також під час створення “піраміди” космічних знімків здійснюється індексація знімків. Вихідний знімок розміром 56х56 пікселів має індекс **1**, на першому рівні “піраміди” космічні знімки мають індекси **111**, **112**, **113**, **114**, а на другому рівні – індекси **11111**, **11112**, **11121**, **11122**, **11211**, **11212**, **11211**, **11222**, **11311**, **11312**, **11321**, **11322**, **11411**, **11412**, **11421**, **11422** і так далі. Такий порядок побудови “піраміди” космічних знімків еквівалентний застосуванню “стрибаючого вікна” під час аналізу космічного знімка. При цьому крок переміщення дорівнює розміру знімка на кожному рівні “піраміди”.

На рис. 2 наведено чотири космічні знімки першого рівня “піраміди” розміром 28х28 пікселів.

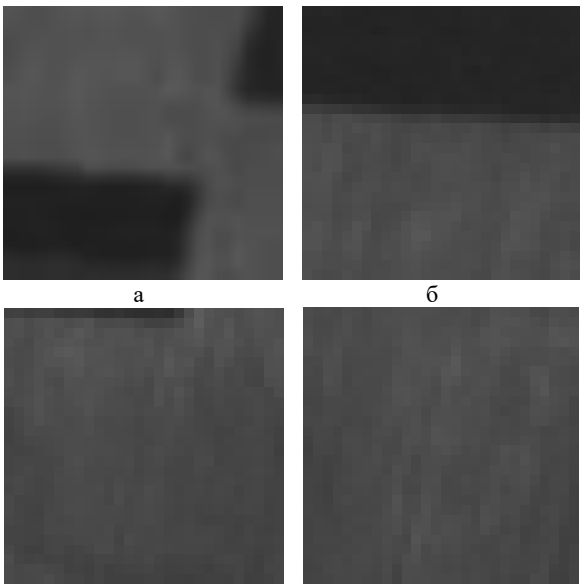


Рис. 2. Елементи вихідного космічного знімка (першого рівня “піраміди”) поля з пшеницею з аномаліями розмірами 28х28 пікселів: перший (а); другий (б); третій (в) і четвертий (г) квадранти

Як видно на рис. 2, аномалії є на трьох знімках першого рівня “піраміди” з чотирьох. На першому знімку дві аномалії, на другому знімку присутня велика аномальна ділянка, яка займає майже половину знімка. На третьому знімку невеличка аномалія є зверху, а на четвертому – аномалій немає. На рис. 3 наведено шістнадцять космічних знімків другого рівня “піраміди” розміром 14х14 пікселів.

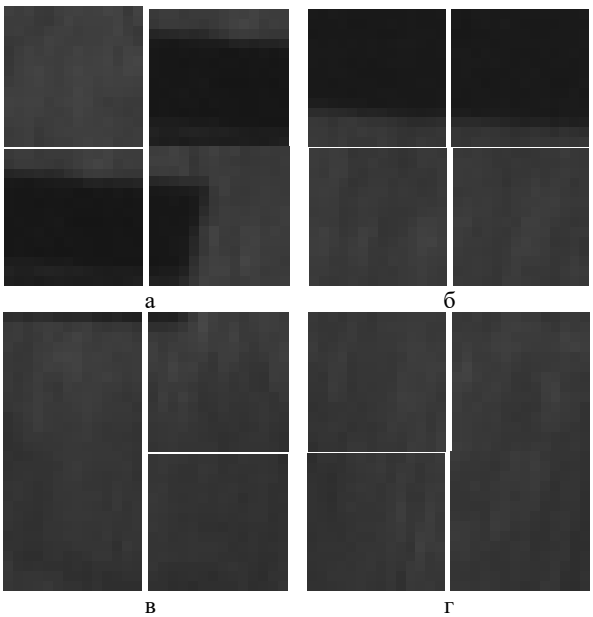


Рис. 3. Елементи вихідного космічного знімка (другого рівня “піраміди”) поля з пшеницею з аномаліями розмірами 14х14 пікселів: перший (а); другий (б); третій (в) і четвертий (г) квадранти космічних знімків першого рівня “піраміди”

Як видно на рис. 3, на трьох (другому, третьому, четвертому) космічних знімках другого рівня “піраміди”, які відповідають першому знімку пер-

шого рівня “піраміди” є аномалії, а на першому космічному знімку аномалій немає. На другому і третьому космічних знімках аномалії займають майже всі знімки другого рівня “піраміди”.

На рис. 3 також видно, що на перших двох космічних знімках другого рівня “піраміди”, що відповідають другому знімку першого рівня, є аномалії. На обох космічних знімках аномалії займають майже всі знімки другого рівня “піраміди”. Але на третьому і четвертому космічних знімках другого рівня “піраміди”, що відповідають другому знімку першого рівня, аномалії відсутн. Також на перших двох космічних знімках другого рівня “піраміди” (див. рис. 3), що відповідають третьому знімку першого рівня “піраміди” є невеличкі аномалії зверху, а третьому і четвертому космічних знімках другого рівня “піраміди” аномалій немає. На всіх космічних знімках другого рівня “піраміди” (див. рис. 3), які відповідають четвертому знімку першого рівня “піраміди” аномалії відсутні. На рис. 4 наведено частка (шістнадцять) космічних знімків третього рівня “піраміди” розміром 7х7 пікселів.

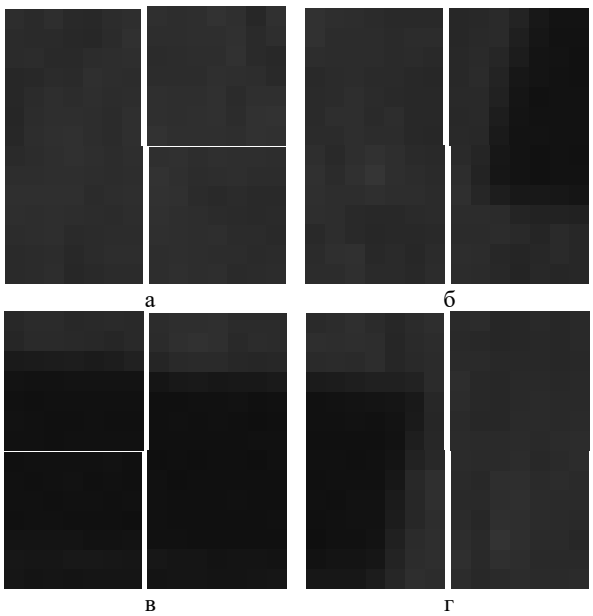


Рис. 4. Елементи вихідного космічного знімка (третього рівня “піраміди”) поля з пшеницею з аномаліями розмірами 7х7 пікселів: перший (а); другий (б); третій (в) і четвертий (г) квадранти чотирьох верхніх лівих космічних знімків другого рівня “піраміди”

Як видно на рис. 4, на всіх космічних знімках третього рівня “піраміди”, які відповідають першому знімку другого рівня аномалій немає. На рис. 4 також видно, що на двох (другому і четвертому) космічних знімках третього рівня “піраміди”, що відповідають другому знімку другого рівня “піраміди” є аномалії. Найбільша аномалія спостерігається на другому космічному знімку. Також необхідно зазначити, що на першому і третьому космічному знімку третього рівня “піраміди” аномалій немає. На всіх космічних знімка третього рівня “піраміди” (див. рис. 4), що відповідають третьому знімку другого рівня “піраміди” є аномалії. Але на першому і другому знімках аномалія займає майже весь знімок, а на третьому і четвертому космічних знімках аномалія займає всі знімки і вони є однорідними, тому можна вважати, що під час їх подальшого аналізу на них аномалій немає. На двох (першому і третьому) космічних знімка третього рівня “піраміди”, що відповідають четвертому знімку другого рівня “піраміди” (див. рис. 4) присутні аномалії. Найбільша аномалія знаходиться на третьому космічному знімку. На другому і четвертому космічних знімків третього рівня “піраміди” аномалій немає.

Проведемо обробку космічних знімків супутника Sentinel-2 поля з пшеницею з аномаліями і без аномалій з використанням фрактального аналізу. Побудуємо “піраміди” фрактальних розмірностей космічних знімків цих полів, а також розглянемо можливість локалізації на них аномальних ділянок.

Побудова “піраміди” фрактальних розмірностей космічних знімків сільськогосподарських земель

Фрактальні розмірності є дробовими величинами, які характеризують складність структури космічного знімка, і їх можна розраховувати за допомогою методу покриття [9], або методу призми [10], які найчастіше застосовують у практичному застосуванні. Під час застосування методу покриття [9] для розрахунку фрактальної розмірності тривимірне подання космічного знімка заповнюється (покривається) кубами певних розмірів. Довжини сторін кубів ε змінюються декілька разів. Для кожної з довжин сторін визначається сумарна кількість кубів $N(\varepsilon)$, що використовувалися для покриття знімка. За отриманими даними будується графік логарифмічної залежності $\log N(\varepsilon)$ від $\log 1/\varepsilon$, який апроксимується за методом найменших квадратів (будується пряма за набором точок). Фрактальна розмірність D визначається, як тангенс кута нахилу отриманої прямої.

Порядок розрахунку фрактальної розмірності з використанням методу призми [10] є подібним до розрахунку за методом покриття. Але замість сумарної кількості кубів $N(\varepsilon)$ використовується площина $P(\varepsilon)$ верхньої грані призми, яка будується за даними яскравості зображення. Під час розрахунку також обираються декілька довжин ε основи призми і будується графік логарифмічної залежності $\log P(\varepsilon)$ від $\log 1/\varepsilon$ (апроксимована пряма). Далі, як і у методі покриття, фрактальна розмірність D визначається, як тангенс кута нахилу отриманої прямої.

Під час побудови “піраміди” фрактальних розмірностей для кожного знімка “піраміди” космічних знімків розраховується середня (мінімальна) фрактальна розмірність з використанням “ковзаючого вікна” розміром 4х4 пікселя. Тобто для кожного “вікна”, що рухається по космічному знімку з кроком переміщення, який дорівнює одиниці, визначається фрактальна розмірність, а далі отримані фрактальні розмірності усереднюють (знаходиться мінімальна фрактальна розмірність). Після цього середні (мінімальні) фрактальні розмірності індексуються і зберігаються у пам’яті комп’ютера. При цьому індексація “піраміди” середніх (мінімальних) фрактальних розмірностей співпадає з індексацією “піраміди” космічних знімків.

Розраховані з використанням методу покриття середні фрактальні розмірності (ФР) для “ковзаючого вікна” розміром 4х4 пікселя і проіндексовані згідно “піраміди” космічних знімків наведено нижче:

- вихідного космічного знімка поля з пшеницею з аномаліями D1 = 2,944;

- першого рівня “піраміди” – D111 = 2,905; D112 = 2,929; D121 = 2,953; D122 = 2,962;

- другого рівня “піраміди” – D11111 = 2,956; D11112 = 2,880; D11121 = 2,897; D11122 = 2,902; D11211 = 2,875; D11212 = 2,910; D11221 = 2,950; D11222 = 2,955; D12111 = 2,936; D12112 = 2,922; D12121 = 2,957; D21122 = 2,969; D21211 = 2,959; D12212 = 2,940; D12221 = 2,959; D12222 = 2,944;

- частки 3 рівня “піраміди” – D1111111 = 2,937; D1111112 = 2,930; D1111121 = 2,965; D1111122 = 2,940; D1111211 = 2,929; D1111212 = 2,889; D1111221 = 2,892; D1111222 = 2,770; D1112111 = 2,788; D1112112 = 2,835; D1112121 = 2,948; D1112122 = 2,956; D1112211 = 2,892; D1112212 = 2,926; D1112221 = 2,905; D1112222 = 2,942.

Для спрощення аналізу отриманих середніх ФР їх краще подати у вигляді таблиці. Для космічного знімка поля з пшеницею з аномаліями “піраміду” середніх ФР наведено у табл. 1.

Під час виявлення аномалій на космічних знімках сільськогосподарських земель будемо проводити порівняння величин середніх ФР на різних рівнях “піраміди” космічних знімків. Можна вважати, що аномалій на космічному знімку немає, якщо на нижчих рівнях “піраміди” середні ФР не відрізняються у першому знаку після коми від середніх ФР на вищому рівні. Аномальні ділянки на космічному знімку присутні, якщо середні ФР розрізняються у першому знаку після коми на різних рівнях “піраміди”.

*Таблиця 1 – “***Піраміда” середніх фрактальних розмірностей космічного знімка поля з пшеницею з аномаліями для “вікна” 4х4 пікселя**

D	1				
1	2,944				
1	1	2			
1	2,905	2,929			
2	2,953	2,962			

11	1	2	**12**	1	2
1	2,956	2,880	1	2,875	2,910
2	2,897	2,902	2	2,950	2,955
21	1	2	**22**	1	2
1	2,936	2,922	1	2,959	2,940
2	2,957	2,969	2	2,959	2,944

11	1	2
1	2,956	2,880
2	2,897	2,902

111	1	2	**112**	1	2
1	2,937	2,930	1	2,929	2,889
2	2,965	2,940	2	2,892	2,770
121	1	2	**122**	1	2
1	2,788	2,835	1	2,892	2,926
2	2,948	2,956	2	2,905	2,942

Особливістю фрактальної розмірності, яка розраховується за методом покриття, є те, що чим однорідніше зображення, тим більше її величина і для однорідних зображень наближається до 3,0.

Як видно у табл. 1, середня ФР вихідного знімка дорівнює D1 = 2,944, що більше більше 2,9 і зробити висновок про наявність на ньому аномалій неможливо. За величиною середньої ФР можна тільки сказати, що космічний знімок майже однорідний. Тобто виявити аномалії на космічному знімку, використовуючи тільки одне значення середньої ФР, не можливо. Аналіз даних у табл. 1 також показує, що на першому рівні “піраміди” середні ФР першого і другого космічних знімків (D111 = 2,905; D112 = 2,929) менше середньої ФР всього вихідного космічного знімка (D1 = 2,944), можна зробити висновок, що свідчить на них є аномалії. Середні ФР третього і четвертого знімків (D121 = 2,953; D122 = 2,962) більше середньої ФР всього вихідного знімка, що свідчить про відсутність на них аномалій. Але для третього знімка середня ФР не значно відрізняється від середньої ФР вихідного знімка. Необхідно зазначити, що меншими величинам середніх ФР відповідають більші аномальні ділянки. Так на першому знімку наявні дві аномалії (рис. 2, а) і для нього середня ФР на першому рівні “піраміди” мінімальна D111 = 2,905 (табл. 1).

Проведемо аналіз середніх ФР космічних знімків другого рівня “піраміди”. Як видно у табл. 1 три середні ФР космічних знімків (D11112 = 2,880; D11121 = 2,897; D11122 = 2,902), що відповідають першому знімку першого рівня “піраміди” менше середньої ФР D111 = 2,905, що свідчить про наявність на них аномалій. Найменші середні ФР відповідають другому і третьому космічним знімкам, на яких аномалії займають майже всі знімки рис. 3, а). Для першого космічного знімку другого рівня “піраміди” середня ФР дорівнює D11111 = 2,956, що більше і D111 = 2,905 і можна сказати, що на ньому аномалій немає.

У табл. 1 також видно, що для двох перших космічних знімків другого рівня “піраміди”, що відповідають другому знімку першого рівня “піраміди” середні ФР (D11211 = 2,875; D11212 = 2,910) менше середньої ФР D112 = 2,929, тобто на них є аномалії (див. рис. 3, б) На цих знімках аномалії займають майже всю їх поверхню. Середні ФР третього і четвертого космічних знімків другого рівня “піраміди” дорівнюють D11221 = 2,950 і D11222 = 2,955, що більше за D112 = 2,929. Такі величини середніх ФР свідчать про відсутність на них аномалій (рис. 3, б).

Середні ФР перших двох космічних знімків другого рівня “піраміди” (див. табл. 1) (D12111 = 2,936; D12112 = 2,922), що відповідають третьому знімку першого рівня “піраміди” менше середньої ФР D121 = 2,953, що відповідає наявності на них аномалій (див. рис. 3, в). Середні ФР третього і четвертого космічного знімку другого рівня “піраміди” більше D121 = 2,953 і аномалій на них немає (рис. 3, в). Середні ФР всіх космічних знімків другого рівня “піраміди”, які відповідають четвертому знімку першого рівня “піраміди” менше D122 = 2,962, але візуально на космічних знімках аномалії не спостерігаються (рис. 3, г). Тобто можна зробити висновок, що на них є особливості, які візуально не помітні. Таким чином, аналіз середніх ФР другого рівня “піраміди” показує, що вони дозволяють локалізувати аномалії, а також виявити особливості, які візуально на космічному знімку не помітні.

За допомого побудови третього рівня “піраміди” середніх ФР можна проводити подальшу локалізацію аномалій на космічному знімку. У табл. 1 для прикладу наведено четверта частина (шістнадцять) середніх ФР третього рівня “піраміди”, яка відповідає першому верхньому зліва квадранту вихідного космічного знімку.

Як видно з даних табл. 1, середні ФР трьох космічних знімків третього рівня “піраміди” (D1111111 = 2,937; D1111112 = 2,930; D1111122 = 2,940), які відповідають першому знімку другого рівня “піраміди”, менше середньої ФР D11111 = 2,956, тобто на них є особливості, які візуально не помітні (рис. 4, а). Одна середня ФР космічного знімка третього рівня “піраміди” (D1111222 = 2,770), що відповідає другому знімку другого рівня “піраміди” менше середньої ФР D11112 = 2,880 і одна середня ФР (D1111212 = 2,889) близька до неї, і така їх поведінка вказує на те, що на них є аномалії (див. рис. 4, б). Також необхідно зазначити, що середні ФР першого і третього космічного знімку третього рівня “піраміди” більші D11112 = 2,880 і на цих знімках аномалії не спостерігаються (рис. 4, б).

Дві середні ФР космічних знімків третього рівня “піраміди” (табл. 1) (D1112111 = 2,788; D1112112 = 2,835), що відповідають третьому знімку другого рівня “піраміди” менше середньої ФР D11121 = 2,897 і на них є аномалії (рис. 4, в). Середні ФР третього і четвертого космічних знімків третього рівня “піраміди” дорівнюють D1112121 = 2,948 та D1112122 = 2,956 і вони більше D11121 = 2,897, тобто такі значення середніх ФР свідчать, що аномалій на них немає. Але на цих космічних знімках є аномалії і вони займають всі ці знімки розміром 7х7 і тому вони не виявляються як аномальні, але ці аномалії були виявлені на вищих рівнях “піраміди” (рис. 4, в).

Середні ФР першого і третього космічних знімків третього рівня “піраміди”, що відповідають четвертому знімку другого рівня “піраміди” дорівнюють D1112211 = 2,892 і D1112221 = 2,905, що менше або близько до середньої ФР D11122 = 2,902, тобто на цих знімках є аномалії (див. рис. 4, г). Середні ФР другого і четвертого космічних знімків третього рівня “піраміди” (див. табл. 1) є більшими за D11122 = 2,902 і на цих знімках аномалій немає (рис. 4, г). Таким чином, застосування “піраміди” середніх ФР дозволяє локалізувати аномалії на космічних знімках, якщо величини середніх ФР космічних знімків на кожному рівні “піраміди” менше D = 2,9, то на них є аномалії.

Розглянемо можливість локалізації аномалій на космічних знімках з використанням “піраміди” мінімальних ФР, яку наведено у табл. 2. У табл. 2 видно, що мінімальна ФР вихідного знімка дорівнює Dм1 = 2,678 (середня ФР D1 = 2,944) і це значення менше D = 2,9 (різниця у першому знаку після коми склала ΔD = 0,3), тобто, якщо мінімальна ФР менше 2,9, то це може свідчити про наявність на космічному знімку аномалії. Така поведінка мінімальної ФР може використовуватися для виявлення аномалій на космічному знімку і у подальшому для їх локалізації.

Таблиця 2 – “**Піраміда” мінімальних фрактальних розмірностей космічного знімка поля з пшеницею з аномаліями для “вікна” 4х4 пікселя**

D	1				
1	2,678				
1	1	2			
1	2,633	2,693			
2	2,868	2,929			
11	1	2	**12**	1	2
1	2,937	2,678	1	2,661	2,763
2	2,791	2,786	2	2,914	2,933
21	1	2	**22**	1	2
1	2,886	2,849	1	2,936	2,916
2	2,907	2,952	2	2,914	2,921
11	1	2			
1	2,937	2,678			
2	2,791	2,786			
111	1	2	**112**	1	2
1	2,937	2,930	1	2,929	2,889
2	2,965	2,940	2	2,892	2,770
121	1	2	**122**	1	2
1	2,788	2,835	1	2,892	2,926
2	2,948	2,956	2	2,959	2,942

З аналізу даних у табл. 2 також видно, що на першому рівні “піраміди” мінімальні ФР першого і другого космічних знімків менші або близькі до мінімальної ФР вихідного космічного знімка і менше D = 2,9 на ΔD = 0,3, тобто на них є аномалії. Для третього і четвертого знімків мінімальні ФР (див. табл. 2) більше мінімальної ФР всього вихідного знімка, але для третього знімка мінімальна ФР не значно, але менше D = 2,9, тобто на ньому є невелика аномалія. Для четвертого знімка мінімальна ФР більше D = 2,9 і на ньому аномалій немає.

Таким чином, під час використання мінімальних ФР для виявлення аномалій на космічних знімках можна вважати, що, якщо величина мінімальної ФР менше D = 2,9, то на космічному знімку є аномалії. Цей висновок підтверджується такою ж поведінкою мінімальних ФР на другому і третьому рівнях “піраміди” (див. табл. 2). Візуально проконтролювати наявність аномалій на космічних знімках можна на рис. 2-4.

Розглянемо яку поведінку мають “піраміди” середніх (мінімальних) ФР під час аналізу космічних знімках без аномалій. На рис. 5 наведено космічний знімок супутника Sentinel-2 (канал b8) поля з пшеницею без аномалій розмірами 56х56 пікселів.

Рис. 5. Космічний знімок супутника Sentinel-2 поля з пшеницею без аномалій розмірами 56х56 пікселів

У табл. 3 наведено “піраміду” середніх ФР для космічного знімка поля з пшеницею без аномалій.

*Таблиця 3 – “***Піраміда” середніх фрактальних розмірностей космічного знімка поля з пшеницею без аномалій для “вікна” 4х4 пікселя**

D	1				
1	2,968				
1	1	2			
1	2,965	2,970			
2	2,964	2,957			
11	1	2	**12**	1	2
1	2,973	2,963	1	2,973	2,966
2	2,965	2,970	2	2,965	2,973
21	1	2	**22**	1	2
1	2,955	2,964	1	2,967	2,938
2	2,952	2,973	2	2,956	2,963
11	1	2			
1	2,973	2,963			
2	2,965	2,970			
111	1	2	**112**	1	2
1	2,955	2,976	1	2,997	2,955
2	2,958	2,978	2	2,950	2,973
121	1	2	**122**	1	2
1	2,946	2,969	1	2,981	2,958
2	2,953	2,943	2	2,974	2,967

На рис. 5 видно, що на космічному знімку поля з пшеницею візуально аномалії не спостерігаються, а у табл. 3 видно, що середня ФР всього знімка дорівнює D1 = 2,968. Така величина середньої ФР більше фрактальної розмірності 2,9 (D > 2,9) і , як зазначалося вище, це може бути ознакою, що на космічному знімку аномалій немає. Аналіз даних табл. 3 показує, що середні ФР на першому рівні “піраміди” (D111 = 2,965; D112 = 2,970 і D121 = 2,964, D122 = 2,957) більше 2,9 і близькі до середньої ФР вихідного космічного знімка D1 = 2,968 і цу підтверджує, що на космічному знімку аномалії немає. Але необхідно зазначити, що середня ФР четвертого знімка є незначно меншою середньої ФР вихідного космічного знімка, що може вказувати на наявність на ньому деяких особливостей, але не аномалій. У табл. 3 також видно, що середні ФР всіх космічних знімків другого рівня “піраміди” більше фрактальної розмірності D = 2,9 і це також підтверджує відсутність на них аномалій. Необхідно зазначити, що на другому рівні “піраміди” також є середні ФР які менше середніх ФР першого рівня “піраміди”, але не аномальні. Тобто на космічних знімках, яким відповідають ці середні ФР, можуть бути деякі особливості. Характер поведінки середніх ФР всіх космічних знімків третього рівня “піраміди” (див. табл. 3) такі ж самі, як і на другому рівні, тобто більше фрактальної розмірності D = 2,9. Таким чином, проведений аналіз дозволяє зробити висновок, що якщо середні ФР на всіх рівнях “піраміди” більше фрактальної розмірності D = 2,9, то на космічному знімку аномалій немає. На деякі особливості на знімку вказують менші середні ФР на нижчих рівнях “піраміди” порівняно з вищими рівнями “піраміди”.

Розглянемо можливість аналізу космічних знімків без аномалій з використанням “піраміди” мінімальних ФР, яку наведено у табл. 4. Аналіз даних у табл. 4 показує, що на першому рівні “піраміди” мінімальні ФР трьох перших знімків більше мінімальної ФР вихідного космічного знімка, що свідчить про відсутність на них аномалій. Для другого і третього знімків першого рівня “піраміди” мінімальні ФР більше D = 2,9, тобто вони більш однорідні. У табл. 4 також видно, що мінімальні ФР всіх знімків на другому і третьому рівнях “піраміди” (крім однієї) більше фрактальної розмірності D = 2,9, тобто на них аномалій немає. Як і для середніх ФР, на деякі особливості на космічному знімку вказують менші мінімальні ФР на нижчих рівнях “піраміди” порівняно з вищими її рівнями.

*Таблиця 4 – “***Піраміда” мінімальних фрактальних розмірностей космічного знімка поля з пшеницею без аномалій для “вікна” 4х4 пікселя**

D	1				
1	2,811				
1	1	2			
1	2,892	2,940			
2	2,907	2,791			
11	1	2	**12**	1	2
1	2,909	2,925	1	2,939	2,928
2	2,926	2,934	2	2,931	2,936
21	1	2	**22**	1	2
1	2,928	2,941	1	2,922	2,803
2	2,925	2,946	2	2,909	2,931
11	1	2			
1	2,909	2,925			
2	2,926	2,934			
111	1	2	**112**	1	2
1	2,927	2,934	1	2,994	2,931
2	2,927	2,934	2	2,937	2,967
121	1	2	**122**	1	2
1	2,917	2,931	1	2,934	2,953
2	2,947	2,937	2	2,937	2,960

Таким чином, результати фрактального аналізу показали, що за допомогою “пірамід” ФР можна здійснювати локалізацію аномалій на космічному знімку і досліджувати на них невеликі особливості.

Висновки

Оцінити негативний стан і структуру сільськогосподарських земель можна з використанням інформації, що отримується з космічних засобів дистанційного зондування Землі. На космічних знімках ділянки землі, що мають ознаки деградації, проявляються у вигляді різноманітних аномалій. Фрактальні розмірності є дробовими величинами, які характеризують складність структури космічного знімка, і їх можна розраховувати за допомогою методу покриття, або методу призми. Розглянуто можливість локалізації аномалій на космічних знімках сільськогосподарських земель з використанням “пірамідального” фрактального аналізу. Під час побудови “піраміди” фрактальних розмірностей для кожного знімка “піраміди” космічних знімків розраховується середня фрактальна розмірність з використанням “ковзаючого вікна”. Тобто для кожного “вікна”, що рухається по космічному знімку з кроком переміщення, який дорівнює одиниці, визначається фрактальна розмірність, а далі отримані фрактальні розмірності усереднюють. Показано, що застосування “пі-

раміди" середніх фрактальних розмір-ностей дозволяє локалізувати аномалії на космічних знімках, якщо величини середніх (мінімальних) фрактальних розмірностей космічних знімків на кожному рівні "піраміди" менше D = 2,9, то на них є аномалії. За допомогою фрактального аналізу космічних знімків без аномалій показано, що, якщо середні (мінімальні) фрактальні розмірності на всіх рівнях "піраміди" більше фрактальної розмірності D = 2,9, то на космічному знімку аномалій немає. На деякі особливості на космічному знімку вказують менші середні (мінімальні) фрактальні розмірності на нижчих рівнях "піраміди" порівняно з вищими рівнями "піраміди". Під час подальших досліджень доцільно розглянути можливість локалізації аномалій на космічних знімках з використанням розрахунку фрактальних розмірностей у "вікні" з розмірами, що дорівнюють розмірам знімків на кожному рівні "піраміди" космічних знімків.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автор підтверджує, що не використовував технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Тараріко О.Г., Сиротенко О.В., Ільєнко Т.В., Кучма Т.Л. Агроекологічний супутниковий моніторинг. К.: Аграр. наука, 2019. 204 с. URL: https://agroeco.org.ua/vydannya/agroekologichnij-suputnikovij-monitoring
2. Yang L., Lu B., Schmidt M., Natesan S. et al. Applications of remote sensing for crop residue cover mapping. Smart Agricultural Technology. 2025. №. 11. P. 1 – 14. DOI: https://doi.org/10.1016/j.atech.2025.1008080
3. Copernicus Europe's eyes on Earth, Sentinel-2. Copernicus Europe's eyes on Earth [Electronic resource]. URL: https://www.copernicus.eu/en/about-copernicus/infrastructure/discover-our-satellites
4. Landsat 8 Bands: Combinations For Satellite Images. URL: https://eos.com/blog/band-combinations-for-landsat-8/
5. Feder J. Fractals. New York: Springer US, 1988. 263 p. DOI: https://doi.org/10.1007/978-1-4899-2124-6
6. Марюшко М.В., Пащенко Р.Е. Фрактальний аналіз космічних знімків SENTINEL-2 для моніторингу сільськогосподарських культур. Радіоелектронні і комп'ютерні системи. 2020. №4(96). С. 34–47. DOI: https://doi.org/10.32620/reks.2020.4.03
7. Пащенко Р.Е., Марюшко М.В. Оцінка стану різних сільськогосподарських культур з використанням фрактального аналізу. Сучасні інформаційні системи. 2023. Т. 7, № 3. С. 81–88. DOI: https://doi.org/10.20998/2522-9052.2023.3.12
8. Пащенко Р.Е., Марюшко М.В. Моніторинг змін стану сільськогосподарських земель за даними фрактального аналізу космічних знімків. Системи управління, навігації та зв'язку. 2021. Вип. 3(65). С. 8–17. DOI: https://doi.org/10.26906/SUNZ.2021.3.008
9. Crownover R.M. Introduction to Fractals and Chaos. London: Jones and Bartlett Publishers, Inc, 1995. 352 p. URL: https://apps.dtic.mil/sti/tr/pdf/ADA210257.pdf
10. Clarke K.C. Computation of the Fractal Dimension of Topographic Surface Using the Triangular Prism Surface Area Method. Computers & Geosciences. 1986. Vol. 12. № 5. P. 713–722. URL: https://www.sciencedirect.com/science/article/abs/pii/0098300486900476

Received (Надійшла) 21.01.2026
Accepted for publication (Прийнята до друку) 18.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Пащенко Руслан Едуардович – доктор технічних наук, професор, старший науковий співробітник відділу дистанційного зондування Землі, Інститут радіофізики та електроніки імені О.Я. Усикова НАН України, Харків, Україна;
Ruslan Pashchenko – Doctor of Technical Sciences, Professor, Senior research scientist of Department remote sensing of the Earth, O.Ya. Usikov Institute for Radio physics and Electronics of the NAS of Ukraine, Kharkov, Ukraine,
e-mail: r.paschenko@i.ua; ORCID Author ID: https://orcid.org/0000-0001-6218-0324;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=58126357500.

Localization anomalies on the spaces pictures of agricultural earths by construction "pyramid" of fractals dimensions

Ruslan Pashchenko

Abstract. With the use information that turns out from the remote sensing of Earth spaces facilities, it is possible to estimate the negative state of agricultural earths which show up as the varied anomalies. The **subject** of the study in the article is estimation possibility localization anomalies on spaces pictures with the use "pyramid" of fractals dimensions. The **object** of the study are agricultural earths spaces pictures with anomalies and without anomalies which are got from a satellite Sentinel-2. **The aim** is to assess possibility localization anomalies on the agricultural earths spaces pictures with use fractal analysis. **The following results were obtained.** Possibility localization anomalies is considered on agricultural earths spaces pictures with the use "pyramidal" fractal analysis. Creation "pyramid" of spaces pictures is carried out due to the division the base space picture on the less sizes pictures which in subsequent are indexed. During construction "pyramid" of fractals dimensions for every picture "pyramid" of spaces pictures a middle (minimum) fractal dimension settles accounts with the use of "sliding window", that "window", that moves on a space picture with the moving step, which equals to unit. **Conclusions**. Application "pyramid" of middle (minimum) fractals dimensions allows to localize anomalies on spaces pictures, if sizes middle (minimum) fractals dimensions spaces pictures at every level "pyramid" less D=2,9, there are anomalies on them. By the fractal analysis of spaces pictures without anomalies showed that, if middle (minimum) fractals dimensions at all levels "pyramid" more fractal dimension D=2,9, on the space picture anomalies it is not. On some features on a space picture specify less middle (minimum) fractals dimension at lower levels "pyramid" comparatively with the higher levels "pyramid".

Keywords: monitoring the state of agricultural earths; space pictures; anomaly; fractal dimension.

Автомобільний, річковий, морський та авіаційний транспорт

УДК 656.7.052:629.7 **doi: 10.26906/SUNZ.2026.2.012**

Ю. В. Баістов[1], О. В. Сальник[1], О. Ю. Дроль[1], С. В. Мельник[1], Л. В. Грушенкова[2]

[1] Харківський національний університет Повітряних Сил ім. І. Кожедуба, Харків, Україна
[2] Науково-дослідний інститут Воєнної розвідки, Київ, Україна

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ УПРАВЛІННЯ ПОВІТРЯНИМ РУХОМ В УМОВАХ ДИНАМІЧНОЇ ПОВІТРЯНОЇ ОБСТАНОВКИ

Анотація. Предметом вивчення в статті є існуючі методи управління повітряним рухом. **Метою** є аналіз існуючих методів управління повітряним рухом в умовах динамічної повітряної обстановки. **Завдання:** здійснити аналіз існуючих підходів до управління повітряним рухом; виокремити групи методів управління повітряним рухом за схожим принципом дії; провести порівняльну характеристику сучасних методів управління повітряним рухом; провести оцінку методів управління повітряним рухом за ключовими критеріями. Використовуваними **методами** є: аналітичні та емпіричні методи порівняльного дослідження. Отримано такі **результати.** Проведено комплексний аналіз сучасних методів управління повітряним рухом в умовах динамічної повітряної обстановки. Виокремлено та розглянуто шість основних груп методів управління повітряним рухом. Встановлено їх основні принципи роботи, їх основні переваги та недоліки та рівень ефективності роботи методів в умовах динамічної повітряної обстановки. Проведено оцінку виокремлених груп методів управління повітряним рухом за такими ключовими критеріями, як адаптивність до змін повітряної обстановки, оперативність, масштабованість та рівень їх автоматизації. Доведено, що подальший розвиток систем управління повітряним рухом має бути спрямований на підвищення рівня автоматизації, забезпечення масштабованості, оперативності прийняття рішень та ефективної інтеграції великої кількості повітряних об'єктів повітряного простору. **Висновки.** Отримані результати можуть бути використані при розробці та удосконаленні методів повітряного руху. Напрямом подальших досліджень є розробка методу управління повітряним рухом з урахуванням сучасних умов динамічної повітряної обстановки.

Ключові слова: авіаційна технологія, повітряний об'єкт, управління повітряним рухом, метод управління повітряним рухом, динамічна повітряна обстановка.

Вступ

Постановка проблеми у загальному вигляді. В умовах стрімкого розвитку авіаційних технологій, які в останній час активно застосовуються як в цивільній, так і військовій авіації, а також в рятувальних операціях, вантажних перевезеннях та сільському господарстві, та як результат такого застосування в умовах постійного зростання інтенсивності повітряного руху, питання ефективного управління рухом повітряних об'єктів є дуже актуальним [1–3]. Особливої актуальності це набуває в умовах, коли ситуація в повітряному просторі постійно змінюється в часі, тобто в умовах динамічно змінного операційного середовища, та коли зростають навантаження та складність такої обстановки [4]. Отже, динамічна повітряна обстановка характеризується [4]:

– високою інтенсивністю польотів;

– можливим регіональним військовим конфліктом (різнорідні повітряні атаки);

– появою розвідувальних та ударних безпілотних літальних апаратів (БпЛА);

– можливою дією засобів радіоелектронної боротьби та радіоелектронної протидії;

– швидкоплинною зміною кількості, типів і траєкторій повітряних об'єктів;

– забороною на використання окремих ділянок повітряного простору тощо.

Досвід останніх воєнних конфліктів показав, що така повітряна обстановка є характерною для сучасних умов ведення бойових дій та управління повітряним рухом. Зокрема, в зонах з інтенсивним використанням різними повітряними об'єктами (як пілотованих, так і БпЛА), коли ситуація постійно змінюється та потребує оперативного об'єктивного аналізу та прийняття управлінських рішень. Така обстановка є також характерною під час кризових ситуацій, коли незаплановано відбувається різке зростання невизначеності, відбувається перевантаження інформаційних каналів, обмежений час на прийняття рішення та при цьому є нагальна потреба одночасно враховувати велику кількість взаємопов'язаних факторів в режимі реального часу [5].

Характерною особливістю і вимогою сьогодення є інтеграція різнорідних користувачів повітряного простору як цивільної, так і військової авіації в єдину систему. Це суттєво ускладнює процеси планування, координації польотів повітряних об'єктів та забезпечення безпеки їх польотів. Питанню складності спільного використання повітряного простору сьогодні приділяється багато уваги [6–8].

Традиційні підходи до управління повітряним рухом, які були розроблені для пілотованої авіації, на сьогодні вже не забезпечують достатньої ефективності в тих умовах, що склалися [9]. Отже, важливість розроблення нових підходів визначається необхідністю гарантування безпеки повітряним об'єктам цивільної авіації та ефективного забезпечення виконання бойових завдань повітряними об'єктами військової авіації.

Для виконання бойових завдань проблема є більш критичною. Адже радіолокаційні системи мають обмеження щодо виявлення малопомітних цілей, роботи в умовах радіоелектронної протидії та інтеграції даних із різнорідних джерел. Крім того, вони не забезпечують достатньої гнучкості для підтримки прийняття рішень у режимі реального часу в умовах високої динаміки повітряної обстановки.

Отже, динамічно змінюване операційне середовище з наявністю у повітряному просторі як цивільних, так і військових повітряних об'єктів, потребує високоточного, оперативного та адаптивного управління. При такому спільному використанні повітряного простору потрібна чітка узгодженість різних процедур, пріоритетів та обмежень.

Таким чином, постає завдання щодо розробки та вдосконалення методів управління повітряним рухом, які здатні ефективно функціонувати в умовах динамічно змінюваного операційного середовища, неоднорідної та багатокористувацької повітряної обстановки. При цьому забезпечуючи безпеку повітряним об'єктам цивільної авіації та ефективне забезпечення виконання бойових завдань повітряними об'єктами військової авіації, пропускну здатність та оперативність прийняття рішень.

Аналіз останніх досліджень і публікацій.

Аналіз сучасних наукових досліджень і публікацій [10, 11] свідчить, що розглянута проблема управління повітряним рухом в умовах динамічної повітряної обстановки на сьогоднішній день активно досліджується в контексті інтеграції нових типів повітряних об'єктів, зокрема БпЛА та систем міської повітряної мобільності (Urban Air Mobility, UAM).

В роботах міжнародних організацій та дослідницьких програм, таких як ICAO, EUROCONTROL, SESAR, [12–14], розглядаються концепції інтегрованого управління повітряним простором. Ці концепції передбачають спільне використання повітряного простору як пілотованою, так і безпілотною цивільною та військовою авіацією. В [12–14] підкреслюється необхідність розвитку ефективної взаємодії цивільної та військової авіації та гнучкого використання ними повітряного простору (Flexible Use of Airspace, FUA) як ключового елемента сучасних систем управління повітряним рухом.

В останній час виокремлено окремий напрям досліджень, який пов'язаний з інтеграцією БпЛА в уже існуючі системи управління повітряним рухом. В роботах [15, 16] доведено, що стрімке зростання кількості БпЛА, яке в найближчому майбутньому вже перевищить кількість пілотованої авіації, розпочало створювати значне навантаження на систему управління повітряним рухом. Отже, таке застереження вимагає розробки нових моделей та підходів щодо безконфліктної координації та розведення повітряних об'єктів.

Важливим напрямом є також концепція динамічного управління повітряним простором (Dynamic Airspace Management, DAM) [17]. DAM передбачає адаптивну зміну конфігурації повітряних маршрутів в реальному часі залежно від поточної повітряної обстановки. Ця концепція вже активно використовується в діяльності EUROCONTROL, стандартах ICAO та програмі SESAR.

Також наукові дослідження останніх років [18, 19] вказують на необхідність використання цифрових технологій, зокрема штучного інтелекту, великих даних та інтегрованих комунікаційних систем при управлінні повітряним рухом. Це може забезпечити ситуаційну обізнаність та підтримку прийняття рішень у режимі реального часу, особливо в умовах динамічно змінюваної повітряної обстановки.

Отже, аналіз останніх досліджень і публікацій показує, що сучасні методи управління повітряним рухом вимушено еволюціонують у напрямі інтеграції, цифровізації та динамічного адаптивного управління. Водночас залишаються відкритими питання підвищення ефективності одночасного управління повітряним рухом військової та цивільної авіації, як пілотованої, так і безпілотної, в умовах динамічної повітряної обстановки, що обґрунтовує актуальність подальших досліджень у даній сфері.

Мета статті – провести детальний аналіз існуючих методів управління повітряним рухом в умовах динамічної повітряної обстановки.

Основна частина

Аналіз показав, що сучасні методи управління повітряним рухом умовно можна поділити на шість основних груп. Такий розподіл залежить від принципів побудови, рівня автоматизації, масштабованості та адаптивності до змін повітряної обстановки.

До першої групи було віднесено усі ***традиційні (класичні) методи управління повітряним рухом (традиційне диспетчерське управління) (Air Traffic Control, ATC)***, які були сформовані в другій половині XX століття і які досі використовуються у більшості країн як базові. Група цих методів базується на централізованому диспетчерському управлінні, при якому ключові рішення щодо руху повітряних об'єктів приймає людина (диспетчер), а екіпаж виконує його вказівки, дотримуючись встановлених процедур [20, 21].

Група традиційних (класичних) методів управління повітряним рухом передбачає [20, 21]:

– фіксовану структуру повітряного простору – коли увесь простір умовно поділяється на сектори, границі яких є статичними, та коли кожен з секторів обслуговується окремим диспетчером;

– заздалегіть визначені маршрути руху повітряних об'єктів – коли повітряні об'єкти переміщаються у повітряному просторі по визначеним повітряним трасам, відхилення від яких є обмеженим;

– стандартизовані процедури – коли використовуються уніфіковані схеми усіх процесів, наприклад, заходу повітряного об'єкта на посадку;

– ешелонування – коли безпека польотів забезпечується шляхом витримування мінімально допустимих інтервалів як у вертикальній, так і у горизонтальній площині.

Але попри високу надійність роботи традиційних методів, вони мають низку суттєвих недоліків, а саме:

– низьку адаптивність – тобто система управління повітряним рухом, побудована на класичних методах, слабо реагує на будь-які зміни у трафіку, погодних умовах тощо та на виникнення надзвичайних та кризових ситуацій;

– неоптимальність маршрутів руху повітряних об'єктів – адже фіксовані повітряні траси не завжди є найкоротшими та/або найекономнішими;

– обмеження пропускної здатності – адже через жорсткі інтервали часу між повітряними об'єктами суттєво знижується ефективність використання повітряного простору;

– перевантаження диспетчерів як людський фактор – адже зі збільшенням інтенсивності руху повітряних об'єктів диспетчер обробляє великий обсяг вхідної інформації в реальному часі, що суттєво збільшує ризик його помилок.

Отже, традиційні (класичні) методи управління повітряним рухом забезпечують високий рівень безпеки завдяки централізації та стандартизації, проте їхня жорстка структура обмежує ефективність у сучасних умовах зростання трафіку у повітряному просторі та робить їх недієвими в умовах динамічної повітряної обстановки.

До другої групи було віднесено усі ***методи управління повітряним рухом на основі траєкторій (Trajectory-Based Operations, TBO).*** Ця група методів є одним із ключових напрямів розвитку сучасних систем управління повітряним рухом (Air Traffic Management, ATM).

Основною концепцією роботи методів цієї групи є планування та управління польотами повітряних об'єктів з урахуванням чотиривимірних траєкторій (4D-траєкторій), де три виміри – це просторові координати (широта, довгота та висота), а четвертим виміром є час (точка проходження повітряного об'єкту у конкретний момент часу) [22, 23].

На відміну від попередньої групи (коли диспетчер реагує вже після виникнення ситуації), методи даної групи орієнтовані на прогнозування польоту повітряних об'єктів та управління на випередження (проактивне управління) повітряним простором.

Група методів управління повітряним рухом на основі траєкторій передбачає [22, 23]:

– чотиривимірне траєкторне планування – коли для польоту кожного повітряного об'єкта формується узгоджена з іншими траєкторія. При такому узгодженні враховується маршрут польоту, профіль висоти та часові обмеження;

– оперативний обмін даними в реальному часі – коли взаємодія через цифрові канали зв'язку відбувається між тими, хто задіяний у даному процесі (повітряним об'єктом, диспетчерськими центрами, авіакомпаніями тощо);

– інтеграцію з автоматизованими системами – коли для вирішення завдань планування та управління польотами повітряних об'єктів використовуються алгоритми оптимізації, штучний інтелект, моделі прогнозування тощо;

– прогнозування конфліктів – коли система управління повітряним рухом не тільки завчасно визначає можливі конфлікти прокладених повітряних трас, а й пропонує рішення для їх вирішення.

Але попри підвищення пропускної здатності повітряного простору, зменшення затримок, оптимізацію ресурсів та зниження навантаження на диспетчерів, методи другої групи мають низку суттєвих недоліків та обмежень, а саме:

– залежність від точності прогнозів – адже при наявності неточностей в отриманих метеоданих, похибок даних характеристик польоту або помилок у поведінці екіпажу може призвести до відхилень від запланованої траєкторії польоту;

– залежність від стану розвитку та якості інформаційної інфраструктури – адже для якісної роботи такої системи необхідні високонадійні канали зв'язку, чітка синхронізація даних між усіма учасниками та кібербезпека;

– чутливість до невизначеності, так званих форс-мажорних ситуацій – наприклад ситуація, яка на сьогодні склалася в Україні є гарним прикладом для неможливості використання методів даної групи;

– складність впровадження – адже, це потребує дороговартісної модернізації як наземної інфраструктури, так і бортових систем.

Але досвід міжнародних організацій показує, що методи управління повітряним рухом на основі траєкторій поступово інтегруються у практику управління повітряним рухом, доповнюючи класичні підходи до управління повітряним рухом.

До третьої групи було віднесено усі ***методи управління потоками повітряного руху (Air Traffic Flow Management, ATFM).*** Ця група методів сфокусована на збалансуванні попиту на польоти повітряних об'єктів та пропускної здатності повітряного простору й можливостей наземної інфраструктури.

Основною відмінністю методів цієї групи є робота на стратегічному, тобто планування польотів за добу і більше, та передтактичному, тобто планування за кілька годин до польоту, рівнях.

Група методів управління потоками повітряного руху передбачає [24, 25]:

– балансування попиту та пропускної здатності – коли відбувається постійний аналіз кількості запланованих польотів, можливостей секторів повітряного простору та наземної інфраструктури, усі обмеження та приймає рішення чи не виникне перевантаження;

– централізоване планування – коли управління потоками повітряного руху здійснюється спеціалізованими центрами;

– регулювання повітряних потоків – коли застосовуються вимушені міри, наприклад, затримка повітряного об'єкту на землі, перенаправлення маршрутів, слотування тощо, у випадку коли пронозується або відбувається перевищення пропускної здатності;

– використання моделей для прогнозування – коли для запобігання перевантаженню повітряного простору, ще до виникнення такої ситуації, завчасно аналізуються трафік, погодні умови, завантаженість секторів тощо.

Але попри такі переваги як можливість запобігання перевантаженню повітряного простору, змен-

шення затримок та підвищення загальної ефективності системи управління повітряним рухом, методи третьої групи мають низку суттєвих недоліків та обмежень, а саме:

– залежність від точності прогнозів – адже ефективність роботи методів напряму залежить від якості даних, моделей та коректної координації всіх учасників системи;

– обмежена оперативність у реальному часу – адже група методів працює наперед і не завжди реагує на раптові зміни та форс-мажорні ситуації;

– затримки на окремих маршрутах – адже оптимізація системи в цілому може призвести до індивідуальних затримок повітряних об'єктів.

На сьогоднішній день методи управління потоками повітряного руху вже є невід'ємною частиною систем управління повітряним рухом і ведуться розробки щодо їх інтегрування з методами управління повітряним рухом на основі траєкторій.

До наступної четвертої групи належать ***методи управління повітряним рухом в умовах інтеграції БпЛА***. Поява та виокремлення даної групи зумовлені стрімким збільшенням кількості використання БпЛА як у військовій сфері, так і для цивільних потреб та інтеграцією БпЛА у загальний повітряний простір.

Отже, така інтеграція БпЛА в єдиний повітряний простір разом із пілотованою авіацією стала потребувати нових підходів до управління повітряним рухом, адже попередньо розглянуті групи методів управління повітряним рухом не враховували особливості застосування БпЛА, а саме: малі та гранично малі висоті польоту БпЛА, можливість одночасно дуже великої їх кількості перебування в повітряному просторі та високу динамічність повітряної обстановки.

Отже, у зв'язку з появою у повітряному просторі повітряних об'єктів з переліченими особливостями, розпочали активно розвиватися такі спеціалізовані системи управління повітряним рухом, як UTM (Unmanned Traffic Management) та U-space [26].

Методи даної групи мають наступні особливості [26, 27]:

– високий рівень автоматизації – коли управління повітряним рухом здійснюється переважно без участі диспетчера, а із застосуванням елементів штучного інтелекту, алгоритмів оптимізації та/або цифрових платформ;

– децентралізоване або гібридне управління – коли рішення можуть прийматися не одним центром прийняття рішень, а розподіленими системами або навіть самими операторами БпЛА;

– активне застосування цифрових сервісів та обміну даними в режимі реального часу – коли використовуються онлайн-платформи планування польотів повітряних об'єктів, геозони та цифрова ідентифікація БпЛА.

Концепція розвитку методів даної групи сфокусована на наступних напрямах:

– стратегічному та тактичному розведенні повітряних об'єктів з метою планування маршрутів польоту та уникнення конфліктів та подій як у реальному часі так і заздалегідь;

– виявлення та уникнення конфліктів з метою автоматичного визначення ризику зіткнення повітряних об'єктів та коригування траєкторій їх руху;

– координації великої кількості учасників повітряного руху з метою одночасного управління сотнями, а то й тисячами БпЛА у спільному повітряному просторі.

Але попри такі переваги як можливість масштабування при великій кількості БпЛА, зменшення навантаження на диспетчерів та ефективне використання низьковисотного повітряного простору, методи четвертої групи мають низку суттєвих недоліків та обмежень, а саме:

– висока щільність повітряного трафіку – адже такий підхід потребує нових алгоритмів управління повітряного руху з метою уникнення конфліктів та оптимізації маршрутів руху;

– повна залежність від цифрової інфраструктури – адже зв'язок, навігація та питання кібербезпеки є критично важливими для стабільного функціонування систем і безпечного виконання польотів повітряних об'єктів;

– регуляторні обмеження – адже законодавство з використання таких спеціалізованих систем управління повітряним рухом ще перебуває на стадії розробки;

– проблеми інтеграції з іншими системами управління повітряним рухом – адже необхідною є взаємодія з працюючими системами традиційного диспетчерського управління та управління потоками повітряного руху.

На сьогоднішній день розвиток методів четвертої групи відбувається в рамках як міжнародних ініціатив (Eurocontrol, International Civil Aviation Organization, European Union Aviation Safety Agency), так і за участю Украерорух.

Отже, методи управління повітряним рухом в умовах інтеграції БпЛА є актуальним та новим етапом розвитку систем управління повітряним рухом, який дозволить в майбутньому ефективно використовувати повітряний простір великої кількості як пілотованої, так і безпілотної авіації. Водночас даний підхід потребує високого рівня технологічної зрілості та узгодженого регулювання.

До наступної п'ятої групи належать ***методи динамічного управління повітряним простором (Dynamic Airspace Management, DAM)***. До виокремлення даної групи призвело зростання інтенсивності повітряного руху, необхідність підвищення ефективності використання повітряного простору, а також потреба швидкого реагування на динамічні зміни у повітряному просторі, коли учасниками є як пілотовані, так і безпілотні повітряні об'єкти.

Через ці вимоги виникла нагальна потреба у розробці та впровадженні нових підходів до управління повітряним рухом, які будуть базуватися на принципах гнучкості, адаптивності та роботи у режимі реального часу. Все це закладено в концепції методів динамічного управління повітряним простором, які передбачають динамічну конфігурацію

повітряного простору залежно від поточної повітряної обстановки.

Методи даної групи мають наступні особливості [28, 29]:

– адаптивність повітряного простору – коли секторів повітряного простору змінюються в залежності від інтенсивності повітряного руху та, як результат, навантаження диспетчерів;

– динамічна маршрутизація польотів повітряних об'єктів – коли маршруту польоту оптимізуються в режимі реального часу, враховуючи реальну повітряну обстановку;

– орієнтація на актуальну інформацію – коли управління повітряним рухом базується на постійному моніторингу трафіку, погодних умов тощо;

– оперативне реагування на загрози та обмеження – коли система управління повітряним рухом оперативно змінює конфігурацію повітряного простору у відповідь на виникнення надзвичайних та кризових ситуацій.

Концепція розвитку методів даної групи сфокусована на наступних напрямах:

– динамічній конфігурації секторів обслуговування (Dynamic Airspace Configuration, DAC) з метою оптимального розподілу навантаження між диспетчерськими центрами;

– динамічній перебудові повітряного простору (Dynamic Airspace Reconfiguration, DAR) з метою адаптації до змін інтенсивності руху повітряних об'єктів;

– інтеграції з концепцією гнучкого використання повітряного простору (Flexible Use of Airspace, FUA) з метою ефективного управління повітряним рухом одночасно як цивільною, так і військовою авіацією.

Але попри такі переваги як підвищення пропускної здатності повітряного простору, зниження навантаження на диспетчерські центри, підвищення рівня безпеки, методи п'ятої групи мають низку недоліків та обмежень, які характерні четвертій групі та додатково:

– підвищені вимоги до координації між учасниками повітряного руху – адже необхідна тісна взаємодія між всіма користувачами повітряного простору та різними органами управління повітряним рухом.

Методи динамічного управління повітряним простором є наступним етапом еволюції систем управління повітряним рухом, що дозволяють перейти від статичних до гнучких моделей організації повітряного простору.

До останньої, шостої групи, умовно можна віднести ***методи використання цифрових технологій у системах управління повітряним рухом***.

На сьогоднішній день сучасні системи управління повітряним рухом вже активно інтегрують цифрові рішення, які засновані на елементах штучного інтелекту, технологіях великих даних, інтегрованих інформаційно-обчислювальних системах, хмарних обчисленнях і цифрових платформах.

Методи даної групи мають наступні особливості [30, 31]:

– автоматизація процесів управління повітряним рухом – коли більша частина процесів управління повітряним рухом передається автоматизованим системам управління;

– підтримка прийняття рішень – коли цифрові технології допомагають диспетчерам оцінювати обстановку, моделювати сценарії розвитку подій для прийняття рішення;

– забезпечення ситуаційної обізнаності – коли на основі інтеграції інформації з різнорідних джерел системи управління повітряним рухом формують "повну картину" про поточний стан повітряної обстановки;

– обробка великих обсягів даних в режимі реального часу – коли цифрові технології допомагають диспетчерам обробляти великі обсяги вхідної інформації.

Концепція розвитку методів даної групи сфокусована на наступних напрямах:

– впровадженні інтелектуальних систем підтримки прийняття рішень;

– використанні хмарних технологій;

– інтеграції технологій великих даних;

– розвитку єдиних цифрових екосистем управління повітряним рухом.

Особливо необхідно відзначити застосування технологій штучного інтелекту у системах управління повітряним рухом.

Це дозволяє: оптимізувати маршрути польоту, прогнозувати розвиток повітряної обстановки та планувати управління потоками повітряного руху, моделювати та виявляти можливі конфлікти між повітряними об'єктами.

Але попри такі переваги як зменшення навантаження на диспетчерів, підвищення рівня безпеки та ефективності управління повітряних рухом, методи п'ятої групи мають низку та обмежень, а саме [30, 31]:

– залежність результату роботи від якості та повноти вхідних даних;

– складність, а в деяких випадках, неможливість інтеграції з існуючими системами управління повітряним рухом;

– високі вимоги до захисту даних та кіберзахисту;

– складність у підготовці персоналу з питань роботи з новими технологіями.

Отже, методи управління повітряним рухом з використанням цифрових технологій є новим та важливим етапом розвитку систем управління повітряним рухом, що забезпечує перехід від відомих методів управління повітряним рухом до більш ефективних, інтелектуальних та адаптивних методів управління повітряним рухом.

Детальний аналіз методів управління повітряним рухом дозволив провести порівняльну характеристику досліджених сучасних методів, яка наведена в табл. 1.

В таблиці зазначено принцип, за яким працює кожна група методів, їх основні переваги та недоліки та рівень ефективності роботи методів в умовах динамічної повітряної обстановки.

Таблиця 1 – **Порівняльна характеристика сучасних методів управління повітряним рухом**

Метод управління	Основний принцип	Переваги	Недоліки	Ефективність у динамічній обстановці
Традиційне диспетчерське управління	Централізоване управління, фіксовані маршрути та сектори	Висока надійність, перевірені процедури	Низька гнучкість, перевантаження диспетчера	Низька
Trajectory-Based Operations (TBO)	Управління на основі 4D-траєкторій	Точність прогнозування, зменшення конфліктів	Залежність від якості даних	Середня
Air Traffic Flow Management (ATFM)	Балансування попиту і пропускної здатності	Оптимізація потоків, зменшення затримок	Обмежена реакція в реальному часі	Середня
UTM / U-space	Автоматизоване управління БпЛА	Масштабованість, інтеграція дронів	Високі вимоги до цифрової інфраструктури	Висока
Dynamic Airspace Management (DAM)	Динамічна конфігурація повітряного простору	Гнучкість, адаптивність, ефективність	Складність реалізації	Висока
AI-based, DT-based управління	Використання штучного інтелекту та Big Data	Швидкість аналізу, підтримка рішень	Потреба у великих даних, ризики помилок	Дуже висока

За результатами аналізу можна зробити такі висновки:

– методи традиційного диспетчерського управління є надійними, але недостатньо адаптивними та гнучкими;

– методи управління повітряним рухом на основі траєкторій забезпечує високу точність прогнозування, але залежить від якості та точності вхідних даних;

– методи управління потоками повітряного руху є ефективними для стратегічного планування, але обмежені у реальному часі;

– методи динамічного управління повітряним простором та методи використання цифрових технологій у системах управління повітряним рухом в умовах динамічної повітряної обстановки забезпечують найкращу адаптивність та ефективність.

Проведений аналіз показав, що в умовах динамічної зміни повітряної обстановки ефективні методи управління повітряним рухом мають забезпечувати: адаптивність до змін повітряної обстановки, оперативність та масштабованість. Іншим критерієм для оцінки методів управління повітряним рухом є рівень їх автоматизації. В табл. 2 наведено оцінку методів управління повітряним рухом за цими ключовими критеріями.

Таблиця 2 – **Оцінка методів управління повітряним рухом за ключовими критеріями**

Метод управління	Адаптивність до змін повітряної обстановки	Оперативність	Масштабованість	Рівень автоматизації
Традиційне диспетчерське управління	Низька	Середня	Низька	Низький
Trajectory-Based Operations (TBO)	Середня	Середня	Середня	Середній
Air Traffic Flow Management (ATFM)	Середня	Низька	Висока	Середній
UTM / U-space	Висока	Висока	Висока	Високий
Dynamic Airspace Management (DAM)	Висока	Висока	Висока	Високий
AI-based, DT-based управління	Дуже висока	Дуже висока	Дуже висока	Дуже високий

Отже, проведений порівняльний аналіз існуючих методів управління повітряним рухом в умовах динамічної повітряної обстановки показав, що насьогодні сучасні методи перебувають на етапі трансформації, яка спрямована на підвищення адаптивності до змін повітряної обстановки, оперативності їх роботи та можливості до масштабування, а також автоматизації та інтеграції великої кількості повітряних об’єктів.

Висновки

У статті проведено комплексний аналіз сучасних методів управління повітряним рухом в умовах динамічної повітряної обстановки. Виокремлено та розглянуто шість основних груп методів управління повітряним рухом. Встановлено їх основні принципи роботи, їх основні переваги та недоліки та рівень ефективності роботи методів в умовах динамічної повітряної обстановки. Проведено оцінку виокремлених груп методів управління повітряним рухом за такими ключовими критеріями, як адаптивність до змін повітряної обстановки, оперативність, масштабованість та рівень їх автоматизації.

Встановлено, що методи традиційного диспетчерського управління, попри їх надійність, не відповідають сучасним вимогам щодо гнучкості, адаптивності та ефективності. Натомість перспективними є методи динамічного управління повітряним простором, методи управління повітряним рухом в умовах інтеграції БпЛА та методи використання цифрових

технологій у системах управління повітряним рухом, зокрема штучного інтелекту і великих даних.

Доведено, що подальший розвиток систем управління повітряним рухом має бути спрямований на підвищення рівня автоматизації, забезпечення масштабованості, оперативності прийняття рішень та ефективної інтеграції великої кількості повітряних об'єктів повітряного простору. Це дозволить підвищити безпеку польотів і ефективність використання повітряного простору в умовах його навантаження.

Отримані результати можуть бути використані при розробці та удосконаленні методів повітряного руху. Напрямом подальших досліджень є розробка методу управління повітряним рухом з урахуванням сучасних умов динамічної повітряної обстановки.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Для початкового пошуку літератури та формування структури огляду був використаний генеративний інструмент штучного інтелекту ChatGPT 5, який допоміг систематизувати приблизно 100 джерел. Остаточний аналіз літератури та написання рукопису були виконані автором самостійно.

Список літератури

1. Державне підприємство обслуговування повітряного руху України. Річні звіти. URL : https://uksatse.ua/index.php?act=Part&CODE=376
2. Wandelt S., Antoniou C., Birolini S., Delahaye D., Dresner M., Fu X., Gössling S., Hong S.-J., Odoni A. R., Zanin M., Zhang A., Zhang H., Zhang Y., Sun X. Status quo and challenges in air transport management research. *Journal of the Air Transport Research Society*. 2024. Vol. 2. Article 10001. DOI: https://doi.org/10.1016/j.jatrs.2024.100014
3. Global Air Navigation Plan 2016-2030. Montreal : ICAO, 2016. 142 p. URL: https://www.icao.int/sites/default/files/global-airnavigation/9750_5ed_en.pdf
4. Журід В., Тягній В., Васін І., Яковлєв Р. Планування та виконання польотів БПЛА в умовах змінного операційного середовища. *Повітряна міць України*. 2026. № 2 (9). С. 133–144. https://doi.org/10.33099/2786-7714-2025-2-9-133-144
5. Supporting European Aviation. Disruption and crisis management. URL :https://www.eurocontrol.int/disruption-and-crisis-management
6. Supporting European Aviation. A practical approach to civil-military interoperability. URL: https://www.eurocontrol.int/article/practical-approach-civil-military-interoperability?utm_source=chatgpt.com
7. Davies L., Vagapov Yu., Grout V., et al. Review of Air Traffic Management Systems for UAV Integration into Urban Airspace. *2021 28th International Workshop on Electric Drives: Improving Reliability of Electric Drives (IWED)*. 2021. DOI: https://doi.org/10.1109/IWED52055.2021.9376343
8. Khadmiry B. Airspace Integration Strategies for Safe and Efficient UAV Operations. *The Journal of Unmanned System Technology*. 2022. Vol. 10. No. 3. URL: https://ojs.unsysdigital.com/index.php/just/article/view/1210
9. Case R. P., Hupy J. P. Methods for GIS-Driven Airspace Management: Integrating Unmanned Aircraft Systems, Advanced Air Mobility, and Crewed Aircraft in the NAS. Drones. 2026. Vol. 10(2). No. 82. DOI: https://doi.org/10.3390/drones10020082
10. Schuchardt B. I., Chan W. N., Swieringa K. A., Uijt de Haag M. Special issue on urban air mobility: research on aircraft, infrastructure, operation, and public acceptance. *CEAS Aeronautical Journal*. 2025. Vol. 16. DOI: https://doi.org/10.1007/s13272-025-00875-6.
11. He Z., Wang Z., Li L. Urban Air Mobility: A Review of Recent Advances in Communication, Management, and Sustainability. *Electrical Engineering and Systems Science. Systems and Control*. DOI: https://doi.org/10.48550/arXiv.2510.18235
12. ICAO (International Civil Aviation Organization). URL : https://www.icao.int/safety/UA/Pages/UTM-Guidance.aspx
13. Supporting European Aviation. New integrated ATM/ U-space services and capabilities will meet the airspace integration challenges of Urban Air Mobility. URL: https://www.eurocontrol.int/article/
14. Sesar Joint Undertaking. A new concept of operation to integrate drones with traditional aviation. URL : https://www.sesarju.eu/sesar-solutions/development-integrated-u-space-atm-conops
15. Aposporis P. A review of global and regional frameworks for the integration of UAS into air traffic management. *Transportation Research Interdisciplinary Perspectives*. 2024. Vol. 24. DOI: https://doi.org/10.1016/j.trip.2024.101064.
16. Balázs B., Vicsek T., Somorjai G. et al. Decentralized traffic management of autonomous drones. *Swarm Intell*. 2025. Vol. 19. P. 29–53. DOI: https://doi.org/10.1007/s11721-024-00241-y.
17. Xu Q., Pang Y., Liu Y. Dynamic airspace sectorization with machine learning enhanced workload prediction and clustering. *Journal of Air Transport Management*. 2024. Vol. 121 (1):102683. DOI: https://doi.org/10.1016/j.jairtraman.2024.102683
18. Supporting European Aviation. Digitalisation and AI in air traffic control: balancing innovation with the human element. URL : https://www.eurocontrol.int/article/
19. NOMMON. Research Projects. BigData4ATM. Aviation. URL :https://www.nommon.es/research-projects/bigdata4atm/?utm_source=chatgpt.com
20. Cook A. European Air Traffic Management: Principles, Practice and Research. 2016. 279 p. ISBN-10: 1138255769. URL: https://www.routledge.com/European-Air-Traffic-Management-Principles-Practice-and-Research/Cook/p/book/9781138255760
21. Arblaster M. Air Traffic Management: Economics, Regulation and Governance. 2018. 286 p. URL: https://www.amazon.com/Air-Traffic-Management-Regulation-Governance/dp/0128111186
22. Gardi A., Marino M., Ramasamy S., Sabatini R., Kistan T. 4-Dimensional Trajectory Optimization Algorithm for Air Traffic Management Systems. Proceedings of the IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). Sacramento, CA, USA, 2016. DOI: https://doi.org/10.1109/DASC.2016.7778083.
23. Gardi A., Sabatini R., Kistan T. Multi-Objective 4D Trajectory Optimisation for Integrated Avionics and Air Traffic Management Systems. Proc. of the IEEE Trans. on A&ES, 2018. DOI: https://doi.org/10.1109/TAES.2018.2849238

24. Kistan T., Gardi A., Sabatini R., Ramasamy S., Batuwangala E. An evolutionary outlook of air traffic flow management techniques. *Progress in Aerospace Sciences*. 2017. Vol. 88. P. 15–42. DOI: https://doi.org/10.1016/j.paerosci.2016.10.001.
25. Aditya V., Aswin D. S., Dhaneesh S. V., Chakravarthy S., Kumar B. S., Venkadavarahan M. A review on air traffic flow management optimization: trends, challenges, and future directions. *Discover Sustainability*. 2024. Vol. 5. № 1. DOI: https://doi.org/10.1007/s43621-024-00781-7.
26. Capitán C., Pérez-León H., Capitán J., Castano A., Ollero A. Unmanned aerial traffic management system architecture for U-space in-flight services. *Applied Sciences*. 2021. Vol. 11. № 9. Article 3995. DOI: https://doi.org/10.3390/app11093995.
27. AirHub Knowledge Series: Understanding U-space and UTM for Drone Operators. URL : https://www.airhub.app/resources/news/understanding-u-space-and-utm-for-drone-operators
28. Rui G., Peng C. Dynamic air route open-close problem for airspace management. *Tsinghua Science and Technology*. 2007. Vol. 12. No. 6. P. 647–651. DOI: https://doi.org/10.1016/S1007-0214(07)70169-4.
29. Kulkarni S., Ganesan R., Sherry L. Dynamic airspace configuration using approximate dynamic programming: intelligence-based paradigm. *Transportation Research Record: Journal of the Transportation Research Board*. 2012. Vol. 2266. No. 1. P. 35–42. DOI: https://doi.org/10.3141/2266-04
30. Li L. A review of data science and artificial intelligence applications in air transportation systems. *Artificial Intelligence for Transportation*. 2025. Vol. 2. Art. 100023. DOI: https://doi.org/10.1016/j.ait.2025.100023.
31. Schier-Morgenthal S., Abdellaoui R., Metz I. C. Introducing digital air-traffic controllers for urban-air mobility to ensure safe and energy-efficient flight operations. *CEAS Aeronautical Journal*. 2015. Vol. 16. P. 339–352. DOI: https://doi.org/10.1007/s13272-024-00802-1

Received (Надійшла) 29.12.2025
Accepted for publication (Прийнята до друку) 01.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Баістов Юрій Вікторович – викладач кафедри інженерно-авіаційного забезпечення факультету авіаційного транспорту інституту цивільної авіації Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Yurii Baistov – Lecturer of the Department of Engineering and Aviation Support, Faculty of Air Transport, Institute of Civil Aviation, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: Yurii_Baistov@gmail.com; ORCID Author ID: https://orcid.org/0009-0007-2356-6532.

Сальник Олег Вікторович – старший науковий співробітник науково-дослідної лабораторії факультету професійної військової та післядипломної освіти Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Oleh Salnyk – Senior Researcher, Scientific Research Department of Faculty of Professional Military and Postgraduate Education of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: Oleh_Salnyk@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-2688-1198.

Дроль Олександр Юрійович – старший викладач кафедри тактики та загальновійськових дисциплін Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Oleksandr Drol – Senior Instructor of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: Oleksandr_Drol@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-5472-208X.

Мельник Сергій Вікторович – викладач кафедри бойового застосування радіотехнічного озброєння факультету радіотехнічних військ ППО Харківського національного університету Повітряних Сил імені І. Кожедуба, Харків, Україна;
Serhii Melnyk – Lecturer at the Department of Combat Application of Radio Engineering Weapons, Faculty of Radio Engineering Troops of Air Defense of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: Serhii_Melnyk@gmail.com; ORCID Author ID: https://orcid.org/0009-0006-9107-3169.

Грушенкова Людмила Володимирівна – старший науковий співробітник НДІ воєнної розвідки, Київ, Україна;
Liudmyla Hrushenkova – Senior Research Scientist of Defence Intelligence Research Institute, Kyiv, Ukraine;
e-mail: Liudmyla_Hrushenkova@gmail.com; ORCID Author ID: ORCID ID: https://orcid.org/0009-0005-4339-1376.

Analysis of existing air traffic management methods in a dynamic air situation

Yurii Baistov, Oleh Salnyk, Oleksandr Drol, Liudmyla Hrushenkova

Abstract. The subject of the article is existing methods of air traffic management. The **aim** is to analyze existing air traffic management methods in dynamic air traffic situations. **Tasks**: to analyze existing approaches to air traffic management; to identify groups of air traffic management methods based on a similar principle of action; to conduct a comparative analysis of modern air traffic management methods; to evaluate air traffic management methods based on key criteria. The **methods** used include analytical and empirical methods of comparative research. The **following results were obtained**. A comprehensive analysis of modern air traffic management methods in dynamic air conditions has been carried out. Six main groups of air traffic management methods have been identified and considered. Their basic principles of operation, their main advantages and disadvantages, and the level of efficiency of the methods in dynamic air conditions have been established. The identified groups of air traffic management methods have been assessed according to such key criteria as adaptability to changes in the air situation, efficiency, scalability, and the level of their automation. It has been proven that further development of air traffic management systems should be aimed at increasing the level of automation, ensuring scalability, speed of decision-making, and effective integration of a large number of air objects in the airspace. **Conclusions.** The results obtained can be used to develop and improve air traffic methods. The direction of further research is the development of an air traffic management method that takes into account modern conditions in the dynamic air environment.

Keywords: aviation technology, air object, air traffic management, air traffic management method, dynamic air situation.

УДК 004.62:629.3 **doi: 10.26906/SUNZ.2026.2.020**

І. І. Скляров[1], М.О. Геревич[2]

[1] Харківський національний автомобільно-дорожній університет, Харків, Україна
[2] Ужгородський національний університет, Ужгород, Україна

МЕТОД ПРОГНОЗУВАННЯ ТЕХНІЧНОГО СТАНУ ТРАНСПОРТНИХ ЗАСОБІВ ІЗ ВИКОРИСТАННЯМ ТЕХНОЛОГІЙ ВЕЛИКИХ ДАНИХ

Анотація. Актуальність. Актуальність роботи зумовлена необхідністю підвищення ефективності сервісних технологій в автомобільній галузі в умовах зростання обсягів телематичних, діагностичних та експлуатаційних даних, що супроводжують функціонування сучасних транспортних засобів. Для сучасних автомобілів характерні складна структура технічних систем, наявність великої кількості взаємопов'язаних компонентів, різнорідність режимів експлуатації та підвищений ризик виникнення відмов, що ускладнює своєчасне оцінювання їх технічного стану. Існуючі підходи до прогнозування технічного стану автомобілів здебільшого не забезпечують комплексного врахування потокових сенсорних даних, історії технічного обслуговування та експлуатаційних параметрів, що ускладнює обґрунтоване прийняття сервісних рішень. Тому актуальною є розробка методу прогнозування технічного стану автомобілів на основі Big Data для підтримки процесу своєчасного виявлення ризикових станів і оптимізації технічного обслуговування. **Метою даної роботи** є розробка методу прогнозування технічного стану автомобілів на основі технологій Big Data шляхом інтеграції потокових сенсорних даних, історії технічного обслуговування та експлуатаційних параметрів для своєчасного виявлення ймовірних відмов, підвищення точності оцінювання технічного стану транспортних засобів і оптимізації сервісних рішень. **Об'єктом дослідження** процес прогнозування технічного стану автомобілів у системах сервісного обслуговування на основі аналізу великих обсягів різнорідних даних. **Предметом дослідження** методи, моделі та засоби прогнозування технічного стану автомобілів на основі Big Data шляхом інтеграції сенсорних, експлуатаційних і сервісних даних. **Результати.** У роботі розглянуто задачу прогнозування технічного стану автомобілів на основі Big Data з метою підвищення ефективності сервісних технологій в автомобільній галузі. Запропоновано метод, що передбачає інтеграцію експлуатаційних, сенсорних і сервісних даних у межах єдиного аналітичного контуру для оцінювання ризику виникнення відмов. Реалізацію методу виконано мовою Python у середовищі Google Colab із використанням відкритого набору даних SCANIA Component X. У процесі дослідження виконано підготовку даних, усунення витоку цільової змінної, побудову прогнозної моделі, налаштування порога класифікації та формування сервісних рішень. Отримані результати підтвердили високу якість прогнозування та придатність запропонованого підходу до використання в задачах підтримки технічного обслуговування транспортних засобів.

Ключові слова: Big Data; прогнозування технічного стану; автомобільна галузь; машинне навчання; аналіз даних; технічне обслуговування; транспортні засоби; класифікація; бінарний класифікатор; Google Colab; Python.

Вступ

Сучасний етап розвитку автомобільної галузі характеризується глибокою цифровою трансформацією, що охоплює не лише процеси проєктування та виробництва транспортних засобів, а й системи їх технічного супроводу, сервісного обслуговування та підтримки після продажу. Умови зростання складності автомобільних систем, поширення електронних блоків керування, сенсорних мереж, телематичних платформ та підключених сервісів зумовлюють безперервне накопичення значних обсягів різнорідних даних про функціонування транспортних засобів. Такі дані формуються в процесі експлуатації автомобілів, технічного обслуговування, діагностики окремих вузлів і агрегатів, взаємодії з сервісною інфраструктурою та користувачами. У цьому контексті технології Big Data набувають особливого значення як інструмент інтеграції, зберігання, оброблення та інтелектуального аналізу великих масивів інформації з метою підвищення ефективності сервісних технологій в автомобільній галузі.

Традиційні підходи до оцінювання технічного стану автомобілів ґрунтуються переважно на регламентованих інтервалах технічного обслуговування, результатах періодичної діагностики або експертних оцінках фахівців. За таких умов прийняття рішень щодо обслуговування часто має реактивний характер, тобто виконується після виявлення несправності або у межах заздалегідь встановлених часових чи інтервалів пробігу. Однак індивідуальні режими експлуатації транспортних засобів, відмінності в дорожніх, кліматичних і навантажувальних умовах, а також варіативність поведінки водія суттєво впливають на темпи зношування компонентів і ризики виникнення відмов. Унаслідок цього регламентні підходи не завжди забезпечують належну точність прогнозування технічного стану, що призводить або до передчасного виконання сервісних робіт, або до запізнілого виявлення критичних змін у функціонуванні автомобіля.

Одним із перспективних напрямів розв'язання цієї проблеми є перехід до моделей обслуговування з використанням прогнозування, заснованих на аналізі фактичних експлуатаційних даних. Використання Big Data у поєднанні з методами інтелектуального аналізу даних і машинного навчання відкриває можливість виявлення прихованих закономірностей у поведінці технічних систем, ідентифікації процесів деградації та формування обґрунтованих прогнозів щодо майбутнього стану автомобіля. Особливої актуальності це набуває у сфері сервісних технологій, де своєчасне виявлення ознак потенційної відмови дозволяє оптимізувати графіки технічного обслуговування, знижувати експлуатаційні витрати, підвищувати надійність транспортних засобів і покращувати якість сервісу для кінцевого споживача.

Актуальність дослідження посилюється також тим, що на практиці доступ до повномасштабних реальних наборів даних автомобільної галузі часто є обмеженим через комерційну конфіденційність, неоднорідність форматів зберігання, неповноту спостережень і складність отримання тривалих часових рядів для великої кількості транспортних засобів. У зв'язку з цим обґрунтованим підходом є використання синтетичних датасетів, які відтворюють структурні, статистичні та причинно-логічні властивості реальних даних і дають змогу проводити відтворювані експериментальні дослідження. Формування такого набору даних у контексті Big Data дозволяє змоделювати значний обсяг записів, різнорідність джерел інформації, наявність шумів, пропусків, дисбалансу класів і залежностей між експлуатаційними режимами та технічними відмовами.

Метою роботи є розробка методу прогнозування технічного стану автомобілів на основі технологій Big Data шляхом інтеграції потокових сенсорних даних, історії технічного обслуговування та експлуатаційних параметрів для своєчасного виявлення ймовірних відмов, підвищення точності оцінювання технічного стану транспортних засобів і оптимізації сервісних рішень.

Основна частина

Для розв'язання задачі прогнозування технічного стану автомобілів у межах розвитку сервісних технологій автомобільної галузі необхідно врахувати сучасні підходи до аналізу телематичних, діагностичних та експлуатаційних даних, визначити їхні функціональні можливості, переваги й обмеження, а також обґрунтувати доцільність розроблення власного методу на основі технологій Big Data. У зв'язку з цим доцільним є аналіз наукових досліджень, присвячених використанню методів машинного навчання, інтелектуального аналізу даних і прогнозного обслуговування у задачах оцінювання технічного стану транспортних засобів.

У статті [1] розглянуто актуальну науково-прикладну задачу прогнозування технічного стану автомобілів на основі технологій Big Data з метою підвищення ефективності сервісних технологій в автомобільній галузі. У роботі запропоновано метод прогнозування технічного стану автомобілів, який базується на інтеграції потокових даних із сенсорів, історії технічного обслуговування та експлуатаційних параметрів у єдиному аналітичному контурі. У роботі проаналізовано наукові публікації з позицій прикладних сценаріїв і типів методів машинного навчання, а також обговорюються відкриті проблеми, серед яких обмеженість доступних даних, залежність більшості підходів від розмічених вибірок, потреба в поєднанні кількох джерел даних і необхідність підвищення інтерпретації моделей.

У статті [2] автори розглядають задачу прогнозного технічного обслуговування в автомобільній галузі на основі даних бортових сенсорів і пропонують конвеєр PREPIPE для прогнозування стану засмічення кисневого датчика дизельного двигуна. У роботі використано часові ряди сигналів, зібраних з бортового блоку керування двигуном, а сам підхід охоплює вибір інформативних сигналів, перетворення часових рядів у ознаки, відбір найважливіших характеристик, урахування історичних ознак і порівняння різних класифікаційних моделей, зокрема з архітектурами глибокого навчання. Автори показують, що якісна попередня обробка та інженерія ознак є критично важливими для досягнення високої точності прогнозування. Водночас дослідження зосереджене на окремому вузлі автомобіля, хоча було б логічним використовувати ширший підхід.

У роботі [3] автори розглядають сучасний стан застосування методів машинного навчання для прогнозного технічного обслуговування автомобілів і діагностики технічного стану транспортних засобів. У роботі підкреслено, що профілактика є важливим напрямом розвитку автомобільної галузі, оскільки дає змогу підвищити якість сервісу, своєчасно виявляти ризики відмов і зменшувати ймовірність критичних несправностей. Водночас наголошено, що в автомобільному секторі існують суттєві обмеження, пов'язані з недостатньою доступністю сенсорів, складністю отримання необхідних даних і обмеженістю окремих проєктних рішень. У зв'язку з цим методи машинного навчання розглядаються як перспективний інструмент для аналізу навіть обмежених за обсягом даних і побудови моделей оцінювання технічного стану автомобілів.

У роботі [4] автори узагальнюють сучасні напрями застосування штучного інтелекту, машинного та глибокого навчання в транспортних системах. У роботі показано, що машинне навчання використовується для керування трафіком, автономного транспорту, інтелектуального паркування, оптимізації громадського транспорту, логістики, безпеки та моніторингу інфраструктури. Автори наголошують, що ці технології забезпечують прийняття рішень, дають змогу обробляти великі обсяги даних у реальному часі, прогнозувати тенденції та підвищувати стійкість транспортних систем. Водночас у статті окреслено ключові обмеження таких рішень, зокрема дефіцит і неоднорідність даних, недостатню узагальнюваність моделей, високі обчислювальні витрати, а також етичні й регуляторні проблеми, пов'язані з прозорістю, упередженістю та конфіденційністю даних.

У статті [5] автори здійснюють систематизований огляд застосування штучного інтелекту для діагностики несправностей транспортних засобів і прогнозного технічного обслуговування. У роботі підкреслено, що традиційні підходи, засновані на правилах і даних бортової діагностики, мають певні обмеження, оскільки переважно орієнтовані на виявлення вже наявних відмов, тоді як сучасні автомобілі потребують більш гнучких і прогностично орієнтованих засобів аналізу. Автори показують, що методи штучного інтелекту, зокрема машинного та глибокого навчання, дають змогу аналізувати дані з різних сенсорів, виявляти аномалії, прогнозувати деградацію технічного стану та підвищувати точність діагностики різних підсистем автомобіля, серед яких двигун, підвіска, шини, трансмісія та гальмівна система. Особливу увагу приділено підходам, заснованим на знаннях, аналізу даних з багатьох сенсорів, а також поєднанню різних

інтелектуальних методів для досягнення більш надійної й універсальної діагностичної підтримки. Водночас у роботі наголошено на проблемах гетерогенності даних, складності узагальнення моделей, залежності від якості вхідної інформації та потребі в комплексних рішеннях для діагностики транспортного засобу в цілому, а не лише його окремих вузлів.

У статті [6] автори розглядають проблему формалізації процесу розробки моделей профілактичного технічного обслуговування в умовах великих обсягів даних, міждисциплінарної взаємодії та складної організації робіт. У роботі зазначено, що зростання доступності даних створює нові можливості для профілактичного обслуговування, однак сам процес побудови таких моделей залишається складним, трудомістким і потребує координації між фахівцями предметної області, аналітиками даних, IT-фахівцями та розробниками програмного забезпечення. Для подолання цих труднощів автори пропонують еталонну модель розробки рішень профілактичного обслуговування, побудовану на основі CRISP-DM і структуровану за чотирма фазами, що охоплюють повний життєвий цикл моделі: від опису компонента і сценарію використання до розгортання, моніторингу та повторного використання результатів. Особливу увагу приділено архітектурі роботи з даними, яка включає збір сенсорної інформації, попередню обробку, псевдонімізацію, централізоване зберігання в сховище даних, підготовку даних, моделювання, оцінювання та виконання моделі в хмарному або бортовому середовищі. Водночас стаття має переважно процесно-організаційний характер і зосереджується на еталонній схемі розробки моделей профілактичного обслуговування, а не на конкретному методі прогнозування технічного стану автомобіля.

У статті [7] автори здійснюють огляд підходів до моніторингу технічного стану транспортних засобів і діагностики несправностей для традиційних, електричних та автономних автомобілів. У роботі підкреслено, що автомобільна галузь досі не має уніфікованої структури, яка б поєднувала різні методи діагностики з урахуванням масштабованості, адаптивності та ефективності в різних типах транспортних засобів. Автори систематизують сучасні підходи обробки даних, акцентують увагу на ролі сучасних сенсорів, IoT, штучного інтелекту, багатосенсорної інтеграції та аналітики великих даних, а також пропонують концептуальну модель для раннього виявлення несправностей. У статті окремо наголошено, що методи штучного інтелекту та Big Data стають ключовими засобами для аналізу великих потоків сенсорних даних, виявлення аномалій і зниження ризику раптових відмов.

Варто зазначити, що в контексті обробки даних сенсорами велике значення мають інтелектуальні методи енергозбереження в сенсорних мережах. І також є доцільним розглянути дослідження за цим напрямком. В роботі [8] розглянуто застосування штучних нейронних мереж для підвищення енергоефективності безпровідних сенсорних мереж. Автор підкреслює, що обмеженість енергетичних ресурсів є однією з головних проблем таких мереж, а тому актуальним є використання інтелектуальних інструментів для зниження витрат на передавання даних, кластеризації вузлів, прогнозування сенсорної інформації та оптимізації маршрутизації. У роботі показано, що нейронні мережі добре узгоджуються з природою сенсорних мереж завдяки можливості паралельної обробки, розподіленого зберігання даних, автоматизованої класифікації та роботи з шумними даними. Хоча стаття присвячена безпровідним сенсорним мережам, а не безпосередньо автомобільній галузі, її результати добре підтримують ідею використання сенсорних потоків, інтелектуальної обробки та прогнозних моделей у задачах оцінювання технічного стану автомобілів. Водночас робота має ширший інфраструктурний характер і не розглядає інтеграцію сервісної історії, експлуатаційних параметрів та автомобільних телематичних даних у єдиному контурі прогнозування.

Робота [9] присвячена підвищенню енергоефективності безпровідних сенсорних мереж в умовах обмежених ресурсів вузлів і необхідності тривалого автономного функціонування. У роботі підкреслено, що енергоспоживання таких мереж залежить не лише від апаратної конфігурації, а й від способу організації обміну даними, топології мережі, частоти передавання та впливу зовнішнього середовища. Хоча робота присвячена безпровідним сенсорним мережам, а не безпосередньо автомобільній галузі, вона підтверджує загальний принцип, за яким складні системи з великою кількістю сенсорів потребують інтелектуального аналізу даних, адаптивного керування та прогнозних механізмів для підвищення ефективності функціонування.

У статті [10] подано огляд сучасних підходів до діагностики несправностей і моніторингу технічного стану складних інженерних систем. Автори наголошують, що зі зростанням складності технічних об’єктів підвищуються вимоги до надійності, безпеки та своєчасного виявлення процесів деградації, а тому методи діагностики несправностей та моніторингу на основі даних набувають критичного значення для раннього виявлення дефектів, прогнозування залишкового ресурсу та планування технічного обслуговування. У роботі узагальнено результати шістнадцяти досліджень, присвячених застосуванню штучного інтелекту, глибокого навчання, цифрових двійників, аналізу вібраційних і струмових сигналів, моніторингу даних з сенсорів та методів роботи з малими й незбалансованими наборами даних. Окремо підкреслено, що сучасні рішення орієнтуються на раннє виявлення несправностей, підвищення точності діагностики, зменшення простоїв і підтримку прогнозного технічного обслуговування в умовах обмежених або шумних даних. Водночас стаття має широкий міжгалузевий характер і не зосереджується безпосередньо на автомобільній сервісній інфраструктурі чи інтеграції телематичних, сервісних та експлуатаційних даних автомобіля.

Проведений аналіз розглянутих праць засвідчив, що сучасні дослідження підтверджують високу ефективність підходів, заснованих на Big Data, машинному навчанні та інтелектуальному аналізі сенсорних даних, для задач діагностики несправностей і прогнозування технічного стану транспортних засобів. Водночас більшість наявних робіт або зосереджена на окремих вузлах автомобіля, або має оглядовий характер, не забезпечуючи цілісної інтеграції

потокових сенсорних даних, історії технічного обслуговування та експлуатаційних параметрів у межах єдиного методу. Це дає підстави стверджувати, що розроблення власного методу прогнозування технічного стану автомобілів на основі Big Data є обґрунтованим і актуальним.

Нехай множина автомобілів, для яких здійснюється моніторинг технічного стану, задається як

$$V = \{v_1, v_2, \ldots, v_N\}, \quad (1)$$

де N – кількість автомобілів у вибірці. Для кожного автомобіля v_i у процесі експлуатації формується множина різнорідних даних, яка включає потокові сенсорні дані, історію технічного обслуговування та експлуатаційні параметри. Тоді інтегрований опис автомобіля можна подати у вигляді

$$D_i = \{S_i, H_i, E_i\}, \quad (2)$$

де S_i – потокові сенсорні дані, H_i – історія технічного обслуговування, E_i – експлуатаційні параметри. Потокові сенсорні дані для автомобіля v_i задаються часовим рядом

$$S_i(t) = \left(s_{i1}(t), s_{i2}(t), \ldots, s_{im}(t)\right), \quad (3)$$

де m – кількість сенсорних параметрів, $s_{ij}(t)$ – значення j-го сенсорного параметра в момент часу *t*. До таких параметрів можуть належати температура двигуна, тиск мастила, рівень вібрації, оберти двигуна, напруга акумулятора, швидкість, витрата пального та інші показники. Історія технічного обслуговування подається вектором

$$H_i = \left(h_{i1}, h_{i2}, \ldots, h_{ip}\right), \quad (4)$$

де p – кількість ознак сервісної історії, серед яких можуть бути кількість попередніх ремонтів, тривалість інтервалу між сервісним обслуговуванням, кількість зафіксованих несправностей, дата останнього технічного обслуговування, типи виконаних сервісних робіт. Експлуатаційні параметри задаються як

$$E_i = \left(e_{i1}, e_{i2}, \ldots, e_{iq}\right), \quad (5)$$

де q – кількість параметрів експлуатації, зокрема стиль водіння, тип дорожнього покриття, інтенсивність використання, кліматичні умови, середнє навантаження, режим руху. Оскільки сенсорні дані мають потоковий характер, для прогнозування технічного стану використовується механізм ковзного часового вікна. Для автомобіля v_i на інтервалі часу $[t-\Delta t, t]$ формується підмножина даних

$$W_i(t) = \{S_i(\tau) \mid \tau \in [t-\Delta t, t]\}, \quad (6)$$

де Δt – довжина часового вікна. На основі часового вікна для кожного сенсорного параметра обчислюються статистичні та динамічні ознаки. Для j параметра вони можуть бути визначені таким чином:

- середнє значення

$$\mu_{ij}(t) = \sum_{k=1}^{K} s_{ij}\left(t_k\right) \Big/ K, \quad (7)$$

- стандартне відхилення

$$\sigma_{ij}(t) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(s_{ij}(t_k) - \mu_{ij}(t)\right)^2}, \quad (8)$$

- лінійний тренд зміни параметра

$$\beta_{ij}(t) = \sum_{k=1}^{K}(t_k - \bar{t})\left(s_{ij}(t_k) - \bar{s}_{ij}\right) \Big/ \sum_{k=1}^{K}(t_k - \bar{t})^2, \quad (9)$$

де K – кількість спостережень у часовому вікні, $\bar{t}$ – середнє значення часу у вікні, $\bar{s}_{ij}$ – середнє значення *j*-го параметра у вікні.

Для виявлення аномальних режимів роботи визначається частота перевищення допустимого порогу:

$$a_{ij}(t) = \frac{1}{K}\sum_{k=1}^{K} I\left(s_{ij}(t_k) > s_j^{crit}\right), \quad (10)$$

де $I()$ – індикаторна функція, s_j^{crit} – критичне значення *j*-го параметра. Після цього формується узагальнений вектор ознак технічного стану автомобіля:

$$X_i(t) = \left[F_i^{(S)}(t), H_i, E_i\right] \quad (11)$$

де $F_i^{(S)}(t)$ – вектор ознак, сформований із сенсорних часових рядів у вікні $[t-\Delta t, t]$. Отже, $X_i(t)$ є інтегрованим описом технічного стану автомобіля в момент часу t. Ціль методу полягає у прогнозуванні ймовірності виникнення відмови або потреби в технічному обслуговуванні у майбутньому часовому горизонті $[t, t+\Delta p]$, де Δp – горизонт прогнозування. Тоді цільова змінна задається як

$$y_i(t) = \begin{cases} 1, & \text{в } [t, t+\Delta p] \text{ відмова або потрібне ТО}, \\ 0, & \text{інакше}. \end{cases} \quad (12)$$

Задача прогнозування формалізується як задача бінарної класифікації, у якій необхідно побудувати відображення

$$f : X_i(t) \to \hat{y}_i(t) \quad (13)$$

де $\hat{y}_i(t)$ – прогнозований клас технічного стану. Для оцінювання ризику відмови використовується ймовірнісна модель

$$P_i(t) = P\left(y_i(t) = 1 \mid X_i(t)\right), \quad (14)$$

У найпростішому випадку така модель може бути реалізована логістичною функцією:

$$P_i(t) = \frac{1}{1 + \exp\left(-\left(w_0 + \sum_{r=1}^{M} w_r x_{ir}(t)\right)\right)}, \quad (15)$$

де M – розмірність вектору ознак, $x_{ir}(t)$ – r-та ознака для автомобіля, v_i, w_r – вагові коефіцієнти моделі. Правило прийняття рішення щодо технічного стану має такий вигляд:

$$\hat{y}_i(t) = \begin{cases} 1, & \text{якщо } P_i(t) \geq \theta, \\ 0, & \text{якщо } P_i(t) < \theta. \end{cases} \quad (16)$$

де θ – порогове значення, що визначається експериментально.

Для підвищення чутливості методу до процесів деградації вводиться інтегральний індекс технічного ризику, який враховує вплив сенсорних, сервісних та експлуатаційних факторів:

$$R_i(t) = \alpha R_i^{(S)}(t) + \beta R_i^{(H)} + \gamma R_i^{(E)}, \quad (17)$$

де $R_i^{(S)}(t)$ – ризик, оцінений за сенсорними даними, $R_i^{(H)}$ – ризик на основі історії технічного обслуговування, $R_i^{(E)}$ – ризик, зумовлений умовами експлуатації, α, β, γ – вагові коефіцієнти, для яких виконується така умова:

$$\alpha + \beta + \gamma = 1. \quad (18)$$

Сенсорна складова ризику може бути визначена як зважена сума нормованих ознак:

$$R_i^{(S)}(t) = \sum_{j=1}^{m} \lambda_j z_{ij}(t), \quad (19)$$

де $z_{ij}(t)$ – нормоване значення j-ї сенсорної ознаки, λ_j – коефіцієнт важливості ознаки, $\sum_{j=1}^{m} \lambda_j = 1$. Сервісна складова ризику може бути задана, наприклад, через кількість попередніх відмов і тривалість після останнього технічного обслуговування:

$$R_i^{(H)} = \eta_1 \frac{n_i^{fail}}{n_{\max}{}^{fail}} + \eta_2 \frac{d_i^{serv}}{d_{\max}{}^{serv}}, \quad (20)$$

де n_i^{fail} – кількість попередніх відмов, d_i^{serv} – пробіг або час після останнього технічного обслуговування, $\eta_1 + \eta_2 = 1$, а експлуатаційна складова ризику визначається як

$$R_i^{(E)} = \sum_{l=1}^{q} \rho_l e_{il}^*, \quad (21)$$

де e_{il}^* – нормовані експлуатаційні параметри, ρ_l – їх вагові коефіцієнти, $\sum_{l=1}^{q} \rho_l = 1$. Тоді кінцеве правило сервісного рішення можна подати як

$$u_i(t) = \begin{cases} u_1, & \text{якщо } R_i(t) < \delta_1, \\ u_2, & \text{якщо } \delta_1 \le R_i(t) < \delta_2, \\ u_3, & \text{якщо } R_i(t) \ge \delta_2, \end{cases} \quad (22)$$

де u_1 – продовження експлуатації без втручання, u_2 – призначення додаткової діагностики, u_3 – рекомендація щодо термінового технічного обслуговування; δ_1, δ_2 – пороги прийняття рішення.

Отже, запропонований метод включає послідовність етапів: збір та інтеграцію різнорідних даних, формування часових вікон, обчислення статистичних і динамічних ознак, побудову прогнозної моделі для визначення ймовірності відмови, обчислення інтегрального індексу ризику та формування сервісного рішення.

На відміну від підходів, що спираються лише на окремі сенсорні сигнали або на вузькоспеціалізовані дані, запропонований метод враховує багатофакторну природу технічного стану автомобіля та дає змогу підвищити обґрунтованість прогнозування в умовах Big Data.

Для перевірки працездатності запропонованого методу використано середовище Google Colab і мову Python. Реалізацію виконано із застосуванням бібліотек NumPy, Pandas, Scikit-learn та Matplotlib, що забезпечило повний цикл оброблення даних, формування ознак, навчання моделі та оцінювання результатів.

Для експериментальної перевірки запропонованого методу було обрано відкритий набір даних SCANIA Component X Dataset [11], оскільки він найбільшою мірою відповідає предметній області дослідження та містить реальні багатовимірні дані про технічний стан транспортних засобів, їх експлуатаційні характеристики та сервісно-орієнтовані мітки.

Для формування робочого масиву даних було використано шість файлів набору:

– test_operational_readouts.csv,
– test_specifications.csv,
– train_tte.csv,
– validation_operational_readouts.csv,
– validation_specifications.csv,
– validation_labels.csv.

На першому етапі було проаналізовано структуру таблиць і склад їхніх атрибутів, після чого виконано поєднання файлів за спільними ідентифікаційними ознаками. Операційні дані, що містили показники функціонування компонентів у часовій динаміці, були інтегровані зі параметрами специфікацій транспортних засобів, а також із цільовими мітками, що характеризують технічний стан. У результаті злиття було сформовано єдиний узгоджений датасет, придатний для подальшого машинного аналізу. Після об'єднання даних було виконано їх первинну перевірку, в межах якої оцінено розмір отриманої вибірки, структуру ознак та наявність пропущених значень. Сформований масив виявився неоднорідним за повнотою заповнення, що відповідає природі реальних технічних даних. Подальша підготовка полягала у виділенні рядків, для яких наявна цільова мітка, а також у приведенні задачі до бінарної постановки: нормальний технічний стан було позначено як клас 0, а наявність ризику відмови – як клас 1. Такий підхід дозволив узгодити експеримент із розробленим методом прогнозування технічного стану в термінах своєчасного виявлення небезпечних станів. Для забезпечення коректності обчислювального експерименту зі складу ознак було вилучено змінні, що могли прямо або опосередковано містити інформацію про цільовий клас і призводити до витоку цільової змінної.

Окремо було виконано фільтрацію ознак із надмірною часткою пропусків, після чого набір було поділено на числові та категоріальні параметри. Це створило основу для подальшої побудови прогнозної моделі та дало змогу перейти до етапу оцінювання якості методу на очищених і структурованих даних.

Основні результати експерименту свідчать, що після усунення витоку цільової змінної та оптимізації порога класифікації запропонований метод забезпечив високу якість прогнозування технічного стану автомобілів. Матриця помилок на рис. 1 показує, що модель правильно ідентифікувала 38202 випадки нормального технічного стану та 925 випадків ризику відмови.

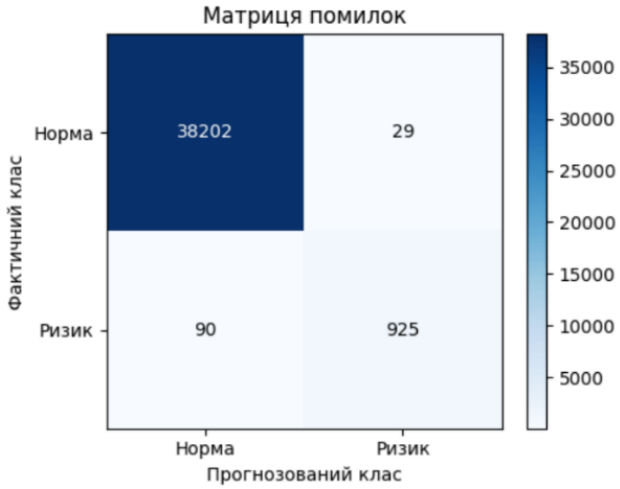


Рис. 1. Матриця помилок

Водночас зафіксовано 29 хибнопозитивних рішень, коли справний стан було віднесено до ризикового, і 90 хибнонегативних рішень, коли ризиковий стан залишився невиявленим. Такі результати дають підстави стверджувати, що модель характеризується дуже високою точністю розпізнавання нормального стану та достатньо високою здатністю виявляти ризикові режими. Особливо важливим є те, що кількість хибнопозитивних рішень залишається незначною, що знижує ймовірність необґрунтованого сервісного втручання.

Отримані значення ROC-AUC = 0.9962 та AP = 0.9674 на рис. 2 та рис. 3 відповідно підтверджують високу роздільну здатність моделі та її здатність відокремлювати ризикові стани від нормальних навіть в умовах істотного дисбалансу класів.

Це свідчить про те, що запропонований метод є не лише придатним для класифікації, а й ефективним для ранжування об’єктів за рівнем ризику, що має особливе значення для задач сервісної аналітики.

Аналіз графіка залежності метрик від порога, який представлений на рис. 4 показав, що використання стандартного порога не є оптимальним для даної задачі. Підібране значення порога на рівні 0.15 забезпечило більш збалансоване співвідношення між точністю та повнотою виявлення ризикових станів.

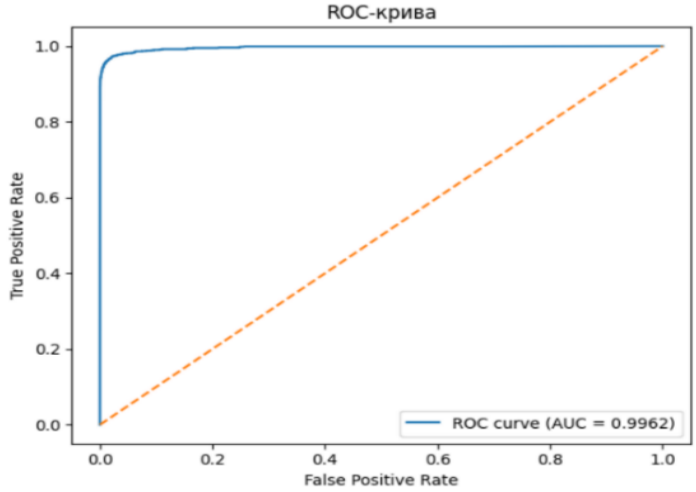


Рис. 2. ROC крива

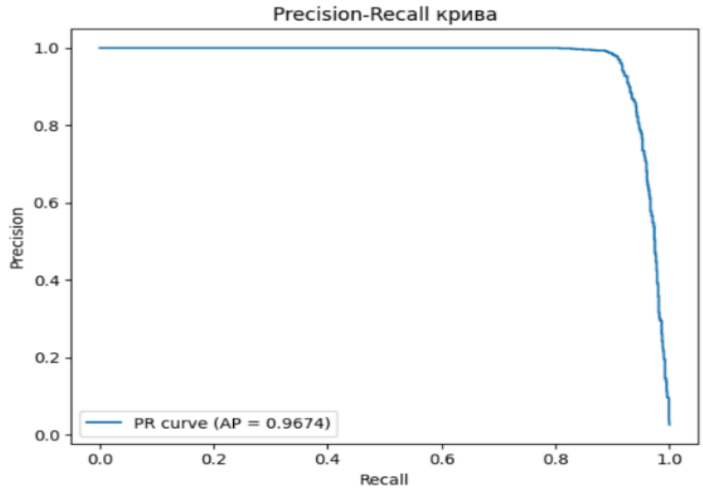


Рис. 3. Precision-Recall крива

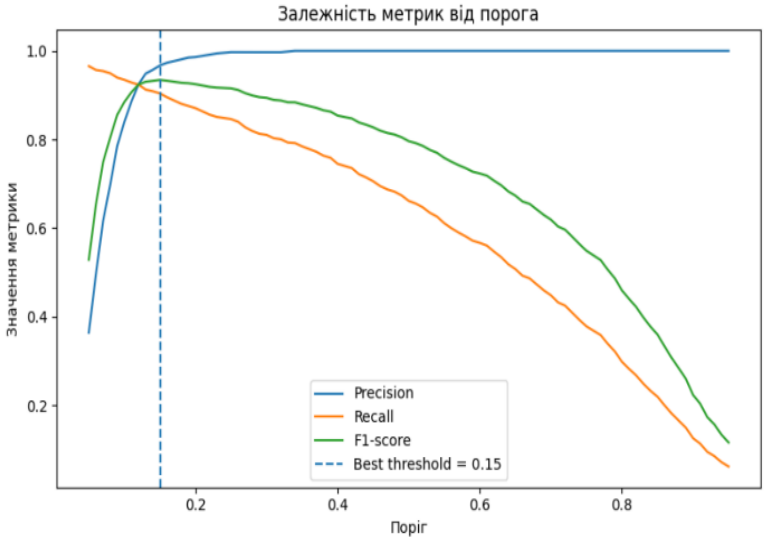


Рис. 4. Залежність метрик від порога

Саме завдяки цьому вдалося зменшити кількість пропущених відмовних випадків і підвищити практичну цінність моделі для прийняття сервісних рішень.

Додатково встановлено, що найбільший внесок у формування прогнозу мають ознаки груп 459, 158, 837, 167, 309 і 397, що представлені на рис.5, які можна розглядати як найбільш інформативні параметри технічного стану в межах досліджуваного набору даних.

Це підтверджує, що після очищення набору ознак модель спирається на реальні експлуатаційні характеристики, а не на службові або цільові змінні.

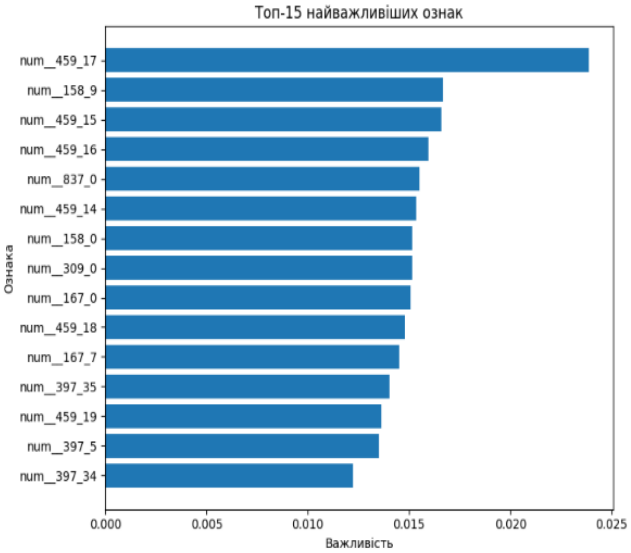


Рис. 5. Важливі ознаки

Таким чином, результати експерименту підтверджують працездатність запропонованого методу та його придатність для задач прогнозування технічного стану автомобілів на основі великих даних. Практичне значення отриманих результатів полягає у можливості своєчасного виявлення ризикових станів, зниження ймовірності раптових відмов і підвищення обґрунтованості рішень щодо технічного обслуговування.

Висновки

Розроблено метод прогнозування технічного стану автомобілів, який базується на інтеграції сенсорних, експлуатаційних та сервісних даних і дозволяє формувати оцінку ризику відмови з подальшим перетворенням її на сервісне рішення.

Метод формалізовано математично, описано його основні етапи та реалізовано засобами Python у середовищі Google Colab. В якості датасету використано SCANIA Component X, на базі якого сформовано єдиний робочий набір даних для перевірки працездатності підходу. У процесі моделювання виконано підготовку даних, усунення витоку цільової змінної, відбір інформативних ознак, побудову моделі для прогнозу та налаштування порога класифікації. Отримані результати показали високу якість прогнозування: модель забезпечила надійне відокремлення нормальних станів від ризикових, а також продемонструвала високу придатність до ранжування об’єктів за рівнем ризику. Аналіз матриці помилок, ROC- та PR-характеристик підтвердив, що запропонований метод може бути використаний для підтримки рішень щодо продовження експлуатації, додаткової діагностики або термінового технічного обслуговування.

Практичне значення роботи полягає в тому, що запропонований підхід може бути основою для побудови інтелектуальних сервісних систем в автомобільній галузі, орієнтованих на своєчасне виявлення ризикових технічних станів, зниження імовірності раптових відмов і підвищення обґрунтованості технічного обслуговування.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Andreas Theissler, Judith Pérez-Velázquez, Marcel Kettelgerdes, Gordon Elger. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety. Volume 215. 2021. 23 p. https://doi.org/10.1016/j.ress.2021.107864.
2. Danilo Giordano, Flavio Giobergia, Eliana Pastor, Antonio La Macchia, Tania Cerquitelli, Elena Baralis, Marco Mellia, Davide Tricarico. Data-driven strategies for predictive maintenance: Lesson learned from an automotive use case. Computers in Industry.Volume 134.2022. https://doi.org/10.1016/j.compind.2021.103554.
3. M. Jain, D. Vasdev, K. Pal, V. Sharma. Systematic literature review on predictive maintenance of vehicles and diagnosis of vehicle's health using machine learning techniques. Computational Intelligence, 38. 2022. p. 1990–2008. https://doi.org/10.1111/coin.12553.
4. Siavash Saki, Mohsen Soori. Artificial intelligence, machine learning and deep learning in advanced transportation systems, a review. Multimodal Transportation. Volume 5, Issue 1. 2026. 25 p. https://doi.org/10.1016/j.multra.2025.100242.
5. Md Naeem Hossain, Md Mustafizur Rahman, Devarajan Ramasamy. Artificial Intelligence-Driven Vehicle Fault Diagnosis to Revolutionize Automotive Maintenance: A Review. CMES - Computer Modeling in Engineering and Sciences.Volume 141, Issue 2.2024.P. 951-996. https://doi.org/10.32604/cmes.2024.056022.
6. Sielaff Lennard, Lucke Dominik, Wolf Yannic. A Reference Model for Predictive Maintenance Model Development. Procedia CIRP. Volume 130. 2024. P. 1537-1542. https://doi.org/10.1016/j.procir.2024.10.279.
7. Md Naeem Hossain, Md Mustafizur Rahman, Devarajan Ramasamy. Advances in intelligent vehicular health monitoring and fault diagnosis: Techniques, technologies, and future directions. Measurement. Volume 253, Part B. 2025. https://doi.org/10.1016/j.measurement.2025.117618.
8. Дяченко В.О. Інтелектуальні підходи енергозбереження у безпровідних сенсорних комп'ютерних мережах. Системи управління, навігації та зв'язку, т. 4 (62), 2020. P. 114-118. https://doi.org/10.26906/SUNZ.2020.4.114.
9. Harmash V., Diachenko V., Mikhal O., Znaidiuk V. Energy-Saving Method In Wireless Sensor Networks // Control, Navigation and Communication Systems, v.2 (80), 2025. P.54–58. https://doi.org/10.26906/SUNZ.2025.2.054.
10. Li Y, Wang T, Noman K, Li B. Advanced Fault Diagnosis and Health Monitoring Techniques for Complex Engineering Systems: 2nd Edition. Sensors. 2025; 25(22):7054. https://doi.org/10.3390/s25227054.
11. SCANIA Component X Dataset: A Real-World Multivariate Time Series Dataset for Predictive Maintenance. https://doi.org/10.58141/1w9m-yz81.

Received (Надійшла) 09.01.2026
Accepted for publication (Прийнята до друку) 25.03.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Скляров Ілля Ігорович – аспірант кафедри комп'ютерних наук і інформаційних систем, Харківський національний автомобільно-дорожній університет, Харків, Україна;
Illia Skliarov – PhD student, Department of Computer Science and Information Systems, Kharkiv National Automobile and Highway University, Kharkiv, Ukraine;
e-mail: is.ilyasklyarov@gmail.com; ORCID Author ID: http://orcid.org/0009-0001-9116-5315.

Геревич Михайло Олександрович – доктор філософії, доцент кафедри теорії та історії держави і права, Ужгородський національний університет, Ужгород, Україна;
Mykhailo Herevych – PhD, Associate Professor of Department of Theory and History of State and Law, Uzhhorod, Ukraine;
e-mail: mykhailo.herevych@uzhnu.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-0842-2828.

Method for predicting the technical condition of vehicles based on Big Data

Illia Skliarov, Mykhailo Herevych

Abstract. Relevance. The relevance of the study is determined by the need to improve the efficiency of service technologies in the automotive industry under conditions of growing volumes of telematics, diagnostic, and operational data accompanying the functioning of modern vehicles. Modern automobiles are characterized by a complex structure of technical systems, the presence of a large number of interconnected components, the heterogeneity of operating modes, and an increased risk of failures, which complicates the timely assessment of their technical condition. Existing approaches to forecasting the technical condition of automobiles mostly do not ensure the comprehensive consideration of streaming sensor data, maintenance history, and operational parameters, which complicates the substantiated decision-making process regarding service actions. Therefore, the development of a method for forecasting the technical condition of automobiles based on Big Data is relevant for supporting the timely detection of risky conditions and optimizing maintenance. **Purpose of the article** is to develop a method for forecasting the technical condition of automobiles based on Big Data technologies through the integration of streaming sensor data, maintenance history, and operational parameters for the timely detection of probable failures, improvement of the accuracy of assessing the technical condition of vehicles, and optimization of service decisions. **Object of research** is the process of forecasting the technical condition of automobiles in service maintenance systems based on the analysis of large volumes of heterogeneous data. **Subject of research** is methods, models, and tools for forecasting the technical condition of automobiles based on Big Data through the integration of sensor, operational, and service data. **Research results**. The study addresses the problem of forecasting the technical condition of automobiles based on Big Data in order to improve the efficiency of service technologies in the automotive industry. A method is proposed that involves the integration of operational, sensor, and service data within a unified analytical framework for assessing the risk of failures. The method was implemented in Python in the Google Colab environment using the open SCANIA Component X dataset. In the course of the study, data preparation, elimination of target variable leakage, construction of a predictive model, adjustment of the classification threshold, and formation of service decisions were carried out. The obtained results confirmed the high quality of forecasting and the suitability of the proposed approach for use in tasks supporting vehicle maintenance.

Keywords: Big Data, technical condition forecasting, automotive industry, machine learning, data analysis, maintenance, vehicles, classification, binary classifier, Google Colab, Python.

UDC 621.3 **doi: 10.26906/SUNZ.2026.2.028**

Vitalii Breslavets, Igor Yakovenko, Juliya Breslavets, Vitalii Voronets

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine

SURFACE ELECTRONIC STATES AT THE INHOMOGENEOUS INTERFACE SEMICONDUCTOR - DIELECTRIC

Abstract. The **subject matter** is the analysis processes and conditions for the generation (amplification) of electromagnetic oscillations in the submillimeter range by creating surface electronic states at an inhomogeneous interface between media, which is realized for structures such as metal – semiconductor – dielectric (M-S-D), where the localization sizes of surface states are within the range of approximately 10^{-4} cm. The **aim** is the possibility of conducting theoretical and experimental studies based on the proposed physical model for the emergence of surface electronic states at an inhomogeneous boundary of solid bodies, in conditions where the amplitude of the irregularities is much smaller than their period. Parameters of the lateral pulsed electromagnetic field, induced currents, and characteristics of semiconductor devices are established within which the mode of amplification of the intrinsic oscillations of the surface electron layer of the semiconductor structure is observed. The **objectives** are: the mechanisms of electron interaction at the interface between the conductor and dielectric, where the inhomogeneities are either random or periodic. As a result of this interaction, the electron concentration exponentially decreases with distance from the boundary. The **methods** used are: the methods of the theory of small perturbations (Rayleigh's method) to determine the spectrum of surface electronic states under conditions where the amplitude of irregularities is much smaller than their period. The following **results** are obtained: A dispersion equation for the spatial harmonics of electrons at the inhomogeneous boundary of a conducting solid body is derived. By the method of successive approximations for the small parameter, its solution is determined, and it is shown that in limiting cases—long-wavelength and short-wavelength—the localization sizes of electrons at the boundary have similar orders of magnitude for both periodic and random surfaces. **Conclusion.** The work provides quantitative estimates for the radiation energy values – the energy loss values of charged particle flows induced by an external electromagnetic field, leading to the excitation of surface electromagnetic oscillations in structures with boundary inhomogeneities in the presence of surface electronic states. The results show that the radiation energy for structures like metal – dielectric – semiconductor (M-S-D) lies in the range of ≈ (10^{-7} – 10^{-8}) *Wt*, which is detectable by modern microwave radiation receivers (approximately 10^{-10} *Wt*).

Keywords: Rayleigh's method; inhomogeneous boundary; surface electronic states; dispersion equation; spatial harmonics; induced current; surface oscillations.

Introduction

Mastering the submillimeter and short-wavelength parts of the millimeter electromagnetic wave ranges is one of the most relevant tasks in modern radio physics. These ranges are crucial for many technical applications: communication systems, radar, radionavigation, and computing, as well as for studying the impact of external electromagnetic fields on the performance of equipment (electromagnetic compatibility (EMC) tasks). Exciting oscillations in this range requires the creation of corresponding electromagnetic radiation sources [1, 2].

Modern technology makes it possible to create conductive solid-state structures: semiconductors with two-dimensional (2D) electron gas, superlattices, as well as films and structures like metal-dielectric-semiconductor (MDS), etc. In their formation, the electronic properties of the boundary between media play a key role.

Thus, surface electronic states lead to the formation of a two-dimensional (2D) electron gas and the appearance of surface oscillations in the submillimeter range. The presence of surface electronic states at the interface between media allows the transformation of the energy of charged particle flows into energy of surface oscillations. The generation and amplification mechanisms of surface oscillations are based on Cherenkov, transition, and braking radiation effects [3, 4].

A large number of works have been devoted to the study of surface electronic states, where the main focus was on the study of electronic states arising on the surface of crystals due to the limitation of the crystal lattice or, in other words, the disruption of the periodic potential [5, 6]. Depending on the chosen physical model, Tamm states, arising from the change in the potential at the crystal-vacuum boundary, and Shockley states, caused by the break in atomic bonds at the boundary, are distinguished [7].

The two models mentioned above do not cover all problems related to surface states. Another scenario is possible when a charged particle moves in the field of a constant, rather than a periodic, potential, but its motion is limited in one direction by an inhomogeneous surface, which represents an infinitely high potential barrier.

It is known that if the boundary is homogeneous, surface states do not arise. However, for an inhomogeneous boundary, the issue of quantum surface states has not been sufficiently explored. This work investigates the possibility of surface electronic states arising due to small periodic or random inhomogeneities at the boundary of a solid body. One-dimensional roughness was used as the object of study.

The problem is solved under the condition that the scales of the inhomogeneities are small compared to the wavelengths existing in the structure. The mathematical apparatus used is based on the representation of surface

roughness as small disturbances, the influence of which is taken into account in the boundary conditions.

Results

Let us consider the electronic states in a half-space $y > y_0(x)$, bounded by a potential barrier $U(x,y)$:

$$U(x,y): \begin{array}{ll} U(x,y)=\infty, & y \le y_0(x); \\ U(x,y)=0, & y > y_0(x), \end{array} \quad (1)$$

where $y_0(x)$ - is the function describing the form of the boundary between media. In this work, we limit ourselves to considering the boundary as an infinitely high potential barrier, with roughness depending on one coordinate X. The eigenfunctions $\Psi(x,y,z)$ and eigenvalues of the electron energy are determined by solving the Schrödinger equation:

$$\Delta\Psi + \frac{2m}{\hbar^2}\left[E - U(x,y)\right]\Psi = 0, \quad (2)$$

and boundary conditions at the surface and infinity. At the surface $y = y_0(x)$, boundary conditions are of two types [8]:

$$\Psi(y_0(x)) = 0 \quad (3)$$

$$\vec{n}\vec{\nabla}\Psi\big|_{y=y_0(x)} = 0; \quad \vec{\nabla} = \vec{i}\frac{\partial}{\partial x} + \vec{j}\frac{\partial}{\partial y} + \vec{k}\frac{\partial}{\partial z}, \quad (4)$$

where $\vec{n}$ is the normal vector to the surface $y = y_0(x)$:

$$n_x = -\frac{\frac{\partial y_0}{\partial x}}{\sqrt{\left(\frac{\partial y_0}{\partial x}\right)^2 + 1}}; \; n_y = -\frac{1}{\sqrt{\left(\frac{\partial y_0}{\partial x}\right)^2 + 1}}; \; n_z = 0, \quad (5)$$

The conditions correspond to the zero particle flux density (4) and particle density (3).

This work considers two types of boundary roughness: periodic $y_0(x) = \zeta_0 \cos(Gx)$; $d = 2\pi/G$ – period of roughness) and rough $y_0(x) = \zeta(x)$, where $\zeta(x)$ is a random function. In the case of a periodically rough boundary, the wave function $\Psi(x,y,z)$ looks as follows:

$$\Psi(x,y,z) = \sum_{n=-\infty}^{\infty} A_n \exp\left[i(k_x + nG)x + ik_y y + ik_z z\right], \quad (6)$$

where $\vec{k}(k_x, k_y, k_z)$ is the wave vector of the electron.

From the Schrödinger equation (2), the relationship between E и $\vec{k}$ is as follows:

$$k_{yn}^2 = \frac{2mE}{\hbar^2} - (k_x + Gn)^2 - k_z^2 \quad (7)$$

The boundary condition (4) establishes the connection between k_x, k_{yn} and k_z, thereby defining the dispersion $E = E(\vec{k})$.

To solve the Schrödinger equation with boundary condition (4), we use perturbation theory [9], assuming that the amplitude of the roughness is small compared to its period ($\zeta_0 k_x << 1 \ll \lambda$). This allows us to limit the analysis to the harmonics $n = -1, 0, 1$, where the amplitude of the harmonic A_0 is maximum.

Substituting expression (6) into equation (4), we get the following dispersion relation:

$$k_{y0} = -\frac{1}{4}\zeta_0^2\left(\frac{\left[k_{y-1}^2 - G(k_x - G)\right]\left(k_{y0}^2 + Gk_x\right)}{k_{y-1}} + \frac{\left[k_{y1}^2 + G(k_x + q)\right]\left(k_{y0}^2 - Gk_x\right)}{k_{y1}}\right). \quad (8)$$

We will solve equation (8) using the method of successive approximations for the small parameter ζ_0: $k_{y0} = k_{y0}^{(0)} + \delta k_{y0} + \ldots$

If $\zeta_0 = 0$, then $k_{y0}^{(0)} = 0$ and

$$k_x^2 = \frac{2mE}{\hbar^2} - k_z^2. \quad (9)$$

The next approximation yields:

$$\delta k_{y0} = -\frac{1}{4}(\zeta_0 k_x G)^2\left(\frac{1}{k_{y1}} + \frac{1}{k_{y-1}}\right); \quad (10)$$

$$\delta E = \frac{\hbar^2 \delta k_{y0}^2}{2m}; \quad k_{y\pm1}^2 = -G(G \pm 2k_x). \quad (11)$$

In the case when $k_x << q$ is small, equation (10) gives:

$$\delta k_{y0} = \frac{1}{2}i(\zeta_0 k_x)^2 G; \quad k_{y1} = k_{y-1} = iG. \quad (12)$$

Solution (12) defines the localized electronic states near the surface with energy

$$E = \frac{\hbar^2}{2m}\left[k_z^2 + k_x^2\left((1 - \frac{1}{4}\zeta_0^2 G^2 k_x^2\right)\right]. \quad (13)$$

From equation (12), it is evident that the spatial localization length of the electron wave function $R = i \,/\left(\delta k_{y0}\right)$ decreases exponentially as the wave vector k_x increases. Thus, the electron concentration also decreases exponentially with distance from the boundary, forming a surface electron layer.

Surface inhomogeneities most effectively affect the electronic states in resonance conditions, when the wave vectors of adjacent harmonics traveling in opposite directions along the axis X coincide ($k_{y0} = k_{y-1}$). In this case, $k_x = G/2 \equiv k_r$ and from equation (10), we obtain:

$$\delta k_{y0}^2 = -\zeta_0^2 k_r^2; \quad (14)$$

$$E = \frac{\hbar^2}{2m}\left[k_z^2 + k_r^2(1 - \zeta_0^2 k_r^2)\right]. \quad (15)$$

Equation (14) has the following solutions:

$$\mathrm{Re}\,\delta k_{y0} = 0; \quad \mathrm{Im}\,\delta k_{y0} = \zeta_0 k_r^2 \tag{16}$$

These correspond to a highly localized surface state. Thus, electronic surface states exist in the region $k_x \le G/2 - (\mathrm{Im}\,k_{y0\pm1} > 0)$.

In the region $k_x > G/2$, surface states typically do not arise. In this case, both δk_{y0} and E take on complex values. In the region $k_x >> G$ the equation (8) has the solution:

$$\delta k_{y0} = \frac{(-1+i)\zeta_0^2 (k_x G)^{3/2}}{\sqrt{2}}; \tag{17}$$

$$\delta E = -i\frac{\hbar^2 \zeta_0^4 (k_x G)^3}{2m}. \tag{18}$$

Thus, the quantum states are quasi-stationary, meaning they have a lifetime $\Psi \sim e^{-t/\tau}$:

$$\tau = \frac{2m}{\hbar^2 \zeta_0^4 (k_x G)^3} \tag{19}$$

Now, let us define the mechanisms of the formation of surface electronic states when the surface roughness is random. Let the shape of the boundary surface be given by a random function $y = \zeta(x)$). Assume that $\zeta(x)$ is a stationary homogeneous process with an average value $\overline{\zeta(x)} = 0$, and its statistical properties are described by the correlation function:

$$\overline{\zeta(x)\zeta(x')} = \zeta_0^2 W(x - x') \tag{20}$$

We assume, as in the case of periodic roughness, that the amplitude of the deviation of the random function from its mean value is small, i.e., $\frac{\partial \zeta}{\partial x} << 1 \ll \lambda$. In this case, to solve the Schrödinger equation with boundary condition (4), we can use the standard procedure to determine the field over a statistically rough surface $\zeta(x)$ [8]. After performing the necessary calculations, we obtain an expression that defines the particle spectrum:

$$k_y = -\zeta_0^2 \int_{-\infty}^{\infty} \frac{d\chi_x}{k_y}\Big[k_x(\chi_x - k_x) - k_y^2\Big] \times \\ \times\Big[\chi_x(k_x - \chi_x) - \chi_y^2\Big] W(k_x - \chi_x), \tag{21}$$

where $\chi_y^2 = k_x^2 - \chi_x^2$; $k_y^2 = \frac{2mE}{\hbar^2} - k_x^2 - k_z^2$, $W(\vec{k})$ is the Fourier transform of the correlation function (20), which subsequently takes the Gaussian form:

$$W(k_x - \chi_x) = \frac{l}{2\sqrt{\pi}} \exp\left(-\gamma^2 k_x^2 L^2\right). \tag{22}$$

Here, ℓ is the correlation $\gamma = \frac{1}{2} - \frac{\chi_x}{2} k_x$.

From equation (21), it follows that when $\zeta = 0$, we have: $k_y = 0$.

Substituting into the right-hand side of equation $k_{y0} = 0$ (21) and performing integration over the angles, we obtain δk_y, in the first approximation:

$$\delta k_y = \frac{2i\zeta_0^2 k_x l}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{d\gamma \gamma^{3/2} \exp(-\gamma^2 k_x^2 L^2)}{\sqrt{\gamma - 1}}. \tag{23}$$

The solution to equation (23) can be analytically evaluated in two limiting cases: $k_x l << 1$ and $k_x l >> 1$.

In the long-wavelength limit $k_x l << 1$, the value (23) is determined by expanding the integrand in terms of the small parameter $k_x l$:

$$\delta k_y = 2i\zeta_0^2 k^2 \Big/ \sqrt{\pi} l. \tag{24}$$

The solution to equation (24) corresponds to a localized surface state with energy:

$$E = \frac{\hbar^2}{2m}\Big[k_z^2 + k_x^2(1 - 4\zeta_0^4 k_x^2 / \pi l)\Big]. \tag{25}$$

The solution to equation (23) in the long-wavelength limit implies that near the rough surface of a solid, there exist surface electronic states with a non-quadratic dispersion law. The electron wave function in this case takes the form:

$$\overline{\Psi} \sim \exp\left(-\frac{2\zeta_0^2 k_x^2}{\sqrt{\pi} l} y\right). \tag{26}$$

In the short-wavelength limit ($k_x l >> 1$), the solution to equation (23) becomes:

$$\delta k_y = \frac{\Gamma(5/4)(1-i)}{\sqrt{\pi}} \frac{\zeta_0^2 k_x^3}{(k_x l)^{3/2}}, \tag{27}$$

where $\Gamma(x)$ is the Gamma function. Expression (26) describes a quasi-stationary, localized electronic state near the surface with a characteristic lifetime:

$$\tau = \frac{\pi m l^3}{\Gamma^2(5/4)\hbar \zeta_0^4 k_x^3} \tag{28}$$

Thus, small surface inhomogeneities, representing an infinitely high potential barrier, lead to the formation of surface electronic states whose function exponentially decays with distance from the boundary. The expressions that describe localized states near a rough boundary and those describing states near a periodically rough surface are analogous. In the first case, the characteristic size is the correlation length, and in the second case, it is the period of the surface inhomogeneities.

Analysis of the results obtained

The implementation of the effects mentioned can be realized, for example, at the boundary between a semiconductor and a dielectric. The boundary may have natural roughness or a periodic structure in the form of dislocation mismatches, as well as by creating an artificial periodic relief. According to the results obtained, electrons will be localized near the boundary in a layer of thickness R, since $\Psi \sim e^{-y/R}$.

In the case when the period of the surface inhomogeneity is of the order of several micrometers (microns) $a = 10^{-5}\,sm$, a value accessible by lithographic methods for structure formation, and the ratio between the amplitude of the inhomogeneity ς_0 and the wavelength $(\lambda = 1/k)$ $\varsigma_0 k \approx 0.1$, the electrons will be localized in a layer of thickness $R \approx 10^{-4}$ in the resonant case. In the long-wavelength limit, the thickness of this layer will be an order of magnitude larger.

Let us provide quantitative estimates of the conditions for the resonant (Cherenkov) interaction of surface oscillations of the surface electronic states with the flows of charged particles induced by external electromagnetic radiation, i.e., the possibilities for their generation (amplification) in current semiconductor devices used in radar and communication systems [2].

The frequency of surface plasmons for typical values of semiconductor structures used in modern radio electronics is $\omega_s \approx 10^9 - 10^{11} s^{-1}$.

The drift velocity of carriers for fields in the range E of electric and H of magnetic field strengths, affecting the semiconductor structure with surface electronic states emission, is $E < 100\frac{kV}{m};\ H < 600\frac{A}{m}$.

Therefore, the conditions for resonant interaction between waves v_f and particles (equality of the wave phase velocity and the drift velocity of the induced current $v_f = \omega_s / q \approx v_{dp}$ are satisfied for millimeter (submillimeter) wavelengths, corresponding to the size of the localization of surface electronic states $R \approx 10^{-4}$ sm.

Let us now provide quantitative estimates of the radiation energy ΔW_{rad}, i.e., the energy losses of the flow of charged particles induced by external electromagnetic fields, in exciting surface electromagnetic oscillations in structures such as metal-dielectric-semiconductor (MDS) with surface electronic states [3]. For electric fields with a field strength $E_0 \approx 10 - 50\frac{kv}{m}$ in the region of reversible failures, the pulse duration is $\Delta t_{imp} \approx 500\,ns$.

The concentration of carrier currents and their drift velocities lie in the range [4]:

$$n_b \approx 10^{10} - 10^{12}\,sm^{-3} \quad v_0 \approx 10^5 - 10^7\,sm/s.$$

The radiation energy of the own oscillations of solid-state MOS structures lies in the range $\approx (10^{-7} - 10^{-8})vm$ GHz, and thus, with the sensitivity of modern microwave radiation receivers [3] (from $10^{-10} Wt$), it is easily detectable.

Thus, the proposed physical model of the conditions for the generation (amplification) of oscillations by creating surface electronic states at an inhomogeneous interface between media is feasible for MDS structures, since the sizes of their localization are in the range of a few centimeters $R \approx 10^{-4}\ sm$.

Conclusions

1. The results obtained indicate that periodic (random) inhomogeneities at the boundary between two media lead to the appearance of surface electronic states, with the concentration of electrons (wave function) exponentially decreasing with distance from the boundary.

2. It should be noted that in the limiting cases – long-wavelength and short-wavelength – the localization sizes of the electrons have the same order of magnitude for both periodic and random surfaces. The most effective interaction occurs when the de Broglie wavelength of the electron is comparable to the characteristic size of the inhomogeneity and the Bragg reflection condition is met. If the period of the surface inhomogeneities is a few micrometers, electrons will localize in a layer of thickness $R \approx 10^{-4}$ in the resonant case, and in the long-wavelength limit, in a layer an order of magnitude thicker.

3. The proposed physical model of the formation of surface electronic states at an inhomogeneous boundary can be realized in MDS structures, which creates opportunities for the generation (amplification) of oscillations in the submillimeter range, since the localization layer sizes are within a few centimeters.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCE

1. Serkov O., Breslavets V., Breslavets J., Yakovenko I. Excitation of own oscillations in semiconductor components of radio products under the exposure of third-party electromagnetic radiation. *Advanced Information Systems*. 2022. Vol. 6, No. 1. P. 124–128. DOI: https://doi.org/10.20998/2522-9052.2022.1.20
2. Serkov O.A., Breslavets V.S., Breslavets Y.V., Yakovenko I.V. Mechanisms of the influence of external electromagnetic radiation on the performance of communication equipment. *Systems of Control, Navigation and Communication*. 2022. Vol. 2, No. 68 (2022). P. 129–133. DOI: https://doi.org/10.20998/2522-9052.2022.1.20
3. Serkov O., Breslavets V., Breslavets J., Yakovenko I. Excitation of magnetoplasma oscillation in semiconductor structures by fluxes of charged particles. *Advanced Information Systems*. 2021. Vol. 5, No. 3. P. 18–21. DOI: https://doi.org/10.20998/2522-9052.2021.3.03
4. Serkov O., Breslavets V., Dzubenko A., Yakovenko I. Excitation of surface vibrations of semiconductor structures exposed to external electromagnetic radiation. *Advanced Information Systems*. 2019. Vol. 2, No. 3. P. 142–146. DOI: https://doi.org/10.20998/2522-9052.2018.3.25

5. Potylitsyn A.P. Transition radiation and diffraction radiation. Similarities and differences. *Nuclear Instruments and Methods in Physics Research Section B Beam Interactions with Materials and Atoms*. 1998. Vol. 145, P. 67. DOI: https://doi.org/10.1016/S0168-583X(98)00384-X
6. Rule D.W., Fiorto R.B., Kimura W.D. Noninterceptive beam diagnostics based on diffraction radiation. AIP Conf. Proc. 1997. Vol. 590. P. 510–517. DOI: https://doi.org/10.1063/1.52327
7. Fiorito R.B., Rule D.W. Diffraction radiation diagnostics for moderate to high energy beams. *Nuclear Instruments and Methods in Physics Research Section B Beam Interactions with Materials and Atoms*, Vol. 173(1). P. 67–82. DOI: https://doi.org/10.1016/S0168-583X(00)00066-5
8. Shilliday T.S. and Vaccaro J. (Editors). Physics of Failure in Electronics. Vol. 5, RADS Series in Reliability, Rome Air Development Center. 1966. Also AD. 655397. URL: https://apps.dtic.mil/sti/tr/pdf/AD0637529.pdf
9. Queisser H.J. Failure Mechanisms in Silicon Semiconductors. Final Report Contract AF 30 (602)-2556. Rome Air Development Center, Report No. RADC-TDR-62-533. 1963. https://apps.dtic.mil/sti/tr/pdf/AD0297033.pdf

Received (Надійшла) 08.12.2025
Accepted for publication (Прийнята до друку) 01.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Бреславець Віталій Сергійович – кандидат технічних наук, доцент, професор кафедри систем інформації, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Vitalii Breslavets – Candidate of Technical Sciences, Associate Professor, Professor of Information Systems Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e mail: bres123@ukr.net; ORCID Author ID: https://orcid.org/0000-0002-9954-159X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57204843959.

Яковенко Ігор Володимирович – доктор фізико-математичних наук, професор, професор кафедри систем інформації Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Igor Yakovenko – Doctor of Physical and Mathematical Sciences, Professor, Professor of Information Systems Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e mail: yakovenko60iv@ukr.net; ORCID: https://orcid.org/0000-0002-0963-4347;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=6601985070.

Бреславець Юлія Віталіївна – асистент кафедри систем інформації, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Juliya Breslavets – Assistant of Information Systems Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: juliettar941@gmail.com; ORCID: https://orcid.org/0000-0003-4530-8028.

Воронець Віталій Миколайович – доктор філософії, доцент кафедри «Системи інформації», Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Vitalii Voronets – PhD, Associate Professor, Department of Information Systems, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: Vitalii.Voronets@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-7793-3824;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=58986447800.

Поверхні електронні стани на неоднородному кордоні півпровідник – діелектрик

В. С. Бреславець, Ю. В. Бреславець, І. В. Яковенко, В. М. Воронець

Анотація. Об'єктом дослідження є процес аналізу умов для генерації (посилення) електромагнітних коливань субміліметрового діапазону шляхом створення поверхневих електронних станів на неоднорідному кордоні розділу середовищ, що реалізується для структур метал – напівпровідник – діелектрик (МДП), коли розміри локалізації поверхневих станів знаходяться в межах $R \approx 10^{-4}$ см. **Мета дослідження** –можливість постановки теоретичних та експериментальних досліджень на основі запропонованої фізичної моделі виникнення поверхневих електронних станів на неоднорідному кордоні розділу твердих тіл, в умовах коли амплітуда нерівності набагато менша за її період. Встановлено параметри стороннього імпульсного електромагнітного поля, наведених струмів та характеристик напівпровідникових приладів у рамках яких спостерігається режим посилення власних коливань поверхневого електронного шару напівпровідникової структури. Методи дослідження: методи теорії малих обурень (метод Релея) при визначенні спектра поверхневих електронних станів в умовах коли амплітуда нерівності набагато менша від її періоду. **Отримані результати.** Отримано дисперсійне рівняння для просторових гармонік електронів на неоднорідній межі твердого тіла, що проводить. Методом послідовних наближень за малим параметром визначено його рішення та показано, що у граничних випадках – довгохвильовому та короткохвильовому – розміри локалізації електронів на кордоні мають однакові порядки величин як для періодичної поверхні, так і для випадкової. **Висновки.** У роботі наведено кількісні оцінки величин енергії випромінювання - величини втрат енергії потоків заряджених частинок, наведених зовнішнім електромагнітним полем, на збудження поверхневих електромагнітних коливань у структурах з неоднорідностями межі середовищ за наявності поверхневих електронних станів. Вони показують, що величина енергії випромінювання структур типу метал – діелектрик - напівпровідник лежить у діапазоні $\approx (10^{-7} – 10^{-8})$ *вт*, тобто. при чутливості сучасних приймачів НВЧ випромінювання (10^{-10} *вт*) цілком виявлена.

Ключові слова: метод Релея; неоднорідна межа; поверхневі електронні стани; дисперсійне рівняння; просторові гармоніки; наведений струм; поверхневі коливання.

УДК 629.42:004.032.26 **doi: 10.26906/SUNZ.2026.2.033**

М. Й. Заполовський, М. В. Мезенцев, М. В. Оліфір

Національний технічний університет "Харківський політехнічний інститут", Харків, Україна

МАТЕМАТИЧНА МОДЕЛЬ ТА ПРОГРАМНО-АЛГОРИТМІЧНІ КОМПОНЕНТИ ДЛЯ СИНТЕЗУ СИСТЕМИ КЕРУВАННЯ КОВЗАННЯМ ЧАСТОТИ НАПРУГИ ЖИВЛЕННЯ ТАД

Анотація. Актуальність. При синтезі оптимальних систем керування постають ряд задач, як із розробленням моделей об'єкта управління, так і вибору методу оптимізації з подальшою реалізацією та дослідженнями отриманих законів у вигляді програмно-апаратної реалізації. **Об'єкт дослідження**: програмні компоненти для система керування тяговим електроприводом дизель-поїзда. **Мета статті**: створення програмно-алгоритмічних компонентів для системи керування тяговим приводом дизель-поїзда на основі синтезу законів управління ковзанням частоти напруги живлення тягового асинхронного двигуна.. **Результати дослідження**. У статті запропоновано підхід до розв'язання задач синтезу систем керування на основі рішення загальної задачі Лагранжа та розроблено методику знаходження невизначених множників Лагранжа. На основі математичної моделі системи керування отримані аналітичні співвідношення для обчислення оптимальних керуючих впливів. Ці співвідношення інтегровані в комп'ютерну модель електромеханічної системи, розгорнуту в MATLAB/Simulink, що дозволило провести цикл віртуальних випробувань. **Висновки.** Реалізація даного підходу дозволяє розширити можливості проєктування оптимальних систем керування електроприводом змінного струму транспортних засобів. Результати моделювання демонструють, що запропонований алгоритм, реалізований у вигляді функціональних блоків Simulink, забезпечує стійку роботу приводу в усьому діапазоні тягової характеристики з оптимізацією витрати енергії. Розроблені моделі та алгоритми є готовим інструментарієм для впровадження в мікроконтролери сучасних систем автоматики тягового електроприводу.

Ключові слова: програмно-алгоритмічні компоненти; тяговий електропривод; дизель-поїзд; метод Лагранжа; комп'ютерне моделювання; функціональні блоки.

Вступ

Постановка проблеми. Згідно з результатами досліджень [1, 2], перспективним напрямом розвитку тягового електроприводу для залізничного транспорту, зокрема дизель-поїздів, є використання асинхронних двигунів змінного струму та спеціалізованих систем керування. Ключовим аспектом при створенні таких приводів є розробка ефективних алгоритмів керування, спрямованих на оптимізацію експлуатаційних показників, зокрема зниження енерговитрат. Сучасний підхід до реалізації таких систем передбачає використання програмно-апаратних компонентів, що дозволяють реалізувати складні алгоритми керування в реальному часі.

Одним із поширених підходів до керування асинхронним електроприводом є підтримання сталості відношення напруги живлення до її частоти (U/f = = const). Сучасна реалізація даного підходу потребує створення цифрових модулів керування, зокрема програмованих контролерів із спеціалізованими обчислювальними блоками. При цьому кругова частота напруги живлення формується відповідно до виразу: $\omega = \omega\_к + s$, де $\omega\_к$ – кругова швидкість обертання колеса небуксуючої колісної пари; s – ковзання частоти напруги живлення ТАД, яке визначається як різниця між частотою напруги статора та електричною круговою частотою обертання ротора [3].

Як зазначається в роботах [4–6], перспективним напрямом дослідження та розробки систем керування є використання моделей, орієнтованих на специфіку функціонування ТАД при регулюванні частоти ковзання. Особливу актуальність набуває створення програмних компонентів для реалізації алгоритмів керування ковзанням, зокрема у вигляді спеціалізованих бібліотек блоків для середовищ моделювання типу MATLAB/Simulink. Таким чином, актуальним завданням є синтез управлінь ковзанням частоти напруги живлення тягового асинхронного двигуна та їх програмна реалізація.

Регулювання цієї складової відповідно до синтезованого закону керування дозволяє оптимізувати величину напруги живлення ТАД та її частоту. Сучасна реалізація таких алгоритмів передбачає розробку програмних модулів, здатних у реальному часі обчислювати оптимальні параметри керування, забезпечуючи формування необхідного тягового моменту при мінімізації енергетичних витрат. Використання комп'ютерних компонентів дозволяє створити уніфіковану платформу для моделювання, верифікації та впровадження алгоритмів керування для сучасних систем керування рухомим складом.

Аналіз останніх досліджень і публікацій. Для об'єктів керування, що описуються системами диференціальних рівнянь до п'ятого порядку, ефективним інструментом синтезу керувань є методи варіаційного числення, зокрема розв'язок загальної задачі Лагранжа або використання принципу максимуму Понтрягіна [7]. У випадку систем вищого порядку може застосовуватись метод аналітичного конструювання регуляторів за критерієм узагальненої роботи. Однак аналіз методів варіаційного числення виявляє проблематику ідентифікації та визначення функцій, що формують структуру допоміжного функціоналу, що потребує розробки спеціалізованих підходів та моделей для їх коректного знаходження. Оскільки досліджуваний об'єкт керування може бути адекватно описаний системою диференціальних рівнянь не вище п'ятого порядку [6], для синтезу керувань доцільно застосувати метод варіаційного числення шляхом розв'язку загальної задачі Лагранжа. При розв'язку даної задачі використовується допоміжний функціонал виду [8]:

$$J = \int_{t_0}^{T} [G + \sum_{j=1}^{n} \lambda_j(t)\varphi_j(t)]dt \ , \qquad (1)$$

де G – функціонал, який мінімізується; $\lambda_j(t)$ – невизначені множники Лагранжа; $\varphi_j(t)$ – диференційні рівняння, що описують об'єкт керування.

Визначення оптимальних керувань здійснюється через розв'язок рівнянь Ейлера-Лагранжа з використанням сформованого допоміжного функціоналу. Таким чином, для успішного синтезу керувань необхідно мати математичну модель мінімізуємого функціоналу та систему диференціальних рівнянь, що описують динаміку об'єкта керування.

Для експериментальної перевірки ефективності синтезованих керувань пропонується використання сучасних інструментів комп'ютерного моделювання, зокрема проведення імітаційного моделювання в середовищі MATLAB/Simulink з розробкою відповідних програмних компонентів та функціональних блоків, що реалізують запропоновані алгоритми керування [6]. Також передбачається створення бібліотеки спеціалізованих блоків для моделювання систем керування тяговим електроприводом з урахуванням специфіки регулювання ковзання та частоти живлення ТАД.

Метою роботи є створення програмно-алгоритмічних компонентів для системи керування тяговим приводом дизель-поїзда на основі синтезу законів керування ковзанням частоти напруги живлення тягового асинхронного двигуна. Для досягнення поставленої мети передбачається використання апарату варіаційного числення шляхом розв'язання загальної задачі Лагранжа.

Основні завдання дослідження включають:

– синтез аналітичних співвідношень для оптимального керування, спрямованих на мінімізацію енергетичних витрат під час розгону рухомого складу;

– розробку програмних компонентів та функціональних блоків для реалізації отриманих алгоритмів керування в середовищі MATLAB/Simulink;

– створення комплексної імітаційної моделі системи керування електроприводом з метою експериментального дослідження якісних характеристик системи;

– Валідацію ефективності запропонованого підходу шляхом аналізу результатів моделювання замкнутої системи керування.

Наукова новизна роботи полягає в розробці архітектури програмно-алгоритмічних компонентів системи керування, що реалізують оптимізаційні алгоритми на основі методу Лагранжа для забезпечення енергоефективної роботи тягового електроприводу.

Основний матеріал

Для синтезу системи керування тяговим електроприводом дизель-поїзда застосовано математичну модель асинхронного двигуна у синхронній системі координат. Відмінною особливістю запропонованого підходу є визначення оптимальних керуючих впливів шляхом синтезу закону зміни ковзання, що забезпечує мінімізацію енергетичних витрат у процесі розгону рухомого складу.

Експериментальне дослідження синтезованої системи керування проводилось на основі математичної моделі ТАД у синхронній системі координат, що була реалізована у вигляді спеціалізованих програмних компонентів у середовищі MATLAB/ Simulink. Для параметризації моделі використано експлуатаційні характеристики об'єкта керування, наведені в [5, 6].

Математична модель ТАД у синхронній системі координат (*d*-*q*), що обертається зі швидкістю ω_s, описується системою рівнянь електричних кіл статора та ротора [4]:

$$\begin{cases} \frac{d\Psi_{s1}}{dt} = U_{s1} - a_s\Psi_{s1} + a_s k_R \Psi_{r1} + \Psi_{s2}\omega_s; \\ \frac{d\Psi_{s2}}{dt} = U_{s2} - a_s\Psi_{s2} + a_s k_R \Psi_{Y2} - \Psi_{s1}\omega_s; \\ \frac{d\Psi_{r1}}{dt} = -a_r\Psi_{r1} + a_r k_s \Psi_{s1} + (\omega_s - \omega)\Psi_{r2}; \\ \frac{d\Psi_{r2}}{dt} = -a_r\Psi_{r2} + a_r k_s \Psi_{s2} - (\omega_s - \omega)\Psi_{r1}; \\ M = \frac{3}{2} p \frac{k_R}{\sigma L_S}(\Psi_{r1}\Psi_{s2} - \Psi_{s1}\Psi_{r2}); \\ \frac{d\omega}{dt} = \frac{p}{J}(M - M_C), \end{cases} \qquad (2)$$

де ω_S – кутова частота напруги живлення статора ТАД при p = 1; ω – кутова швидкість ротора; Ψ_{i1}, Ψ_{s2}, Ψ_{r1}, Ψ_{r2}, U_{s1}, U_{s2} – відповідно проекції векторів потоків статора і ротора та напруги живлення на осі координат; k_s і k_r – коефіцієнти електромагнітного зв'язку відповідно статора та ротора;

$L_s = L_m + L_{\sigma s}$, $L_r = L_m + L_{\sigma r}$, $k_s = \frac{L_m}{L_s}$, $k_r = \frac{L_m}{L_r}$,

$\sigma = 1 - k_r k_S = 1 - \frac{L_m^2}{L_S L_r}$, $T_s = \frac{L_s}{r_1}$, $T_r = \frac{L_r}{r_2}$, $a_s = \frac{1}{\sigma T_S}$,

$a_r = \frac{1}{\sigma T_r}$; L_r, L_m, L_s, r_1, r_2 — параметри схеми заміщення ТАД; T_s і T_r – відповідно електромагнітна постійна часу статора та ротора; σ – повний коефіцієнт розсіювання; p – постійний коефіцієнт ТАД (кількість пар полюсів); J – момент інерції.

Змінні по осям U_{s1} і U_{s2} визначаються рівняннями (для випадку синусоїдального виду напруги на статорі ТАД):

$$U_{s1} = U_M \cos[(\omega_s - \omega_K)t + \varphi_K];$$

$$U_{s2} = U_M \sin[(\omega_s - \omega_K)t + \varphi_K],$$

де U_M – амплітуда першої гармоніки фазної напруги на статорі ТАД; ω_K – кутова швидкість обертання осей координат; φ_K – початкова фаза напруги живлення.

У випадку синхронного обертання осей $\omega_s = \omega_K$ і початкова фаза φ_K = 0, то функції за-

дання напруги – це постійне значення напруги живлення величиною U_M по осі S_1 і нульове по осі S_2.

Рівняння для моделювання електромагнітного моменту:

$$M = \frac{3}{2} p \frac{k_R}{\sigma L_S} (\Psi_{r1}\Psi_{s2} - \Psi_{s1}\Psi_{r2}). \quad (3)$$

де σ – постійне значення коефіцієнта, яке визначаються електричними параметрами схеми заміщення ТАД. Систему рівнянь (2) і рівняння (3) використано для розроблення комп'ютерної моделі в середовищі пакету MATLAB/Simulink для проведення дослід-жень та отримання якісних характеристик роботи систем керування в процесі моделювання.

Для подальших перетворень представимо систему (2) в загальному виді:

$$\dot{\Psi}_{s1} = -a_{11}\Psi_{s1} + a_{12}\Psi_{r1} + \Psi_{s2}\omega_s + U_м;$$
$$\dot{\Psi}_{s2} = -a_{21}\Psi_{s2} + a_{22}\Psi_{r2} - \Psi_{s1}\omega_s;$$
$$\dot{\Psi}_{r1} = -a_{31}\Psi_{r1} + a_{32}\Psi_{s1} + \Psi_{r2}(\omega_s - \omega);$$
$$\dot{\Psi}_{r2} = -a_{41}\Psi_{r2} + a_{42}\Psi_{s2} - \Psi_{r1}(\omega_s - \omega);$$
$$\dot{\omega} = a_2(M - M_c);$$
$$M = a_1(\Psi_{r1}\Psi_{s2} - \Psi_{s1}\Psi_{r2}),$$

де $a_{11} = a_{21} = a_s$; $a_{12} = a_{22} = a_s k_R$; $a_{31} = a_{41} = a_r$; $a_{32} = a_{42} = a_r k_s$; $a_1 = \frac{3}{2} p \frac{k_R}{\sigma L_S}$; $a_2 = \frac{p}{J}$.

Або в традиційному виді (фазових координат):

$$\dot{X}_1 = -a_{11}X_1 + a_{12}X_3 + X_2\omega_s + U_м;$$
$$\dot{X}_2 = -a_{21}X_2 + a_{22}X_4 - X_1\omega_s;$$
$$\dot{X}_3 = -a_{31}X_3 + a_{32}X_1 + X_4(\omega_s - X_5);$$
$$\dot{X}_4 = -a_{41}X_4 + a_{42}X_2 - X_3(\omega_s - X_5); \quad (4)$$
$$\dot{X}_5 = a_2(M - M_c);$$
$$M = a_1(X_2X_3 - X_1X_4).$$

Розглянемо випадок використання одного із найпоширенішого закону керування електроприводу змінного струму, згідно якого задання напруги живлення ТАД та її частоти ω_s формується як *U/f* = *const* $= U_2$, для рішення задачі знаходження оптимальних управлінь в процесі розгону дизель-поїзда до заданої швидкості за певний час згідно заданого функціоналу шляхом рішення загальної задачі Лагранжа. В якості управлінь використаємо частоту ковзання U_1 і закон керування $U/f = U_2$. Напругу U_M (напругу живлення ТАД) представимо як:

$$U_M = U_2(X_5 + U_1). \quad (5)$$

В процесі синтезу управлінь розглядається задача мінімізації енергетичних витрат в процесі розгону дизель-поїзда до заданої швидкості. Для цього використано рівняння швидкості руху *V*.

За відсутності боксування швидкість руху пропорційна кутовій швидкості обертання ротора двигуна ω. Рівняння руху має вигляд:

$$\frac{d\omega}{dt} = \frac{p}{J_T}(M_T - M_C), \quad (6)$$

де M_T – тяговий момент дизель-поїзда; J_T – момент інерції дизель-поїзда; M_C – момент опору руху; $J_T = mR_k^2$; m – маса дизель-поїзда; R_K – радіус колеса. Тяговий момент дизель-поїзда (утворюють чотири ТАД):

$$M_T = 4\, i\, M_d, \quad (7)$$

де i – передавальне число редуктора. Момент опору руху згідно правил тягових розрахунків:

$$M_c = R_K\, W_0\, G_H = 9.8\, m\, R_K (1.1 + 0.12V). \quad (8)$$

При розрахунковому завантажені:

$$M_C = 1334.0 + 2.25\,\omega. \quad (9)$$

Оскільки тяговий момент формується на виході редуктора, то і швидкість обертання ротора ω_p в математичній моделі повинна враховуватися на виході редуктора в залежності від його передавального числа. В результаті рівняння руху має вигляд:

$$\frac{d\omega_p}{dt} = \frac{pi}{J_T}(4iM_d - 1334.3 - 2.25\omega_p) =$$
$$= 0.0028\, M_d - 0.00043\,\omega_p - 0.254. \quad (10)$$

В результаті перетворень співвідношень (4) система рівнянь об'єкта керування в загальному плані для рішення задачі оптимізації буде мати вигляд:

$$\dot{X}_1 = -a_{11}X_1 + a_{12}X_3 + X_2(X_5 + U_1) + U_2(X_5 + U_1);$$
$$\dot{X}_2 = -a_{21}X_2 + a_{22}X_4 - X_1(X_5 + U_1);$$
$$\dot{X}_3 = -a_{31}X_3 + a_{32}X_1 + X_4U_1; \quad (11)$$
$$\dot{X}_4 = -a_{41}X_4 + a_{42}X_2 - X_3U_1;$$
$$\dot{X}_5 = a_5M_d - a_6X_5 - a_7,$$

де $M_d = a_1(X_2X_3 - X_1X_4)$; $a_5 = 0.0028$; $a_7 = 0.254$; $a_6 = 0.00043$.

Рішення задачі розглянемо для знаходження управлінь U_1 і U_2 при їх одночасній зміні. Система рівнянь для рішення загальної задачі Лагранжа має вигляд:

$$\dot{X}_1 + a_{11}X_1 - a_{12}X_3 - X_2(X_5 + U_1) - U_2(X_5 + U_1) = 0;$$
$$\dot{X}_2 + a_{21}X_2 - a_{22}X_4 + X_1(X_5 + U_1) = 0;$$
$$\dot{X}_3 + a_{31}X_3 - a_{32}X_1 - X_4U_1 = 0;$$
$$\dot{X}_4 + a_{41}X_4 - a_{42}X_2 + X_3U_1 = 0; \quad (12)$$
$$\dot{X}_5 - a_5M_d + a_6X_5 + a_7 = 0;$$
$$M_d = a_1(X_2X_3 - X_1X_4).$$

Допоміжний функціонал:

$$J = \int_{t_0}^{T} [G + \sum_{j=1}^{5} \lambda_j(t)\varphi_j(t)]dt, \quad (13)$$

де $G = K(X_5 + U_1)^2$, K – відповідне значення константи (*const*) закону керування.

Згідно методу рішення загальної задачі Лагранжа, допоміжне рівняння H для знаходження управління U_1, має вигляд:

$$H = K(X_5 + U_1)^2 +$$
$$+ \lambda_1(\dot{X}_1 + a_{11}X_1 - a_{12}X_3 - X_2(X_5 + U_1) - U_2(X_5 + U_1)) +$$
$$+ \lambda_2(\dot{X}_2 + a_{21}X_2 - a_{22}X_4 + X_1(X_5 + U_1)) +$$
$$+ \lambda_3(\dot{X}_3 + a_{31}X_3 - a_{32}X_1 - X_4U_1) +$$
$$+ \lambda_4(\dot{X}_4 + a_{41}X_4 - a_{42}X_2 + X_3U_1) +$$
$$+ \lambda_5(\dot{X}_5 - a_8(X_2X_3 - X_1X_4) + a_6X_5 + a_7). \quad (14)$$

На основі функції H (14) відносно функцій U_1, X_j, λ_j складаються рівняння Ейлера ($j = \overline{1,5}$):

$$\frac{\partial H}{\partial U_i} - \frac{d}{dt}\frac{\partial H}{\partial \dot{U}_i} = 0. \quad (15)$$

Тут в рівнянні (15) під U_i розуміються функції U_j, X_j, λ_j. В результаті отримаємо систему рівнянь за кількістю невідомих $U_j(t)$, $X_j(t)$, $\lambda_j(t)$, $j = \overline{1,n}$. Використовуючи співвідношення (13) знаходимо управління U_1:

$$U_1 = \frac{\lambda_1(X_2 + U_2) - \lambda_2X_1 + \lambda_3X_4 - \lambda_4X_3}{2K}. \quad (16)$$

Система диференційних рівнянь для знаходження фазових змінних X_j згідно (12) для ТАД з електричними параметрами та коефіцієнтами

$L_m = 0.09172$ (мГн); $L_r = 0.092819$ (мГн);

$L_s = 0.093331$ (мГн); $R_r = 0.0676$ (Ом);

$R_s = 0.0831$ (Ом); $p/J = 0.000051$; $i = 3.69$; p = 3;

$$k_s = \frac{L_m}{L_s} = 0.9827;\ k_r = \frac{L_m}{L_r} = 0.9882;$$

$$\sigma = 1 - k_sk_r = 0.0289;\ T_s = \frac{L_s}{R_s} = 1.123;\ T_r = \frac{L_r}{R_r} = 1.373;$$

$$a_s = \frac{1}{\sigma T_s} = 30.81;\ a_r = \frac{1}{\sigma T_r} = 25.20;\ a_{11} = a_s = 30.81;$$

$$a_{32} = a_{42} = a_rk_s = 24.76;\ a_{21} = a_s = 30.81;$$

$$a_{42} = a_rk_s = 24.76;\ a_{31} = a_r = 25.20;\ a_{12} = a_sk_r = 30.45;$$

$$a_{22} = a_{12} = 30.45;\ a_{41} = a_{31} = 25.20;\ a_1 = \frac{3}{2}p\frac{k_R}{\sigma L_S} =$$

$$= 1648.67;\ a_2 = p/J = 0.000051;\ a_5 = 0.0028;$$

$$a_6 = 0.00043;\ a_7 = 0.254;\ a_8 = 4.62$$

має вигляд:

$$\dot{X}_1 + 30.81X_1 - 30.45X_3 - X_2(X_5 + U_1) - U_2(X_5 + U_1) = 0;$$

$$\dot{X}_2 + 30.81X_2 - 30.45X_4 + X_1(X_5 + U_1) = 0;$$

$$\dot{X}_3 + 25.20X_3 - 24.76X_1 - X_4U_1 = 0;$$

$$\dot{X}_4 + 25.20X_4 - 24.76X_2 + X_3U_1 = 0; \quad (17)$$

$$\dot{X}_5 - 0.0028M_d + 0.00043X_5 + 0.254 = 0;$$

$$M_d = 1648.67(X_2X_3 - X_1X_4).$$

Система диференційних рівнянь для знаходження множників Лагранжа λ_j:

$$\dot{\lambda}_1 = 30.81\lambda_1 + \lambda_2(X_5 + U_1) - 24.76\lambda_3 + 4.62\lambda_5X_4;$$

$$\dot{\lambda}_2 = -\lambda_1(X_5 + U_1) + 30.81\lambda_2 - 24.76\lambda_3 - 4.62\lambda_5X_3;$$

$$\dot{\lambda}_3 = -30.45\lambda_1 + 25.20\lambda_3 + \lambda_4U_1 - 4.62\lambda_5X_2; \quad (18)$$

$$\dot{\lambda}_4 = -30.45\lambda_2 - \lambda_3U_1 + 25.20\lambda_4 + 4.62\lambda_5X_1;$$

$$\dot{\lambda}_5 = -\lambda_1(X_2 + U_2) + \lambda_2X_1 + 0.00043\lambda_5.$$

Аналіз системи рівнянь (18) та результатів імітаційного моделювання демонструє принципову неможливість отримання аналітичного розв'язку даної системи. Обчислювальна складність зумовлена наявністю в структурі моделі контуру з позитивним зворотним зв'язком, що призводить до нестійкості системи та виключає можливість прямого визначення множників Лагранжа.

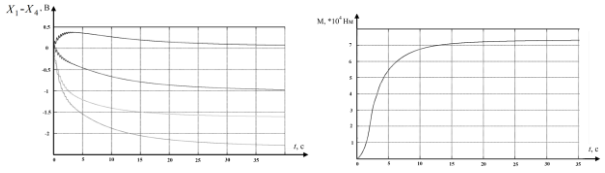
Одночасно аналітичне дослідження підтверджує, що формування керуючих впливів згідно з рівнянням (16) забезпечує кореляцію між динамікою зміни ковзання та темпом розгону рухомого складу. Зокрема, з рівняння (16) випливає обернено пропорційна залежність між величиною ковзання та коефіцієнтом розгону K, що відповідає фізичній сутності процесів керування тяговим електроприводом.

У зв'язку з виявленими обчислювальними обмеженнями для вирішення задачі оптимізації запропоновано перехід до синтезу субоптимальних керувань. На основі рівнянь (17) розроблено структурну схему моделі в середовищі MATLAB, яка пройшла процедуру верифікації на предмет адекватності. Експериментальні дослідження включали порівняльний аналіз різних стратегій керування: – розімкненої системи з законом *U/f = const* – замкнутої системи із зворотним зв'язком за частотою напруги живлення ТАД.

Для подолання обчислювальних складнощів, пов'язаних із системою (18), запропоновано метод ідентифікації множників Лагранжа на основі аналізу динаміки електромагнітних процесів, що описуються системою диференціальних рівнянь (17). Оскільки диференційні рівняння (18) знаходились на основі рівнянь (17), використовуючи часткові похідні згідно методики рішення загальної задачі Лагранжа, то можливо стверджувати, що і характер протікання процесів λ_j будуть аналогічними, як і при рішенні системи (17), де немає проблем з їх моделюванням.

На рис. 1 приведено графіки перехідних процесів при розгоні дизель-поїзда, як результати моделювання системи рівнянь (17) – фазових координат (потокозчеплень) $X_1 - X_4$ (а) та графік зміни тягового моменту дизель-поїзда (б).

Проведене моделювання дозволило провести кількісну та якісну оцінку динамічних характеристик системи. Аналіз перехідних процесів показує, що отримані залежності мають неперервний характер із чітко вираженою експоненційною компонентою, тобто вони можуть бути представлені у вигляді суми експонент та відповідних констант.



а б

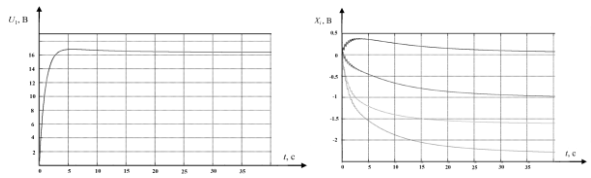
Рис. 1. Графіки зміни потокозчеплень (а) та тягового моменту ТАД (б)

Для визначення оптимальних керуючих впливів згідно з рівнянням (16) було застосовано емпіричний підхід до знаходження множників Лагранжа (λ_j) із системи (18). Методологія базувалася на наступному алгоритмі: 1. Використання апріорної інформації про діапазони зміни фазових координат у рівняннях (17). 2. Врахування обмежень на величину керуючого впливу (ковзання). 3. Ітеративний пошук оптимальних значень шляхом комбінації аналітичних методів та методів перебору. 4. Валідація результатів на основі аналізу енерговитрат.

Такий підхід знаходить підтвердження в науковій літературі [8, 9] і дозволяє ефективно визначати параметри системи на основі експертних оцінок та аналізу фазових траєкторій.

Виявлення експоненційного характеру перехідних процесів дозволило спростити процедуру синтезу субоптимальних управлінь. В результаті рішення задачі знаходження управлінь (закону зміни ковзання U_1) за умови використання темпу розгону дизель-поїзда та забезпечення при цьому мінімальних затрат енергії забезпечується безпосередньо апроксимацією функції, яку отримано в процесі моделювання.

На рис. 2 приведено графік функції закону зміни ковзання U_1 та відповідні (характерні) фазові координати X_i.

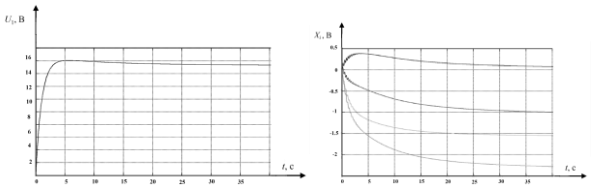


а б

Рис. 2. Графік зміни ковзання U_1 (а) та фазових координат X_i (б)

Задання функції закону зміни ковзання U_1 в процесі моделювання та при технічній реалізації САР можливо у вигляді аперіодичної ланки першого порядку.

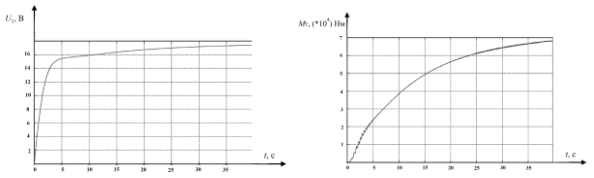
На рис. 3 приведено графік функції закону зміни ковзання U_1 (а) та відповідні (характерні) фазові координати X_i (б) за умови формування U_1 згідно співвідношення (16) та фіксованому значені U/f = 14 при наступних значеннях λ_j: λ_1 = 36; λ_2 = 60; λ_3 = -120; λ_4 = -120. Процеси знаходяться в робочому діапазоні: електрична кругова частота ω_s = 261/1 (1/сек) (41.6 Гц), напруга живлення U_M = 917.5(В) при моделюванні на 50 сек. розгону дизель-поїзда.



а б

Рис. 3. Графік зміни ковзання U_1 (а) та фазових змінуих X_i (б)

На рис. 4 приведено графік функції закону зміни ковзання U_1 (а) та графік зміни тягового моменту M_T (б) за умови формування U_1 згідно співвідношення (16) та формуванні співвідношення U/f = var згідно експоненціального закону при наступних значеннях λ_j: λ_1 = 36; λ_2 = 60; λ_3 = - 60; λ_4 = - 60. Процеси знаходяться в робочому діапазоні: в кінці розгону електрична кругова частота ω_s = 368.9 (1/сек) (58.7 Гц), напруга живлення U_M = 1022.0 (В). При цьому значення згідно експоненціального закону *U/f* змінювалось від 0 до 17.35 (В*с).



а б

Рис. 4. Графік зміни ковзання U_1 (а) та тягового моментуM_T (б)

В табл. 1 приведено результати дослідження запропонованих математичних моделей для синтезу систем керування тяговим електроприводом змінного струму дизель-поїзда за допомогою математичних моделей, які реалізовані в пакеті MATLAB. Розглянуто варіанти формування управлінь (ковзання частоти ТАД U_1) як з використанням невизначених множників Лагранжа λ_j (п/п. 1-10, 14), так і формування управлінь при заданій формі відповідно до експоненціального закону (п/п. 11 - 13). Результати п/п. 14 – формування управлінь (ковзання частоти ТАД U_1) з використанням невизначених множників Лагранжа λ_j та формуванні співвідношення *U/f* = *var* згідно експоненціального закону.

П/п. 1 – 8 відображають результати формування управління U_1 в можливому робочому діапазоні функціонування електроприводу в процесі розгону

дизель-поїзда (від 1.16 Гц до 2.8 Гц) при використанні отриманого співвідношення на основі інформації фазових координат X_j та значень λ_j множників Лагранжа. Приведені результати досліджень характеризують можливий процес формування управлінь (ковзання частоти ТАД U_1) згідно отриманих аналітичних співвідношень як для випадку *U/f = const*, так і для випадку *U/f = var* за умови забезпечення оптимізації енергетичних показників в процесі розгону.

Таблиця 1 – **Результати дослідження**

№ п/п	λ_1	λ_2	λ_3	λ_4	U_1, 1/сек	U / f В*сек	ω_s, 1/сек	U_M, В	Примітка
1	36	60	-120	-120	17.32	18	611.4	1753.0	
2	36	60	-120	-120	17.36	16	360.1	917.5	
3	36	60	-120	-120	17.42	14	261.1	582.1	Рис. 2
4	18	60	-120	-120	11.72	18	373.9	1072.0	
5	9	60	-120	-120	9.0	18	298.9	856.6	
6	9	60	-120	-120	9.0	19	341.4	1033.0	
7	3	60	-120	-120	7.29	19	269.8	816.3	
8	3	60	-120	-120	7.30	20	307.2	978.2	
9	36	60	-60	-60	16.37	18	461.1	1322.0	
10	36	60	-60	-60	16.39	16	349.3	890.0	
11					16.0	16	347.2	884.5	
12					18.0	16	354.1	902.1	
13					20.0	16	360.6	918.8	
14	36	60	-60	-60	16.43	Експоненціальний	368.9	1022.0	Рис. 4

Висновки

1. Розроблено комплекс програмних компонентів у середовищі MATLAB/Simulink для дослідження систем керування тяговим асинхронним електроприводом дизель-поїздів. Створено спеціалізовані бібліотеки блоків, що дозволяють моделювати динамічні характеристики системи та оптимізувати її енергоефективність.

2. Розроблено програмні модулі для синтезу оптимальної системи автоматичного регулювання електроприводу, що реалізують алгоритми формування керувань ковзанням частоти напруги живлення тягового асинхронного двигуну на основі модифікованого закону:

U/f=const.

3. Реалізовано програмні компоненти для чисельної реалізації методу варіаційного числення, що дозволяють вирішувати задачі синтезу керувань для об’єктів, що описуються системами диференціальних рівнянь п’ятого порядку.

4. Створено програмну реалізацію математичної моделі тягового електроприводу в синхронній системі координат у вигляді спеціалізованих блоків Simulink, що враховують ступінь завантаження дизель-поїзда.

5. Розроблено програмні модулі для реалізації методу Лагранжа, що включають блоки формування квадратичного функціоналу якості та алгоритми його мінімізації.

6. Створено спеціалізовані обчислювальні процедури для визначення невизначених множників Лагранжа, що реалізують стабілізуючі алгоритми для усунення проблем нестійкості диференціальних рівнянь.

7. Розроблено альтернативні програмні модулі формування керувань ковзанням частоти ТАД, що реалізують як методи на основі множників Лагранжа, так і експоненціальні закони керування.

8. Інтегровано аналітичні співвідношення для синтезу керувань у вигляді програмних блоків, що враховують номінальне навантаження дизель-поїзда для різних режимів роботи.

9. Розроблено комплекс програмних засобів верифікації моделі, що включають модулі побудови графіків фазових змінних та порівняльних таблиць ефективності різних стратегій керування.

10. Створено імітаційну модель системи керування, що дозволяє експериментально підтвердити її працездатність та енергоефективність при різних режимах розгону.

11. Розроблено архітектуру програмної платформи для подальших досліджень, що забезпечує можливість інтеграції методів експертних оцінок та машинного навчання для визначення оптимальних параметрів системи керування.

Список літератури

1. Horstmann D. 100 Jahre Entwicklung der Antriebstechnik für elektrische Bahnen. Teil 2 / D. Horstmann, R. Wagner, W-D. Weigel. *Elektrische Bahnen*. 2003. No 7. P. 338-345. URL: https://www.researchgate.net/publication/

2. Bose B. K. Modern Power Electronics and AC Drives / B. K. Bose. Prentice Hall PTR: Prentice-Hall Inc., 2002. 712 p. URL: https://www.academia.edu/41701330/MoDERN_PowER_ELECTRONICS_AND_AC_DRIVES
3. Volkov A. V., Kosenko I. A. Asynchronous motor drive based on self-excited current inverter with switched-off thyristors and provided with predicting relay and vector regulation of stator current. Elektrotekhnika. 2008. No 10. P. 6–17. URL: https://link.springer.com/article/10.3103/S1068371208100039
4. Носков В. И., Дмитриенко В. Д., Заполовский Н. И., Леонов С. Ю. Моделирование и оптимизация систем управления и контроля локомотивов. Х. : Транспорт Украины, 2003. 248 с. URL: https://scholar.google.com/scholar?hl=en&as_sdt=0,5&cluster=8949436925584089778
5. Заполовський М. Й., Мезенцев М. В., Скородєлов, В. В. Математична модель для синтезу управлінь електроприводом змінного струму. Системи управління, навігації та зв'язку. Вип. 5(57). Полтава: Нац. ун-т «Полтавська політехніка ім. Ю. Кондратюка», 2019. С. 16–21. DOI: https://doi.org/10.26906/SUNZ.2019.5.016
6. Заполовський М. Й., Мезенцев М. В., Баленко О. І., Оліфір М. В. Оптимізаційна модель тягового асинхронного електроприводу дизель-поїзда та її дослідження. *Системи управління, навігації та зв'язку*. Вип. 3. Полтава: Нац. ун-т «Полтавська політехніка ім. Ю. Кондратюка», 2023. С. 45–53. DOI: https://doi.org/10.26906/SUNZ.2023.3.045
7. Толочко О. І. Моделювання електромеханічних систем. Математичне моделювання систем асинхронного електроприводу. Київ : НТУУ «КПІ», 2016. 150 с. URL: https://ela.kpi.ua/server/api/core/bitstreams/6fc21e95-34d2-4a4a-bb67-f53d0a327a4e/content
8. Моклячук М. П. Варіаційне числення. Екстремальні задачі. К. : Видавничо-поліграфічний центр "Київський університет", 2009. 380 с. URI: https://mechmat.knu.ua/wp-content/uploads/2020/05/var-book-2010.pdf
9. Носков В.І., Гавриленко С.Ю., Скородєлов В.В., Гейко М.В. Розробка методу оцінки показників тягових електропередач автономних локомотивів. *Системи управління, навігації та зв'язку*. Вип. 3 (73). Полтава: Нац. ун-т «Полтавська політехніка ім. Ю. Кондратюка», 2023. С. 54–57. DOI: https://doi.org/10.26906/SUNZ.2023.3.054

Received (Надійшла) 15.12.2025
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Заполовський Микола Йосипович - кандидат технічних наук, професор, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет України «Харківський політехнічний інститут, Харків, Україна;
Mykola Zapolovskyi – Candidate of Technical Sciences, Professor, Professor, Department of Computer Engineering and Programming, National Technical University of Ukraine "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: zapolovsky@email.ua; ORCID Author ID: http://orcid.org/0000-0002-7617-9700;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57194577942.

Мезенцев Микола Вікторович – кандидат технічних наук, доцент, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет України «Харківський політехнічний інститут, Харків, Україна;
Mykola Mezentsev – Candidate of Technical Sciences, Associate Professor, Professor, Department of Computer Engineering and Programming, National Technical University of Ukraine "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: mykola.mezentsev@khpi.edu.ua ORCID Author ID: https://orcid.org/0000-0001-7834-2797
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57210418332.

Оліфір Максим Вікторович – аспірант кафедри комп'ютерної інженерії та програмування, Національний технічний університет України «Харківський політехнічний інститут, Харків, Україна;
Maksym Olifir – PhD student, Department of Computer Engineering and Programming, National Technical University of Ukraine "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Maksym.Olifir@cs.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0008-6956-1208.

Mathematical model and software-algorithmic components for the synthesis of a slip control system of the supply-voltage frequency of a traction asynchronous motor

Mykola Zapolovskiy, Mykola Mezentsev, Maksym Olifir

Abstract. Relevance. In the synthesis of optimal control systems, a number of tasks arise, including the development of models of the control object, the choice of an optimization method, and the subsequent implementation and investigation of the obtained control laws in the form of software-hardware solutions. **Object of research:** software components for the control system of a diesel-train traction electric drive. **Purpose of the article:** to create software-algorithmic components for the control system of a diesel-train traction drive based on the synthesis of control laws for slip of the supply-voltage frequency of a traction asynchronous motor. **Research results.** The article proposes an approach to solving problems of control-system synthesis based on the general Lagrange problem and develops a method for determining the unknown Lagrange multipliers. Based on the mathematical model of the control system, analytical relations for calculating optimal control actions are obtained. These relations are integrated into a computer model of the electromechanical system deployed in MATLAB/Simulink, which made it possible to perform a cycle of virtual tests. **Conclusions.** Implementing this approach expands the possibilities for designing optimal control systems for AC traction drives of vehicles. The simulation results demonstrate that the proposed algorithm, implemented as Simulink functional blocks, ensures stable drive operation throughout the entire traction-characteristic range with optimized energy consumption. The developed models and algorithms are ready-to-use tools for implementation in microcontrollers of modern automation systems of traction electric drives.

Keywords: software-algorithmic components; traction electric drive; diesel train; Lagrange method; computer simulation; functional blocks.

Інформаційні технології

УДК 004.9 **doi: 10.26906/SUNZ.2026.2.040**

С. В. Бондаренко[1], В. О. Мартовицький[1], Н. М. Бологова[1], В. Г. Рикун[2]

[1] Харківський національний університет радіоелектроніки, Харків, Україна
[2] Харківський національний університет Повітряних Сил ім. І. Кожедуба, Харків

ПЛАНУВАННЯ ЗАДАЧ У БАГАТОПРОЦЕСОРНИХ СИСТЕМАХ НА ОСНОВІ ГІБРИДНИХ МЕТОДІВ

Анотація. Стрімкий розвиток багатопроцесорних та розподілених обчислювальних систем обумовлює необхідність підвищення ефективності планування задач у гетерогенних середовищах. Об'єктом дослідження є процес планування задач у багатопроцесорних системах з неоднорідними обчислювальними ресурсами. Метою роботи є підвищення ефективності формування розкладу шляхом розробки гібридного підходу, що поєднує класичні евристичні алгоритми та методи навчання з підкріпленням. У роботі проведено аналіз існуючих методів планування задач, включаючи спискові евристики та підходи на основі машинного навчання. Запропоновано функціональну модель планування, яка інтегрує алгоритми HEFT і CPOP з агентом навчання з підкріпленням, реалізованим за схемою Actor–Critic із використанням алгоритму Proximal Policy Optimization. Розроблено математичну модель середовища планування, що враховує структуру DAG, матрицю очікуваного часу виконання та показники балансування навантаження. Для експериментальної перевірки створено потокову фабрику генерації різноманітних сценаріїв та реалізовано векторизоване середовище навчання. Результати експериментального дослідження показали, що запропоновані гібридні методи забезпечують стабільне зменшення makespan до 5-7% та 13-15% та середнього часу завершення задач порівняно з базовими евристиками, при контрольованому впливі на показники балансування навантаження. Отримані результати підтверджують можливість керованого компромісу між мінімізацією часу виконання та рівномірністю використання ресурсів. Наукова новизна роботи полягає у розробці узагальненої функціональної моделі гібридного планування задач, що забезпечує адаптивне коригування евристичних алгоритмів на основі навчання з підкріпленням у гетерогенному багатопроцесорному середовищі. Практична значущість полягає у можливості застосування запропонованого підходу в системах високопродуктивних, хмарних та розподілених обчислень для підвищення продуктивності та ефективності використання ресурсів.

Ключові слова: планування задач; багатопроцесорні системи; гібридні методи; навчання з підкріпленням; проксимальна оптимізація політики; критичний шлях на процесорі; балансування навантаження.

Вступ

Постановка проблеми. Стрімкий розвиток інформаційних технологій та зростання обсягів даних, що потребують обробки в режимі реального часу, призвели до повсюдного впровадження багатопроцесорних та розподілених обчислювальних систем. Сучасні архітектури, від хмарних (Cloud) та туманних (Fog) обчислень до вбудованих систем реального часу, вимагають ефективного управління ресурсами для забезпечення високої продуктивності та надійності. Ключовим елементом такого управління є планування задач (task scheduling), яке визначає порядок виконання завдань та їх розподіл між доступними процесорами. Як зазначають дослідники у свіжих оглядах, ефективність планування безпосередньо впливає на загальну пропускну здатність системи, час відгуку та енергоспоживання [1].

Планування задач у багатопроцесорних системах належить до класу NP-важких (NP-hard) задач комбінаторної оптимізації. Це означає, що зі збільшенням кількості завдань та процесорів складність знаходження оптимального розв'язку зростає експоненціально, що робить неможливим використання методів повного перебору для систем реальної розмірності. Особливої гостроти проблема набуває в системах реального часу, де порушення часових обмежень (deadlines) може призвести до критичних збоїв у роботі всієї інфраструктури [2]. Традиційні підходи до планування, такі як статичні евристики (наприклад, Round Robin, First-Come-First-Served) або класичні алгоритми спискового планування, часто виявляються неефективними в умовах динамічного навантаження та гетерогенності ресурсів. Водночас сучасні вимоги до обчислювальних систем стають багатоцільовими (multi-objective): окрім мінімізації часу виконання (makespan), необхідно враховувати вартість обчислень, балансування навантаження та енергоефективність. Наприклад, останні дослідження підкреслюють важливість одночасної оптимізації вартості та енергоспоживання в середовищах хмарних та граничних обчислень (Cloud-Edge computing), що вимагає застосування більш адаптивних інтелектуальних підходів [3].

Для подолання обмежень класичних методів в останні роки активно розробляються алгоритми, що базуються на метаевристиках (генетичні алгоритми, ройовий інтелект) та методах машинного навчання. Проте, як свідчить аналіз методів планування в системах реального часу, кожен окремий підхід має свої недоліки: евристики можуть "застрягати" в локальних оптимумах, а методи глибокого навчання потребують значних ресурсів для тренування моделей [4]. Саме тому науковий інтерес зміщується в бік гібридних методів (hybrid methods), які поєднують сильні сторони різних підходів. Гібридизація дозволяє компенсувати недоліки одного методу перевагами іншого, наприклад, використовуючи евристики для початкової ініціалізації популяції в генетичних алгоритмах або застосовуючи локальний пошук для покращення результа-

тів глобальної оптимізації. Дослідження показують, що багатоцільові гібридні алгоритми здатні забезпечити кращий баланс між часом пошуку розв'язку та його якістю порівняно з монолітними підходами [5].

Незважаючи на значний прогрес у цій галузі, проблема розробки універсальних та ефективних гібридних методів планування залишається відкритою. Існує потреба в детальному дослідженні механізмів поєднання різних алгоритмічних стратегій для адаптації до динамічних змін у багатопроцесорних середовищах. Це дослідження спрямоване на розробку та аналіз підходу до планування задач на основі гібридних методів, що дозволить підвищити ефективність використання ресурсів багатопроцесорних систем.

Аналіз останніх досліджень і публікацій. Проблема ефективного розподілу обчислювального навантаження у багатопроцесорних системах є предметом ґрунтовних наукових досліджень протягом останніх десятиліть. Аналіз літературних джерел дозволяє стверджувати, що підходи до її вирішення еволюціонували від класичних математичних моделей до складних адаптивних систем штучного інтелекту. Фундаментальну основу для розуміння процесів планування закладено в теорії масового обслуговування (Queuing Theory). Як зазначається у роботі [6], використання моделей черг дозволяє оцінити базові характеристики продуктивності системи, такі як середній час очікування задачі та коефіцієнт завантаження процесорів. Проте, класичні ймовірнісні моделі часто ідеалізують реальні умови функціонування, не враховуючи гетерогенність сучасних обчислювальних вузлів та складні залежності між задачами.

Для більш детального моделювання часових характеристик та синхронізації процесів у паралельних системах дослідники часто звертаються до апарату мереж Петрі. Зокрема, у дослідженні [7] показано, що цей математичний інструмент дозволяє ефективно виявляти потенційні блокування (deadlocks) та конфлікти доступу до ресурсів ще на етапі проектування розкладу. Однак, зі зростанням кількості вузлів та завдань, простір станів мережі Петрі вибухоподібно зростає, що ускладнює її використання для динамічного планування в режимі реального часу (online scheduling). Існуючі методи планування традиційно поділяють на статичні та динамічні. Статичні підходи, розглянуті у роботі [8], передбачають, що характеристики всіх завдань (час виконання, обсяг даних, директивні строки) відомі заздалегідь. Це дозволяє використовувати точні математичні методи або алгоритми на основі графа завдань (DAG – Directed Acyclic Graph) для побудови оптимального розкладу ще до початку виконання програм. Хоча такі методи гарантують детермінованість, вони виявляються безпорадними в умовах невизначеності, характерної для сучасних хмарних та туманних обчислень. У спробі подолати обмеження статичних методів та високу обчислювальну складність точних алгоритмів, значна частина досліджень була спрямована на розробку евристичних алгоритмів. Як зазначається в огляді методів та засобів планування [9], популярні евристики спискового планування (List Scheduling), такі як HEFT (Heterogeneous Earliest Finish Time) або CPOP (Critical Path on a Processor), забезпечують прийнятний компроміс між якістю розкладу та часом його генерації. Ці алгоритми базуються на пріоритезації завдань, часто використовуючи поняття критичного шляху в графі задач.

Розвиток цього напрямку призвів до появи рангових підходів. У роботі [10] досліджено застосування алгоритмів рангового підходу, де кожній задачі присвоюється певний ранг на основі її обчислювальної ваги та комунікаційних витрат. Це дозволяє більш гнучко керувати чергою виконання, проте більшість таких алгоритмів є «жадібними» (greedy) за своєю природою. Вони приймають локально оптимальні рішення на кожному кроці, що часто призводить до субоптимального глобального результату, особливо в системах з високим ступенем гетерогенності ресурсів.

Окремий клас проблем становить планування у розподілених системах реального часу, де порушення часових обмежень є неприпустимим. Дослідження [11] підкреслює, що в таких системах критерій оптимізації зміщується з мінімізації загального часу виконання (makespan) на максимізацію кількості завдань, виконаних до настання дедлайну. Традиційні евристики часто не здатні ефективно балансувати між цими суперечливими цілями, що спонукало науковців до пошуку більш досконалих методів оптимізації.

Недоліки детермінованих евристик призвели до широкого застосування метаевристичних методів, натхненних природними процесами. Узагальнюючий аналіз методів планування [12] виділяє генетичні алгоритми (GA), алгоритми імітації відпалу (Simulated Annealing) та ройовий інтелект (PSO, ACO) як потужні інструменти глобальної оптимізації. Ці методи здатні знаходити розв'язки, близькі до оптимальних, у величезних просторах пошуку, не вимагаючи повної інформації про систему. Проте, як вже згадувалося в попередніх оглядах [5], їхня головна вада – повільна збіжність та висока чутливість до налаштування гіперпараметрів, що робить їх проблематичними для використання у високодинамічних середовищах без додаткових модифікацій. Таким чином, аналіз класичних та метаевристичних методів демонструє, що жоден з «чистих» підходів не забезпечує універсального рішення для сучасних багатопроцесорних систем, які характеризуються динамічністю, гетерогенністю та багатокритеріальністю. Це створює передумови для звернення до методів машинного навчання та, зрештою, до гібридних архітектур, здатних адаптуватися до змін у навантаженні.

З появою концепцій інтелектуального управління ресурсами (Intelligent Resource Management), вектор наукових пошуків змістився в бік методів машинного навчання (Machine Learning, ML), зокрема навчання з підкріпленням (Reinforcement Learning, RL). На відміну від статичних евристик, RL-агенти здатні навчатися оптимальної стратегії планування через постійну взаємодію із середовищем, отримуючи "винагороду" за ефективні рішення. У роботі [13] пропонується підхід на основі ієрархічного глибокого навчання з підкріпленням (Hierarchical Deep Reinforcement Learning), який дозволяє декомпозувати складну задачу багатоцільового планування на підзадачі меншої розмірності. Це значно підвищує здатність системи адаптуватися до

непередбачуваних змін у потоці завдань, що є критичним для динамічних середовищ.

Особливу увагу в контексті планування залежних задач привертають графові нейронні мережі (Graph Neural Networks, GNN). Оскільки більшість паралельних програм моделюються як направлені ациклічні графи (DAG), використання GNN дозволяє ефективно витягувати просторові ознаки та залежності між вершинами графа. Дослідники [14] демонструють ефективність поєднання GNN з мультиагентним навчанням з підкріпленням (MARL) для задач розміщення навантаження в граничних обчисленнях (Edge Computing). Такий підхід дозволяє враховувати топологію мережі та мінімізувати затримки передачі даних. Аналогічний напрямок розвивається у роботі [15], де представлена архітектура ScheduleNet. Вона дозволяє агентам "навчитися" розв’язувати проблеми планування, оперуючи безпосередньо структурою графа завдань, що забезпечує кращу генералізацію на нові, раніше не бачені типи навантажень.

Практичне впровадження інтелектуальних методів знаходить своє відображення і в популярних системах оркестрації контейнерів. Зокрема, у дослідженні [16] розглядається вдосконалення стандартного планувальника Kubernetes за допомогою методів глибокого навчання та RL. Автори підтверджують, що інтеграція навчених моделей у реальні кластери дозволяє суттєво знизити час очікування подів та покращити утилізацію кластера порівняно зі стандартними алгоритмами пакування (bin packing). Також існують спроби вдосконалення самих нейромережевих архітектур для специфічних задач багатопроцесорної обробки, як показано у [17], де модифікована нейронна мережа використовується для прискорення збіжності процесу пошуку розкладу. Втім, попри значні переваги, методи на основі чистого машинного навчання мають суттєві обмеження: вони вимагають значних обчислювальних ресурсів для тренування, страждають від проблеми "холодного старту" та можуть демонструвати нестабільну поведінку при різкій зміні розподілу вхідних даних. Саме тому сучасна наукова думка схиляється до гібридних методів, які інтегрують швидкість та надійність евристик із адаптивністю штучного інтелекту. Наприклад, використання евристик для початкового розподілу або як "страхувального механізму" для RL-агента дозволяє нівелювати ризики та підвищити загальну надійність системи. Для верифікації таких складних гібридних моделей необхідні спеціалізовані інструменти, і, як зазначено в роботі [18], розробка адекватних систем моделювання є невід’ємною частиною процесу створення нових алгоритмів планування.

Постановка проблеми. Проведений аналіз літературних джерел дозволяє виявити суттєве протиріччя. З одного боку, сучасні багатопроцесорні системи вимагають планування, яке є одночасно швидким (як статичні евристики) та адаптивним до динамічних змін (як методи машинного навчання). З іншого боку, існуючі "чисті" підходи не здатні повною мірою задовольнити обидві вимоги одночасно: евристики дають локальні оптимуми та не адаптуються, а методи RL є ресурсомісткими та складними у навчанні. **Актуальною науковою проблемою** є розробка методології планування, яка б ефективно поєднувала різні класи алгоритмів у єдину гібридну структуру. Більшість існуючих гібридних рішень фокусуються на конкретних вузьких сценаріях (наприклад, тільки хмарні обчислення або тільки енергоефективність) і недостатньо досліджують загальні принципи взаємодії компонентів гібридної системи для досягнення балансу між часом планування та якістю розкладу (makespan, load balancing).

Таким чином, проблемою дослідження є недостатня ефективність існуючих методів планування задач у гетерогенних багатопроцесорних системах в умовах змінного навантаження. Вирішення цієї проблеми потребує розробки нового гібридного підходу, який інтегрує евристичні методи з адаптивними механізмами, забезпечуючи підвищення продуктивності системи без надмірних накладних витрат. Це обумовлює необхідність створення функціональної моделі такого планування та проведення експериментальних досліджень її ефективності.

Основний матеріал

Для досягнення поставленої мети та вирішення задач дослідження необхідно чітко визначити методологічну базу. Об’єктом дослідження є процес планування обчислювальних задач у багатопроцесорних системах з неоднорідними ресурсами. Предметом дослідження виступають гібридні методи та алгоритми, що забезпечують оптимізацію розподілу навантаження та мінімізацію часу виконання завдань. Основою для проведення експериментів та верифікації запропонованого підходу є розроблена математична модель, яка дозволяє формалізувати параметри обчислювального середовища та структуру завдань. Загальна структура математичної моделі, що відображає взаємозв'язок між параметрами вхідних даних, структурою графа задач (DAG), моделлю платформи та метриками ефективності, наведена на рис. 1.

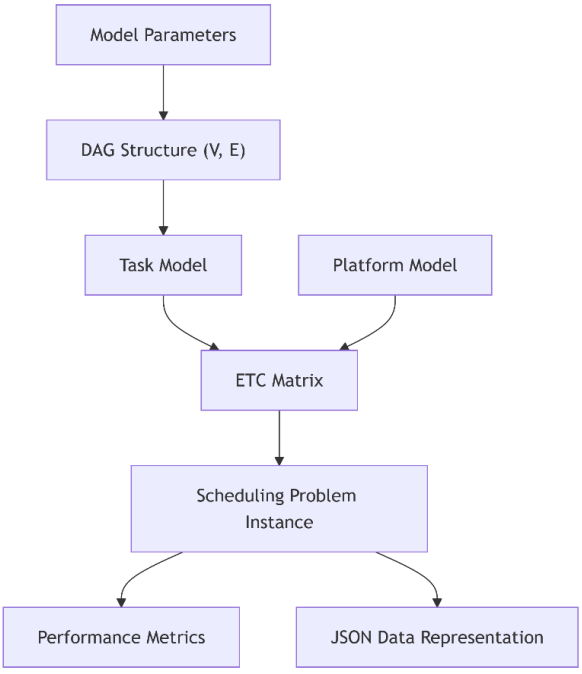


Рис. 1. Загальна структура математичної моделі

У даній роботі задача планування розглядається для моделі workflow, яка представлена у вигляді орієнтованого ациклічного графа (DAG):

$$G = (V, E), \quad (1)$$

де $V = \{v_1, v_2, \ldots, v_n\}$ – множина задач (вузлів графа), а E⊆V×V – множина залежностей між задачами. Наявність ребра $(v_i, v_j) \in E$ означає сувору технологічну залежність: задача v_j може розпочатися лише після повного завершення задачі v_i та отримання необхідних даних. Кожна задача $v_j \in V$ в моделі описується кортежем:

$$v_i = (id_i, level_i, type_i), \quad (2)$$

де id_i – унікальний ідентифікатор задачі, $type_i$ – тип задачі (обчислювальна, об'єднання даних або введення/виведення), а $level_i$ – топологічний рівень задачі у DAG. Топологічний рівень є критично важливим параметром для рангових евристик і визначається рекурсивно:

$$level_i = \begin{cases} 0, & \text{if } pred(v_i) = \varnothing, \\ \max_{v_j \in pred(v_i)}(level_j) + 1, & \text{otherwise}, \end{cases} \quad (3)$$

де $pred(v_i)$ – множина безпосередніх попередників задачі v_i. Процес формування параметрів задач та платформи візуалізовано на рис. 2.

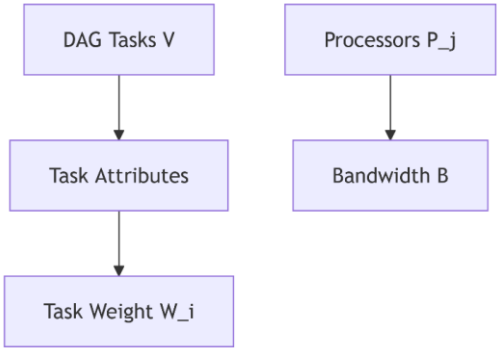


Рис. 2. Формування параметрів задач і платформи

Обчислювальна складність кожної задачі W_i не є статичною величиною, а моделюється як функція від її типу та рівня в графі, що дозволяє імітувати реальні сценарії, де складність обробки зростає по мірі проходження даних через конвеєр:

$$W_i = f(type_i, level_i). \quad (4)$$

Функція $f(\cdot)$ задається наступним чином:

$$W_i = \begin{cases} \alpha \cdot g(level_i), & \text{if } type_i = compute, \\ \beta \cdot g(level_i), & \text{if } type_i = merge, \\ \gamma, & \text{if } type_i = io, \end{cases} \quad (5)$$

де α,β,γ – коефіцієнти складності для різних типів задач, а $g(level_i)$ – монотонно зростаюча функція.

Паралельна обчислювальна система (платформа) моделюється множиною гетерогенних процесорів $P = \{p_1, p_2, \ldots, p_m\}$. Кожен процесор p_j характеризується коефіцієнтом відносної продуктивності $speed_j > 0$. Ключовим елементом моделі є матриця очікуваного часу виконання (Expected Time to Compute, ETC):

$$ETC = [ETC(i, j)]_{n \times m}, \quad (6)$$

де елемент *ETC(i,j)* визначає час виконання задачі v_i на процесорі p_j. Для наближення до реальних умов вводиться випадкове збурення $\epsilon_{i,j}$:

$$ETC(i, j) = \frac{w_i}{speed_j}(1 + \epsilon_{i,j}). \quad (7)$$

Окрім обчислень, модель враховує комунікаційні витрати. Кожне ребро $(v_i, v_j) \in E$ асоціюється з обсягом даних $data\_size_{i,j}$. Час передачі даних $Comm_{i,j}$ залежить від пропускної здатності каналу B та розміщення задач:

$$Comm_{i,j} = \begin{cases} \frac{data\_size_{i,j}}{B}, & \text{if } p(i) \neq p(j), \\ 0, & \text{if } p(i) = p(j), \end{cases} \quad (8)$$

де *p(i)* та p(j) – процесори, на які призначено задачі v_i та v_j відповідно. Для оцінки ефективності розроблених методів планування використовується набір метрик, взаємозв'язок яких показано на рис. 3.

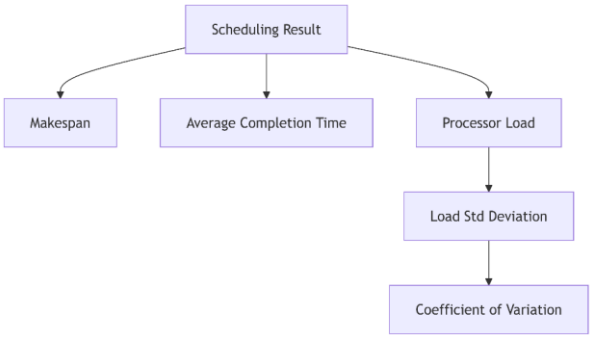


Рис. 3. Метрики оцінювання

Часові обмеження виконання визначаються двома параметрами: часом можливого початку $EST(v_i)$ та часом завершення $FT(v_i)$.

$$EST(v_i) = \max_{v_j \in pred(v_i)}(FT(v_j) + Comm_{i,j}), \quad (9)$$

$$FT(v_i) = EST(v_i) + EST(i, p(i)). \quad (10)$$

Основним критерієм оптимізації є Makespan (загальний час виконання workflow):

$$Makespan = \max_{v_i \in V} FT(v_i). \quad (11)$$

Додатковою метрикою, що дозволяє оцінити загальну затримку виконання незалежно від критичного шляху, є середній час завершення задач:

$$AvgCompletionTime = \frac{1}{|V|}\sum_{v_i \in V} FT(v_i). \quad (12)$$

Для аналізу якості використання ресурсів вводиться метрика балансування навантаження. Навантаження процесора p_j є сумою часів виконання всіх призначених на нього задач:

$$Load_j = \sum_{v_i \in V_j} ETC(i, j). \quad (13)$$

Якість розподілу оцінюється через середнє навантаження:

$$\overline{Load} = \frac{1}{m}\sum_{j=1}^{m} Load_j. \quad (14)$$

та його стандартне відхилення:

$$LB_{std} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(Load_j - \overline{Load})^2}. \quad (15)$$

Для можливості порівняння результатів на різних масштабах систем використовується нормалізований коефіцієнт варіації:

$$LB_{cv} = \frac{LB_{std}}{\overline{Load}}.$$

Менше значення LB_{cv} свідчить про більш рівномірний розподіл навантаження, що є індикатором ефективності гібридного методу планування.

Для проведення експериментального дослідження та верифікації розроблених методів, описану математичну модель необхідно трансформувати у формат, придатний для обробки програмним комплексом. У рамках даної роботи для серіалізації вхідних даних було обрано формат JSON через його наочність, розширюваність та легкість інтеграції з різними мовами програмування. Розроблена структура даних описує повний контекст одного експериментального сценарію, включаючи конфігурацію платформи, матрицю часу виконання та структуру графа завдань.

Структура вхідного файлу складається з трьох ключових блоків, що безпосередньо відображають компоненти математичної моделі: *processors* (модель платформи), ETC (матриця очікуваного часу виконання) та DAG (структурна модель workflow).

Блок processors являє собою масив об'єктів, що описує гетерогенний кластер обчислювальних вузлів *P*. Кожен об'єкт містить унікальний ідентифікатор *id* та коефіцієнт швидкодії *speed*, який відповідає параметру $speed_j$ у математичній моделі. Це дозволяє моделювати системи з різним ступенем гетерогенності, варіюючи значення швидкості процесорів.

Блок ETC реалізує матрицю *ETC(i,j)* у вигляді двовимірного масиву. Рядки масиву відповідають окремим завданням v_i, а стовпці – процесорам p_j. Значення у клітинці *[i][j]* визначає прогнозований час виконання *i*-ї задачі на j-му процесорі. Така структура дозволяє зберігати попередньо обчислені часові характеристики, що включають змодельовані випадкові збурення $\epsilon_{i,j}$.

Найбільш складною частиною структури є блок DAG, який описує граф завдань G(V,E) та складається з трьох підоб'єктів: dag_meta, tasks та edges.

Об'єкт dag_meta містить метадані, необхідні для налаштування генераторів навантаження та аналізу результатів. Сюди входять:

type – тип топології графа (наприклад, layered_fork_join);

num_tasks – загальна кількість вершин |V|;

max_level – глибина графа;

avg_parallelism – середня ширина графа, що впливає на потенціал розпаралелювання;

communication_model – параметри пропускної здатності мережі B.

Масив tasks відповідає множині вершин V. Кожен елемент масиву описує окрему задачу та містить поля *id* (ідентифікатор), *level* (топологічний рівень, необхідний для рангових евристик) та *type*. Поле type може приймати значення compute, merge або io, що визначає, яку саме формулу обчислення ваги W_i буде застосовано до даної задачі.

Масив edges описує множину дуг E та комунікаційні залежності. Кожен об'єкт зв'язку визначає задачу-попередника (from), задачу-наступника (to) та обсяг даних data_size, що передається між ними. Цей параметр використовується для розрахунку комунікаційної затримки $Comm_{i,j}$ у випадку, якщо зв'язані задачі плануються на різні процесори.

Загальний вигляд представлення даних у форматі JSON, що використовується в експериментах, наведено нижче:

Така організація даних забезпечує повну відповідність між теоретичною моделлю та її програмною реалізацією, дозволяючи проводити відтворювані експерименти для оцінки ефективності запропонованих гібридних методів планування.

Розроблена математична модель визначає задачу планування як проблему пошуку відображення множини задач V на множину процесорів P з метою мінімізації загального часу виконання workflow (*Makespan*) та забезпечення рівномірного розподілу навантаження між обчислювальними ресурсами (LB_{std}).

З урахуванням NP-складної природи задачі планування, а також гетерогенності сучасних багатопроцесорних систем, у даній роботі пропонується функціональна модель планування на основі гібридних методів, яка поєднує класичні евристичні алгоритми та методи глибокого навчання з підкріпленням (Deep Reinforcement Learning, DRL).

Функціональна модель системи планування формалізується у вигляді кортежу:

$$M = \langle E, S, A, H, \pi_\theta, R, L \rangle, \quad (17)$$

де V (Environment) – модель обчислювального середовища; S (State Space) – простір станів середовища; A (Action Space) – простір дій агента; H (Heuristic Guidance) – модуль евристичної підтримки, що використовується для маскування недопустимих дій та спрямування пошуку; π_θ (Policy) – політика RL-агента, що реалізована на базі FFN та оптимізується алгоритмом PPO у схемі Actor–Critic; R (Reward Function) – функція винагороди; L (Local post-processing) – необов'язковий модуль локального покращення (post-processing) готового розкладу.

Компонент E безпосередньо базується на математичній моделі і включає орієнтований ациклічний граф задач G=(V,E), матрицю очікуваного часу виконання ETC, характеристики гетерогенних процесорів P та модель комунікаційних витрат. Таким чином, середовище повністю визначає допустимі стани та обмеження планування. Стан системи на кожному кроці прийняття рішення t подається у вигляді вектора ознак фіксованої розмірності:

$$s_t = [\mathrm{DAG}_{\mathrm{embedding}}, \mathrm{P}_{\mathrm{status}}, \mathrm{Task}_{\mathrm{candidates}}], \quad (18)$$

де $\mathrm{DAG}_{\mathrm{embedding}}$ відображає агреговані характеристики поточного стану workflow (частка виконаних задач, оцінка довжини критичного шляху), $\mathrm{P}_{\mathrm{status}}$ – нормалізований вектор поточного завантаження процесорів:

$$P_{status} = [Load_1, Load_2, \ldots, Load_m], \quad (19)$$

а $\mathrm{Task}_{\mathrm{candidates}}$ описує множину готових до виконання задач, для яких виконані всі залежності $pred(v_i)$. Формування стану ґрунтується на метриках навантаження $Load_j$, визначених формулою (13), та оцінці критичного шляху DAG.

Простір дій A у загальному випадку визначається як вибір пари (v_i, p_j), де v_i – задача з множини готових, а p_j – процесор, на який вона призначається. Однак у запропонованій гібридній моделі RL-компонент використовується як спеціалізований модуль керування, тому структура простору дій A уточнюється залежно від схеми гібридизації.

У схемі Heuristic-guided RL для алгоритму HEFT (HG-RL–HEFT) евристика використовується для формування та ранжування множини готових задач:

$$Ready(t) = \{v \in V \mid pred(v) \subseteq Done(t)\}. \quad (20)$$

Із цієї множини формується підмножина $TopK(t)$ з k задач з найвищим пріоритетом. RL-актор на кроці t вибирає лише індекс задачі:

$$a_t \in \{1, \ldots, k\}, \quad (21)$$

тоді як вибір процесора виконується детерміновано за класичним правилом HEFT:

$$p^*(v) = \arg\min_{p_j \in P} EFT(v, p_j), \quad (22)$$

де EFT обчислюється на основі EST/FT із математичної моделі.

Тобто, для HG-RL–HEFT простір дій має вигляд:

$$A^{HG} = \{1, \ldots, k\}, \quad (v, p) = (TopK[a_t], \arg\min_{p_j} EFT(TopK[a_t], p_j)). \quad (23)$$

Таким чином, агент навчається коригувати вибір задачі, спираючись на глобальний стан системи, тоді як перевірка допустимості та розрахунок часових характеристик делегуються евристиці.

Для схеми RL-guided heuristic у випадку алгоритму CPOP (RL–GH–CPOP) агент керує параметризацією евристики.

Дія визначається як:

$$a_t = (p_{cp}, \mathrm{mode}), \quad (24)$$

де $p_{cp} \in P$ – процесор критичного шляху, а $mode \in \{soft, hard\}$ – режим обробки критичного шляху.

Далі застосовується стандартний алгоритм CPOP з фіксованими параметрами. Відповідно, простір дій має вигляд:

$$A^{RG} = P \times \{soft, hard\}. \quad (25)$$

Архітектура агента реалізується за схемою Actor–Critic із використанням алгоритму Proximal Policy Optimization (PPO).

- Актор (Actor) апроксимує політику $\pi_\theta(a/s)$, яка приймає стан s_t і повертає розподіл ймовірностей над допустимими діями, структура яких визначається обраною схемою гібридизації.

- Критик (Critic) оцінює функцію цінності поточного стану $V_\phi(s)$, прогнозуючи очікувану сумарну винагороду (Value function), що використовується для обчислення переваги (Advantage) у процесі навчання та дозволяє зменшити дисперсію градієнта.

Мережа реалізується у вигляді багатошарової нейронної мережі типу Feed-Forward Network (FFN) з 2–3 прихованими шарами по 128–256 нейронів та функцією активації ReLU.

Для гарантування допустимості рішень та прискорення збіжності навчання застосовується механізм евристичного маскування дій (Heuristic Guidance). Модифікована політика має вигляд:

$$\pi_{guided}(a \mid s) = Softmax(Logits(s) + M(s)), \quad (26)$$

де маска $M(s)$ приймає значення ∞ для недопустимих дій, що порушують топологічні обмеження DAG, та 0 для допустимих.

Функція винагороди формується на основі цільових метрик математичної моделі та визначається як:

$$r_t = \alpha \cdot (Makespan_{est}(t-1), -Makespan_{est}(t)) - \beta \cdot LB_{std}(t), \quad (27)$$

де $Makespan\_est(t)$ – евристична оцінка нижньої межі завершення workflow, яка може бути задана як $max\ \{FT_{current}, length_{CP}\}$, LB_{std} – стандартне відхилення навантаження процесорів, обчислене за формулою (15), а α, β – вагові коефіцієнти значущості метрик.

Перший доданок стимулює дії, що зменшують прогнозний час завершення всього workflow. Другий доданок (штраф) базується на стандартному відхиленні навантаження LB_{std}, запобігаючи дисбалансу системи.

Така функція винагороди стимулює одночасне зменшення часу виконання та покращення балансування навантаження.

Після побудови початкового розкладу може бути застосований необов'язковий модуль локального покращення $L(\cdot)$, який виконує обмежену кількість локальних перетворень (Move, Swap, Critical-path focusing), не порушуючи топологічних обмежень графа.

Типові локальні оператори:

Move: перемістити задачу v на інший процесор p', якщо це зменшує $FT(v)$ або зменшує критичний "хвіст" розкладу.

Swap: обміняти процесори двох задач v_i, v_j (за умови збереження допустимості за залежностями та ресурсами).

Critical-path focusing: обмежити пошук задачами, що належать до критичного шляху поточного розкладу або найближчого оточення CP.

У загальному вигляді процес побудови розкладу описується композицією:

$$Scheduler = L(S_{base}(G, P, ETC, Comm, \pi_\theta)), \quad (28)$$

де S_{base} – базовий гібридний планувальник (HG-RL–HEFT або RL–GH-CPOP).

Таким чином, запропонована функціональна модель забезпечує чітке розмежування ролей між евристичними та інтелектуальними компонентами, поєднуючи детермінованість і швидкодію класичних алгоритмів з адаптивністю методів навчання з підкріпленням. Це створює основу для експериментальної перевірки ефективності гібридного підходу.

Для перевірки ефективності запропонованої функціональної моделі планування задач у багатопроцесорних системах на основі гібридних методів було проведено серію експериментальних досліджень, спрямованих на оцінювання якості сформованих розкладів за ключовими часовими та ресурсними метриками. Метою експерименту було встановлення ступеня покращення класичних евристичних алгоритмів HEFT та CPOP шляхом їх інтеграції з методами глибокого навчання з підкріпленням у схемах HG-RL-HEFT та RL-GH-CPOP відповідно.

Програмна реалізація гібридних методів виконана мовою Python із використанням сучасних бібліотек машинного навчання та моделювання середовища. Для побудови та навчання нейромережевих моделей використовувалася бібліотека PyTorch (torch v2.10.0+cu128), яка забезпечує ефективну роботу з тензорними обчисленнями та підтримку апаратного прискорення. Формування середовища навчання з підкріпленням реалізовано на базі Gymnasium v1.2.3, що дозволило інтегрувати математичну модель (розділ 4) у стандартний RL-інтерфейс. Для реалізації алгоритму Proximal Policy Optimization (PPO) використовувалися бібліотеки Stable-Baselines3 v2.7.1 та SB3-Contrib v2.7.1, що забезпечують перевірені механізми Actor-Critic навчання та стабілізацію процесу оновлення політики.

Навчання моделей проводилося в середовищі з підтримкою GPU-прискорення, що дозволило забезпечити достатню швидкість обробки великої кількості епізодів. Апаратна конфігурація включала багатоядерний процесор (6 ядер), 32 ГБ оперативної пам'яті та графічний прискорювач NVIDIA GeForce RTX 5070 з підтримкою CUDA 13.1. Така конфігурація є типовою для сучасних дослідницьких задач машинного навчання та забезпечує відтворюваність експерименту. Для формування навчального потоку використовувалася спеціалізована фабрика генерації екземплярів задач, яка формує неперервний потік різноманітних DAG-графів із варіацією кількості задач, рівнів графа, середньої ширини, коефіцієнтів складності та гетерогенності процесорів. Базовий екземпляр передбачав генерацію workflow із базовою кількістю 100 задач, максимальною глибиною 10 рівнів, 6 процесорами та випадковими варіаціями гетерогенності, пропускної здатності каналу та шуму ETC-матриці. При цьому активовано механізми семплювання кількості задач (від 50 до 300), пропускної здатності (логарифмічно-однорідний розподіл у діапазоні 0.5–100.0), числа процесорів (2, 4, 8, 16), а також варіацій гетерогенності та коефіцієнтів складності. Такий підхід забезпечує формування неперервного потоку максимально різноманітних екземплярів, що запобігає переадаптації (overfitting) політики до конкретного типу графа та підвищує її узагальнювальну здатність

Фабрика інтегрується безпосередньо до середовища навчання через потоковий механізм генерації екземплярів, що дозволяє на кожному епізоді формувати новий DAG із незалежною конфігурацією. Таким чином, агент взаємодіє не з фіксованим набором сценаріїв, а з безперервною стохастичною послідовністю задач, що суттєво підвищує узагальнювальну здатність моделі. Для прискорення навчання використовувався механізм векторизованого середовища (vectorized environments). Одночасно запускалося по 6 паралельних екземплярів середовища для кожної гібридної схеми (n_envs_hg = 6, n_envs_rg = 6), що дозволяло накопичувати rollout-фрагменти з кількох незалежних сценаріїв паралельно. Така організація значно зменшує дисперсію градієнтної оцінки та підвищує стабільність алгоритму PPO, особливо в умовах високої варіативності графів.

Навчання моделі HG-RL-HEFT виконувалося з використанням GPU-прискорення, тоді як модель RL-GH-CPOP тренувалася на CPU. Це обумовлено різною обчислювальною складністю кроків взаємодії із середовищем: у схемі HG-RL-HEFT агент приймає рішення на кожному кроці побудови розкладу, тоді як у RL-GH-CPOP дія виконується одноразово на епізод (rg_action_frequency = "once").

На відміну від навчального режиму, для тестування та фінальної оцінки моделей використовувався окремий типовий сценарій із фіксованою конфігурацією параметрів та заздалегідь визначеною кількістю екземплярів. Це забезпечує коректне порівняння результатів між базовими евристиками та їх RL-реалізаціями в однакових умовах.

Сам процес навчання відбувався у режимі взаємодії агента з середовищем, де кожен епізод відповідав побудові повного розкладу для згенерованого DAG. Після завершення епізоду обчислювалися метрики makespan та балансування навантаження, які формували інтегральну функцію винагороди (формула (27)). Оновлення параметрів політики здійснювалося за алгоритмом PPO із використанням міні-пакетного градієнтного спуску.

З метою забезпечення відтворюваності результатів у всіх модулях, де це не порушувало логіку навчання, було зафіксовано значення random seed. Це дозволило мінімізувати вплив стохастичних

факторів та забезпечити стабільність експериментальних результатів.

У процесі дослідження було навчено декілька моделей із різними гіперпараметрами та тривалістю навчання. Відбір фінальної моделі для кожної гібридної схеми здійснювався на основі інтегральної оцінки якості розкладів на валідаційній вибірці. При цьому враховувався не лише мінімальний досягнутий показник оцінки, а й стабільність та відтворюваність отриманих результатів на інших наборах даних.

Для схеми RL-GH-CPOP спостерігалося стабільне та відтворюване покращення порівняно з базовою евристикою як на контрольних точках із найкращим значенням інтегральної оцінки (близько 0.0157), так і на більш стабільних контрольних точках з показником близько 0.065. Обидва варіанти демонстрували узгоджене покращення на різних тестових сценаріях.

Натомість для HG-RL-HEFT найкращі контрольні точки з мінімальним значенням оцінки (приблизно 0.148) часто характеризувалися нестійкою поведінкою: покращення спостерігалося лише на окремих (першому) тестових сценаріях, тоді як на інших екземплярах результати були близькими до базової евристики або навіть гіршими. Використання більш стабільної контрольної точки (близько 0.067) дозволило отримати повністю відтворюване та стабільне покращення на всіх тестових екземплярах, що і стало підставою для вибору фінальної моделі.

Таким чином, у межах даного дослідження критерій відбору фінальної моделі враховував компроміс між мінімальним значенням інтегральної оцінки та стабільністю узагальнення на нові сценарії.

Необов'язковий модуль локального покращення (L, Local post-processing) у межах даного експериментального дослідження за замовчуванням було вимкнено. Додаткові експерименти показали, що його внесок у покращення метрик є незначним (у межах статистичної похибки), тому основна увага зосереджена на ефекті саме гібридної інтеграції евристики та RL-компонента. Таким чином, організація експерименту забезпечує:

- навчання на максимально різноманітному потоці екземплярів;
- незалежне тестування на типовому сценарії;
- відтворюваність результатів;
- коректне порівняння базових та гібридних методів.

Налаштування алгоритму навчання з підкріпленням для обох гібридних схем виконувалося на основі алгоритму Proximal Policy Optimization (PPO), реалізованого в межах бібліотеки Stable-Baselines3. Для моделі RL-GH-CPOP використовувалися такі гіперпараметри: коефіцієнт навчання learning_rate = 0.0003, коефіцієнт дисконтування $\gamma = 0.99$, параметр GAE $\lambda = 0.95$, коефіцієнт ентропійної регуляризації ent_coef = 0.01, коефіцієнт функції цінності vf_coef = 0.5, обмеження норми градієнта max_grad_norm = 0.5, параметр обрізання clip_range = 0.2, кількість кроків збору траєкторії n_steps = 2, розмір міні-пакета batch_size = 6 та кількість епох оптимізації n_epochs = 2. Невелике значення n_steps обумовлене специфікою схеми RL-GH-CPOP, у якій агент приймає рішення лише один раз на епізод (rg_action_frequency = "once"), а подальше формування розкладу виконується детермінованою евристикою CPOP.

Для моделі HG-RL-HEFT застосовувалися аналогічні базові параметри PPO (learning_rate = 0.0003, $\gamma = 0.99$, $\lambda = 0.95$, ent_coef = 0.01, vf_coef = 0.5, clip_range = 0.2, max_grad_norm = 0.5), однак із суттєво більшими значеннями n_steps = 72, batch_size = 72 та n_epochs = 3. Це пов'язано з тим, що у схемі HG-RL-HEFT агент приймає рішення на кожному кроці побудови розкладу, тобто кількість взаємодій із середовищем у межах одного епізоду є значно більшою. Відповідно, збільшення rollout-фрагмента дозволяє стабілізувати оцінку переваги (advantage) та зменшити дисперсію градієнта при оновленні політики.

Гібридні конфігурації також відрізнялися параметрами інтеграції з евристичними модулями. Для RL-GH-CPOP використовувалося обмеження topk = 5 та два режими роботи з критичним шляхом (soft, hard), а вагові коефіцієнти функції винагороди становили $\alpha = 1.0$ та $\beta = 0.1$. Для HG-RL-HEFT застосовувалося topk = 8 із меншою вагою штрафу за дисбаланс ($\beta = 0.05$), що дозволяло агенту агресивніше оптимізувати makespan із помірним урахуванням балансування навантаження. В обох випадках використовувалася стратегія маскування допустимих дій ready_only та нормалізація завантаження процесорів у векторі стану.

Тривалість експерименту визначалась планом навчання, що передбачав 10 контрольних точок (checkpoints), по 12 екземплярів для оцінювання на кожній контрольній точці, із фіксованим seed = 12345. Така схема дозволила порівнювати динаміку якості моделі на різних етапах навчання та обирати фінальний варіант не лише за мінімальним значенням інтегральної оцінки, а й за стабільністю результатів.

Обрана конфігурація експериментального плану є результатом компромісу між статистичною надійністю оцінювання та обчислювальними витратами. Кількість паралельних середовищ n_envs=6 визначалася апаратними можливостями платформи: при меншій кількості не повністю використовувалися доступні ресурси процесора та GPU, тоді як збільшення до 8–12 середовищ призводило до суттєвого (у 1.5–2 рази) уповільнення навчання через накладні витрати синхронізації та зростання затримок обробки rollout-фрагментів. Відповідно, кількість екземплярів на контрольну точку було обрано кратною 12, що дозволяє ефективно розподіляти сценарії між 6 паралельними середовищами та забезпечує коректне агрегування результатів.

Використання 10 контрольних точок у межах одного експерименту є достатнім для формування моделі з інтегральною оцінкою на рівні близько 0.05 та прийнятною стабільністю узагальнення. Практичні спостереження показали, що контрольна точка з оптимальним співвідношенням мінімального зна-

чення оцінки та стабільності результатів зазвичай з'являється у межах перших десяти ітерацій. Подальше збільшення кількості контрольних точок лише розширює вибірку, але не призводить до якісно нового покращення політики, водночас суттєво збільшуючи час експерименту.

Середній час навчання моделі HG-RL-HEFT становить 25–32 хвилини, тоді як для RL-GH-CPOP – 1–2 хвилини через значно меншу кількість взаємодій із середовищем. Тестування усіх пар контрольних точок займає додатково 8–10 хвилин, що в сумі формує типовий час одного повного експерименту близько 40 хвилин. Збільшення кількості екземплярів на контрольну точку до 24 дозволило б дещо підвищити статистичну стабільність оцінювання, проте практичні результати показали, що це не суттєво впливає на загальну тенденцію метрик, натомість подвоює обчислювальні витрати.

Таким чином, обрана конфігурація експериментального плану забезпечує оптимальний баланс між тривалістю експерименту та якістю отриманої оцінки, дозволяючи в стислі строки перебирати варіації ключових гіперпараметрів (n_steps, batch_size, n_epochs) та знаходити їх раціональне співвідношення, яке забезпечує як прийнятний рівень інтегральної оцінки, так і стабільність узагальнення.

Після завершення навчання було проведено порівняння базових евристик HEFT та CPOP із відповідними гібридними модифікаціями на послідовності типових тестових сценаріїв. Агреговані результати для типового сценарію з 20 незалежних екземплярів наведено у табл. 1.

Таблиця 1 – **Порівняння агрегованих метрик базових та гібридних методів**

Метод	Makespan	Avg Completion Time	Avg Load	LB Std	LB CV
HEFT	57.1236	26.9557	22.1507	3.2682	0.1478
HG-RL-HEFT	53.3066	24.9436	22.3979	4.6743	0.2181
Δ (%)	+6.68	+7.46	−1.12	−43.03	−47.55
CPOP	64.8530	29.4950	22.7164	3.5333	0.1568
RL-GH-CPOP	55.3996	25.4111	23.2109	3.7997	0.1690
Δ (%)	+14.58	+13.85	−2.18	−7.54	−7.82

З наведених результатів видно, що для обох гібридних схем спостерігається суттєве зменшення makespan. Для HG-RL-HEFT середнє скорочення становить 6.68 %, тоді як для RL-GH-CPOP – 14.58 %. Аналогічна тенденція спостерігається і для середнього часу завершення задач (Average Task Completion Time), де покращення становить відповідно 7.46 % та 13.85 %. Значне покращення спостерігається і для Average Task Communication Time, але ця метрика він не є суттєвою для даної моделі.

Водночас для метрик балансування навантаження (Avg Load , LB Std та LB CV) спостерігається певне погіршення порівняно з базовими евристиками. Зокрема, у випадку HG-RL-HEFT стандартне відхилення навантаження зросло на 43.03 % (варіюється у діапазоні 35-45%), а коефіцієнт варіації – на 47.55 % (зазвичай 36-48%). Для RL-GH-CPOP зростання є менш вираженим (7–8 %). На фоні відчутного погіршення стандартного відхилення та коефіцієнту варіацій, середнє навантаження процесорів залишається у нормі з погіршенням лише на 1-2%. Ці погіршення пояснюються тим, що функція винагороди орієнтована насамперед на мінімізацію makespan, а балансування навантаження виступає як допоміжний штрафний компонент із меншою вагою β. Таким чином, отримані результати підтверджують, що гібридна інтеграція RL-компонента дозволяє суттєво скоротити загальний час виконання workflow, навіть якщо це супроводжується помірним зростанням нерівномірності завантаження процесорів. З практичної точки зору така поведінка є очікуваною, оскільки мінімізація критичного шляху часто потребує концентрації задач на швидших процесорах.

Графічний аналіз результатів дозволяє оцінити характер змін показників на рівні окремих тестових екземплярів та простежити стабільність отриманих покращень. На рис. 4 представлено порівняння значень Makespan для всіх чотирьох методів: HEFT, HG-RL-HEFT, CPOP та RL-GH-CPOP.

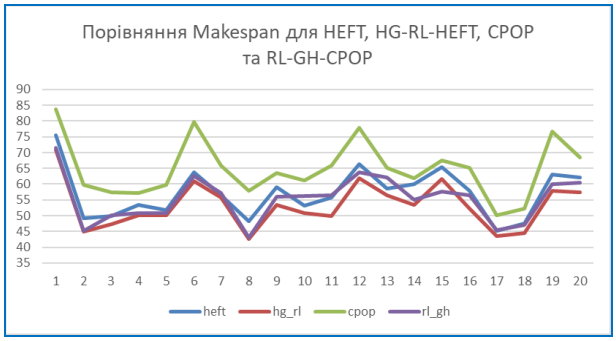


Рис. 4. Порівняння Makespan для HEFT, HG-RL-HEFT, CPOP та RL-GH-CPOP

З графіка видно, що обидві гібридні модифікації систематично зменшують загальний час виконання workflow порівняно з відповідними базовими евристиками. При цьому RL-GH-CPOP демонструє найбільш виражене зниження makespan, що узгоджується з агрегованим покращенням на рівні близько 14–15 %. HG-RL-HEFT забезпечує помірне, але стабільне скорочення (приблизно 6–7 %). Візуально лінії гібридних методів у більшості точок розташовані нижче відповідних базових алгоритмів, що свідчить про узгоджений характер покращення на різних екземплярах, а не лише про локальні виграші.

На рис. 5 наведено порівняння середнього часу завершення задач (Average Task Completion Time) для всіх чотирьох методів.

Форма кривих загалом повторює тенденцію makespan, що є очікуваним, оскільки зменшення довжини критичного шляху та оптимізація розподі-

лу задач скорочують не лише максимальний час завершення, але й середній час виконання задач. Найбільше зниження спостерігається для RL-GH-CPOP, де покращення перевищує 13 %. HG-RL-HEFT також демонструє стабільне зменшення показника. Важливо, що жоден із гібридних методів не демонструє систематичного погіршення середнього часу завершення, що підтверджує ефективність інтеграції RL-компонента.

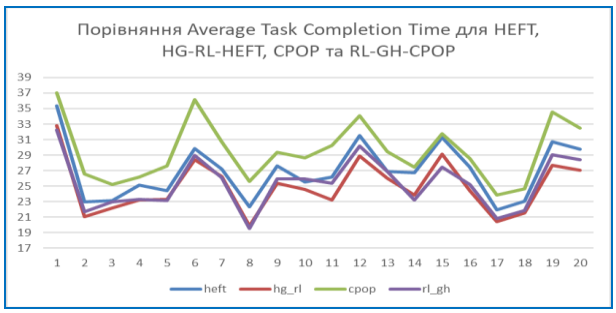


Рис. 5. Порівняння Average Task Completion Time для HEFT, HG-RL-HEFT, CPOP та RL-GH-CPOP

Average Task Communication Time також повторює ту саму тенденцію, але тут значення для базових евристик (HEFT, CPOP) повністю рівні, а у rl-методів окрім значного зниження, спостерігається невеликий шум, при чому HG-RL-HEFT трохи гірший за RL-GH-CPOP приблизно на 0,1-0,2%.

На рис. 6 подано порівняння середнього навантаження процесорів (Average Load).

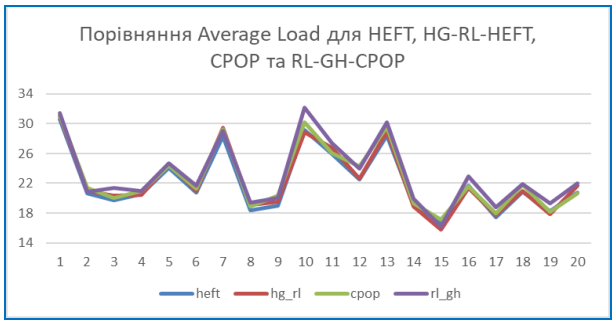


Рис. 6. Порівняння Average Load для HEFT, HG-RL-HEFT, CPOP та RL-GH-CPOP

На відміну від makespan, значення Average Load для всіх методів є близькими. Це пояснюється тим, що загальний обсяг обчислень у межах одного workflow є сталим, а зміни стосуються переважно розподілу задач між процесорами. Невелике зростання середнього навантаження у гібридних методів (близько 1–2 %) є наслідком концентрації задач на швидших процесорах для скорочення критичного шляху. Візуально графік не демонструє різких відхилень або аномальних піків, що свідчить про збереження загальної стабільності розподілу обчислювальної роботи.

Метрики балансування навантаження характеризуються значно більшою дисперсією значень між екземплярами. З цієї причини результати для Standard Deviation–based Load Balance та Coefficient of Variation–based Load Balance представлено у вигляді попарних графіків, що дозволяє уникнути накладання ліній та покращити інтерпретацію тенденцій. На рис. 7 наведено порівняння стандартного відхилення навантаження (Standard Deviation–based Load Balance) для HEFT та HG-RL-HEFT.

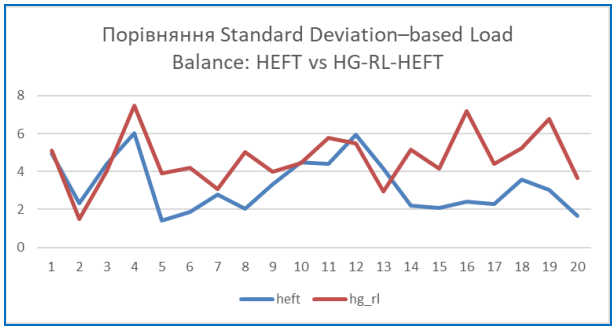


Рис. 7. Порівняння Standard Deviation–based Load Balance: HEFT vs HG-RL-HEFT

Графік демонструє, що HG-RL-HEFT у багатьох випадках має вищі значення стандартного відхилення навантаження порівняно з базовою евристикою. Це узгоджується з агрегованим зростанням показника приблизно на 40 %. Така поведінка пояснюється тим, що мінімізація makespan досягається за рахунок більш інтенсивного використання окремих (швидших) процесорів, що природно призводить до зростання нерівномірності розподілу задач.

На рис. 8 подано аналогічне порівняння для CPOP та RL-GH-CPOP.

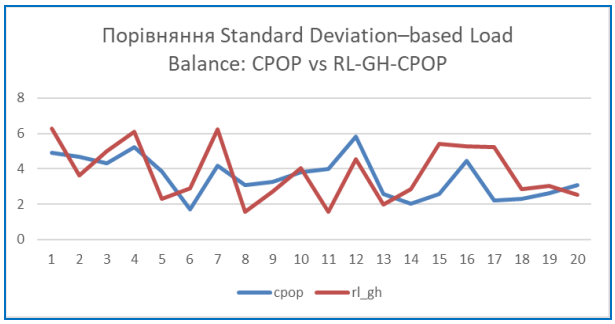


Рис. 8. Порівняння Standard Deviation–based Load Balance: CPOP vs RL-GH-CPOP

У цьому випадку збільшення дисбалансу є значно менш вираженим (близько 7–8 %). Лінії методів часто розташовані близько одна до одної, а пікові значення не мають різких відхилень. Це свідчить про те, що RL-GH-CPOP зберігає балансувальні властивості базової евристики, водночас забезпечуючи істотне зниження makespan.

Аналогічна структура результатів спостерігається для коефіцієнта варіації навантаження (Coefficient of Variation–based Load Balance). На рис. 9 наведено порівняння HEFT та HG-RL-HEFT.

Тут зростання показника є незначним та має стабільний характер, без різких піків. Це свідчить про більш збалансований компроміс між мінімізацією часу виконання та рівномірністю навантаження.

Таким чином, експериментальні результати підтверджують, що гібридні методи забезпечують

стабільне зменшення makespan та середнього часу завершення задач, при цьому характер впливу на балансування навантаження залежить від обраної схеми гібридизації (інтеграції RL-компонента).

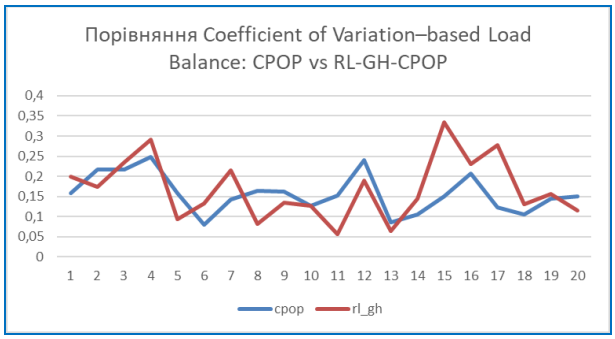


Рис. 9. Порівняння Coefficient of Variation–based Load Balance: CPOP vs RL-GH-CPOP

Висновки

У роботі розроблено модель планування задач у багатопроцесорних системах, що базується на гібридному підході та поєднує класичні евристичні алгоритми спискового планування (HEFT, CPOP) з адаптивним компонентом навчання з підкріпленням на основі алгоритму PPO. Запропонована модель формалізує взаємодію між середовищем, простором станів і дій, евристичним модулем та політикою агента, а також передбачає можливість регулювання компромісу між мінімізацією часу виконання задач і балансуванням навантаження за допомогою параметрів функції винагороди. Для перевірки ефективності підходу було реалізовано експериментальну інфраструктуру, що використовує потокову генерацію графів задач (DAG) та векторизоване середовище навчання. Така організація експерименту дала змогу сформувати узагальнену політику планування для широкого спектра сценаріїв із різною кількістю задач, різним рівнем гетерогенності процесорів та варіативністю комунікаційних параметрів, а також забезпечити стабільність і відтворюваність отриманих результатів.

Проведене експериментальне дослідження показало, що інтеграція компонентів навчання з підкріпленням у класичні евристичні методи планування дозволяє досягти систематичного зменшення часу виконання розкладу (makespan) та середнього часу завершення задач. Найбільш стабільне покращення продемонструвала схема RL-GH-CPOP, тоді як підхід HG-RL-HEFT забезпечує більш агресивне скорочення часу виконання за рахунок деякого зростання нерівномірності навантаження. Це підтверджує можливість керованого компромісу між часовими показниками виконання та ефективністю використання обчислювальних ресурсів.

Практична значущість запропонованого підходу полягає у можливості його застосування в гетерогенних багатопроцесорних середовищах, зокрема у високопродуктивних обчислювальних системах, хмарних та розподілених платформах, де важливо забезпечити мінімальний час виконання workflow за умови раціонального використання обчислювальних ресурсів.

Отримані результати свідчать, що гібридні методи планування, які поєднують евристичні алгоритми та навчання з підкріпленням, є перспективним напрямом розвитку систем керування обчислювальними ресурсами та можуть стати основою для подальших досліджень у сфері інтелектуального планування в складних обчислювальних середовищах.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Shaima Rahim et al. A survey of machine learning-driven task scheduling approaches for multiprocessor systems. *Journal of Systems Architecture*. 2025. Vol. 171. DOI: https://doi.org/10.1016/j.sysarc.2025.103628
2. Xu Jian et al. Real-time scheduling of parallel tasks with tight deadlines. *Journal of Systems Architecture*. 2020. Vol. 108. DOI: https://doi.org/10.1016/j.sysarc.2020.101742
3. Wenfan Zhang, Haijiao Ou. Reinforcement learning based multi objective task scheduling for energy efficient and cost effective cloud edge computing. *Scientific Reports*. 2025. DOI: https://doi.org/10.1038/s41598-025-25666-1
4. Пасічник М. Ю., Зайцев В. Г. Методи диспетчеризації завдань у системах реального часу. *Комп'ютерно-інтегровані технології: освіта, наука, виробництво*. 2024. №56. DOI: https://doi.org/10.36910/6775-2524-0560-2024-56-04
5. Gaurav Agarwal, et al. Multiprocessor task scheduling using multi-objective hybrid genetic Algorithm in Fog–cloud computing. *Knowledge-Based Systems*. 2023. Volume 272. DOI: https://doi.org/10.1016/j.knosys.2023.110563
6. Kusay Hameed Al-Salami, Zaid Taha Sawadi. Task Scheduling for Multiprocessor Systems Using Queuing Theory. *Computer Engineering and Intelligent Systems*. 2016. Vol. 7, No. 2. URL: https://www.iiste.org/Journals/index.php/CEIS/article/view/28602
7. Зайцев В. Г., Цибаєв Є. І. Оцінка часових характеристик задач в багатопроцесорних системах реального часу з використанням сіток Петрі. *Управління розвитком складних систем*. 2020. № 42. С. 43–50. DOI: https://doi.org/10.32347/2412-9933.2020.42.43-50
8. Wafa Hantom et al. A Survey on Scheduling Algorithms in Real-Time Systems. *International Journal of Computer Science and Network Security*. 2022. Vol. 22, No. 4. DOI: https://doi.org/10.22937/IJCSNS.2022.22.4.80
9. Люлька А. В. Методи та засоби планування обчислювальних завдань в комп'ютерній системі: робота на здобуття кваліфікаційного ступеня магістра: спец. 123 - комп'ютерна інженерія. Тернопільський національний технічний університет імені Івана Пулюя. 2024. 74 с. URL: https://elartu.tntu.edu.ua/handle/lib/48105
10. Голубничий Д. Ю., Головченко О. С. Застосування алгоритмів рангового підходу при плануванні розподілу задач в багатопроцесорних системах *Радіотехніка*. 2024. Вип. 219. С. 16–27. DOI: https://doi.org/10.30837/rt.2024.4.219.02

11. Станко П., Охремчук О., Саламатіна Д., Свердлова Д. Оптимізація планування завдань в розподілених обчислювальних системах реального часу. *Наукоємні технології*. 2023. №4(60). С. 386–393. DOI: https://doi.org/10.18372/2310-5461.60.18267
12. Головченко О. С. Методи планування розподілу задач в багатопроцесорних системах : поясн. записка до кваліф. роботи здобувача вищої освіти на другому (магістерському) рівні, спец. 123 Комп'ютерна інженерія. М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. 2025. 87 с. URL: https://openarchive.nure.ua/handle/document/32427
13. Yoni Birman et al. Hierarchical Deep Reinforcement Learning Approach for Multi-Objective Scheduling with Varying Queue Sizes. arXiv. 2020. DOI: https://doi.org/10.48550/arXiv.2007.09256
14. Yihong Li et al. Task Placement and Resource Allocation for Edge Machine Learning: A GNN-based Multi-Agent Reinforcement Learning Paradigm. arXiv. 2023. DOI: https://doi.org/10.48550/arXiv.2302.00571
15. Junyoung Park, et al. ScheduleNet: Learn to solve multi-agent scheduling problems with reinforcement learning. arXiv. 2021. DOI: https://doi.org/10.48550/arXiv.2106.03051
16. Zheng Xu, et al. Enhancing Kubernetes Automated Scheduling with Deep Learning and Reinforcement Techniques for Large-Scale Cloud Computing Optimization. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.07905
17. Xiao Fang Li. Simulation on Task Scheduling for Multiprocessors Based on Improved Neural Network. *Applied Mechanics and Materials*. 2014. Vol. 513-–517. P. 2293–2296. DOI: https://doi.org/10.4028/www.scientific.net/AMM.513-517.2293
18. Олександренко М. А. Система для моделювання багатопроцесорних алгоритмів планування : пояснювальна записка до кваліфікаційної роботи здобувача вищої освіти на першому (бакалаврському) рівні, спеціальність 123 Комп'ютерна інженерія. М-во освіти і науки України, Харків. нац. ун-т радіоелектроніки. 2025. 55 с. DOI: https://openarchive.nure.ua/handle/document/32703

Received (Надійшла) 22.12.2025
Accepted for publication (Прийнята до друку) 01.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Бондаренко Сергій Вікторович – студент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Serhii Bondarenko –student, Department of Electronic Computing Machines, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: serhii.bondarenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0002-8255-1764.

Мартовицький Віталій Олександрович – кандидат технічних наук, доцент**,** доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vitalii Martovytskyi – Candidate of Technical Sciences, Associate Professor, Associate Professor of Electronic Computing Machines Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: vitalii.martovytskyi@nure.ua; ORCID ID: https://orcid.org/0000-0003-2349-0578;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57196940070.

Бологова Наталія Миколаївна – кандидат технічних наук, доцент**,** доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Nataliia Bolohova – Candidate of Technical Sciences, Associate Professor, Associate Professor of Electronic Computing Machines Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: nataliia.bolohova@nure.ua; ORCID ID: https://orcid.org/0000-0001-8927-0055;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57203140922&origin=resultslist.

Рикун Володимир Георгійович – кандидат технічних наук, доцент**,** доцент Харківського національного університету Повітряних Сил ім. І. Кожедуба, Харків, Україна;
Volodymyr Rykun – Candidate of Technical Sciences, Associate Professor, Associate Professor of Ivan Kozhedub Kharkiv National Air Force University,Kharkiv, Ukraine;
e-mail: Volodymyr_Rykun@gmail.com; ORCID ID: https://orcid.org/0000-0003-0162-1178.

Tasks scheduling in multi-processor systems based on hybrid methods

Serhii Bondarenko, Vitalii Martovytskyi, Nataliia Bolohova, Volodymyr Rykun

Abstract. Relevance. The rapid development of multiprocessor and distributed computing systems necessitates improving task scheduling efficiency in heterogeneous computing environments. Efficient resource utilization and reduction of execution time are critical challenges for modern high-performance and cloud computing systems. **Object of study:** task scheduling processes in heterogeneous multiprocessor systems. **Purpose of the article:** to improve scheduling efficiency by developing a hybrid approach that combines classical heuristic algorithms with reinforcement learning methods. **Research results.** The paper analyzes existing task scheduling approaches, including list heuristics and machine learning-based methods. A functional scheduling model integrating the HEFT and CPOP algorithms with a reinforcement learning agent based on the Actor–Critic architecture using the Proximal Policy Optimization algorithm is proposed. A mathematical model of the scheduling environment considering DAG structure, expected time to compute matrix, and load balancing metrics is developed. Experimental evaluation demonstrates that the proposed hybrid approach reduces makespan by 5–7% and 13–15% and improves average task completion time compared to baseline heuristics while maintaining balanced resource utilization. **Conclusions.** The proposed hybrid scheduling model enables adaptive adjustment of heuristic algorithms through reinforcement learning in heterogeneous multiprocessor environments and demonstrates potential for improving performance and resource utilization in high-performance, cloud, and distributed computing systems. **Scope of application of the obtained results:** high-performance computing systems, cloud infrastructures, and distributed computing environments for improving task scheduling efficiency and resource utilization.

Keywords: task scheduling, multiprocessor systems, hybrid methods, reinforcement learning, PPO (Proximal Policy Optimization), CPOP (Critical Path on a Processor), load balancing.

УДК 004.8:004.056.55:004.93 **doi: 10.26906/SUNZ.2026.2.052**

В. А. Висоцька[1,2], Л. В. Чирун[1,3], О. О. Лаврут[4], Т. В. Лаврут[4], Р. В. Романчук[1]

[1] Національний університет «Львівська політехніка», Львів, Україна
[2] Харківський національний університет внутрішніх справ, Харків, Україна
[3] Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна
[4] Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ ДІПФЕЙКІВ НА ОСНОВІ ГЛИБИННОГО НАВЧАННЯ ТА МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ ДЛЯ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Анотація. Актуальність. Стрімкий розвиток технологій глибинного навчання сприяв появі високоякісного синтетичного медіаконтенту (діпфейків), що становить суттєву загрозу для інформаційної безпеки, цифрової довіри та медіапростору. Сучасні методи детекції діпфейків, які базуються на аналізі окремих модальностей (відео, аудіо або тексту), часто не забезпечують достатньої точності та узагальнюваності, що обумовлює необхідність розроблення мультимодальних підходів. **Об'єкт дослідження.** Процеси виявлення синтетичного медіаконтенту (діпфейків) у цифровому інформаційному середовищі. **Мета статті.** Розробка ефективного методу виявлення діпфейків на основі мультимодального аналізу з використанням моделей глибинного навчання та attention-механізмів. **Результати дослідження.** У роботі запропоновано інформаційну технологію виявлення діпфейків, що базується на комплексній обробці відео-, аудіо- та текстових даних. Розроблено узагальнений пайплайн, який включає попередню обробку медіаконтенту, виділення ознак для кожної модальності, мультимодальну інтеграцію та класифікацію. Для підвищення ефективності застосовано трансформерні архітектури з використанням механізмів self-attention і cross-attention, що дозволяють моделювати внутрішньо- та міжмодальні залежності. Проведені експериментальні дослідження на публічних датасетах продемонстрували, що запропонований підхід забезпечує підвищення точності виявлення діпфейків до 0,95 та F1-міри до 0,925, що перевищує результати одномодальних моделей. **Висновки.** Отримані результати підтверджують доцільність використання мультимодального підходу та attention-механізмів для задачі виявлення діпфейків. Запропонована інформаційна технологія забезпечує підвищену точність, інтерпретованість та може бути використана у системах інформаційної безпеки, цифрової криміналістики та автоматизованого аналізу медіаконтенту. Перспективи подальших досліджень пов'язані з оптимізацією обчислювальної складності моделей та адаптацією до обробки потокових даних у реальному часі.

Ключові слова: кібербезпека, діпфейк, мультимодальний аналіз, глибинне навчання, трансформери, attention-механізм, інформаційна безпека, синтетичний медіаконтент, комп'ютерний зір, обробка аудіо, машинне навчання.

Вступ

Стрімкий розвиток технологій штучного інтелекту, зокрема глибинного навчання, призвів до появи нових інструментів створення синтетичного медіаконтенту [1]. Одним із найбільш відомих прикладів таких технологій є діпфейки – реалістичні підроблені відео-, аудіо- або зображення, створені за допомогою нейронних мереж [2]. Хоча подібні технології можуть використовуватися у сфері розваг, кіноіндустрії або освіти, їх активне поширення також створює серйозні ризики для інформаційної безпеки, суспільної довіри та цифрової ідентичності. Зокрема, діпфейки можуть використовуватися для маніпуляції громадською думкою, поширення дезінформації, шахрайства або дискредитації окремих осіб [3]. Відомі методи виявлення діпфейків здебільшого зосереджені на аналізі окремих модальностей, таких як відео або аудіо. Проте такі підходи мають обмежену ефективність, оскільки не враховують складні міжмодальні залежності, які можуть містити ключову інформацію про фальсифікацію [3–6]. Наприклад, невідповідність між рухами губ і звуковою доріжкою або семантичні розбіжності в тексті можуть бути важливими індикаторами синтетичного походження контенту.

У зв'язку з цим особливої актуальності набуває проблема виявлення синтетично згенерованого контенту. Традиційні методи аналізу медіаданих, які базуються лише на одному типі інформації (наприклад, візуальному або аудіальному), часто виявляються недостатньо ефективними для протидії сучасним алгоритмам генерації діпфейків. Це зумовлює необхідність застосування більш комплексних підходів, серед яких важливе місце займає мультимодальний аналіз, що поєднує обробку різних типів даних – відео, аудіо, тексту та метаданих. Мультимодальні підходи у поєднанні з методами глибинного навчання дозволяють виявляти невідповідності між різними модальностями контенту, наприклад між рухами губ та аудіодоріжкою, мімікою обличчя та мовленням, або часовими характеристиками сигналів [7–9]. Завдяки використанню згорткових, рекурентних та трансформерних нейронних мереж стає можливим автоматичне вилучення складних ознак та побудова високоточних моделей детекції діпфейків. Тому актуальним та перспективним є розроблення інформаційної технології виявлення діпфейків на основі мультимодального аналізу та глибинного навчання, а також визначення їх ефективності та перспектив застосування для автоматичної детекції синтетичного медіаконтенту [10–12].

Постановка проблеми. Активний розвиток інформаційних технологій, зокрема методів глибинного навчання та генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео-, аудіо- та текстові

матеріали, які складно відрізнити від автентичних. Поширення діпфейків становить серйозну загрозу для інформаційної безпеки, функціонування систем управління в складних інформаційних середовищах, а також для забезпечення цивільної безпеки, оскільки може використовуватися для маніпуляції громадською думкою, дезінформації, соціальної інженерії та дискредитації осіб чи організацій.

Особливої актуальності проблема набуває в умовах цифровізації суспільства, розвитку систем зв'язку та масового поширення мультимедійного контенту через інформаційно-комунікаційні мережі. У таких умовах виникає необхідність створення ефективних автоматизованих систем виявлення синтетичного контенту як складової інтелектуальних систем моніторингу та аналізу інформаційних потоків.

Існуючі підходи до виявлення діпфейків переважно базуються на аналізі окремих типів даних (відео, аудіо або тексту), що обмежує їх ефективність у реальних умовах. Такі методи не враховують складні міжмодальні залежності між різними джерелами інформації, які є важливими індикаторами фальсифікації (наприклад, невідповідність між рухами губ і аудіосигналом або семантичні розбіжності у тексті). Крім того, сучасні моделі часто мають обмежену узагальнюваність, високу чутливість до шумів і компресії, а також недостатню інтерпретованість результатів.

У контексті розвитку інтелектуальних систем управління та забезпечення безпеки інформаційних процесів виникає необхідність розроблення нових підходів до виявлення діпфейків, які б забезпечували комплексний аналіз мультимедійних даних, враховували міжмодальні зв'язки та забезпечували високу точність і надійність функціонування в умовах реального часу. Таким чином, актуальною науково-прикладною проблемою є розроблення інформаційної технології виявлення діпфейків на основі мультимодального аналізу з використанням методів глибинного навчання, що дозволить підвищити ефективність систем інформаційної безпеки, автоматизованого аналізу медіаконтенту та управління інформаційними потоками в складних системах.

Аналіз останніх досліджень і публікацій. У сучасному інформаційному суспільстві цифровий медіаконтент відіграє важливу роль у формуванні громадської думки, поширенні інформації та комунікації. Стрімкий розвиток технологій штучного інтелекту та глибинного навчання сприяв появі нових методів генерації синтетичного контенту, серед яких особливе місце займають діпфейки [12–15]. Діпфейк-технології дозволяють створювати реалістичні підроблені відео, аудіо та зображення, які складно відрізнити від оригінальних матеріалів. Такі можливості можуть використовуватися не лише у сфері розваг, кіновиробництва чи віртуальної реальності, але й у протиправних цілях, зокрема для поширення дезінформації, маніпулювання громадською думкою, фінансового шахрайства та дискредитації осіб або організацій. Зростання кількості та якості діпфейків створює серйозні виклики для інформаційної безпеки, медіадовіри та цифрової ідентичності. Традиційні методи аналізу медіаданих, які ґрунтуються на дослідженні лише однієї модальності (наприклад, відео або аудіо), часто не здатні ефективно виявляти сучасні підробки, створені за допомогою складних нейронних мереж [15–18]. У зв'язку з цим особливої актуальності набуває застосування мультимодального аналізу, що передбачає одночасне використання кількох джерел інформації, таких як відеоряд, аудіосигнал, текстові дані та метадані. Поєднання мультимодального підходу з методами глибинного навчання відкриває нові можливості для підвищення точності та надійності систем автоматичного виявлення діпфейків.

З розвитком методів глибинного навчання, зокрема генеративних моделей (GAN, автоенкодери, дифузійні моделі), проблема створення діпфейків стала однією з ключових загроз інформаційній безпеці. У зв'язку з цим активно розвиваються методи їх автоматичного виявлення, які базуються на аналізі візуальних, аудіо та текстових ознак. Ранні підходи до виявлення діпфейків зосереджувалися переважно на аналізі візуальних артефактів, таких як неприродні рухи обличчя, артефакти компресії та некоректне освітлення або тіні. Перші ефективні візуальні методи (Unimodal approaches) базувалися на згорткових нейронних мережах (CNN), які навчалися виявляти просторові особливості зображень $F_v = CNN(X_v)$. Ці підходи показали високу точність на контрольованих датасетах (наприклад, FaceForensics++), проте мають обмеження: чутливість до компресії, низька узагальнюваність та неможливість врахування часових залежностей. Для врахування часової динаміки використовувались 3D-CNN та CNN з LSTM $Z_v = LSTM\big(CNN(X_v)\big)$.

Аудіоаналіз став окремим напрямом, особливо для виявлення синтетичних голосів:

$$F_a = CNN_{audio}\big(Spectrogram(X_a)\big),$$

або $F_a = LSTM(X_a)$. Ці методи здатні виявляти неприродні спектральні характеристики та артефакти синтезу мовлення. Однак вони не враховують відео та не можуть виявляти візуально-акустичні невідповідності. З розвитком трансформерів (наприклад, BERT) з'явилися методи аналізу текстових компонентів а основі семантичних підходів:

$$F_t = Transformer(X_t).$$

Такі моделі дозволяють аналізувати семантичні невідповідності та виявляти штучно згенерований текст. Однак вони рідко використовуються ізольовано для детекції діпфейків.

Останні роки характеризуються стрімким розвитком моделей глибокого навчання, зокрема архітектур на основі механізму уваги (attention) та трансформерів. Вперше трансформер було представлено у роботі [1], де запропоновано архітектуру, що повністю базується на self-attention і дозволяє ефективно моделювати довгострокові залежності без використання рекурентних або згорткових мереж [2].

1. Attention-механізми та трансформери.
2. Мультимодальні трансформери.
3. Cross-modal attention та моделі MulT.
4. Трансформери у комп'ютерному зорі та відео.

Механізм уваги став ключовим компонентом сучасних моделей, оскільки дозволяє адаптивно визначати важливість різних частин вхідних даних. У роботі [3] показано, що attention значно покращує якість моделей комп'ютерного зору та дозволяє будувати ефективні гібридні архітектури. Трансформери забезпечують паралельну обробку даних, ефективне моделювання контексту та масштабованість до великих датасетів. Завдяки цим властивостям вони стали стандартом у NLP, CV та аудіообробці. Сучасні дослідження активно зосереджені на мультимодальних моделях, які інтегрують різні типи даних (відео, аудіо, текст). У роботі [4] представлено систематичний огляд мультимодальних трансформерів, де описано основні підходи до злиття модальностей: раннє об'єднання (early fusion); пізнє об'єднання (late fusion); ієрархічні attention-механізми; cross-attention. Особливу роль відіграє cross-attention, який дозволяє одній модальності "звертати увагу" на іншу, формуючи узгоджені представлення.

Однією з ключових робіт у цій області є [5], де запропоновано механізм спрямованої попарної cross-modal attention. Основні переваги підходу: робота з неузгодженими (unaligned) даними; моделювання залежностей між різними часовими шкалами; покращення точності класифікації мультимодальних сигналів. З появою Vision Transformer (ViT) трансформери стали активно застосовуватися у задачах комп'ютерного зору [6]. Як зазначено у Springer-огляді, вони поступово конкурують із CNN, особливо у великих датасетах [2]. У відеоаналізі трансформери моделюють часову динаміку; використовують attention для виділення важливих кадрів; дозволяють реалізувати explainability через Grad-CAM та attention maps.

Сучасні дослідження демонструють, що найбільш ефективними є мультимодальні методи (табл. 1), які інтегрують кілька джерел інформації [7–12].

Таблиця 1 - **Порівняння підходів**

Підхід	Переваги	Недоліки
Візуальний (CNN)	Простота, швидкість	Низька узагальнюваність
Аудіо	Виявлення синтетичного голосу	Ігнорує відео
Текст	Семантичний аналіз	Обмежене застосування
Мультимодальний	Найвища точність	Висока складність

1. Просте об'єднання (Fusion):

$$Z_{fusion} = [F_v \parallel F_a \parallel F_t].$$

Недоліками є відсутність міжмодальної взаємодії та слабка адаптивність.

2. Attention-based підходи (Self-attentio та Cross-attention):

$$Z_m = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V,\ Z_{v\leftarrow a} = \text{attention}(Z_v, Z_a).$$

Ці методи дозволяють моделювати залежності між модальностями та виявляти невідповідності (наприклад, губи ≠ звук).

Сучасні підходи використовують контрастивні функції втрат (контрастивне навчання та representation learni):

$$L = -\log\frac{\exp\left(sim(z_i, z_i^+)\right)}{\sum_j \exp\left(sim(z_i, z_j)\right)}.$$

Це дозволяє розділяти real та fake представлення, та покращувати узагальнення моделі.

Незважаючи на значний прогрес, існуючі роботи мають ряд обмежень як недостатня інтеграція модальностей, відсутність пояснюваності, проблеми узагальнення та висока обчислювальна складність, особливо для трансформерів [15-18]. Багато моделей використовують просту конкатенацію без глибокої взаємодії. Більшість моделей є "black-box". Моделі часто погано працюють на нових типах діпфейків. Враховуючи недоліки існуючих робіт, у даному дослідженні запропоновано підхід, який:

– використовує multimodal transformers;
– інтегрує self- та cross-attention;
– застосовує адаптивне об'єднання ознак;
– забезпечує інтерпретованість через attention;
– оптимізований для реального часу.

Аналіз літератури показує, що мультимодальні підходи є найбільш перспективними та attention-механізми відіграють ключову роль. Інтеграція різних модальностей дозволяє значно підвищити точність, але необхідні подальші дослідження у напрямку ефективності та узагальнення моделей. У сучасних дослідженнях значна увага приділяється інтерпретації моделей. Основні підходи:

- attention visualization (heatmaps) – показують важливість ознак;
- Grad-CAM – локалізує важливі області у відео/зображеннях;
- аналіз attention-heads – дозволяє зрозуміти роль окремих голів уваги.

Ці методи критично важливі для мультимодальних систем, де складно інтерпретувати взаємодію різних джерел даних. Незважаючи на значний прогрес, залишаються відкриті питання:

- висока обчислювальна складність трансформерів;
- потреба у великих обсягах даних;
- складність інтерпретації attention-механізмів;
- оптимальне злиття мультимодальних даних.

Таким чином, дослідження методів і засобів виявлення діпфейків на основі технологій мультимодального аналізу та глибинного навчання є актуальним науковим і практичним завданням, спрямованим на підвищення рівня достовірності цифрової інформації та забезпечення інформаційної безпеки. Основним напрямом дослідження є аналіз сучасних методів та засобів виявлення діпфейків на основі технологій мультимодального аналізу та глибинного навчання, а також визначення їх ефективності, переваг і обмежень. У роботі розглядаються підходи до обробки різних типів медіаданих, архітектури нейронних мереж, що застосовуються для детекції синтетичного контенту, та перспективи розвитку систем автоматичного виявлення діпфейків у контексті сучасних викликів інформаційної безпеки.

Формулювання мети статті Метою даного дослідження є розробка ефективного методу виявлення діпфейків на основі мультимодального аналізу з використанням глибинного навчання та attention-механізмів. Для досягнення цієї мети запропоновано комплексний підхід, який поєднує обробку відео, аудіо та текстових даних у єдиній моделі, що забезпечує підвищену точність та інтерпретованість результатів.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1. Проаналізувати сучасний стан розвитку технологій створення діпфейків.

2. Дослідити основні підходи до виявлення синтетичного медіаконтенту.

3. Розглянути принципи та методи мультимодального аналізу при обробці медіаданих.

4. Проаналізувати застосування моделей глибинного навчання для детекції діпфейків.

5. Визначити переваги та обмеження існуючих методів виявлення діпфейків.

6. Оцінити перспективи розвитку систем автоматичного виявлення діпфейків на основі мультимодальних підходів.

Об’єкт дослідження – процеси виявлення синтетичного медіаконтенту (діпфейків) у цифровому інформаційному середовищі.

Предмет дослідження – методи та засоби виявлення діпфейків, що базуються на використанні мультимодального аналізу та алгоритмів глибинного навчання.

У роботі представлено математичну формалізацію запропонованого підходу, описано архітектуру моделі, проведено експериментальні дослідження та здійснено аналіз отриманих результатів. Отримані результати підтверджують доцільність використання мультимодального підходу для задачі виявлення діпфейків та відкривають перспективи для подальших досліджень у цьому напрямі.

Основний матеріал

Активний розвиток технологій глибинного навчання, зокрема генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео, аудіо та текстові матеріали, які складно відрізнити від автентичних. Незважаючи на потенційні позитивні застосування, діпфейки становлять серйозну загрозу для інформаційної безпеки, зокрема через можливість маніпуляції громадською думкою, поширення дезінформації та підрив довіри до цифрових медіа. Стрімкий розвиток технологій глибинного навчання, зокрема генеративних моделей, призвів до появи високоякісного синтетичного медіаконтенту, відомого як діпфейки. Такі технології дозволяють створювати реалістичні відео, аудіо та текстові матеріали, які складно відрізнити від автентичних. Незважаючи на потенційні позитивні застосування, діпфейки становлять серйозну загрозу для інформаційної безпеки, зокрема через можливість маніпуляції громадською думкою, поширення дезінформації та підрив довіри до цифрових медіа.

Процес виявлення діпфейків на основі мультимодального аналізу передбачає комплексну обробку кількох типів даних (відео, аудіо, тексту або метаданих) та їх подальший аналіз із використанням моделей глибинного навчання. Такий підхід дозволяє виявляти невідповідності між різними модальностями та підвищувати точність детекції синтетичного контенту. Загальний пайплайн системи подамо у вигляді послідовності етапів.

1. Збір та підготовка даних.

2. Попередня обробка медіаконтенту.

3. Виділення ознак (Feature Extraction) для кожної модальності.

4. Мультимодальна інтеграція даних.

5. Класифікація за допомогою моделей глибинного навчання.

6. Оцінювання якості моделі та верифікація результатів.

7. Прийняття рішення щодо наявності діпфейку та інтерпретація результатів.

На першому етапі здійснюється збір медіаданих, які можуть містити як справжній, так і синтетично згенерований контент. Джерелами даних можуть бути відеоплатформи, соціальні мережі, відкриті датасети або власні колекції медіафайлів. Отримані дані проходять попередню підготовку, що включає очищення, нормалізацію та анотацію даних для подальшого навчання моделей. Нехай вхідний медіаконтент представлено у вигляді множини модальностей $X = \{X_v, X_a, X_t\}$, де X_v– відеодані, X_a – аудіодані, X_t – текстові дані (наприклад, транскрипція мовлення).

На етапі «Попередня обробка медіаконтенту» виконується розділення медіафайлу на окремі модальності:

- відео – виділення кадрів, детекція та вирівнювання обличчя, нормалізація зображень;

- аудіо – екстракція аудіодоріжки, фільтрація шумів, перетворення у спектрограми або інші ознаки;

- текст (за наявності) – транскрипція мовлення та обробка текстових даних.

Попередня обробка включає нормалізацію, сегментацію та виділення необхідних фрагментів $\widetilde{X_m} = P_m(X_m)$, де $P_m(\cdot)$ – функція попередньої обробки для модальності m. Тоді:

$$\tilde{X} = \{\widetilde{X_v}, \widetilde{X_a}, \widetilde{X_t}\}.$$

Метою цього етапу є підготовка структурованих даних, придатних для подальшого аналізу. Для кожної модальності здійснюється автоматичне виділення релевантних ознак за допомогою моделей глибинного навчання для:

- відео – згорткові нейронні мережі (CNN) для аналізу візуальних характеристик обличчя, міміки та артефактів генерації;

- аудіо – рекурентні або згорткові моделі для аналізу спектральних та часових характеристик голосу;

- тексту – мовні моделі для аналізу змісту та узгодженості мовлення.

Для кожної модальності застосовується модель глибинного навчання для отримання ознак.

$$F_m = f_m(\widetilde{X_m}; \theta_m),$$

де f_m– модель виділення ознак (наприклад CNN, RNN, Transformer), θ_m– параметри моделі. Отримуємо множину ознак $F = \{F_v, F_a, F_t\}$.

Далі відбувається об'єднання ознак, отриманих із різних модальностей. Мультимодальна інтеграція може здійснюватися за допомогою різних стратегій:

- рання інтеграція (early fusion) – об'єднання ознак на початковому етапі;
- пізня інтеграція (late fusion) – поєднання результатів окремих моделей;
- гібридна інтеграція – комбінування кількох підходів.

Ознаки з різних модальностей об'єднуються в єдиний вектор ознак $F_{fusion} = \Phi(F_v, F_a, F_t)$, де $\Phi(\cdot)$ – функція мультимодальної інтеграції (конкатенація, attention або інші механізми). Наприклад: $F_{fusion} = [F_v \parallel F_a \parallel F_t]$, де $\parallel$ – операція конкатенації. Цей етап «Мультимодальна інтеграція даних» дозволяє враховувати взаємозв'язки між відео-, аудіо- та текстовими даними.

Інтегровані мультимодальні ознаки подаються на вхід моделі класифікації, яка визначає, чи є медіаконтент справжнім або синтетичним. Для цього використаємо глибокі нейронні мережі, трансформерні архітектури та ансамблі моделей. Модель навчається на попередньо розмічених даних для розпізнавання характерних ознак діпфейків.

Інтегрований вектор ознак подається на класифікатор $\hat{y} = g(F_{fusion}; \theta_c)$, де g – класифікаційна модель, θ_c – параметри моделі.

Ймовірність того, що контент є діпфейком:

$$P(y = 1|X) = \sigma(WF_{fusion} + b),$$

де W – матриця ваг, b – зсув, σ – сигмоїдна функція активації.

Після класифікації здійснюється оцінка точності роботи моделі. Для цього використовуються метрики якості, такі як точність (accuracy), повнота (recall), точність передбачення (precision) та F1-міра. За потреби модель додатково оптимізується та перенавчається.

Для навчання моделі використовується функція втрат, наприклад бінарна крос-ентропія:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)],$$

де y_i– істинна мітка, $\hat{y}_i$ – передбачене значення.

Параметри моделі оптимізуються за допомогою градієнтного спуску $\theta = \theta - \eta\nabla_\theta L$, де η – коефіцієнт навчання.

На фінальному етапі система формує висновок щодо автентичності медіаконтенту. Фінальне рішення визначається пороговою функцією:

$$y = \begin{cases} 1, P(y = 1|X) > \tau, \\ 0, \text{otherwise}, \end{cases}$$

де 1 – діпфейк, 0 – автентичний контент, τ – поріг класифікації. Результати можуть супроводжуватися поясненнями або візуалізацією виявлених підозрілих ділянок, що підвищує прозорість роботи системи. Метою експериментів було оцінити ефективність запропонованого мультимодального пайплайну для виявлення діпфейків. Зокрема, перевірялась здатність:

1. Виявляти синтетичні відео та аудіо (deepfake) з високою точністю.
2. Використовувати self- та cross-attention для покращення узгодженості між модальностями.
3. Об'єднувати мультимодальні ознаки у компактне представлення для класифікації.

Для навчання та тестування використовувалися публічні мультимодальні датасети:

1. FaceForensics++ – відео з реальними та синтетичними обличчями:
https://justusthies.github.io/posts/faceforensics++/,
https://github.com/ondyari/faceforensics,
https://www.kaggle.com/datasets/xdxd003/ff-c23,
https://www.kaggle.com/datasets/greatgamedota/faceforensics.

2. DeepfakeDetection Audio-Visual Dataset (DFAVD) – аудіо-відео пари з синхронізацією мовлення та обличчя:
https://github.com/vcbsl/audio-visual-deepfake/,
https://www.kaggle.com/datasets/elin75/localized-audio-visual-deepfake-dataset-lav-df,
https://www.kaggle.com/datasets/elin75/localized-audio-visual-deepfake-dataset-lav-df/code, https://github.com/controlnet/av-deepfake1m, https://cisaad.umbc.edu/data-sets/.

3. TTS-Deepfake Dataset – синтетичні голоси з відповідними текстовими транскриптами
https://data.mendeley.com/datasets/h4zbs27tkr/2,
https://github.com/YMLLG/SpeechFake,
https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset,
https://zenodo.org/records/6560159,
https://huggingface.co/datasets/garystafford/deepfake-audio-detection.

Попередня обробка:

- для відео ресайз до 224×224, нормалізація пікселів.
- для аудіо спектрограма $\in R^{T\times F}$.
- для тексту токенізація та embedding через BERT (d=768).

Формалізація мультимодальних ознак:

$$F_v \in R^{T_v \times d_v},\ F_a \in R^{T_a \times d_a}, F_t \in R^{T_t \times d_t}.$$

Архітектура моделі:

- Feature Extraction – CNN для відео, 1D-CNN + LSTM для аудіо, BERT для текста.
- Self-Attention – кожна модальність оброблялась трансформером:

$$Z_m = SelfAttn(F_m), m \in \{v, a, t\}.$$

- Cross-Attention: інтеграція між модальностями:

$$Z_{v\leftarrow a} = CrossAttn(Z_v, Z_a),$$

$$Z_{v\leftarrow t} = CrossAttn(Z_v, Z_t).$$

- Fusion (зважена сума на основі уваги):

$$Z_{fusion} = \sum_m \alpha_m \widetilde{Z_m}.$$

- Classification (MLP з бінарним виходом):

$$\hat{y} = \sigma(WZ_{fusion} + b).$$

Параметри навчання:

- Оптимізатор: Adam ($\eta = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$).
- Batch size 32 та кількість епох 50.
- Функція втрат комбінована:

$$L_{total} = L_{BCE} + \lambda_1 L_{consistency} + \lambda_2 L_{align}.$$

Метрики оцінки (Accuracy, Precision, Recall, F1-score):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, P = \frac{TP}{TP+FP},$$
$$R = \frac{TP}{TP+FN}, F1 = \frac{2 \cdot P \cdot R}{P+R},$$

де TP, TN, FP, FN – відповідно: істинно позитивні, істинно негативні, хибно позитивні, хибно негативні класифікації.

Результати експериментів наведені у табл. 2.

Таблиця 2- **Результати експериментів**

Модель	Acc	P	R	F1
Відео лише (CNN)	0.88	0.85	0.83	0.84
Аудіо лише (LSTM)	0.81	0.79	0.77	0.78
Відео та Аудіо (Fusion)	0.91	0.89	0.87	0.88
Відео, Аудіо та Текст (Fusion з Cross-Attention)	0.95	0.93	0.92	0.925

Cross-attention підкреслює розсинхронізацію між відео та аудіо. Self-attention виділяє аномальні ділянки обличчя та губ. Мультимодальність покращує точність. Інтеграція відео, аудіо та тексту забезпечує додаткову інформацію для виявлення subtle deepfake артефактів. Self- та cross-attention критичні для узгодженості: без cross-attention F1-score падає на ~4%. Attention-based зважене об'єднання ознак виявилося ефективнішим за просту конкатенацію (Fusion strategy). Обмеження є втому, що часова складність $O(T^2 \cdot d)$ для відео великої довжини. Модель чутлива до якості аудіо (шум може знижувати recall). З адаптивним streaming inference можливе використання в соцмережах та системах безпеки.

Запропонований мультимодальний пайплайн продемонстрував високу точність та узгодженість між модальностями. Використання self- та cross-attention дозволяє виявляти синтетичні артефакти, які неможливо помітити при використанні однієї модальності. Attention-based fusion забезпечує компактне та інформативне представлення, що підвищує точність класифікації. Подальше вдосконалення: оптимізація обчислювальної складності та адаптація до потокових даних.

Наведемо графік на рис. 1, що відображає результати точності, precision, recall та F1-score для різних конфігурацій моделей у задачі виявлення діпфейків. Можна побачити, що мультимодальна модель з відео, аудіо та текстом (Fusion з Cross-Attention) демонструє найвищі показники по всіх метриках, що підтверджує ефективність інтеграції різних модальностей та attention-механізмів.

Графік на рис. 2 показує, що інтеграція всіх трьох модальностей (відео, аудіо, текст) з Cross-Attention дає найвищий F1-score (0.925). Heatmap уваги (Attention) на рис. 3 демонструє важливість різних ознак у часових кроках; тепліші ділянки відповідають більш значущим ознакам для класифікації діпфейків.

Графік на рис. 4 показує, що кожен рядок (Time step) – окремий момент часу (кадр/аудіо-фрейм/токен), кожен стовпець (Feature) – окрема ознака в латентному просторі та колір (інтенсивність) – важливість ознаки (attention weight).

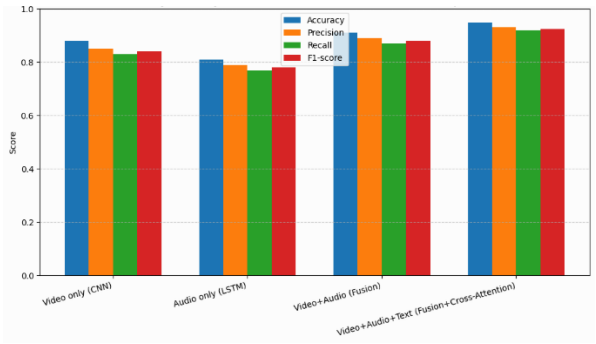


Рис. 1. Результати мультимодальних моделей для виявлення діпфейків

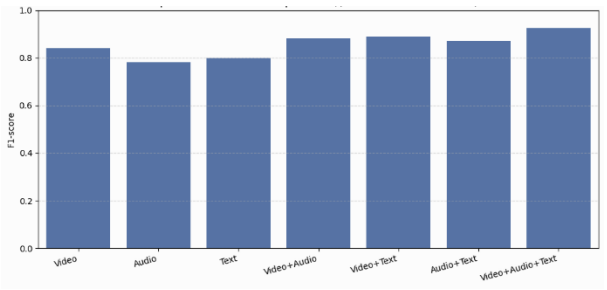


Рис. 2. Порівняння F1-score по модальностях та їх комбінаціях

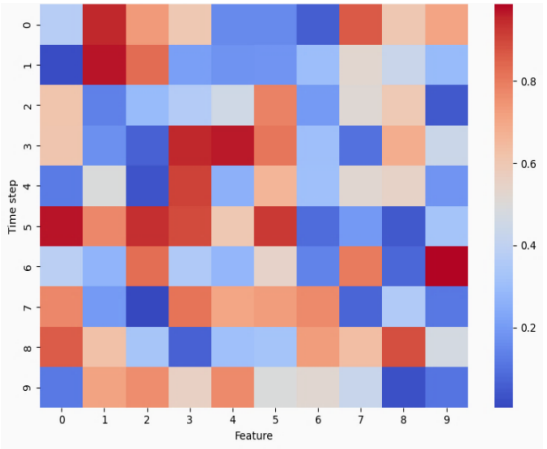


Рис. 3. Heatmap уваги (Attention) мультимодальних моделей для виявлення діпфейків

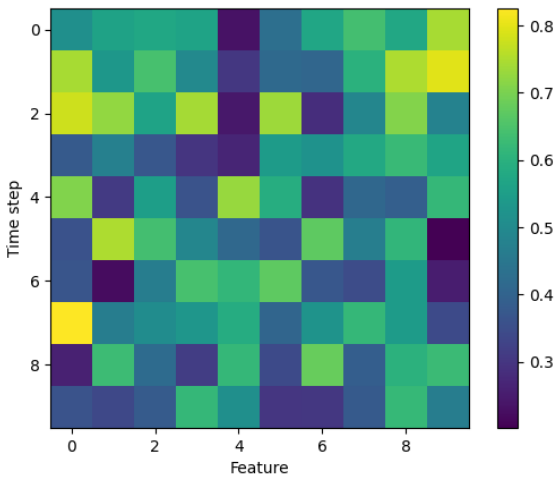


Рис. 4. Комбінований графік attention

На рис. 5-7 подано окремі attention heatmaps для кожної модальності, які демонструють інтерпретованість моделі. Рис. 5 показує окремі attention heatmaps для кожної модальності. Рис. 6 оказує, на які просторово-часові ознаки (наприклад, міміка, рух губ) модель звертає увагу. Рис. 7 підкреслює важливі частотні та часові патерни (інтонація, паузи).

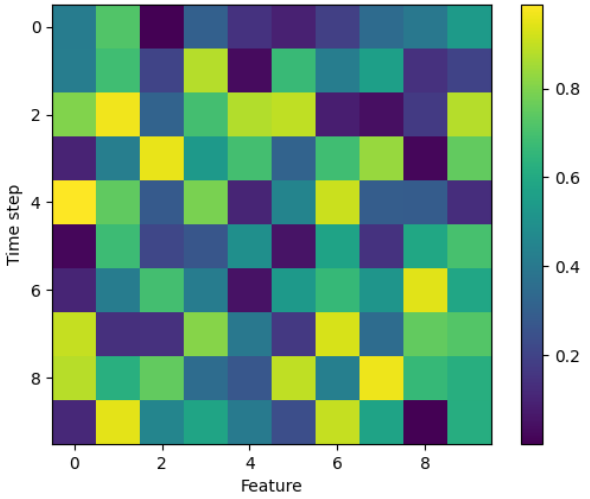


Рис. 5. Attention Heatmap для відео

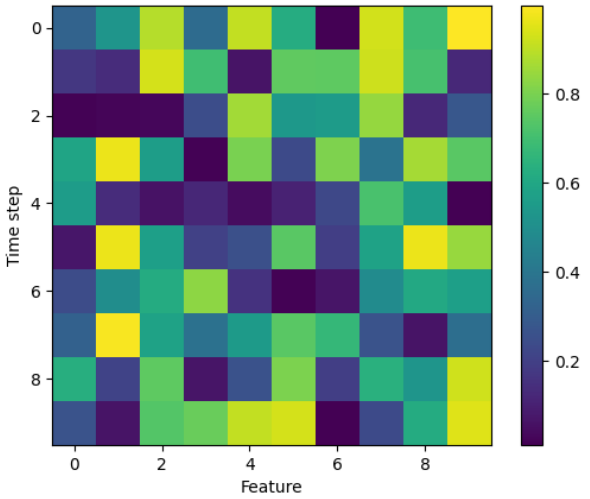


Рис. 6. Attention Heatmap для audio

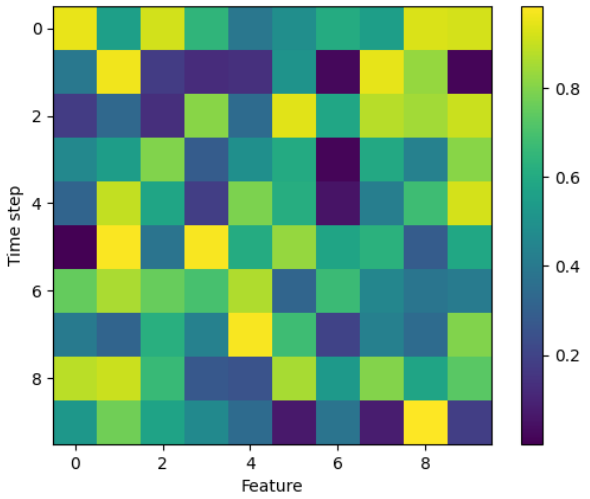


Рис. 7. Attention Heatmap для text

Світлі ділянки (жовті/зелені) означають, що модель сильно фокусується на відповідних ознаках як, наприклад, рух губ (відео), голосові переходи (аудіо) та ключові слова (текст). Темні ділянки (сині/фіолетові) – це менш важливі ознаки, наприклад, шум або нерелевантні дані. Нерівномірність карти уваги означає, що модель вибірково “дивиться” на критичні моменти, бо діпфейки часто містять локальні артефакти.

На рис. 8 графік показує 3D-візуалізацію комбінованого attention, де вісь X – ознаки (features), вісь Y – часові кроки, вісь Z – рівень уваги.

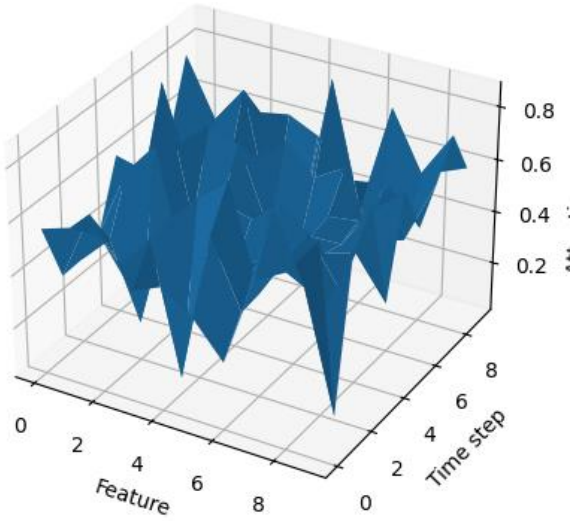


Рис. 8. 3D-візуалізацію Attention Heatmap для відео для трьох модальностей (відео, аудіо, текст)

Комбінований attention можна формально описати як:

$$A_{fusion} = \frac{1}{3}(A_v + A_a + A_t),$$

де A_v, A_a, A_t – матриці уваги для відео, аудіо та тексту.

Це дозволяє інтегрувати важливість ознак з різних модальностей, виявляти міжмодальні невідповідності та підвищувати інтерпретованість моделі.

Візуалізація attention підтверджує, що модель ефективно виділяє релевантні ознаки, мультимодальний підхід забезпечує більш повне представлення даних та cross-attention дозволяє виявляти міжмодальні невідповідності, характерні для діпфейків.

На рис. 9 ОС-крива показує залежність TPR (True Positive Rate) – чутливість,та FPR (False Positive Rate) – рівень хибних спрацювань. Діагональна лінія – випадковий класифікатор. Отримано значення AUC≈–0.48.

ROC-крива та AUC визначається як:

$$TPR = \frac{TP}{TP+FN}, FPR = \frac{FP}{FP+TN},$$

$$AUC = \int_0^1 TPR(FPR)\ d(FPR).$$

Grad-CAM на рис. 10 відображає області кадру, на які модель звертає увагу. Яскраві області – це найбільш важливі для рішення, наприклад, область рота (синхронізація) та очі (мікроекспресії). Темні області – це менш значущі.

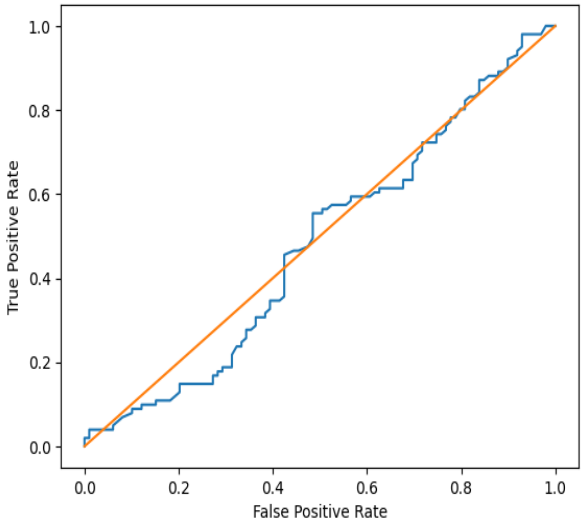


Рис. 9. ROC-крива з AUC

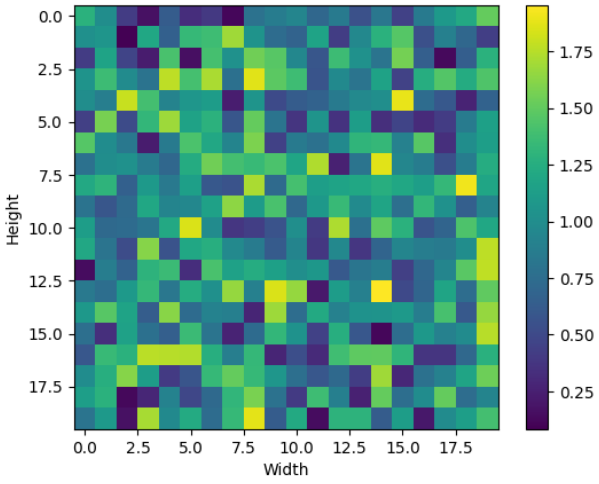


Рис. 10. Grad-CAM візуалізація для відео

Формалізація Grad-CAM:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k),$$

де A^k – feature maps, α^c_k– ваги важливості:

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}.$$

ROC-крива підтверджує ефективність класифікаційної моделі. AUC є ключовою метрикою якості, яку необхідно при наступних дослідженнях покращити. Grad-CAM забезпечує інтерпретованість. Модель фокусується на критичних ділянках обличчя, що характерно для задачі виявлення діпфейків

Висновки

У статті розглянуто проблему виявлення діпфейків як одну з актуальних задач сучасної інформаційної безпеки та цифрової медіааналітики. Запропоновано мультимодальний підхід до детекції синтетичного контенту, який базується на поєднанні відео-, аудіо- та текстових даних із використанням методів глибинного навчання. Розроблено математично обґрунтований пайплайн, що включає етапи попередньої обробки даних, виділення ознак за допомогою нейронних мереж, застосування механізмів self-attention і cross-attention, а також мультимодального об’єднання ознак. Особливу увагу приділено використанню трансформерних архітектур для моделювання внутрішньо- та міжмодальних залежностей. Проведені експериментальні дослідження продемонстрували, що запропонований підхід забезпечує підвищення точності виявлення діпфейків порівняно з одномодальними моделями. Отримані результати підтверджують ефективність інтеграції різних типів даних та доцільність використання attention-механізмів для підвищення інтерпретованості моделі. Запропонований підхід може бути використаний у системах автоматизованого аналізу медіаконтенту, цифрової криміналістики та протидії інформаційним загрозам.

У даному дослідженні розроблено та проаналізовано підхід до виявлення діпфейків на основі мультимодального аналізу із застосуванням сучасних методів глибинного навчання, зокрема трансформерних архітектур та attention-механізмів. У ході роботи сформовано повний математично обґрунтований пайплайн обробки даних, що включає:

– формалізацію мультимодальних вхідних даних (відео, аудіо, текст);

– попередню обробку та сегментацію;

– виділення ознак за допомогою глибинних нейронних мереж;

– застосування self-attention для моделювання внутрішніх залежностей;

– використання cross-attention для інтеграції міжмодальних зв’язків;

– мультимодальне об’єднання ознак;

– класифікацію та оптимізацію моделі;

– формування фінального рішення в реальному часі.

Отримані експериментальні результати показали, що використання мультимодального підходу забезпечує суттєве підвищення якості виявлення діпфейків. Зокрема, модель, яка інтегрує відео, аудіо та текстові модальності із застосуванням механізмів cross-attention, досягла найкращих показників точності (Accuracy ≈ 0.95) та F1-міри (≈ 0.925), що перевищує результати моделей, які використовують лише одну модальність. Проведений аналіз attention-механізмів показав, що модель здатна ефективно фокусуватися на критичних ділянках даних, зокрема аномаліях міміки та руху губ у відео, синтетичних артефактах у голосі тасемантичних невідповідностях у тексті. Візуалізація attention-карт та їх тривимірне представлення підтвердили, що мультимодальна інтеграція дозволяє виявляти складні міжмодальні залежності та невідповідності, які є характерними для діпфейків і не можуть бути виявлені при аналізі окремих модальностей.

Запропонований підхід має такі основні переваги:

– підвищена точність завдяки інтеграції різних джерел інформації;

– здатність виявляти приховані та складні патерни діпфейків;

– висока інтерпретованість результатів через attention-механізми;

– можливість застосування в реальному часі.

Водночас дослідження має певні обмеження, серед якихвисока обчислювальна складність моделей трансформерного типу, залежність якості результатів від синхронізації модальностей та чутливість до шумів у аудіо та відео даних.

Подальші напрями досліджень можуть включати:

– оптимізацію архітектури для зменшення обчислювальних витрат;

– використання більш ефективних варіантів attention (sparse, linear attention);

– розширення набору модальностей (наприклад, біометричні або поведінкові дані);

– застосування методів explainable AI для глибшого аналізу рішень моделі;

– адаптацію моделі до умов потокової обробки великих обсягів даних.

Запропонований мультимодальний підхід є ефективним та перспективним рішенням для задачі виявлення діпфейків і може бути використаний у практичних системах інформаційної безпеки, медіааналізу та цифрової криміналістики.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser Ł., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
2. Xu, Y., Wei, H., Lin, M. et al. Transformers in computational visual media: A survey. Comp. Visual Media 8, 33–62 (2022). https://doi.org/10.1007/s41095-021-0247-3
3. Hafiz, A. M., Parah, S. A., & Bhat, R. U. A. (2021). Attention mechanisms and deep learning for machine vision: A survey of the state of the art. arXiv preprint arXiv:2106.07550. https://doi.org/10.48550/arXiv.2106.07550
4. Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113-12132. https://doi.org/10.1109/TPAMI.2023.3275156
5. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 6558-6569). https://doi.org/10.48550/arXiv.1906.00295
6. Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications, 241, 122666. https://doi.org/10.48550/arXiv.2306.07303
7. Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. Journal of Imaging, 9(6), 122. https://doi.org/10.3390/jimaging9060122
8. Raza, M. A., & Malik, K. M. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 993-1000). https://doi.org/10.3390/info17040347
9. Erokhin, D., & Komendantova, N. (2026). A Review of Tools and Technologies to Combat Deepfakes. Information, 17(4), 347. https://doi.org/10.3390/info17040347
10. Nailwal, S., Singhal, S., Singh, N. T., & Raza, A. (2023, November). Deepfake detection: A multi-algorithmic and multimodal approach for robust detection and analysis. In 2023 international conference on research methodologies in knowledge management, artificial intelligence and telecommunication engineering (RMKMATE) (pp. 1-8). IEEE. https://doi.org/10.1109/RMKMATE59243.2023.10369155
11. Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M., & Ghag, K. (2024). A multimodal framework for deepfake detection. arXiv preprint arXiv:2410.03487. https://doi.org/10.48550/arXiv.2410.03487
12. Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520
13. Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: a survey of deep learning techniques. Social Network Analysis and Mining, 13(1), 101. https://doi.org/10.1007/s13278-023-01104-w
14. Sedik, A., Faragallah, O. S., El-sayed, H. S., El-Banby, G. M., El-Samie, F. E. A., Khalaf, A. A., & El-Shafai, W. (2022). An efficient cybersecurity framework for facial video forensics detection based on multimodal deep learning. Neural Computing and Applications, 34(2), 1251-1268. https://doi.org/10.1007/s00521-021-06416-6
15. Vysotska, V., Smelyakov, K., Chupryna, A., Darahan, D., Torubara, O., & Shyshymenko, O. (2025). Social engineering in Ukraine: Threats and intelligent detection approaches. In CEUR Workshop Proceedings (Vol. 4110, pp. 317-331). https://ceur-ws.org/Vol-4110/paper24.pdf
16. Tan, D., Yang, Y., Niu, C., Li, S., Yang, D., & Tan, B. (2025). A review of deep learning based multimodal forgery detection for video and audio. Discover Applied Sciences, 7(9), 987. https://doi.org/10.1007/s42452-025-07629-3
17. Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. PeerJ Computer Science, 10, e2037. https://doi.org/10.7717/peerj-cs.2037
18. Vysotska, V., Nazarkevych, M., Vladov, S., Lozynska, O., Markiv, O., Romanchuk, R., & Danylyk, V. (2024). Devising A Method For Detecting Information Threats In The Ukrainian Cyber Space Based On Machine Learning. Eastern-European Journal of Enterprise Technologies, 132(2). 132, Issue 2, p36. https://doi.org/10.15587/1729-4061.2024.317456

Received (Надійшла) 03.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Висоцька Вікторія Анатоліївна – доктор технічних наук, доцент, професор кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна; професор кафедри протидії кіберзлочинності, Харківський національний університет внутрішніх справ, Харків, Україна;
Victoria Vysotska – Doctor of Technical Sciences, Associate Professor, Professor of the Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine; Professor of Counteracting Cybercrime Department, Kharkiv National University of Internal Affairs, Kharkiv, Ukraine;
e-mail: victoria.a.vysotska@lpnu.ua; ORCID Author ID: https://orcid.org/0000-0001-6417-3689;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=24484045400.

Чирун Любомир Вікторович – кандидат технічних наук, доцент, доцент кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна; доцент кафедри комп'ютерних наук, Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна;
Lyubomyr Chyrun – Candidate of Technical Sciences, Associate Professor, Associate Professor, Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine; Associate Professor, Department of Computer Science, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine;
e-mail: lyubomyr.v.chyrun@lpnu.ua; ORCID Author ID: https://orcid.org/0000-0002-9448-1751;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55225672300.

Лаврут Олександр Олександрович – доктор технічних наук, професор, професор кафедри тактико спеціальних дисциплін, Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна;
Oleksandr Lavrut – Doctor of Technical Sciences, Professor, Professor at the Department for Tactical and Special Disciplines Hetman Petro Sahaidachnyi National Army Academy, Lviv, Ukraine;
e-mail: alexandrlavrut@gmail.com; ORCID Author ID https://orcid.org/0000-0002-4909-6723;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217195493.

Лаврут Тетяна Валеріївна – кандидат географічних наук, доцент, старший дослідник, провідний науковий співробітник науково-дослідного відділу Наукового центру Сухопутних військ Національна академія сухопутних військ імені гетьмана Петра Сагайдачного, Львів, Україна;
Tetiana Lavrut – Candidate of Geographical Sciences, Associate Professor, Senior Researcher, Leading Researcher Army Scientific Center of the Hetman Petro Sahaidachnyi National Army Academy, Lviv, Ukraine;
e-mail: lavrut_t_v@i.ua; ORCID Author ID https://orcid.org/0000-0002-1552-9930;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217204350.

Романчук Роман Васильович – аспірант кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна;
Roman Romavchyk – PhD student of the Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine;
e-mail: roman.v.romanchuk@lpnu.ua; ORCID Author ID: https://orcid.org/0009-0004-4352-1073;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=58765557000.

Information technology of deepfake detection based on deep learning and multimodal analysis for intellectual information security systems

Victoria Vysotska, Lyubomyr Chyrun, Oleksandr Lavrut, Tetiana Lavrut, Roman Romavchyk

Abstract. Relevance. The rapid development of deep learning technologies has led to the emergence of high-quality synthetic media content (deepfakes), posing a significant threat to information security, digital trust, and the media space. Modern methods for detecting deepfakes, based on the analysis of individual modalities (video, audio, or text), often lack sufficient accuracy and generalizability, necessitating the development of multimodal approaches. **Object of research.** Processes of detecting synthetic media content (deepfakes) in the digital information environment. **Purpose of the article.** Development of an effective method for detecting deepfakes based on multimodal analysis using deep learning models and attention mechanisms. **Research results.** The paper proposes an information technology for detecting deepfakes based on complex processing of video, audio, and text data. A generalised pipeline has been developed that includes pre-processing of media content, feature extraction for each modality, multimodal integration, and classification. To increase efficiency, transformer architectures using self-attention and cross-attention mechanisms were employed, enabling modelling intra- and intermodal dependencies. Experimental studies on public datasets demonstrated that the proposed approach increases the accuracy of deepfake detection to 0.95 and the F1-measure to 0.925, exceeding the results of single-modal models. **Conclusions.** The results confirm the feasibility of a multimodal approach and attention mechanisms for deepfake detection. The proposed information technology provides increased accuracy and interpretability and can be used in information security systems, digital forensics, and automated media content analysis. Prospects for further research include optimising the computational complexity of models and adapting them for real-time streaming data processing.

Keywords: cybersecurity, deepfake, multimodal analysis, deep learning, transformers, attention mechanism, information security, synthetic media content, computer vision, audio processing, machine learning.

УДК 004. 4 **doi: 10.26906/SUNZ.2026.2.062**

Б. Ю. Вінтенко[1,2], Т. В. Смірнова[3], І. В. Миронець[2], О. А. Смірнов[3], К. О. Буравченко[3]

[1] ПАТ “Науково-виробниче підприємство “Радій”, Кропивницький, Україна
[2] Черкаський державний технологічний університет, Черкаси, Україна
[3] Центральноукраїнський національний технічний університет, Кропивницький, Україна

МЕТОД ОЦІНКИ ФУНКЦІОНАЛЬНОЇ СТІЙКОСТІ КОМП'ЮТЕРНО-ОРІЄНТОВАНИХ ПРОЦЕДУР СИСТЕМИ ПІДТРИМКИ ОПЕРАТИВНОГО ПЕРСОНАЛУ АЕС

Анотація. Актуальність. Атомні електростанції є великими високотехнологічними підприємствами, що містять велику кількість обладнання, складних процесів перетворення енергії та інформаційно-керуючих систем. Послідовності дій керування, що описані в регламентах, на АЕС мають вигляд паперово-орієнтованих або комп’ютерно-орієнтованих процедур. Використання комп’ютерно-орієнтованих процедур надає можливість створення систем підтримки оперативного персоналу, які підвищують надійність та зменшують навантаження на оперативний персонал при виконанні складних операцій. Важливою вимогою до систем підтримки оперативного персоналу, що використовуються на АЕС, є забезпечення відмовостійкості. При виникненні відмови у системи критичного застосування оператору необхідно мати достовірну інформацію, як сильно відмова впливає на працездатність системи та чи здатна система виконувати свої функції, тобто оцінити її функціональну стійкість. Для цього необхідно констатувати не тільки факт відмови, а і сформувати чисельну оцінку рівня функціональної стійкості за відповідною шкалою. **Об’єкт дослідження**: процес функціональної стійкості елементів комп’ютерно-орієнтованої процедури у складі інформаційної системи підтримки оперативного персоналу АЕС. **Мета статті**: розробка методу чисельної оцінки функціональної стійкості елементів інформаційної системи підтримки оперативного персоналу АЕС. **Результати дослідження**. Розглянута оцінка працездатності компонентів комп’ютерно-орієнтованої процедури у складі системи підтримки оперативного персоналу АЕС. Запропоноване використання показників функціональної стійкості в якості числових критеріїв оцінювання. Запропоновано метод обчислення максимального, поточного та критичного рівня функціональної стійкості. **Висновки.** Даний метод відрізняється від відомих методів діагностики не тільки оцінкою факту наявності відмови, а і визначенням кількісної оцінки здатності системи виконувати свої функції. Запропонований метод може бути застосований для діагностування компонентів інформаційних систем критичного застосування, що отримають інформацію від резервованих та диверсифікованих джерел даних.

Ключові слова: чисельна оцінка, функціональна стійкість, інформаційна система, система підтримки прийняття рішень, підтримка оперативного персоналу, АЕС.

Вступ

Постановка проблеми. Атомні електростанції є великими високотехнологічними підприємствами, що містять велику кількість обладнання, складних процесів перетворення енергії та інформаційно-керуючих систем (ІКС). Основний оперативний персонал (ОП), що керує енергоблоком АЕС, знаходиться на блочному щиті керування (БЩУ). Для керування енергоблоком ОП використовує прилади, індикатори, табло, монітори та ключі керування, які знаходяться на БЩУ. Керування здійснюється згідно з технологічними регламентами та інструкціями, виконання вимог яких є обов’язковим для ОП [1].

Послідовності дій керування, що описані в регламентах, на АЕС мають вигляд паперово-орієнтованих або комп’ютерно-орієнтованих процедур (КОП). Використання КОП надає можливість створення систем підтримки оперативного персоналу (СПО), які підвищують надійність та зменшують навантаження на ОП при виконанні складних операцій [2].

СПО, які містять інформацію про різноманітні КОП, можуть працювати в реальному часі. При цьому вони безперервно отримають інформацію від суміжних систем енергоблоку цифровими каналами у вигляді сигналів та формують значення технологічних параметрів. Після цього значення параметрів використовуються для ідентифікації настання умов, при яких ОП має виконати керуючі дії згідно процедур.

Важливою вимогою до СПО, що використовуються на АЕС, є забезпечення відмовостійкості. Для цього застосовується велика кількість методів: самоконтроль, самодіагностика, резервування джерел даних та використання багатоверсійних технологій [3, 4]. Ці методи зменшують імовірність відмови системи та своєчасно інформують оператора про настання такої відмови. Водночас, при виникненні відмови у системи критичного застосування оператору необхідно мати достовірну інформацію, як сильно відмова впливає на працездатність системи та чи здатна система виконувати свої функції, тобто оцінити її функціональну стійкість [5]. Для цього необхідно констатувати не тільки факт відмови, а і сформувати чисельну оцінку рівня функціональної стійкості за відповідною шкалою.

Аналіз останніх досліджень і публікацій. У роботі [1] досліджено організацію процесу керування енергоблоком; розглянуті основні інформаційно-керуючі системи та програмно-технічні комплекси енергоблоку, а також засоби їх взаємодії з оперативним персоналом; розглянута інформаційну модель, яка використовується оперативним персоналом при керуванні енергоблоком; проаналізовано структуру технологічних регламентів та інструкцій з керування енергоблоком і визначити об’єм та складність цих документів; визначено доцільність досліджень щодо покращення процесів керування енергоблоком АЕС. Робота [2] присвячена перехід від використання паперових процедур до комп'ютерних процедур у видах діяльності,

що передбачають взаємодію людини з системами на атомній електростанції, що створює потребу в комп'ютерній системі процедур. Визначено що, щоб надати комп'ютерній системі процедур необхідну інформацію для відображення кроків процедури користувачеві, потрібна особлива увага до формату, який використовується для передачі всіх даних та інструкцій для створення кроків. Процедура повинна бути розбита на основні елементи та відформатована стандартним методом для комп'ютерної системи процедур. Комп'ютерна система процедур забезпечить контекст для кроку для передачі довідкової інформації, запиту рішення або прийняття вхідних даних від користувача. У роботі [3] наведено довідник з критично важливих систем безпеки: простий посібник з функціональної безпеки: IEC 61508 та IEC 61511, й відповідні вказівки, де представлено найновіші відомості про електричні, електронні та програмовані електронні системи, що забезпечують функції безпеки, які захищають працівників та населення від травм або смерті, а також навколишнє середовище від забруднення. У статті [4] запропонована платформа для оцінки ризиків системи цифрового контролю та управління (PRADIC), розроблена Національною лабораторією Айдахо (INL). Як частина цієї структури, було розроблено методологію оцінки програмних систем цифрового контролю та управління у високо значущих для безпеки систем цифрового контролю та управління АЕС. Структура об'єднує три етапи типової оцінки ризиків: якісний аналіз небезпек та кількісний аналіз надійності та наслідків. Кількісно визначені ризики порівняно з відповідними критеріями прийнятності надають цінну інформацію для альтернатив архітектури системи, що дозволяє оптимізувати проектування з точки зору зниження ризиків та економії коштів. Результати показують, що PRADIC є потужним інструментом, здатним виявляти потенційні цифрові можливості виникнення відмов за спільною причиною, оцінювати їх ймовірність та оцінювати їхній вплив на безпеку системи та електростанції. У статті [5] визначено функціональну стійкість, як властивість складних технічних систем. Стаття [6] присвячена розв’язанню наступних завдань: розробці структури даних моделі комп’ютерно-орієнтованої процедури, що відповідає актуальним регламентам та інструкціям; створенню логіко-структурної моделі комп’ютерно-орієнтованої процедури, що буде використовуватися для створення систем підтримки оперативного персоналу АЕС; розробці методу ідентифікації умов входу в процедуру та виконання її кроків. Дана модель, розроблена для вирішення задачі підтримки актуальності, ранньої ідентифікації станів енергоблоку та його обладнання, забезпечення можливості отримання інформації про стан технологічних параметрів з необхідним рівнем надійності та достовірності. Визначено, що запропонована у роботі модель може бути використана для створення систем підтримки оперативного персоналу, що керує складними технологічними об’єктами з використанням визначених процедур та інструкцій. Робота [7] присвячена побудові функціонально стійких розподілених інформаційних систем. У роботі [8] розглянутий досвід належного консерватизму у справах безпеки ядерної промисловості Великої Британії. У роботі [9] визначено у рамках дій з ліквідації порушень нормальної експлуатації (ІЛП) енергоблоку №4 РАЕС наступне: 50 процедур, до 10 правил активації процедури, до 3-х правил завершення процедури, до 10 кроків на процедуру, до 10 підкроків на кожний крок процедури, до 3 умов активації та завершення кроку. У роботах [10, 11] проведено дослідження застосування систем підтримки оперативного персоналу об’єкту критичної інфраструктури при керуванні енергоблоком АЕС з реактором типу ВВЕР-1000 та розроблена модель шляхів отримання вхідних даних комп’ютерної інтелектуальної системи підтримки оперативного персоналу АЕС.

Метою роботи є розробка методу чисельної оцінки функціональної стійкості елементів інформаційної системи підтримки оперативного персоналу АЕС.

Для досягнення поставленої мети, у даному дослідженні необхідно розв’язати наступні завдання:

1. Проаналізувати, які елементи КОП у складі СПО найбільше впливають на функціональну стійкість системи.

2. Розробити метод чисельної оцінки функціональної стійкості елементів СПО.

3. Реалізувати засіб моделювання відмов елементів СПО в об’ємі технологічних регламентів керування АЕС та оцінити вплив цих відмов на показники функціональної стійкості.

1. Модель КОП СПО на основі продукційних правил

У технологічних регламентах, які є основою для створення СПО у даному дослідженні, вказані дії оператора, які мають бути виконані в залежності від вихідних умов та результатів попередніх дій. Елементи даних регламентів розділяються на два основні типи: правила (Rules) та дії (Actions). В термінології інтелектуальних систем дані елементи утворюють систему продукційних правил на основі дерев рішень.

У рамках дослідження [6] було розроблено модель комп'ютерно-орієнтованої процедури СПО на основі графу. Вершинами графа є правила та дії, ребрами – переходи між діями згідно правил. Модель такої процедури формальна описана функцією:

$$P = f(\{R\}, \{A\}, \{L\}), \qquad (1.1)$$

де {R} – множина всіх правил, {A} – множина всіх дій, {L} – сукупність всіх переходів.

На рис. 1.1 приведена структурна схема моделі КОП згідно наведених складових.

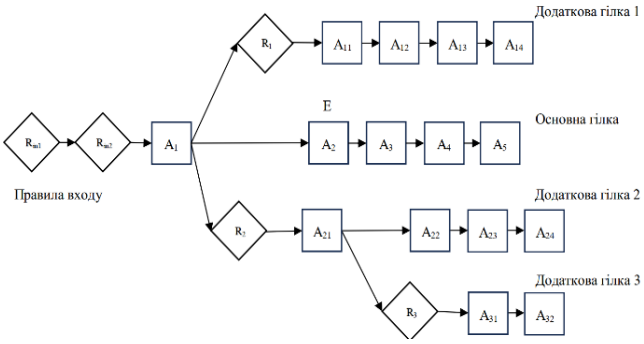


Рис. 1.1. Хід виконання процедури

Спрацювання правил активує гілки дій, які має виконати оператор. Також правилами визначається факт виконання дії оператором. Для визначення факту спрацювання правила використовуються значення технологічних параметрів, визначені в реальному часі, та вагові коефіцієнти. В свою чергу, значення кожного технологічного параметру P визначається на основі набору сигналів, отриманих від суміжних систем, логікою та власними ваговими коефіцієнтами:

$$P = f(\{s\}, l, \{\omega\}), \qquad (1.2)$$

де $\{s\}$ – сигнали, l – логіка (алгоритм) обробки, ω – вагові інформаційні внески.

Наприклад, при відключенні головного циркуляційного насоса (ГЦН), системою нормальної експлуатації формується дискретний сигнал #GCN01_OFF, що спричиняє відображення інформації про дану подію на табло БЩУ. Для КОП відключення насоса **є параметром**, а дискретний **сигнал** є таким, що **формує даний параметр**.

Саме сигнали, отримані від суміжних систем, є визначальними даними для фіксації поточного стану правил процедури керування.

2. Метод визначення ФС КОП

Система підтримки оператора може функціонувати в умовах відсутності певної частини вхідних даних. Для забезпечення даної здатності в даному дослідженні пропонується використовувати резервовані вхідні дані різної природи, а також розробити алгоритм заміщення відсутніх даних іншими. Для роботи такого алгоритму необхідний критерій, при досягненні якого буде відбуватися заміщення. В якості такого критерію пропонується використовувати показник функціональної стійкості розподілених інформаційних систем [7].

Під функціональною стійкістю СПО будемо розуміти здатність даної системи надати оператору інформацію про стан та актуальні кроки керуючих процедур у будь-який момент часу в автоматичному режимі за відсутності або неактуальності частини даних через відмову або виведення обладнання в ремонт.

Стан КОП в системі визначається спрацюванням правил. В свою чергу, визначення факту спрацювання відбувається на основі значень технологічних параметрів, даними для яких є вхідні сигнали.

Розглянемо процес розрахунку функціональної стійкості для різних компонентів даної системи.

КОП можна вважати фрагментом розподіленої інформаційної системи на основі наявності ієрархії окремих компонентів: вхідні сигнали отримуються з різноманітних інформаційних систем; стани параметрів обчислюються на основі значення сигналів, правила спрацьовують на основі станів параметрів і формують інформацію для людини-оператора. Дану ієрархію зобразимо у вигляді графу на рис. 2.1.

Вершини $\{s\}$ даного графу є мажоритованими резервованими наборами сигналів, $\{P\}$ – параметрами, $\{R\}$ – правилами. Кожне ребро означає напрям передачі даних між елементами та має ваговий коефіцієнт ω.

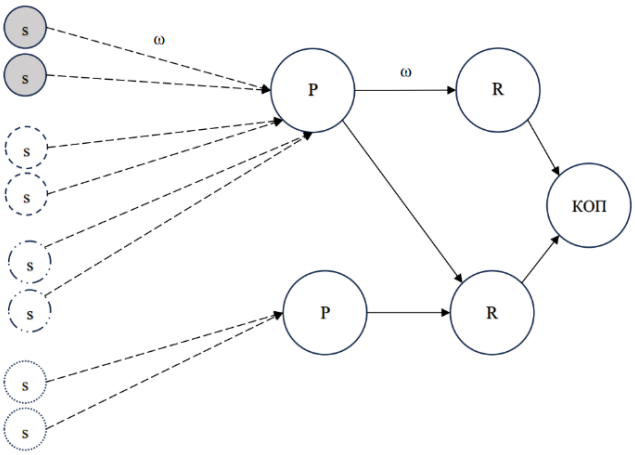


Рис. 2.1. КОП як розподілена інформаційна система

Джерелом негативних впливів на дану систему є можливі недостовірні або відсутні значення сигналів $\{s\}$, що з точки зору функціональної стійкості спричиняє розрив окремих ребр графу, тобто знижує вершинну зв'язність. Проте, вершинна зв'язність не є зручною в якості показника функціональної стійкості, оскільки кожний зв'язок між елементами має свою вагу (ω). Більш зручним є використання критерію імовірнісної зв'язності розподіленої системи, що у теорії ФС характеризує імовірність встановлення зв'язку між парами вузлів. У контексті КОП даний критерій використаємо в якості показника імовірності достовірного визначення стану її елементів.

ФС наборів сигналів. Сигнали, які є вхідною інформацією для КОП, об'єднані в набори з метою резервування. Кожний набір сигналів $\{s\}$ містить інформацію від обраних з загальної множини сигналів, отриманих від зовнішніх джерел – інших ПТК енергоблоку. Після прийому значення сигналів піддаються логічній обробці алгоритмом отримання значення технологічного параметру шляхом голосування або обчислення (усереднення, медіана). У даному дослідженні прийняте використання схеми «M/N», при якому кожний набір містить N сигналів, і для достовірного визначення значення набору необхідні наявність та узгодження M сигналів (рис. 2.2).

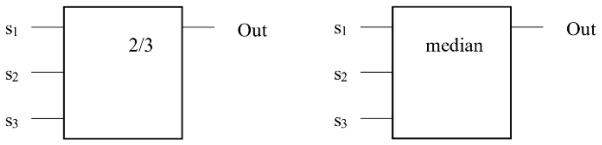


Рис. 2.2. Варіанти обчислення значення набору сигналів

Приймемо, що кожний вхідний сигнал s_i має імовірність наявності значення(безвідмовної роботи) $\mathrm{Rel}(s_i)$. Під **максимальним показником** ФС набору сигналів будемо вважати суму імовірностей безвідмовного стану всіх N сигналів:

$$FS_{max} = \sum_{i=1}^{N} \mathrm{Rel}(s_i), \qquad (2.1)$$

Це значення буде прирівняне до **100% функціональної стійкості** набору сигналів, яка враховуватиме максимальну надмірність при наявності всіх N

сигналів. Для прирівнювання ФС до відсотків обчислимо коефіцієнт пропорційності

$$k_{fs} = \frac{100}{FS_{max}}. \quad (2.2)$$

Достовірне визначення значення з набору сигналів можливе, коли присутні значення *M* сигналів і можливі *M* визначень значень. Якщо у роботі залишаться менше ніж N сигналів, то функціональна стійкість набору знизиться. Проте, якщо у роботі залишаться *M* сигналів з *мінімальним значенням імовірності* безвідмовної роботи - це буде досягнення критичного показника ФС, тобто при подальшій відмові будь-якого елемента отримується гарантована відмова при визначенні значення набору сигналів. Даний рівень визначимо як **критичне значення** функціональної стійкості набору сигналів. Він виражається як сума мінімальних значень імовірностей безвідмовної роботи M сигналів:

$$FS_{critical} = \sum_{i=1}^{M} Rel(s_i), \quad (2.3)$$

де $Rel(s_i)$-імовірності безвідмовної роботи сигналів, відсортовані за зростанням, M≤N.

Поточний рівень ФС визначається як значення імовірності суми подій (визначень стану) з фактично достовірних (наявних) N сигналів. Достовірність сигналу s_i визначається функцією $V(i) \in \{0, 1\}$ (0 – недостовірність, 1 – достовірність):

$$FS_{current} = \begin{matrix} s \subseteq \{i \mid V(i) = 1\} \\ |S| = N \end{matrix} \left(\sum_{i \in S} Rel(s_i) \right). \quad (2.4)$$

Запас функціональної стійкості визначається з різниці між поточною та критичною функціональною стійкістю:

$$FS_{reserve} = FS_{current} - FS_{critical}. \quad (2.5)$$

Усі визначені абсолютні значення ФС можуть бути виражені у відсотках з використанням коефіцієнту, визначеного у 2.2.

За наведеними формулами в один етап визначається функціональна стійкість технологічних параметрів КОП, що формуються зі значень наборів сигналів.

ФС правил КОП. Технологічні параметри використовуються для фіксації стану правил КОП {R} з використанням інформаційних вагових внесків {ω}. Таким чином, умовою можливості визначення стану параметрів та правил є не тільки наявність та узгодженість, а і достатня «вага» наборів даних (рис. 2.3).

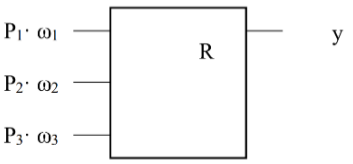


Рис. 2.3. Визначення стану параметрів та правил

Для розрахунку ФС правила кожний технологічний параметр P_i має мати попередньо розраховані показники ФС. Під **максимальним показником** ФС правила за середньою оцінкою будемо вважати **середнє** значення ФС всіх технологічних параметрів, помножених на вагові коефіцієнти:

$$FS(Rule)_{max}^{mean} = \overline{FS(P_i)_{max} \cdot \omega_i}. \quad (2.6)$$

За консервативною оцінкою [8] за **максимальний показник ФС** будемо вважати **максимальне** значення серед ФС технологічних параметрів, помножених на вагові коефіцієнти:

$$FS(Rule)_{max}^{conservative} = \max(FS(P_i)_{max} \cdot \omega_i). \quad (2.7)$$

Це значення становить **100%** функціональної **стійкості** правила**.** Для прирівнювання ФС до відсотків використовується (2.2).

Під **критичним та поточним показниками** ФС правила будемо вважати **середні** (для середньої оцінки) або **мінімальні** (для консервативної оцінки) значення серед відповідних рівнів ФС всіх технологічних параметрів, помножених на вагові коефіцієнти:

– критичний середній:

$$FS(Rule)_{critical}^{mean} = \overline{FS(P_{i_{critical}}) \cdot \omega_i}; \quad (2.8)$$

– поточний середній:

$$FS(Rule)_{current}^{mean} = \overline{FS(P_{i_{current}}) \cdot \omega_i}; \quad (2.9)$$

– критичний консервативний:

$$FS(Rule)_{critical}^{conservative} = \min(P_{i_{critical}} \cdot \omega_i); \quad (2.10)$$

– поточний консервативний:

$$FS(Rule)_{current}^{conservative} = \min(P_{i_{current}} \cdot \omega_i). \quad (2.11)$$

Запас функціональної стійкості визначається з різниці між поточною та критичною функціональною стійкістю згідно (2.5). Усі визначені абсолютні значення ФС можуть бути виражені у відсотках з використанням коефіцієнту, визначеного у (2.2).

ФС КОП та СПО. На основі обчислених показників ФС кожного правила може бути обчислена ФС всієї КОП. В свою чергу, обчислення ФС всіх КОП надає можливість оцінити ФС всієї СПО. Виходячи з того, що у даному дослідженні розглядається створення СПО для АЕС як системи критичного застосування, для визначення ФС КОП та СПО пропонується використовувати консервативну оцінку, яка являє собою найнижче поточне значення ФС серед усіх правил:

$$FS_{proc} = \min(\{FS_R\}); FS_{system} = \min(\{FS_{proc}\}).$$

3. Реалізація оцінки ФС КОП та моделювання відмов

До складу розроблюваного прототипу СПО входить програма OpSupport, що виконує функції власне інформаційної підтримки оператора. Вона оперує наступними даними:

– значення вхідних сигналів, що отримуються мережевим протоколом від серверів ПТК енергоблоку;

– базу даних технологічних параметрів, для кожного з яких визначені кластеризовані сигнали та скрипти розрахунку;

– базу даних правил та дій.

Під час функціонування програми проводиться безперервний розрахунок показників функціональної стійкості для технологічних параметрів (рис. 3.1), правил та дій (рис. 3.2).

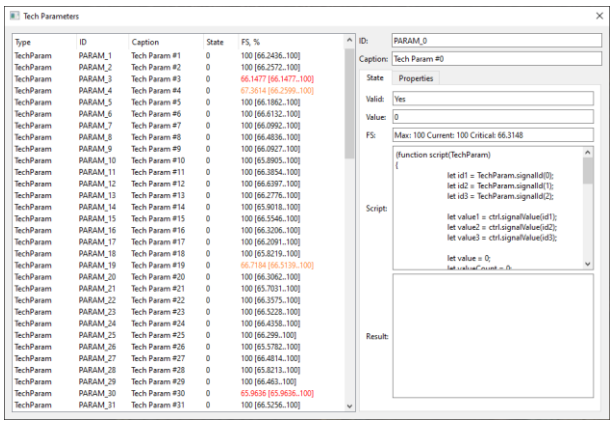


Рис. 3.1. Розрахунок ФС параметрів

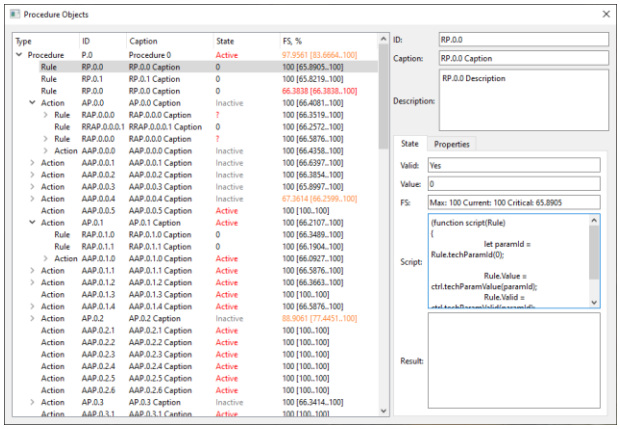


Рис. 3.2. Розрахунок ФС правил та дій

Програма виконує розрахунок ФС на основі показників прогнозованої імовірності надійної роботи сигналів та визначеної логіки мажоритування M/N. Розрахунок, описаний у розділі 2, обчислює максимальний, поточний та критичний рівні ФС у відсотках. Критичний рівень розраховується як сума імовірностей роботи M сигналів, відсортованих за зростанням (тобто мінімальним показником імовірності визначення, при якому можлива робота). Поточний визначається як сума імовірностей визначення значення N сигналів, що знаходяться в роботі в даний момент.

Для дослідження ефективності даних алгоритмів була змодельована база знань СПО з об'ємом, що відповідає документу ІЛП [9]: 50 процедур, до 10 правил активації процедури, до 3-х правил завершення процедури, до 10 кроків на процедуру, до 10 підкроків на кожний крок процедури, до 3 умов активації та завершення кроку. В якості вхідних даних було змодельовано 1000 сигналів, імовірність безвідмовної роботи кожного з яких випадковим чином становить від 0.95 до 1 (рис. 3.3). Для проведення експериментів було реалізовано можливість імітування відмов заданої кількості сигналів. В кожному експерименті була імітована відмова 20 випадкових сигналів та збережено результати зниження ФС. Кожний експеримент проводився 10 разів за однакових умов.

Середнє значення критичного рівня ФС технологічних параметрів, що відмовили, становило **66.23%**. Середнє значення поточного рівня ФС становило **97.96%**. Відповідно, запас ФС у середньому становив **31.03%** (рис. 3.4)**.**

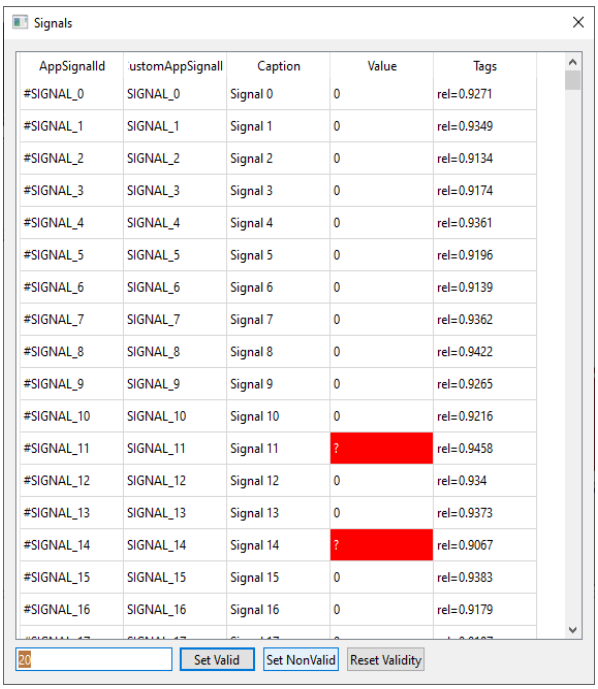


Рис. 3.3. Імітація відмов сигналів

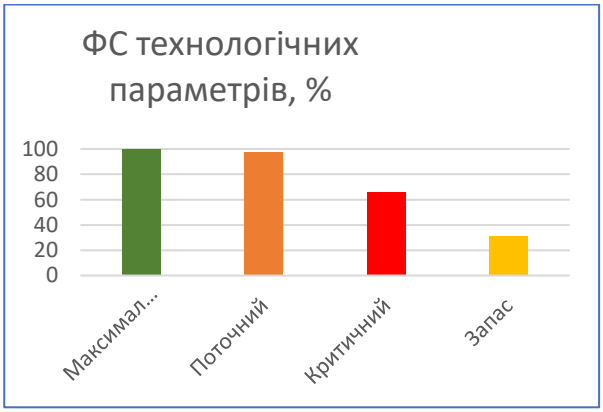


Рис. 3.4. Імітація відмов сигналів

При розрахунку ФС правил, дій та процедур реалізована можливість використання двох оцінок: середньої та консервативної.

Середня оцінка ФС обчислюється як середнє арифметичне відповідного показника ФС дочірніх елементів моделі КОП, тобто дочірніх правил та дій. Консервативна оцінка обирається за принципом мінімальної («найгіршої») оцінки серед усіх наявних. Середня оцінка є більш «чутливою» і надає можливість точніше оцінювати вплив відмов сигналів на працездатність елементів КОП.

Проте у системі критичного застосування, коли система або її частина вважається повністю непрацездатною при повній відмові хоча б одного елемента, більш вірним є використання консервативної оцінки для фіксації відмов.

Під час кожного експерименту також було зафіксовано зниження поточного ФС СПО за середньою та консервативною оцінкою. При консервативній оцінці критичний рівень ФС системи становив **65.78%**, поточний у середньому **74.35%**. При цьому середній запас ФС становив **8.57%** (рис. 3.5).

При середній оцінці критичний рівень ФС становив **75.6%**, поточний – **98.8%**. При цьому середній запас ФС становив у середньому **23.2%** (рис. 3.6).

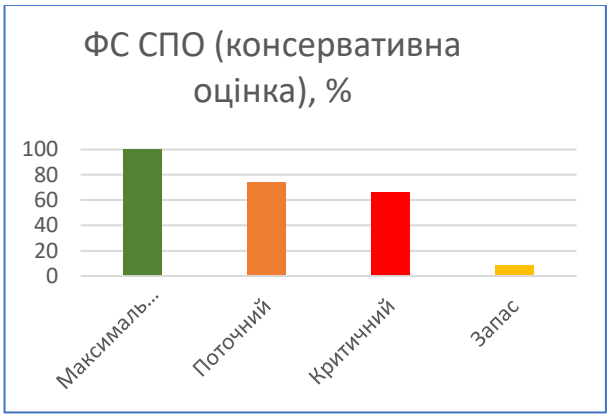


Рис. 3.5. Імітація відмов сигналів (консервативна оцінка)

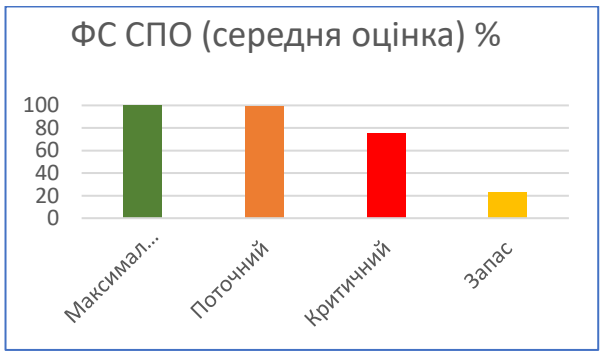


Рис. 3.6. Імітація відмов сигналів (середня оцінка)

Висновки

У дослідженні було запропоновано використання поняття функціональної стійкості для оцінки відмовостійкості елементів комп'ютерно-орієнтованої процедури у реальному часі.

Наведено метод розрахунку максимального, поточного та критичного рівнів функціональної стійкості.

Також було наведено приклад реалізації розрахунку рівнів функціональної стійкості для систем підтримки оперативного персоналу в об'ємі інструкції з ліквідації порушень (ІЛП) АЕС.

При проведеному моделюванні відмов вхідних сигналів було показано значне зниження запасу функціональної стійкості за консервативною оцінкою.

Напрямком подальших досліджень є методи підвищення відмовостійкості компонентів систем підтримки оперативного персоналу, ефективність яких буде оцінена підвищенням показника запасу функціональної стійкості.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Вінтенко Б.Ю., Миронець І.В., Смірнов О.А. Коваленко О.В., Смірнов С.А., Буравченко К.О., Якименко Н.М. «Дослідження інформаційного забезпечення та технологічних регламентів процесів керування критичною інфраструктурою енергоблоку АЕС з реактором типу ВВЕР-1000». Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка». 2024. № 1(25), С. 253–278. DOI: https://doi.org/10.28925/2663-4023.2024.25.253278
2. Bly A., Oxstrand J., Blanc K. Standardized Procedure Content and Data Structure Based on Human Factors Requirements for Computer-Based Procedures. 2015. Conf.: the 9th Nuclear Plant Instrumentation, Control & Human-Machine Interface Technologies topical meeting of the American Nuclear Society. https://www.researchgate.net/publication/287216133
3. Daivd J. Smith, Kenneth G. L. Simpson. The Safety Critical Systems Handbook. A Straightforward Guide to Functional Safety IEC 61508 (2010 Edition), IEC 61511 (2015 Edition) and Related Guidance: Forth Edition. Elsevier Ltd, 2016. https://shop.elsevier.com/books/the-safety-critical-systems-handbook/smith/978-0-12-805121-4#full-description
4. Bao, H., Zhang, H., Shorthill, T., Chen, E., Lawrence, S.: Quantitative evaluation of common cause failures in high safety-significant safety-related digital instrumentation and control systems in nuclear power plants. Reliability Engineering & System Safety 230, 108973 (2023). DOI: https://doi.org/10.1016/j.ress.2022.108973
5. Барабаш О.В., Кравченко Ю. В. Функціональна стійкість – властивість складних технічних систем. Збірник наукових праць Національної академії оборони України. Бюл. № 40. – К.: НАОУ, 2002. – С. 225-229. URL: https://nuou.org.ua/nauka/npub/trudi-unversitetu/
6. Вінтенко, Б.Ю., Миронець, І.В., Смірнов, О.А., Коваленко, О.В., Усік, П.С., Буравченко, К.О., Лисенко, І.А. «Логіко-структурна модель комп'ютерно-орієнтованої процедури системи підтримки оперативного персоналу АЕС». Кібербезпека: освіта, наука, техніка. 2025. Том 2 № 30. С. 413-427, 2025. DOI: https://doi.org/10.28925/2663-4023.2025.30.984
7. Барабаш О.В. Побудова функціонально стійких розподілених інформаційних систем / О.В. Барабаш. – К.: НАОУ, 2004. – 226 с. https://scholar.google.com/scholar?hl=uk&as_sdt=0,5&cluster=14371859060207516977
8. Appropriate conservatism in safety cases: A nuclear industry guide. 2015/ Safety Case Forum. Режим доступу: https://www.nuclearinst.com/media/brheshlg/appropriate-conservatism-in-safety-cases.pdf
9. Інструкція з ліквідації порушень нормальної експлуатації (ІЛП) енергоблоку №4 РАЕС. ВП «Рівненська АЕС», 2024. 286 с. URL: https://energoatom.com.ua/branch/filiia-vp-raes
10. Вінтенко, Б., Миронець, І., Смірнов, О., Коваленко, А., Конопліцька-Слободенюк , О., Смірнова , Т., Константинова, Л. «Дослідження застосування систем підтримки оперативного персоналу об'єкту критичної інфраструктури при керуванні енергоблоком АЕС з реактором типу ВВЕР-1000». *Електронне фахове наукове видання «Кібербезпека: освіта, наука, техніка»,* 2024. № 2(26), С. 6-26. DOI: https://doi.org/10.28925/2663-4023.2024.26.673

11. Вінтенко Б.Ю., Смірнов О.А., Миронець І.В., Смірнова Т.В., Коваленко О.В., Мацуй А.М. «Модель шляхів отримання вхідних даних комп'ютерної інтелектуальної системи підтримки оперативного персоналу АЕС». *Центрально-український науковий вісник. Технічні науки.* 2025. Вип. 11(42), ч. II. С.52-62. DOI: https://doi.org/10.32515/2664-262X.2025.11(42).2.63-69

Received (Надійшла) 28.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Вінтенко Борис Юрійович – аспірант кафедри інформаційної безпеки та комп'ютерної інженерії Черкаський державний технологічний університет, Черкаси, Україна; провідний інженер-програміст КБ АСУ ТП ПАТ "Науково-виробниче підприємство "Радій", Кропивницький, Україна;
Borys Vintenko – PhD Graduate Student of the Department of Information Security and Computer Engineering, Cherkasy State Technological University, Cherkasy, Ukraine, Leading Engineer-Programmer of KB ACS TP PJSC "Radio Scientific and Production Enterprise", Kropyvnytskyi, Ukraine;
e-mail: boris.vintenko@gmail.com; ORCID Author ID: https://orcid.org/0009-0008-3748-0374;

Смірнова Тетяна Віталіївна – кандидат технічних наук, доцент, старший викладач кафедри автоматизації виробничих процесів, Центральноукраїнський національний технічний університет, Кропивницький, Україна;
Tetiana Smirnova Candidate of Technical Sciences, Associate Professor, Senior Lecturer, Department of Automation of Production Processes, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: sm.tetyana@gmail.com; ORCID Author ID: https://orcid.org/0000-0001-6896-0612;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57219108044.

Миронець Ірина Валеріївна – кандидат технічних наук, доцент, доцент кафедри інформаційної безпеки та комп'ютерної інженерії, Черкаський державний технологічний університет, Черкаси, Україна;
Iryna Myronets – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Information Security and Computer Engineering, Cherkasy State Technological University, Cherkasy, Ukraine;
e-mail: i.myronets@chdtu.edu.ua; ORCID Author ID: https://orcid.org/0000-0003-2007-9943;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57208663236.

Смірнов Олексій Анатолійович – доктор технічних наук, професор, завідувач кафедри кібербезпеки та програмного забезпечення, Центральноукраїнський національний технічний університет, Кропивницький, Україна;
Oleksii Smirnov – Doctor of Technical Sciences, Professor, Head of Cybersecurity & Software Academic Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: dr.smirnovoa@gmail.com; ORCID Author ID: https://orcid.org/0000-0001-9543-874X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57208667815.

Буравченко Костянтин Олегович – кандидат технічних наук, доцент, доцент кафедри кібербезпеки та програмного забезпечення, Центральноукраїнський національний технічний університет, Кропивницький, Україна;
Kostiantyn Buravchenko – Candidate of Technical Sciences, Associate Professor, Associate Professor of Cybersecurity & Software Academic Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine;
e-mail: buravchenkok@gmail.com; ORCID Author ID: https://orcid.org/0000-0001-6195-7533;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=58698282700.

Method for assessing the functional stability of computer-oriented procedures of the NPP operational personnel support system

Borys Vintenko, Tetiana Smirnova, Iryna Myronets, Oleksii Smirnov, Kostiantyn Buravchenko

Abstract. Relevance. Nuclear power plants are large high-tech enterprises containing a large amount of equipment, complex energy conversion processes and information and control systems. The control sequences described in the regulations at nuclear power plants take the form of paper-oriented or computer-oriented procedures. The use of computer-oriented procedures makes it possible to create operational personnel support systems that increase reliability and reduce the burden on operational personnel when performing complex operations. An important requirement for operational personnel support systems used at nuclear power plants is to ensure fault tolerance. In the event of a failure in a critical application system, the operator must have reliable information on how much the failure affects the system's performance and whether the system is able to perform its functions, that is, to assess its functional stability. To do this, it is necessary not only to state the fact of failure, but also to form a numerical assessment of the level of functional stability on the appropriate scale. **Object of research:** the process of functional stability of elements of a computer-oriented procedure as part of the information system for supporting operational personnel of a nuclear power plant. **Purpose of the article:** development of a method for numerical assessment of the functional stability of elements of an information system for supporting operational personnel of a nuclear power plant. **Results of the study.** The assessment of the operability of components of a computer-oriented procedure as part of the system for supporting operational personnel of a nuclear power plant is considered. The use of functional stability indicators as numerical evaluation criteria is proposed. A method for calculating the maximum, current and critical levels of functional stability is proposed. **Conclusions.** This method differs from known diagnostic methods not only by assessing the fact of the presence of a failure, but also by determining a quantitative assessment of the system's ability to perform its functions. The proposed method can be used to diagnose components of critical information systems that receive information from redundant and diversified data sources.

Keywords: numerical assessment, functional stability, information system, decision support system, support for operational personnel, nuclear power plant.

УДК 004.93.1 **doi: 10.26906/SUNZ.2026.2.069**

Д. В. Герасимчук, В. М. Федорченко

Харківський національний університет радіоелектроніки, Харків, Україна

ДОСЛІДЖЕННЯ ПРОГРАМНО-АПАРАТНИХ ЗАСОБІВ РОЗПІЗНАВАННЯ МОВИ ЖЕСТІВ У РЕАЛЬНОМУ ЧАСІ

Анотація. Актуальність. Реалізація розпізнавання жестів у реальному часі на доступних обчислювальних платформах (ноутбук/ПК) є важливою для створення інклюзивних інтерфейсів та систем взаємодії людини з комп'ютером. Практичний інтерес становить вибір такого програмно-апаратного конвеєра, який забезпечує прийнятний компроміс між точністю та швидкодією під час роботи з відеопотоком з веб-камери. **Об'єкт дослідження:** програмно-апаратні конвеєри розпізнавання статичних жестів зображення руки у відеопотоці в режимі реального часу. **Мета статті:** розробити прототип системи, що зчитує кадри з веб-камери, та виконати порівняльне дослідження трьох підходів (MediaPipe, OpenCV, YOLOv8n) за показниками якості розпізнавання і швидкодії на платформі класу Intel Core i3 + NVIDIA GeForce MX350. **Результати дослідження.** Реалізовано модульну програмну архітектуру, у якій кожен підхід оформлено як окремий конвеєр із уніфікованим виходом (клас жесту, довіра, затримка). Для MediaPipe використано landmarks-ознаки з подальшою класифікацією, для OpenCV — ознаки форми (контур, Hu-інваріанти, HOG) із SVM, для YOLOv8n — детекцію класу жесту на кадрі. Проведено оцінювання Accuracy/Precision/Recall/F1 і вимірювання FPS/latency для кількох роздільних здатностей; показано, що MediaPipe і YOLOv8n забезпечують близьку якість для статичних жестів, тоді як OpenCV-підхід більш чутливий до освітлення та складності фону, а також має помітні втрати швидкодії на високих роздільностях. **Висновки.** Встановлено, що для ноутбуків класу i3+MX350 найпрактичнішим за співвідношенням «якість/ресурси» є MediaPipe-конвеєр для статичних жестів, тоді як YOLOv8n доцільний у задачах, де потрібна вища стійкість до фону та більший контекст зображення; класичний OpenCV-підхід може бути корисним як легкий базовий варіант, але потребує ретельної нормалізації умов зйомки та доопрацювання сегментації.

Ключові слова: мова жестів, розпізнавання жестів, реальний час, веб-камера, MediaPipe, OpenCV, YOLOv8, комп'ютерний зір.

Вступ

Сучасні інтерфейси дедалі частіше спираються на комп'ютерний зір: жест може замінювати кнопку, голос або клавіатуру в ситуаціях, де контакт із пристроєм небажаний чи неможливий. Для систем реального часу критичним параметром стає не лише правильність розпізнавання, а й час реакції: затримка на рівні сотень мілісекунд сприймається користувачем як «підвисання».

На практиці застосовують три сімейства рішень: landmarks-підходи (наприклад, MediaPipe Hands) [1], класичні конвеєри на ручних ознаках з реалізацією в OpenCV [2] та нейромережеві детектори сімейства YOLO [3], зокрема сучасну реалізацію YOLOv8 [4].

Внесок роботи. У межах роботи:

1) створено прототип, що працює з веб-камерою та дозволяє перемикати режими MediaPipe [1], OpenCV [2] і YOLOv8 [4] в єдиному інтерфейсі;

2) запропоновано легкі ознаки для двох CPU-сценаріїв (landmarks та Hu+HOG) і виконано порівняння з детектором YOLOv8n [4];

3) сформовано компактний набір UkrSL-10 і наведено узгоджені метрики якості та швидкодії.

Аналіз публікацій. У фахових роботах з розпізнавання жестів найчастіше зустрічаються три напрями: геометрія за ключовими точками кисті, класичні конвеєри з сегментацією та ручними дескрипторами, а також моделі глибинного навчання. Узагальнення підходів і проблематики SLR наведено в оглядах [5] та [6].

Для статичних жестів поширені CNN-підходи, що навчаються на зображеннях або попередньо виділених ознаках [7]. Для безперервного розпізнавання та перекладу жестової мови застосовують трансформерні архітектури [8]. Альтернативою landmarks-підходам є позові/ключові моделі на кшталт OpenPose [9]. Базові принципи глибинного навчання систематизовано в [10].

Мета роботи – на одному програмному прототипі порівняти MediaPipe Hands [1], OpenCV підхід [2] та YOLOv8n [4] для розпізнавання жестів із веб-камери в реальному часі. Для цього сформовано UkrSL-10, реалізовано три конвеєри та виконано оцінювання якості (Accuracy/F1) і швидкодії (FPS/latency) за фіксованим протоколом експерименту.

Матеріали та методи

Експериментальну частину виконано для конфігурації ноутбука класу Intel Core i3 + NVIDIA GeForce MX350 (2 ГБ VRAM); характеристики набору даних UkrSL-10 подано в табл. 1.

Таблиця 1 – **Характеристики набору даних UkrSL-10**

Параметр	Значення
Кількість класів (жестів)	10 (статичні)
Кількість учасників	15
Загальна кількість зразків	≈30 000 кадрів
Розділення даних	70% навчання / 15% валідація / 15% тест
Камера та частота	веб-камера 30 кадр/с
Роздільна здатність	640×480 та 1280×720

Прототип написано на Python: OpenCV відповідає за захоплення кадру та базову обробку, MediaPipe Hands повертає landmarks кисті, а YOLOv8 використано як нейромережевий детектор жестів. Швидкодію MediaPipe/OpenCV вимірювали у CPU-режимі в

межах прототипу; для YOLOv8n наведено референсні метрики Ultralytics після експорту в ONNX (640×640) та окремо зазначено роль GPU-прискорення [4].

Якість оцінювали стандартними показниками класифікації (Accuracy, Precision, Recall, F1). Параметри реального часу фіксували як середній час кадру (latency) та FPS для трьох роздільностей вхідного відео: 320×240, 640×480 і 1280×720. Latency трактували як суму препроцесингу, інференсу та постобробки/візуалізації.

Додатково враховували практичні обмеження: пікове використання RAM/VRAM та розмір моделі (для YOLOv8), оскільки саме ці параметри часто визначають можливість розгортання на ноутбуках і вбудованих платформах.

Відеопотік оброблявся послідовно: кадр зчитувався через OpenCV VideoCapture і передавався у вибраний конвеєр. Для MediaPipe виконувалося перетворення BGR→RGB. У OpenCV-варіанті з центральної ROI формували зображення 224×224, переводили у grayscale, застосовували Gaussian blur та бінаризацію Оцу; найбільший контур описували Hu-моментами і HOG. Щоб зменшити «стрибки» класу між кадрами, для MediaPipe та OpenCV використовували голосування більшості у вікні 7 кадрів [2].

Метрики Accuracy/Precision/Recall/F1 обчислювали за загальноприйнятими визначеннями на основі TP/FP/FN/TN. FPS оцінювали як відношення кількості оброблених кадрів до сумарного часу вимірювання; додатково аналізували 95-й перцентиль затримки (p95) (рис. 1).

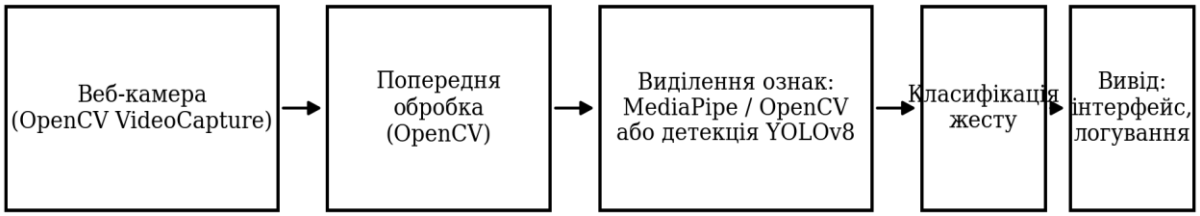


Рис. 1. Узагальнена програмна архітектура системи розпізнавання жестів у реальному часі

Опис алгоритмів розпізнавання.

Підхід 1 (MediaPipe Hands + MLP). MediaPipe виділяє кисть і повертає 21 точку (x, y, z). Ознаки формували як 63-вимірний вектор, нормалізуючи координати відносно зап'ястя та масштабу кисті. Далі застосовували компактний MLP-класифікатор (2 приховані шари), що швидко працює на CPU і легко перенавчається при зміні набору жестів [1].

Підхід 2 (OpenCV + SVM). Використано класичний конвеєр: фіксована ROI → grayscale → згладжування → порогування Оцу → найбільший контур. Для опису форми застосовано Hu-інваріанти, для текстури/градієнтів — HOG; об'єднаний вектор подавали на SVM з RBF-ядром. Метод простий і легкий, але чутливий до сегментації [2].

Підхід 3 (YOLOv8n). YOLOv8n розглядається як детектор, що на одному проході дає рамку руки та клас жесту. Для навчання потрібні дані з bounding boxes; для порівняння швидкодії використано еталон Ultralytics для CPU ONNX (640×640) [4]. У прикладних RTSLR-сценаріях бажано застосовувати прискорення на GPU (зокрема MX350) або оптимізовані бекенди інференсу.

Реалізація програмного забезпечення. Прототип побудовано як модульний конвеєр:

захоплення кадру → легкий препроцесинг → розпізнавання → візуалізація.

Кожен алгоритм оформлено як окремий модуль із однаковим виходом (label, confidence, latency), що спрощує порівняння. Час обробки вимірювали через time.perf_counter() після короткого «прогріву». Для кожного режиму фіксували середню затримку, p95 та FPS. MediaPipe і OpenCV оцінювалися в CPU-режимі; для YOLOv8n використано референсні дані Ultralytics для CPU ONNX [4]. Інтерфейс виводить поточний клас, confidence та службові показники (FPS, latency). Режим роботи задається параметром командного рядка, що робить експерименти відтворюваними.

Результати експериментів та їх обговорення

Вимоги до обчислювальних ресурсів для трьох підходів наведено в табл. 2. Візуальне зіставлення F1-міри подано на рис. 2. Показники швидкодії та затримки для різних роздільностей узагальнено в табл. 3; залежність FPS від роздільної здатності показано на рис. 3, а залежність середньої затримки — на рис. 4. Стійкість конвеєрів до освітлення та складності фону наведено в табл. 4; вплив умов освітлення на F1-міру узагальнено на рис. 5.

Таблиця 2 – **Порівняння якості розпізнавання на тестовій вибірці UkrSL-10**

Метод	Accuracy, %	Precision, %	Recall, %	F1, %
MediaPipe Hands + MLP	95.2	95.6	94.4	95.0
OpenCV (контур+HOG) + SVM	86.4	87.9	83.7	85.7
YOLOv8n	96.8	97.1	96.2	96.6

Таблиця 3 – **Швидкодія та затримка (CPU) для трьох роздільних здатностей (значення в комірках: FPS/мс; для YOLOv8n наведено еталонні дані Ultralytics для входу 640, CPU ONNX)**

Роздільність	MediaPipe (FPS/мс)	OpenCV (FPS/мс)	YOLOv8n (FPS/мс)
320×240	71.4 / 14.0	54.2 / 18.4	–
640×480	66.4 / 15.1	47.6 / 21.0	12.4 / 80.4
1280×720	52.8 / 18.9	54.1 / 18.5	–

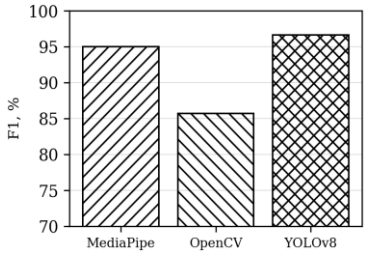


Рис. 2. Порівняння F1-міри для трьох підходів

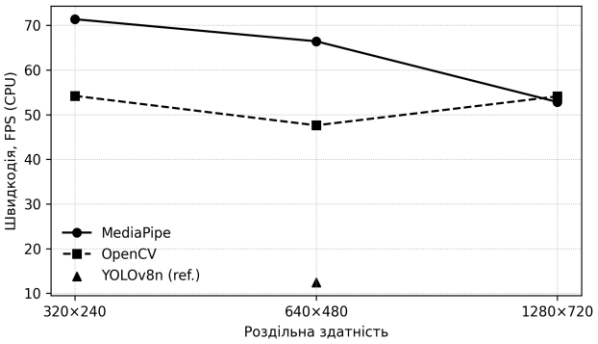


Рис. 3. Залежність швидкодії (FPS) від роздільної здатності (CPU)

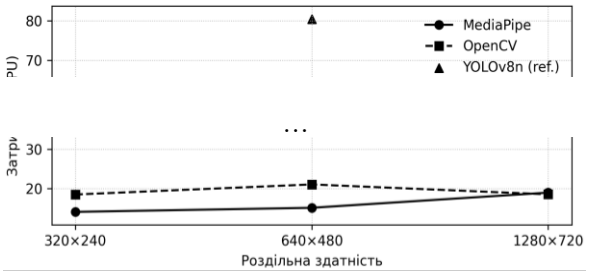


Рис. 4. Залежність середньої затримки (мс) від роздільної здатності (CPU)

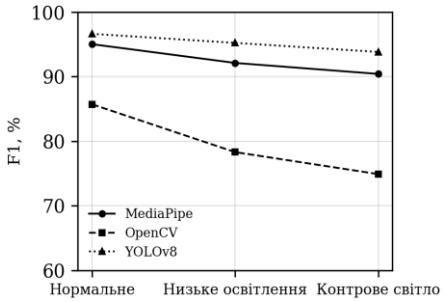


Рис. 5. Залежність F1-міри від умов освітлення

За показниками якості (табл. 2) MediaPipe+MLP і YOLOv8n дають близькі результати для статичних жестів, оскільки перший спирається на стабільні landmarks, а другий використовує піксельний контекст кадру. OpenCV+SVM виступає як базова «легка» лінія: він помітно програє за F1 через залежність від сегментації, але практично не потребує додаткових обчислювальних ресурсів.

Швидкодія (табл. 3, рис. 3 та 4) підтверджує, що MediaPipe та OpenCV можуть працювати у реальному часі на CPU в діапазоні 320×240–1280×720. Для YOLOv8n референс Ultralytics у CPU ONNX (640×640) становить 80.4 мс/кадр (≈12.4 FPS), тому для сценаріїв із вимогами 20–25 FPS доцільно застосовувати прискорення на GPU (MX350) або оптимізований бекенд інференсу.

Стійкість до умов зйомки (табл. 4) показує найбільшу деградацію в OpenCV-підході при низькому/контровому світлі; MediaPipe зберігає стабільність краще, а YOLOv8n має найменший спад на складному фоні. За ресурсами (табл. 5) OpenCV та MediaPipe придатні для систем без дискретного GPU, тоді як нейромережевий детектор вимагає більше пам’яті і виграє при використанні VRAM.

Таблиця 4 – **Стійкість до освітлення та складності фону (F1, %)**

Метод	Нормальне	Низьке освітлення	Контрове світло	Простий фон	Складний фон
MediaPipe+MLP	95.0	92.1	90.4	94.6	91.3
OpenCV+SVM	85.7	78.3	74.9	83.2	76.5
YOLOv8n	96.6	95.2	93.8	96.0	94.7

Таблиця 5 – **Порівняння вимог до обчислювальних ресурсів**

Метод	Параметри моделі, млн	GPU VRAM, ГБ	GPU пам’ять, МБ	RAM, МБ
MediaPipe+MLP	≈0.0	–	–	180
OpenCV+SVM	≈0.0	–	–	140
YOLOv8n	3.2	2.0	320	520

Обмеження дослідження

Дослідження спрямоване на статичні жести; моделювання динаміки (послідовностей) і безперервне розпізнавання фраз не входили до обсягу роботи.

Абсолютні значення швидкодії залежать від ОС, драйверів і бекенда інференсу.

Показник для YOLOv8n подано як еталон Ultralytics (CPU ONNX), тоді як MediaPipe/OpenCV вимірювалися у прототипі Python; тому результати слід інтерпретувати як практичні орієнтири для вибору стеку [4].

Висновки

Створений прототип RTSLR працює з веб-камерою та дозволяє порівнювати три популярні стеки в однакових умовах: MediaPipe, OpenCV і YOLOv8n.

Для UkrSL-10 найкращий баланс «якість/швидкість» у CPU-режимі демонструє MediaPipe+MLP; OpenCV+SVM придатний як максимально легкий варіант, але потребує контрольованого фону/світла; YOLOv8n найстійкіший у складних сценах, однак для стабільного realtime бажано використовувати GPU-прискорення (MX350) [4].

Найближчі напрями розвитку: перехід до динамічних жестів (послідовності), застосування часових моделей (TCN/Transformer) та оптимізація нейромережевого інференсу для вбудованих платформ (ONNX Runtime, квантування).

У практичних системах без дискретного GPU доцільно починати з MediaPipe як базового модуля виділення кисті; за потреби підвищеної робастності або розширення набору класів — переходити до YOLO-підходу з прискоренням.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Zhang F., Bazarevsky V., Vakunov A. et al. MediaPipe Hands: On-device real-time hand tracking. arXiv:2006.10214, 2020. URL: https://arxiv.org/abs/2006.10214
2. Bradski G. The OpenCV Library // Dr. Dobb's Journal. 2000. URL: http://www.drdobbs.com/open-source/the-opencv-library/184404319
3. Redmon J., Farhadi A. YOLO9000: Better, Faster, Stronger // Proc. IEEE CVPR. 2017. P. 7263–7271. DOI: https://doi.org/10.1109/CVPR.2017.690
4. Ultralytics. YOLOv8: документація та бенчмарки швидкодії. 2023–2026. URL: https://docs.ultralytics.com
5. Koller O. Quantitative survey of the state of the art in sign language recognition. arXiv, 2020. URL: https://arxiv.org/abs/2008
6. Subburaj S., Murugavalli S. Survey on sign language recognition in context of vision-based and deep learning // Measurement: Sensors. 2022. Vol. 23. 100385. DOI: https://doi.org/10.1016/j.measen.2022.100385
7. Pigou L., Dieleman S., Kindermans P.-J., Schrauwen B. Sign language recognition using convolutional neural networks // ECCV Workshops (LNCS). 2015. P. 572–578. DOI: https://doi.org/10.1007/978-3-319-16178-5_40
8. Camgoz N. C., Koller O., Hadfield S., Bowden R. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation // Proc. IEEE CVPR. 2020. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/
9. Cao Z., Simon T., Wei S.-E., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields // IEEE TPAMI. 2021. Vol. 43(1). P. 172–186. DOI: https://doi.org/10.1109/TPAMI.2019.2929257
10. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016. URL: https://www.deeplearningbook.org

Received (Надійшла) 15.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Герасимчук Дмитро Вікторович – студент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна; e-mail: dmytro.herasymchuk@nure.ua.

Dmytro Herasymchuk – student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: dmytro.herasymchuk@nure.ua; ORCID Author ID: https://orcid.org/0009-0009-2715-2614.

Федорченко Володимир Миколайович – PhD, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;

Volodymyr Fedorchenko – PhD, Associate Professor? Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: volodymyr.fedorchenko@nure.ua; ORCID Author ID: http://orcid.org/0000-0001-7359-1460.

Research of software and hardware tools for real-time sign language recognition

Dmytro Herasymchuk, Volodymyr Fedorchenko

Abstract. Relevance. Real-time hand-gesture recognition on affordable computing platforms (laptops/PCs) is important for building inclusive human–computer interaction interfaces and assistive technologies. A practical challenge is selecting a processing pipeline that maintains an acceptable trade-off between recognition accuracy and runtime performance when processing a webcam video stream. **Object of research:** software–hardware pipelines for real-time recognition of static hand gestures in a live video stream. **Purpose of the article:** to develop a webcam-based prototype and conduct a comparative study of three approaches (MediaPipe, OpenCV, YOLOv8n) in terms of recognition quality and real-time performance on a platform of the Intel Core i3 + NVIDIA GeForce MX350 class. **Research results.** A modular architecture was implemented where each approach is represented as an independent pipeline with a unified output (gesture class, confidence, latency). The MediaPipe pipeline uses hand landmarks followed by classification; the OpenCV pipeline relies on shape features (contours, Hu moments, HOG) with an SVM classifier; the YOLOv8n pipeline performs single-pass detection/classification of gestures in a frame. The prototype was evaluated using Accuracy/Precision/Recall/F1 and timed using FPS/latency at several input resolutions. The results indicate that MediaPipe and YOLOv8n achieve comparable performance for static gestures, whereas the OpenCV-based solution is more sensitive to illumination changes and complex backgrounds and shows a more pronounced speed drop at higher resolutions. **Conclusions.** For laptops of the i3+MX350 class, the MediaPipe-based pipeline provides the most practical balance between accuracy and computational cost for static gestures, while YOLOv8n is preferable when robustness to background clutter and stronger image context are required. The classical OpenCV pipeline can serve as a lightweight baseline but typically requires careful capture-condition normalization and additional improvements in segmentation to remain stable.

Keywords: sign language, gesture recognition, real-time, webcam, MediaPipe, OpenCV, YOLOv8, computer vision.

УДК 004.056:004.7 **doi: 10.26906/SUNZ.2026.2.073**

Т. М. Деркач[1], Г. В. Головко[1], А. О. Дмитренко[1], Л. А. Клочко[2]

[1] Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна
[2] Геологічне бюро BEG SA, Швейцарія

АНАЛІЗ ЗАГРОЗ І ВРАЗЛИВОСТЕЙ КОМП'ЮТЕРНИХ МЕРЕЖ ТА ОБҐРУНТУВАННЯ КОМПЛЕКСНОГО ПІДХОДУ ДО ЗАБЕЗПЕЧЕННЯ ЇХ КІБЕРБЕЗПЕКИ

Анотація. У статті здійснено комплексний аналіз сучасних загроз та вразливостей комп'ютерних мереж, що виникають у процесі функціонування мережевої інфраструктури в умовах стрімкого розвитку цифрових технологій та зростання кількості кіберзагроз. Особливу увагу приділено дослідженню архітектурних і протокольних вразливостей мережі, зокрема на канальному та транспортному рівнях моделі взаємодії відкритих систем (OSI), які часто стають початковим етапом реалізації складних кібернетичних атак. У роботі проведено систематизацію та класифікацію основних типів мережевих атак за характером впливу, джерелом походження та рівнем мережевої моделі, на якому вони реалізуються. Розглянуто особливості пасивних і активних атак, внутрішніх та зовнішніх загроз, а також їхній вплив на конфіденційність, цілісність і доступність інформаційних ресурсів. Значну увагу приділено аналізу сучасних методів мережевої розвідки та сканування, які використовуються як фахівцями з кібербезпеки для проведення аудиту інформаційних систем, так і потенційними зловмисниками для виявлення вразливостей мережевої інфраструктури. Досліджено механізми пасивної та активної мережевої розвідки, включаючи методи збору інформації з відкритих джерел, сканування хостів, аналіз мережевих портів, ідентифікацію мережевих сервісів, визначення операційних систем та виявлення відомих вразливостей. Встановлено, що використання таких методів дозволяє формувати детальну карту мережевої інфраструктури, що може бути використано для підготовки подальших етапів кібернападу. Окремий розділ дослідження присвячено аналізу архітектурних вразливостей канального рівня, які виникають через відсутність механізмів автентифікації в базових протоколах сімейства IEEE 802. Розглянуто особливості реалізації атак, пов'язаних із маніпуляцією кадрами Ethernet, зокрема атак на таблиці комутації, ARP-спуфінг, VLAN hopping, а також атаки на інфраструктуру DHCP та протокол Spanning Tree. Показано, що експлуатація таких вразливостей може призводити до перехоплення мережевого трафіку, порушення сегментації мережі, підміни маршрутів передачі даних або організації відмови в обслуговуванні. Значну увагу приділено дослідженню атак на відмову в обслуговуванні (Denial of Service) та розподілених атак (Distributed Denial of Service), які належать до найбільш поширених кіберзагроз у сучасних інформаційних системах. Проаналізовано механізми реалізації волюметричних атак, атак транспортного рівня та атак із підсиленням, що використовують уразливості мережевих сервісів для генерації значних обсягів трафіку. Розглянуто роль ботнетів та пристроїв Інтернету речей у формуванні масштабних розподілених атак, здатних суттєво впливати на доступність інформаційних ресурсів. На основі проведеного аналізу обґрунтовано доцільність застосування комплексного підходу до забезпечення безпеки комп'ютерних мереж. Такий підхід передбачає поєднання конфігураційних, криптографічних та організаційних заходів захисту, що реалізуються на різних рівнях мережевої інфраструктури. Показано, що застосування принципу багаторівневого захисту дозволяє підвищити стійкість інформаційних систем до сучасних кіберзагроз, мінімізувати ризики експлуатації вразливостей та забезпечити стабільне функціонування мережевих сервісів. Результати дослідження можуть бути використані під час розроблення політик інформаційної безпеки, проектування захищених корпоративних мереж, проведення аудиту кібербезпеки та навчання фахівців у галузі інформаційних технологій і кіберзахисту.

Ключові слова: комп'ютерні мережі, кібербезпека, мережеві атаки, мережева розвідка, сканування мережі, вразливості мережевих протоколів, DoS-атаки, DDoS-атаки, інформаційна безпека.

Постановка проблеми

Сучасні комп'ютерні мережі є критичною складовою інформаційної інфраструктури державних, промислових та комерційних систем. Водночас зростання складності мережевих архітектур, активне використання хмарних сервісів, Інтернету речей (IoT) та мобільних технологій призводить до підвищення ризику виникнення кіберзагроз. Одними з найбільш критичних є вразливості на канальному та транспортному рівнях, оскільки вони можуть використовуватися як початкові точки проникнення для реалізації складних атак, таких як DoS/DDoS, ARP spoofing, VLAN hopping та інші.

Недостатній контроль доступу, відсутність багатофакторної автентифікації, помилки у конфігурації мережевого обладнання та незашифровані канали передачі даних створюють передумови для витоку інформації, порушення цілісності та доступності ресурсів. В умовах зростання кількості та складності кіберзагроз постає задача систематизації мережевих атак, аналізу методів розвідки та сканування, дослідження архітектурних вразливостей та обґрунтування ефективних механізмів захисту, що включають конфігураційні, криптографічні та організаційні заходи.

Мета та завдання дослідження

Метою дослідження є комплексний аналіз вразливостей комп'ютерних мереж, класифікувати сучасні загрози та атаки, оцінити методи мережевої розвідки та сканування, а також обґрунтувати доцільність застосування багаторівневого підходу до захисту корпоративної мережевої інфраструктури.

Основні завдання дослідження:

1. Проаналізувати сучасний стан проблеми безпеки комп'ютерних мереж.
2. Систематизувати та класифікувати мережеві загрози та атаки за типом, джерелом і рівнем OSI.
3. Дослідити методи мережевої розвідки та сканування, їхню роль у підготовці атак та оцінці захищеності.

4. Проаналізувати механізми DoS/DDoS атак, волюметричних і транспортних атак, а також атак із підсиленням.

5. Обґрунтувати комплексний підхід до захисту мереж, поєднуючи конфігураційні, криптографічні та організаційні заходи.

Аналіз досліджень і публікацій

Проблема захисту комп'ютерних мереж у сучасних умовах розвитку цифрових технологій та зростання частоти кібератак привертає значну увагу наукової спільноти. Аналіз наукових джерел показує, що питання вразливостей мережевої інфраструктури розглядаються як у загальному контексті кібербезпеки, так і в аспекті окремих мережевих протоколів та технологій.

У класичних роботах із мережевих технологій та безпеки, зокрема у працях A. S. Tanenbaum та D. J. Wetherall, розкрито фундаментальні принципи побудови комп'ютерних мереж, що дозволяє зрозуміти природу протокольних механізмів і потенційні зони вразливостей у базових мережевих технологіях [1]. Аналогічно, у виданнях J. F. Kurose і K. W. Ross подано системний опис моделі OSI та стеку TCP/IP, який є основою для подальших досліджень мережевих атак і засобів захисту [2]. Проте такі роботи носять радше освітній характер і не завжди глибоко аналізують специфічні загрози на канальному рівні.

Дослідження W. Stallings спрямовані на прикладні аспекти мережевої безпеки, включаючи криптографічні методи захисту, моделі автентифікації та управління доступом [3]. У цьому контексті підкреслюється важливість криптографії для захисту даних при передачі, однак механізми захисту канального рівня розглядаються побіжно або лише в загальному вигляді.

У дослідженнях R. Anderson висвітлено актуальні концепції безпечної архітектури розподілених систем, у тому числі загрози, що виникають при взаємодії різнорідних мережевих компонентів та сервісів [4]. Він розглядає загальні принципи побудови захищених систем, проте не приділяє достатньої уваги деталям реалізації конкретних атак на канальний рівень.

Публікації, присвячені безпосередньо мережевим атакам і механізмам їх реалізації, включають роботи, що аналізують технології сканування мереж, відмови в обслуговуванні та посилені атаки, зокрема J. M. Stewart та співавтори в CISSP Official Study Guide [5], а також у контексті класифікації DDoS-загроз [6]. Дані праці детально описують тактики та техніки атак, але переважно охоплюють загальні класи загроз без поглибленого розгляду протоколів канального рівня.

Проблематика забезпечення кібербезпеки комп'ютерних мереж є науковим напрямком досліджень також і українських науковців [7–10]. У роботах значна увага приділяється дослідженню архітектурних особливостей комп'ютерних мереж та механізмів їх захисту. У роботах В. Бурячка, Г.Гулака та В. Толубка [11], розглядаються основні принципи забезпечення інформаційної безпеки в комп'ютерних системах, а також класифікація загроз і методів протидії мережевим атакам. Дослідження М. Корченка [12] присвячені аналізу сучасних кіберзагроз, методам виявлення атак та побудові систем захисту інформаційних ресурсів.

У науковій літературі також широко висвітлюються питання мережевої розвідки та сканування. Дослідники відзначають, що етап збору інформації є важливою складовою більшості кібернападів, оскільки дозволяє зловмисникам отримати відомості про структуру мережі, відкриті порти та використовуване програмне забезпечення. Аналіз відповідних методів дозволяє підвищити ефективність систем виявлення вторгнень та аудиту інформаційної безпеки.

Аналіз наукових публікацій свідчить, що ефективне забезпечення кібербезпеки комп'ютерних мереж потребує комплексного підходу, який поєднує технічні, криптографічні та організаційні засоби захисту. Проте подальших досліджень потребують питання вдосконалення методів виявлення атак, аналізу вразливостей мережевих протоколів та підвищення стійкості інформаційних систем до сучасних кіберзагроз.

Виклад основного матеріалу

1. У сучасних умовах стрімкого розвитку інформаційних технологій комп'ютерні мережі стали невід'ємною складовою функціонування державних, промислових і комерційних інформаційних систем. Активне впровадження хмарних сервісів, мобільних технологій, систем Інтернету речей (IoT) та розподілених обчислювальних платформ сприяє підвищенню ефективності обробки та передачі даних, однак водночас призводить до значного зростання кількості кіберзагроз і ускладнення процесу забезпечення інформаційної безпеки.

Однією з ключових проблем сучасної кібербезпеки є збільшення кількості та різноманітності мережевих атак. Дослідження показують, що комп'ютерні мережі можуть зазнавати впливу різних типів загроз, серед яких найбільш поширеними є атаки відмови в обслуговуванні (DoS/DDoS), перехоплення мережевого трафіку, підміна мережевих адрес, несанкціонований доступ до ресурсів та експлуатація вразливостей програмного забезпечення. Суттєвою тенденцією розвитку кіберзагроз є поява гібридних та багаторівневих атак, які поєднують різні методи впливу на інформаційні системи. Сучасні дослідження демонструють, що для виявлення складних мережевих атак дедалі ширше застосовуються технології машинного навчання та штучного інтелекту, які дозволяють аналізувати великі обсяги мережевого трафіку та виявляти аномалії у поведінці систем. Ще однією важливою проблемою є зростання масштабів кіберзагроз для критичної інформаційної інфраструктури. Зокрема, у сучасному кіберпросторі спостерігається тенденція до використання кібератак як інструменту інформаційного та політичного впливу. Отже, аналіз сучасного стану проблеми безпеки комп'ютерних мереж свідчить про постійне зростання складності кіберзагроз і необхідність застосування комплексних мето-

дів їх протидії. Подальші дослідження у цій сфері мають бути спрямовані на вдосконалення методів виявлення мережевих атак, використання технологій штучного інтелекту для аналізу кіберзагроз та підвищення стійкості інформаційних систем до сучасних кібератак.

2. Класифікація загроз та атак на мережеву інфраструктуру. Мережева атака розглядається як сукупність навмисних дій, спрямованих на експлуатацію вразливостей апаратного або програмного забезпечення з метою порушення конфіденційності, цілісності чи доступності інформації. Для побудови ефективної моделі захисту необхідною є систематизація існуючих типів атак.

За характером впливу атаки поділяються на пасивні та активні. Пасивні атаки не передбачають безпосереднього втручання в роботу системи та спрямовані на перехоплення або аналіз трафіку. Основною небезпекою цього класу загроз є їх прихований характер, оскільки вони не порушують цілісність даних і практично не фіксуються стандартними засобами моніторингу. Єдиним ефективним методом протидії є превентивне шифрування каналів зв'язку.

Активні атаки передбачають безпосередню модифікацію інформаційних потоків або порушення роботи сервісів. До них належать маскарад, модифікація повідомлень, повторна передача перехоплених даних та атаки відмови в обслуговуванні. Реалізація таких атак потребує глибокого розуміння принципів функціонування мережевих протоколів та механізмів маршрутизації.

За джерелом загрози атаки поділяються на зовнішні та внутрішні. Внутрішні загрози є особливо небезпечними, оскільки внутрішній трафік часто підлягає менш жорсткому контролю. Саме це зумовлює необхідність сегментації мережі та впровадження механізмів контролю доступу на канальному рівні.

3. Технічний аналіз методів мережевої розвідки та сканування. Етап збору інформації є обов'язковою складовою більшості цілеспрямованих мережевих атак. На цьому етапі зловмисник визначає активні вузли, відкриті порти та версії сервісів, що дозволяє сформувати карту мережі та підготувати подальшу експлуатацію.

У сучасних умовах цифровізації важливого значення набуває дослідження методів мережевої розвідки та сканування. Дані методи застосовуються як у діяльності фахівців із кібербезпеки для оцінювання захищеності інформаційних систем, так і потенційними зловмисниками з метою виявлення вразливостей мережевої інфраструктури. Мережева розвідка є початковим етапом аналізу комп'ютерної мережі, що передує більш складним формам кібернападів або, навпаки, використовується для проведення аудиту безпеки.

У науковій та практичній літературі виділяють два основних типи мережевої розвідки: *пасивну* та *активну*.

Пасивна мережева розвідка передбачає збір інформації без прямої взаємодії з цільовою системою. У межах такого підходу використовуються відкриті джерела інформації, аналіз доменних записів, публічних мережевих сервісів, метаданих та інші методи OSINT (Open Source Intelligence). Основною перевагою цього методу є низька ймовірність виявлення, оскільки безпосередній контакт із досліджуваною системою відсутній.

Активна мережева розвідка, навпаки, передбачає безпосередню взаємодію з мережею або її вузлами шляхом надсилання запитів до серверів, сканування портів, перевірки доступності сервісів тощо. Такий підхід дозволяє отримати більш детальну інформацію, однак може бути зафіксований системами виявлення вторгнень або мережевими засобами моніторингу. Мережеве сканування є ключовим інструментом активної розвідки та використовується для виявлення доступних вузлів мережі, відкритих портів, активних сервісів та конфігурації операційних систем. Основною метою сканування є визначення потенційних точок доступу до мережевої інфраструктури. Серед найбільш поширених методів мережевого сканування виділяють такі.

– Сканування портів (Port Scanning). Цей метод спрямований на визначення відкритих, закритих або фільтрованих портів у мережевих вузлах. Відкриті порти можуть свідчити про наявність активних мережевих сервісів, таких як вебсервери, поштові служби або віддалений доступ. Аналіз відкритих портів дозволяє ідентифікувати потенційні точки проникнення до системи.

– Сканування хостів (Host Discovery). Даний метод використовується для визначення активних пристроїв у мережі. Найчастіше він реалізується за допомогою ICMP-запитів, ARP-сканування або TCP-запитів. Результатом такого аналізу є формування переліку доступних мережевих вузлів.

– Сканування сервісів (Service Scanning). Після виявлення відкритих портів здійснюється аналіз запущених на них сервісів. Це дозволяє визначити тип програмного забезпечення, його версію та конфігурацію. Отримана інформація є важливою для подальшого аналізу потенційних вразливостей.

– Визначення операційної системи (OS Fingerprinting). Метод передбачає аналіз особливостей мережевих відповідей системи з метою встановлення типу операційної системи та її версії. Визначення операційної системи дозволяє звузити коло можливих експлойтів або вразливостей.

– Сканування вразливостей (Vulnerability Scanning). Цей метод полягає у порівнянні отриманої інформації про систему з відомими базами вразливостей. На основі такого аналізу формується перелік потенційних загроз безпеці мережі.

Для реалізації описаних методів використовуються спеціалізовані програмні інструменти, які автоматизують процес аналізу мережевої інфраструктури. Такі системи дозволяють здійснювати комплексне дослідження мережі, включаючи виявлення вузлів, аналіз портів, ідентифікацію сервісів та оцінювання рівня безпеки.

Сучасні інструменти мережевої розвідки забезпечують високу швидкість сканування, підтримку різних протоколів та можливість інтеграції з системами управління інформаційною безпекою (рис. 1).

Їх використання є важливим елементом процесу тестування на проникнення (penetration testing) та аудиту інформаційної безпеки.

Розуміння механізмів мережевого сканування є критично важливим для коректного налаштування систем виявлення вторгнень і фільтрації трафіку.

Рис. 1. Методи мережевої розвідки та сканування

4. Атаки на відмову в обслуговуванні (Denial of Service, DoS) та розподілені атаки відмови в обслуговуванні (Distributed Denial of Service, DDoS) належать до найбільш поширених і небезпечних типів кіберзагроз у сучасному інформаційному середовищі. Їх основною метою є порушення доступності інформаційних ресурсів, мережевих сервісів або обчислювальних систем шляхом навмисного перевантаження інфраструктури великою кількістю запитів чи мережевого трафіку. У результаті таких дій легітимні користувачі не можуть отримати доступ до вебсайтів, серверів, мережевих сервісів або інших інформаційних ресурсів, що призводить до фінансових втрат, зниження довіри до організації та порушення безперервності бізнес-процесів.

Атака типу DoS зазвичай здійснюється з одного джерела або з обмеженої кількості пристроїв, що генерують значний обсяг запитів до цільової системи. У свою чергу, атаки типу DDoS реалізуються за допомогою великої кількості скомпрометованих пристроїв, об'єднаних у так звані ботнети. До складу таких мереж можуть входити тисячі або навіть мільйони заражених комп'ютерів, серверів або пристроїв Інтернету речей (IoT), які координовано надсилають запити до цільового ресурсу. Використання розподіленої інфраструктури значно ускладнює виявлення джерела атаки та підвищує її ефективність. З технічної точки зору атаки на відмову в обслуговуванні можуть реалізовуватися на різних рівнях моделі взаємодії відкритих систем (OSI). Найбільш поширеними є атаки мережевого, транспортного та прикладного рівнів. Кожен із цих типів має свої особливості та механізми впливу на інформаційні системи.

Однією з найбільш поширених категорій є *волюметричні атаки* (volumetric attacks), які спрямовані на перевантаження пропускної здатності мережевого каналу цільової системи. У межах таких атак генерується значний обсяг мережевого трафіку, який перевищує можливості обробки або передачі даних сервером чи мережевим обладнанням. У результаті мережеві канали стають перевантаженими, що призводить до різкого зниження швидкості доступу або повної недоступності сервісу. До цієї категорії належать, зокрема, UDP-флуд атаки, ICMP-флуд та інші види масованого надсилання пакетів.

Іншою важливою категорією є атаки *транспортного рівня*, які спрямовані на виснаження ресурсів серверів або мережевого обладнання шляхом експлуатації механізмів встановлення мережевих з'єднань. Одним із найбільш відомих прикладів є SYN-flood атака, яка використовує особливості встановлення TCP-з'єднання. У процесі такої атаки сервер отримує велику кількість запитів на встановлення з'єднання, але завершення процедури рукостискання (TCP three-way handshake) не відбувається. У результаті сервер змушений утримувати велику кількість напіввідкритих з'єднань, що поступово призводить до вичерпання його обчислювальних ресурсів та неможливості обслуговувати легітимні запити. Окрему категорію становлять *атаки з підсиленням* (*amplification attacks*), які базуються на використанні мережевих сервісів, здатних генерувати відповіді значно більшого обсягу, ніж початковий запит. У такому випадку атакуючий надсилає невеликі запити до відкритих серверів із підробленою IP-адресою жертви. Сервери, у свою чергу, надсилають відповіді значно більшого розміру на адресу цільової системи, що призводить до різкого збільшення обсягу трафіку. Прикладами таких атак є DNS amplification, NTP amplification та інші варіанти використання відкритих мережевих служб для підсилення атаки. Крім зазначених типів, значного поширення набули *атаки прикладного рівня*, спрямовані на перевантаження конкретних вебсервісів або прикладних програм. У межах таких атак генеруються численні запити до вебсторінок, API або баз даних, що призводить до перевантаження серверних проце-

сів. Особливість цього типу атак полягає в тому, що вони часто імітують легітимну поведінку користувачів, що значно ускладнює їхнє виявлення традиційними мережевими засобами захисту.

Важливою тенденцією останніх років є використання *розподілених мереж заражених пристроїв (ботнетів)* для проведення масштабних DDoS-атак (рис. 2).

Рис 2. Класифікації DoS/DDoS атак

Значну частину таких мереж становлять пристрої Інтернету речей, зокрема маршрутизатори, відеокамери спостереження, мережеві накопичувачі та інші пристрої з недостатнім рівнем захисту. Уразливості програмного забезпечення або використання стандартних облікових даних дозволяють зловмисникам отримувати контроль над великою кількістю пристроїв та використовувати їх для генерації шкідливого трафіку.

Зростання масштабів і складності DDoS-атак зумовлює необхідність впровадження комплексних механізмів захисту інформаційних систем. До таких механізмів належать системи виявлення та запобігання вторгненням, фільтрація трафіку на рівні мережевого обладнання, використання балансування навантаження, застосування технологій розподіленої доставки контенту (CDN) та спеціалізованих сервісів захисту від DDoS-атак. Крім того, важливу роль відіграє моніторинг мережевої активності та аналіз аномалій у трафіку, що дозволяє своєчасно виявляти підозрілу активність і запобігати порушенню доступності інформаційних ресурсів.

Отже, атаки на відмову в обслуговуванні та розподілені мережеві атаки становлять серйозну загрозу для сучасних інформаційних систем і мережевої інфраструктури. Їхня ефективність зумовлена можливістю використання великої кількості розподілених джерел трафіку, механізмів підсилення та вразливостей мережевих протоколів. У зв’язку з цим дослідження механізмів реалізації таких атак і розроблення ефективних методів протидії є важливим завданням сучасної кібербезпеки та одним із ключових напрямів забезпечення стабільності функціонування інформаційних систем.

5. Зростання кількості кіберзагроз, ускладнення мережевих архітектур, а також активне використання хмарних сервісів, мобільних технологій і розподілених інформаційних систем зумовлюють необхідність впровадження ефективних механізмів захисту інформаційних ресурсів. У цьому контексті дедалі більшого значення набуває застосування *комплексного підходу до забезпечення безпеки комп’ютерних мереж*, що передбачає поєднання конфігураційних, криптографічних та організаційних заходів захисту.

Доцільність використання комплексного підходу зумовлена тим, що сучасні кіберзагрози характеризуються багаторівневим і багатоетапним характером. Зловмисники використовують комбінацію технічних, програмних і соціальних методів впливу на інформаційні системи, що дозволяє їм обходити окремі механізми захисту. У зв’язку з цим використання лише одного типу засобів безпеки не забезпечує достатнього рівня захищеності мережевої інфраструктури. Ефективна система інформаційної безпеки повинна базуватися на принципі багаторівневого або «глибокого» захисту (defense in depth), коли різні механізми безпеки доповнюють один одного та створюють комплексну систему протидії кіберзагрозам.

Важливим компонентом такого підходу є *конфігураційні заходи безпеки*, які передбачають правильне налаштування мережевого обладнання, серверів, операційних систем та програмного забезпечення. Належна конфігурація мережевих пристроїв дозволяє обмежити несанкціонований доступ до інформаційних ресурсів, контролювати мережевий трафік та запобігати використанню відомих вразливостей. До конфігураційних заходів належать налаштування міжмережевих екранів, сегментація мережі, використання систем виявлення та запобігання вторгненням, контроль доступу до мережевих сервісів, а також регулярне оновлення програмного забезпечення. Правильна конфігурація інфраструктури дозволяє зменшити площу потенційної атаки та мінімізувати ризики експлуатації технічних уразливостей.

Не менш важливим елементом системи захисту є криптографічні механізми, які забезпечують конфіденційність, цілісність і автентичність інформації, що передається мережею. Використання сучасних криптографічних алгоритмів і протоколів дозволяє захистити дані від перехоплення, модифікації або підміни під час передачі між мережевими вузлами. До таких механізмів належать шифрування мережевого трафіку, застосування протоколів захищеного з'єднання, використання цифрових сертифікатів, електронного підпису та інфраструктури відкритих ключів. Криптографічні засоби є особливо важливими в умовах використання відкритих мереж зв'язку, де існує підвищений ризик перехоплення інформації.

Разом із технічними засобами захисту важливу роль відіграють організаційні заходи безпеки, які спрямовані на регулювання процесів використання інформаційних систем і формування культури кібербезпеки в організації. До таких заходів належать розроблення політик інформаційної безпеки, регламентів доступу до інформаційних ресурсів, проведення навчання персоналу, аудит безпеки та контроль дотримання встановлених правил. Значна частина кіберінцидентів пов'язана саме з людським фактором, зокрема використанням слабких паролів, необережним поводженням із конфіденційною інформацією або недостатньою обізнаністю користувачів щодо сучасних кіберзагроз. У зв'язку з цим організаційні заходи є необхідним доповненням до технічних засобів захисту.

Поєднання конфігураційних, криптографічних і організаційних механізмів дозволяє сформувати багаторівневу систему захисту, яка забезпечує протидію різним типам кіберзагроз. У межах такого підходу кожен рівень безпеки виконує власну функцію: конфігураційні механізми обмежують доступ до мережевої інфраструктури, криптографічні засоби захищають інформаційні потоки, а організаційні заходи регулюють діяльність користувачів і забезпечують дотримання політик безпеки. Взаємодія цих елементів створює цілісну систему, здатну ефективно протидіяти як технічним, так і соціальним методам кібератак.

Отже, застосування комплексного підходу до захисту комп'ютерних мереж є обґрунтованим та необхідним у сучасних умовах розвитку інформаційних технологій (рис. 3).

Рис 3. Модель комплексного захисту мережі

Поєднання конфігураційних, криптографічних і організаційних заходів дозволяє підвищити рівень захищеності мережевої інфраструктури, забезпечити надійний захист інформаційних ресурсів та мінімізувати ризики реалізації кіберзагроз. Такий підхід сприяє створенню стійкої системи інформаційної безпеки, здатної ефективно функціонувати в умовах постійної еволюції кіберзагроз та зростання складності сучасних інформаційних систем.

Висновки

У результаті проведеного дослідження було здійснено комплексний аналіз сучасних загроз безпеці комп'ютерних мереж, а також розглянуто основні механізми реалізації мережевих атак та підходи до побудови ефективних систем захисту інформаційної інфраструктури. Отримані результати дозволяють сформулювати низку узагальнених висновків.

По-перше, встановлено, що вразливості комп'ютерних мереж мають системний характер і можуть виникати на різних рівнях мережевої архітектури. Особливо вразливим є канальний рівень моделі OSI, який історично розроблявся для роботи в межах довіреного середовища та не передбачає вбудованих механізмів автентифікації мережевих кадрів. Це створює передумови для реалізації атак, спрямованих на перехоплення трафіку, порушення сегментації мережі та несанкціонований доступ до інформаційних ресурсів.

По-друге, у роботі систематизовано основні типи мережевих загроз та атак на мережеву інфраструктуру. Встановлено, що за характером впливу

вони можуть бути пасивними або активними, а за джерелом походження — внутрішніми та зовнішніми. Пасивні атаки спрямовані переважно на перехоплення та аналіз мережевого трафіку, тоді як активні передбачають модифікацію інформаційних потоків або порушення функціонування сервісів. Така класифікація дозволяє більш системно підходити до розроблення механізмів захисту та формування політик інформаційної безпеки.

По-третє, проаналізовано методи мережевої розвідки та сканування, які є важливим етапом підготовки більшості кібернападів. Встановлено, що процес збору інформації про мережеву інфраструктуру може здійснюватися як пасивними, так і активними методами. Пасивна розвідка базується на використанні відкритих джерел інформації та аналізі доступних мережевих даних, тоді як активна розвідка передбачає безпосередню взаємодію з мережевими вузлами через сканування портів, визначення операційних систем, ідентифікацію сервісів та аналіз потенційних вразливостей. Отримані результати підтверджують, що розуміння механізмів мережевого сканування є важливим для належного налаштування систем моніторингу та виявлення вторгнень.

По-четверте, у дослідженні розглянуто механізми реалізації атак на відмову в обслуговуванні (DoS) та розподілених атак відмови в обслуговуванні (DDoS), які належать до найбільш поширених загроз для сучасних інформаційних систем. Показано, що ефективність таких атак обумовлена можливістю використання розподілених ботнет-мереж, механізмів підсилення трафіку та вразливостей мережевих протоколів. Особливу небезпеку становлять волюметричні атаки, атаки транспортного рівня та атаки прикладного рівня, які можуть призводити до перевантаження мережевих каналів, серверних ресурсів та прикладних сервісів.

По-п’яте, обґрунтовано доцільність застосування комплексного підходу до захисту комп’ютерних мереж, що поєднує конфігураційні, криптографічні та організаційні заходи безпеки. Доведено, що сучасні кіберзагрози мають багаторівневий характер, тому ефективний захист інформаційних систем можливий лише за умови реалізації принципу багаторівневої оборони (defense in depth). Конфігураційні заходи дозволяють мінімізувати технічні вразливості мережевої інфраструктури, криптографічні механізми забезпечують конфіденційність і цілісність інформації під час її передавання, а організаційні заходи сприяють формуванню належної культури інформаційної безпеки та підвищенню рівня підготовки персоналу.

Таким чином, результати дослідження підтверджують, що забезпечення безпеки комп’ютерних мереж потребує системного підходу, який враховує як технічні особливості мережевих протоколів, так і організаційні аспекти управління інформаційною безпекою. Практичне значення отриманих результатів полягає у можливості їх використання під час проєктування та модернізації корпоративних мережевих інфраструктур, а також у процесі розроблення політик кібербезпеки.

Перспективи подальших досліджень полягають у розробленні методів автоматизованого виявлення мережевих атак на основі аналізу мережевого трафіку, використанні технологій машинного навчання для прогнозування кіберзагроз, а також удосконаленні механізмів захисту мережевої інфраструктури в умовах розвитку хмарних обчислень, Інтернету речей та розподілених інформаційних систем.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Tanenbaum A. S., Wetherall D. J. *Computer Networks*. 5th ed. Upper Saddle River: Pearson, 2011. 960 p. URL: https://www.pearson.com/en-us/subject-catalog/p/computer-networks/P200000003188
2. Kurose J. F., Ross K. W. *Computer Networking: A Top-Down Approach*. 9th ed. New York: Pearson, 2025. 864 p. URL: https://www.pearson.com/en-us/subject-catalog/p/computer-networking-a-top-down-approach/P200000013385
3. Stallings W. Network Security Essentials: Applications and Standards. 6th ed. Boston: Pearson, 2017. 464 p. URL: https://www.pearson.com/en-us/subject-catalog/p/network-security-essentials/P200000003180
4. Anderson R. *Security Engineering: A Guide to Building Dependable Distributed Systems*. Hoboken: Wiley, 2020. 1248 p. – URL: https://www.wiley.com/en-us/Security+Engineering%3A+A+Guide+to+Building+Dependable+Distributed+Systems%2C+3rd+Edition-p-9781119642787
5. Stewart, J. M., Chapple, M., Gibson, D. *ISC2 CISSP Certified Inf. Systems Security Prof. Official Study Guide*. Hoboken: Wiley, 2024. 1248 p. URL:https://www.wiley.com/en-us/ISC2+CISSP+Certified+Information+Systems+Security+Professional+Official+Study+Guide%2C+10th+Edition-p-9781394254705?utm_source=copilot.com
6. IEEE Computer Society. *IEEE Standard for Virtual Bridged Local Area Networks (IEEE 802.1Q)*. New York: IEEE, 2018. URL: https://standards.ieee.org/standard/802_1Q-2018.html
7. Деркач Т. М., Лавренко М. Кіберпростір: аналіз загроз та методи захисту. *Innovative Education: Problems and Prospects of Scientific Research*: матеріали L Міжнар. наук.-практ. конф. (4–6 грудня 2024 р.). Stuttgart, 2024. С. 112–115. URL: https://reposit.nupp.edu.ua/handle/PoltNTU/18104
8. Лахно В. та ін. Модель захисту локальної мережі навчального закладу. *Кібербезпека: освіта, наука, техніка*. 2022. № 18. С. 6–23. DOI: https://doi.org/10.28925/2663-4023.2022.18.62
9. Сидоренко В., Максимець А. Модель забезпечення стійкості критичних інформаційних систем в умовах впливу внутрішніх та зовнішніх дестабілізуючих чинників. *Кібербезпека: освіта, наука, техніка*. 2025. № 27. С. 560–571. DOI: https://doi.org/10.28925/2663-4023.2025.27.779

10. Хомчак М. Оцінка ризиків кібербезпеки для вибору хмарного провайдера. *Кібербезпека: освіта, наука, техніка*. 2025. № 27. С. 549–559. DOI: https://doi.org/10.28925/2663-4023.2025.27.773
11. Бурячок В. Л., Гулак Г. М., Толубко В. Б. Інформаційний та кіберпростори: проблеми безпеки, методи та засоби боротьби. Львів «Магнолія-2006», 2024. 448 с. URL:https://magnolia.lviv.ua/wp-content/uploads/2024/04/INF-TA-KIBERPROSTORY-pidruchnyk_zmist.pdf
12. Корченко О. Г., Іванченко Є. В., Бакалінський О. В. та ін. Метод оцінювання рівня підвищення кібербезпеки об'єктів критичної інфраструктури держави. *Наукоємні технології*. 2024. № 61(1). С. 3-20. DOI: https://doi.org/10.18372/2310-5461.61.18509

Received (Надійшла) 22.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Деркач Тетяна Миколаївна – доктор філософії, доцент, доцент кафедри комп'ютерних та інформаційних технологій і систем, Національний університет "Полтавська політехніка імені Юрія Кондратюка", Полтава, Україна
Tetyana Derkach – PhD, Associate Professor, Associate Professor of the Department of Computer and Information Technologies and Systems, National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine
e-mail: tanider@ukr.net ORCID Author ID: https://orcid.org/0000-0001-8062-9105
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57204846765

Головко Геннадій Вячеславович – кандиат технічних наук, доцент, доцент кафедри комп'ютерних та інформаційних технологій і систем, Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Gennadii Golovko – Candidate of Technical Sciences (PhD in Engineering), Associate Professor, Associate Professor of the Department of Computer and Information Technologies and Systems, National University "Yuri Kondratyuk Poltava Polytechnic", Poltava, Ukraine.;
e-mail: GenVGolovko@ukr.net , Контактний тел.: 096-57-40-227, ORCID: http://orcid.org/0000-0002-1745-1321

Дмитренко Андрій Олександрович – доктор філософії, доцент, доцент кафедри будівельних конструкцій, Національний університет "Полтавська політехніка імені Юрія Кондратюка", Полтава, Україна;
Andrii Dmytrenko – PhD, Associate Professor, Associate Professor of Department of building structures, National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine;
e-mail: andmyt@ukr.net; ORCID Author ID: https://orcid.org/0000-0002-8715-7646;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217604473.

Клочко Ліна Андріївна – доктор філософії, інженер-геотехнік, геологічне бюро BEG SA, Швейцарія
Lina Klochko – Ph.D., Geological Bureau BEG SA, Switzerland
e-mail: lina.dmitrenko@gmail.com; ORCID Author ID: http://orcid.org/ 0000-0002-6064-2887;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57217278708.

Threat and vulnerability analysis of computer networks and a comprehensive approach to cybersecurity

Tetyana Derkach, Gennadii Golovko, Andrii Dmytrenko, Lina Klochko

Abstract. The article presents a comprehensive analysis of modern threats and vulnerabilities affecting computer networks in the context of the rapid development of digital technologies and the growing number of cyber threats targeting information infrastructures. Particular attention is paid to the study of architectural and protocol vulnerabilities in computer networks, especially at the data link and transport layers of the Open Systems Interconnection (OSI) model, which are often exploited as initial entry points for sophisticated cyberattacks. The study provides a systematic classification of network threats and attacks according to their nature, origin, and the layer of the network architecture in which they occur. Passive and active attacks, as well as internal and external threats, are analyzed in terms of their impact on the confidentiality, integrity, and availability of information resources. Special emphasis is placed on the analysis of modern network reconnaissance and scanning techniques, which are widely used both by cybersecurity professionals for security auditing and penetration testing and by malicious actors to identify vulnerabilities in network infrastructures. A separate section of the article focuses on the analysis of architectural vulnerabilities at the data link layer caused by the lack of authentication mechanisms in fundamental protocols of the IEEE 802 family. The study describes several common attack techniques based on manipulation of Ethernet frames, including MAC table flooding, Address Resolution Protocol (ARP) spoofing, VLAN hopping, as well as attacks targeting Dynamic Host Configuration Protocol (DHCP) infrastructure and the Spanning Tree Protocol (STP). Considerable attention is also devoted to the analysis of Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks, which remain among the most widespread and dangerous cyber threats affecting modern information systems. The study also highlights the role of botnets and compromised Internet of Things (IoT) devices in the execution of large-scale distributed attacks. The increasing number of poorly secured IoT devices significantly expands the attack surface and enables cybercriminals to generate extremely high traffic volumes capable of disrupting the availability of critical information resources. Based on the conducted analysis, the article substantiates the necessity of implementing a comprehensive approach to computer network security. Such an approach integrates configuration-based, cryptographic, and organizational security measures implemented at different layers of the network infrastructure. The results of the study may be applied in the development of information security policies, the design of secure corporate network infrastructures, cybersecurity auditing, and the training of specialists in the fields of information technology and cybersecurity.

Keywords: computer networks, cybersecurity, network attacks, network reconnaissance, network scanning, protocol vulnerabilities, DoS attacks, DDoS attacks, information security.

УДК 004.94:612.84 **doi: 10.26906/SUNZ.2026.2.081**

О. А. Єрошенко

Харківський національний університет радіоелектроніки, Харків, Україна

МОДЕЛЮВАННЯ ПРОЦЕСІВ ФОРМУВАННЯ ФОСФЕННИХ ОБРАЗІВ У СИСТЕМАХ ВІЗУАЛЬНИХ НЕЙРОІНТЕРФЕЙСІВ

Анотація. Предметом дослідження в статті є алгоритмічні методи обробки візуальної інформації та принципи формування фосфенних образів, адаптовані до нейрофізіологічних особливостей зорової системи, з метою їх використання у системах штучного зору та нейропротезуванні. **Метою роботи** є розроблення комплексного методу симуляції зорового сприйняття шляхом інтеграції покращеного контурного аналізу та математичного моделювання фосфенних мап із урахуванням ретинотопічної організації та імовірнісної деградації елементів стимуляції. У статті **вирішуються такі завдання**: аналіз механізмів виділення ключових ознак зображення за допомогою градієнтних методів; розробка підходу до квантування орієнтації контурів для стабілізації зорового образу; побудова математичної моделі фосфенної мапи як дискретного сітчастого поля з варіативними параметрами вузлів; врахування анатомічної нерівномірності розподілу рецепторів сітківки через динамічну зміну геометрії фосфенів; моделювання технічної нестабільності електродів через ймовірнісні параметри пропуску сигналів. **Використовуються такі методи**: алгоритм детектування меж Кенні з чотирма орієнтаціями градієнта, математичне моделювання на основі двовимірних функцій Гауса з еліптичною деформацією, принципи ретинотопічного картування зорової кори, а також методи стохастичного моделювання для імітації деградації імплантів. **Отримано такі результати**: запропоновано програмну модель фосфенної репрезентації, яка забезпечує адаптивне формування візуального образу залежно від ексцентриситету точок зорового поля; реалізовано механізм просторово-селективної стимуляції, що поєднує високу деталізацію у центрі із дифузним представленням на периферії; продемонстровано ефективність використання еліптичної деформації та квантування кутів для підвищення реалістичності симуляції штучного зору. **Висновки**: Розроблений метод моделювання доводить, що поєднання контурного аналізу з урахуванням індивідуальної ретинотопії та анатомічної варіативності дозволяє створювати інформативні візуальні образи навіть за умов низької роздільної здатності нейропротезів, забезпечуючи надійну теоретичну базу для проектування та налаштування сучасних систем візуального заміщення.

Ключові слова: зір, фосфен, математичне моделювання, стимуляція, візуалізація, модель, штуяний зір, нейропротез.

Вступ

Більшу частину інформації про світ людина отримує саме через зоровий аналізатор. Його пошкодження критично впливає на життєдіяльність людини: виникають складнощі з просторовою орієнтацією, спілкуванням та рутинними справами. Аби допомогти людям із порушеннями зору, науковці у сферах медицини та біоінженерії створюють технології, здатні частково відновити або замістити втрачену функцію [1]. Основою таких технологій є феномен нейропластичності. Мозок здатний адаптуватися, "навчаючи" зорову кору обробляти сигнали, що надходять від слуху або дотику [2]. Такий принцип застосовується у пристроях сенсорної субституції, завдяки яким незрячі отримують візуальні дані через інші канали сприйняття.

Сучасні методи відновлення зору поділяють на два напрямки: неінвазивні (тактильні дисплеї, аудіосистеми); інвазивні (імпланти, нейроінтерфейси).

Попри значний прогрес, технології активно розвиваються, вони стикаються з низкою бар'єрів: низька якість "картинки", дорожнеча, проблеми індивідуальної адаптації та необхідність тривалого звикання користувача до системи.

Метою цієї роботи є моделювання процесу зорової стимуляції, які базується на принципах функціонування кортикальних імплантів.

Матеріали та методи дослідження

Будучи основною сенсорною системою, зоровий аналізатор має критичне значення для формування нашого досвіду та визначає більшість сфер людської життєдіяльності. З огляду на це, втрата зору є масштабною світовою проблемою. Вона кардинально знижує рівень життя величезної кількості людей, що зумовлює гостру потребу в багатогранних методах медичної реабілітації та залученні сучасних технологічних рішень.

Згідно з даними досліджень у сфері біомедичної інженерії, частка візуальної інформації серед усіх зовнішніх стимулів, які аналізує наш мозок, сягає 80–90% [1]. Такі показники яскраво демонструють виняткову функцію зорового апарату в конструюванні картини світу, орієнтуванні в середовищі та маніпуляціях з навколишніми предметами.

Візуальне сприйняття – це надважливий механізм, який відповідає не лише за ідентифікацію кольору, форми та глибини простору, а й за розшифрування комплексних сцен, без яких неможливе прийняття щоденних рішень. На обробку зорових імпульсів виділяється від 30% до 55% кори головного мозку [1], що суттєво більше за обсяг нейронних ресурсів, залучених для слухового чи тактильного аналізу. Подібна структурна спеціалізація мозку є прямим свідченням того, наскільки зір був важливим фактором в еволюції та адаптації людства.

Механізм візуального сприйняття є багатокомпонентним. Він стартує з фіксації світлових променів фоторецепторами на сітківці ока і завершується глибоким аналізом у зоровій корі, де розпізнаються рух, просторові параметри та самі об'єкти. Завдяки такій ієрархічній та багатоетапній будові, зорова система досягає безпрецедентної швидкості й точності обробки даних, яка перевершує можливості будь-яких інших органів чуття.

Крім того, здатність бачити має критичне значення для соціального, емоційного та інтелектуального становлення особистості, насамперед у дитячому віці. Завдяки візуальним подразникам розвиваються абстрактне мислення, мовленнєві навички, увага та пам'ять. Згідно з дослідженнями, дефіцит зору в ранні роки здатний викликати затримку когнітивного та мовного розвитку. Він також ускладнює комунікацію через неможливість зчитувати невербальну інформацію, таку як міміка чи жестикуляція. Щодо дорослих, то збереження зору прямо корелює з їхньою самостійністю, успішністю в професії та психологічним благополуччям, що доводить його необхідність для повноцінного функціонування в соціумі.

Отже, зоровий аналізатор слугує не просто головним каналом отримання даних про світ, а й виступає обов'язковою умовою для формування психоемоційних та когнітивних механізмів, гарантуючи всебічну адаптацію індивіда до умов довкілля.

Втрата або погіршення зорової функції суттєво обмежує життєдіяльність: виникають труднощі з читанням, просторовою орієнтацією, трудовою активністю та спілкуванням. Найбільш гостро ця криза постає в державах із середнім та низьким рівнем економічного розвитку, де близько 90% людей із вадами зору не мають належного доступу до медичної допомоги [2-3]. До того ж сліпота чи слабозорість часто супроводжуються соціальною ізоляцією, депресивними станами та фінансовою залежністю, що ще раз акцентує увагу на потребі у створенні ефективних технологічних та реабілітаційних засобів.

Для України питання захисту офтальмологічного здоров'я є надзвичайно гострим, оскільки проблеми із зором фіксуються приблизно у третини населення [3]. Збільшення кількості патологій тісно пов'язане з поширенням міопії, цукрового діабету та судинних хвороб. Значний негативний вплив має і сучасний стиль життя, зокрема постійне використання цифрових пристроїв. Саме через надмірне напруження очей все частіше діагностуються синдром сухого ока та комп'ютерний зоровий синдром.

Згідно зі статистикою Всесвітньої організації охорони здоров'я (ВООЗ), до 80% усіх порушень зору піддаються лікуванню або профілактиці.

Вік є одним із найвагоміших чинників, що визначає рівень поширення офтальмологічних захворювань. Згідно зі статистикою, 82% повністю незрячих осіб та 65% людей зі зниженим зором належать до вікової категорії від 50 років і старше. Парадоксально, але ця демографічна група становить лише п'яту частину (20%) від загальної кількості населення планети [2, 3]. Водночас до зони підвищеного ризику входять і діти. Як зазначалося раніше, раннє погіршення або втрата зорової функції здатні суттєво гальмувати інтелектуальний прогрес та соціальне становлення дитини.

Підсумовуючи вищезазначене, можна стверджувати, що етіологія зорових дисфункцій має комплексний характер. Вона формується під впливом цілого спектра чинників: вікових змін, спадковості, інфекційних уражень, а також умов соціально-економічного середовища. Для успішного подолання цієї масштабної проблеми необхідна синергія різних напрямків медицини та науки, яка має обов'язково включати своєчасне виявлення патологій, превентивні заходи та активне залучення передових технологічних розробок.

Людський зоровий апарат являє собою високоорганізовану багаторівневу систему. Вона об'єднує в собі як периферичні, так і центральні відділи, тісна і безперервна взаємодія яких гарантує здійснення всього комплексу процесів візуального сприйняття.

Первинним органом зору є око. Воно складається з кількох ключових структур (рис. 1), що взаємодіють на різних етапах обробки світла: рогівка (cornea) відповідає за первинне заломлення; райдужка (iris) регулює кількість потрапляння; кришталик (lens) здійснює додаткове фокусування; сітківка (retina) перетворює сигнал на нервові імпульси.

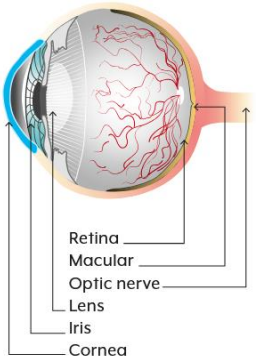


Рис. 1. Будова ока людини

Трансформація світлових подразників у нервові імпульси відбувається в сітківці, після чого згенеровані електричні сигнали транспортуються по зоровому нерву (optic nerve) до хіазми – перехрестя зорових нервів (optic chiasm). Саме в цій анатомічній зоні відділів мозку реалізується часткове переплетення нервових трактів, як продемонстровано на рис. 2.

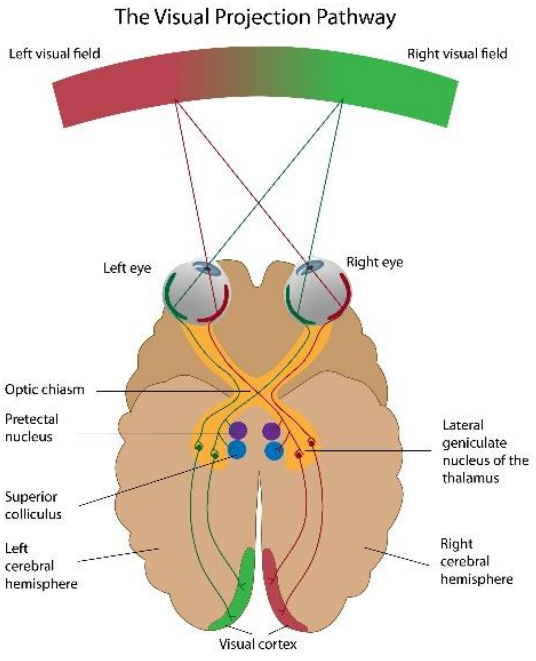


Рис. 2. Структура зорового аналізатору людини

Наступним етапом є передача нервових імпульсів до латерального колінчастого тіла (ЛКТ). Ця специфічна структура таламуса виконує функцію своєрідного фільтра та базового вузла аналізу візуальних даних. Саме в ЛКТ здійснюється класифікація, підсилення та попередня консолідація отриманих сигналів. Завдяки цьому механізму найпріоритетніша сенсорна інформація відокремлюється ще до того, як досягне кори головного мозку [1].

Від ЛКТ нервові волокна (аксони) проходять крізь зорову променистість, досягаючи первинної зорової кори (зона V1), яка локалізована в потилич-ній частці мозку. На цьому етапі стартує глибинний кортикальний аналіз візуальних подразників: нейронна мережа починає ідентифікувати просторове розташування, контури, рівень контрастності та вектори руху об'єктів на вищому когнітивному рівні [1]. Саме за таким алгоритмом центральні відділи зорового апарату фіналізують перетворення зовнішнього фізичного подразника (світла) на наше усвідомлене та суб'єктивне візуальне сприйняття дійсності.

Кодування та передача сигналу

Процес кодування зорових сигналів відбувається у сітківці – делікатному шарі нейрональної тканини. Вона утворена п'ятьма головними класами нейронів (рис. 3). Їхня спільна функція полягає у базовій обробці візуальної картини та спрямуванні згенерованих імпульсів до вищих рівнів зорової системи:

- фоторецептори (палички та колбочки);
- горизонтальні, біполярні, амакринові та гангліозні клітини.

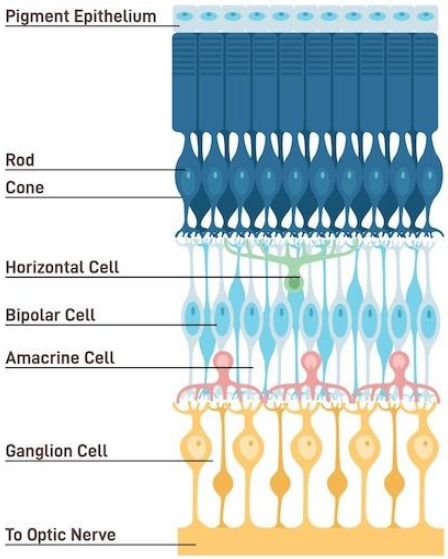


Рис. 3. Структура сітківки

Функція паличок полягає в забезпеченні зорового сприйняття за умов слабкого світла (так званий скотопічний зір). Натомість колбочки є активними переважно під час яскравого денного освітлення (фотопічний зір), відповідаючи за диференціацію кольорів та високу деталізацію візуальної картини [3].

Світлові стимули збуджують фоторецептори, генеруючи імпульси, які далі підлягають проміжній трансформації за участю амакринових, горизонтальних та біполярних нейронів. Після цієї попередньої фільтрації та консолідації інформація надходить до гангліонарних (гангліозних) клітин. Саме з їхніх довгих відростків – аксонів – і формується цілісна структура зорового нерва, яким сигнали прямують до центральної нервової системи.

Механізм фототрансдукції, який полягає у трансформації енергії світла у нервові (електричні) імпульси, стартує в момент захоплення квантів світла клітинами-фоторецепторами. Що запускає каскад фотохімічних перетворень, унаслідок яких виникають коливання мембранного потенціалу рецептора. Як наслідок, відбувається регуляція секреції специфічних нейромедіаторів, що дозволяє транслювати отримані дані до наступних рівнів нейронної мережі [3].

Що стосується сприйняття кольорів, то за їх шифрування на рівні сітківки відповідають три різновиди колбочкових клітин. Як проілюстровано на рис. 4, кожен із цих типів налаштований на сприйняття специфічного спектра світлових хвиль:

- довгих (червоний);
- середніх (зелений);
- коротких (синій).

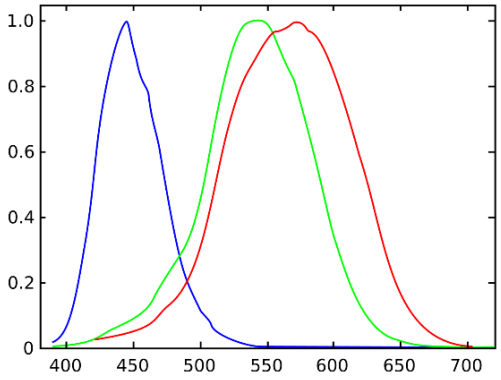


Рис. 4. Функції спектрального відгуку колбочок

Кольорову палітру, яку ми здатні розрізняти, формується завдяки балансу збудження цих трьох різновидів колбочок. За своєю суттю цей біологічний механізм є прямим аналогом колірної системи RGB, що масово застосовується під час кодування цифрової графіки [4, 5].

Фундаментальним принципом функціонування зорового апарату виступає явище ретинотопії. Його сенс полягає в тому, що початкова просторова конфігурація зображення, спроєктована на сітківку ока, не змішується, а переноситься у вигляді своєрідної топографічної карти на всі вищі рівні нейронного аналізу [2].

Завдяки такій специфічній архітектурі гарантується максимально точна трансляція параметрів простору, контрастності та інших якісних характеристик об'єктів. Зрештою, це дає змогу головному мозку конструювати цілісну, глибоко деталізовану та змістовну картину зовнішнього світу.

Головним центром, що відповідає за аналіз та розшифровку візуальних сигналів, є зорова кора (рис. 5), яка локалізується здебільшого в потиличній зоні мозку. Завдяки цій багатокомпонентній нейронній мережі формується повноцінна картина довкілля: ми

отримуємо здатність ідентифікувати предмети, визначати їхні просторові координати та адекватно реагувати на будь-які трансформації навколишнього середовища.

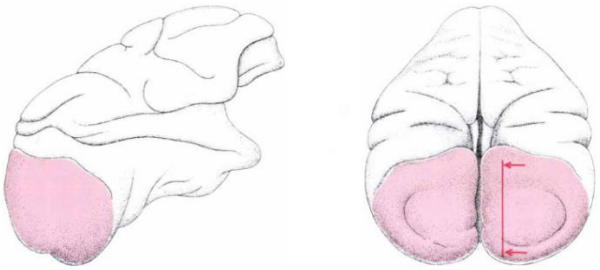


Рис. 5. Позначення первинної зорової кори в мозку людини

Функціонально зорова кора побудована за суворим ієрархічним принципом. Її окремі ділянки мають вузьку спеціалізацію, одні відповідають за примітивний аналіз форм, тоді як інші здійснюють глибоку семантичну обробку побаченого. Така архітектура дає мозку змогу генерувати високодеталізовані візуальні образи, що слугують фундаментом для вищих когнітивних функцій – запам'ятовування, концентрації уваги, процесу навчання та прийняття рішень.

Стартовим майданчиком кортикального аналізу виступає первинна зорова кора (відома як зона V1). Вона анатомічно розміщена вздовж шпорної борозни в потиличній частці та приймає нервові імпульси, що надходять від латерального колінчастого тіла (ЛКТ) таламуса по зорових трактах.

Клітини зони V1 характеризуються наявністю специфічних рецептивних полів. Вони активуються у відповідь на чіткі тригери: рівень контрастності, просторову частоту або певний кут нахилу ліній (горизонтальний, вертикальний чи діагональний). Фактично, ці нейрони здійснюють первинну деконструкцію візуального стимулу, розкладаючи складне зображення на найпростіші базові елементи, такі як текстури та контури об'єктів.

Після початкового етапу в зоні V1 оброблені дані спрямовуються до екстрастріарних (вторинних) відділів зорової кори. Кожна з цих ділянок має чітку функціональну спеціалізацію:

– зона V2: відповідає за розпізнавання складних геометричних форм, вигинів та кутів, а також бере безпосередню участь в оцінці просторової глибини та забезпеченні стереоскопічного бачення;

– зона V3: фокусується на загальній просторовій компоновці сцени та динамічній ідентифікації контурів об'єктів;

– зона V4: виступає головним центром колірного аналізу, який додатково розпізнає специфіку поверхонь та їхні текстури, формуючи сприйняття відтінків;

– зона V5 (або MT): вузькопрофільна ділянка, що відповідає за детекцію руху – вона вираховує швидкість переміщення, вектори та загальну траєкторію фізичних тіл.

У сфері розробки систем штучного бачення саме зорова кора виступає головним об'єктом втручання. Технології кортикального протезування базуються на прямій стимуляції нейронів цієї зони, що провокує появу фосфенів – елементарних світлових спалахів, з яких згодом конструюються спрощені зорові патерни [6-10]. Водночас, спираючись на феномен нейропластичності, зорова кора демонструє виняткову здатність перепрофілюватися на обробку інформації від альтернативних сенсорних каналів, беручи активну участь у крос-модальній інтеграції.

Підсумовуючи, можна стверджувати, що багаторівнева ієрархічна архітектура, високий ступінь пластичності та чітка функціональна спеціалізація перетворюють зорову кору не лише на фундаментальну базу природного сприйняття, а й на головну платформу для створення передових нейроінженерних рішень.

Механізми виникнення фосфенів

Під терміном фосфени в нейрофізіології розуміють специфічні візуальні відчуття у вигляді світючих точок, спалахів чи складних геометричних патернів, які формуються у свідомості людини за повної відсутності зовнішнього світлового подразника.

У природних умовах такі суб'єктивні зорові артефакти найчастіше є наслідком прямого механічного впливу на очне яблуко (зокрема, на сітківку) або ж виступають супутнім симптомом певних патологічних станів, наприклад, зорової аури при мігрені. Проте найбільший інтерес для сучасної нейроінженерії становить можливість генерувати фосфени штучно (рис. 6). Цього досягають за допомогою цілеспрямованого подання електричних імпульсів на структури зорового нерва або безпосередньо на ділянки зорової кори головного мозку [6-9].

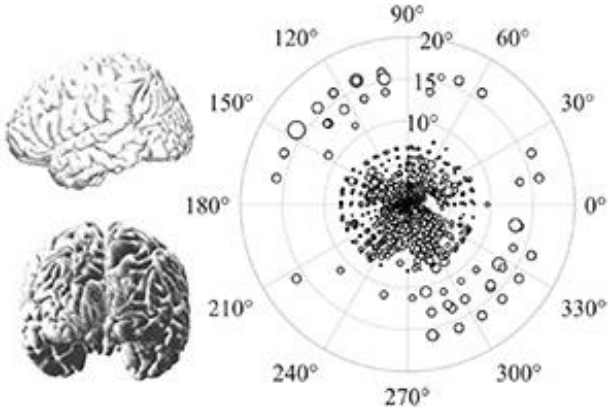


Рис. 6. Схематичне зображення конектому фосфенів

У сфері розробки візуальних протезів феномен фосфенів відіграє центральну роль. Фізіологічне підґрунтя цього явища базується на безпосередньому збудженні нейронних ланцюгів зорового тракту. Зокрема, коли на зорову кору подається електричний імпульс, виникає вогнищева нейрональна активність, яку головний мозок дешифрує як реальні спалахи світла.

Передові системи штучного бачення, такі як ретинальний імплант Argus II або кортикальний протез Orion, використовують сукупність таких штучних фосфенів для конструювання базових візуальних патернів. Згенеровані образи мають низьку роздільну

здатність, проте дають незрячим пацієнтам змогу ідентифікувати загальні обриси предметів, орієнтуватися в просторі та помічати великі перешкоди.

Специфічною рисою штучно викликаних фосфенів є їхня просторова стабільність (статичність). Координати індукованого світлового спалаху в полі зору жорстко прив'язані до анатомічного розташування стимульованих нейронів у кожного конкретного пацієнта. Однак результати актуальних досліджень свідчать про те, що просторові координати фосфенів у суб'єктивному полі зору далеко не завжди лінійно збігаються з фізичною топологією активованих нейронів кори головного мозку [8]. Топографічна невідповідність постає як одна з найгостріших проблем у процесі проєктування та калібрування сучасних нейропротезів.

Слід особливо наголосити: цілеспрямоване збудження поодиноких нервових клітин чи їхніх мікрокластерів рідко призводить до появи чітко прогнозованого геометричного малюнка з фосфенів (рис. 7). Натомість генерується доволі заплутана та нелінійна конфігурація візуальних артефактів. Така специфіка виникнення образів безпосередньо детермінується унікальною мікроархітектурою нейронних мереж кожного окремого пацієнта [6, 8].

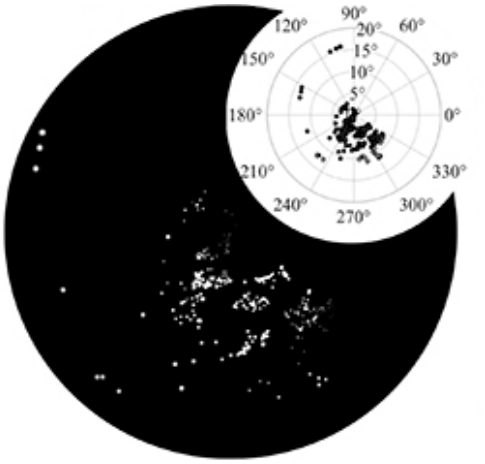


Рис. 7. Мапа фосфенів, придатних до активації

З огляду на це, проблема точного зіставлення штучних фосфенів із конкретними локусами зорової кори перетворюється на одне з найпріоритетніших завдань під час конструювання нейроінтерфейсів. Для забезпечення стабільного та функціонально придатного зорового сприйняття розробникам необхідно здійснювати глибоке ретинотопічне картографування та гнучко налаштовувати параметри стимуляції.

Ще однією вагомою перешкодою виступає гетерогенність нейрональної відповіді. Кортикальний шар налічує мільйони нервових клітин, кожна з яких володіє унікальними рецептивними полями, налаштованими на специфічні тригери (наприклад, вектори руху або орієнтацію ліній у просторі). Поточні технології електростимуляції через імпланти здебільшого збуджують цілі нейронні кластери, а не ізольовані клітини. Як наслідок, індуковані фосфени втрачають чіткість, стаючи розмитими та структурно деформованими [6, 8].

Базові параметри електростимуляції – такі як амплітуда струму, частота та тривалість імпульсу – безпосередньо детермінують морфологію фосфенів, змінюючи їхню інтенсивність світіння, габарити та конфігурацію (рис. 8) [9]. Суб'єктивне сприйняття цих візуальних артефактів суттєво відрізняється, пацієнти можуть ідентифікувати їх як чіткі крапки, яскраві спалахи, смуги, кільцеподібні структури або ж дифузні плями [7]. Крім того, індуковані образи здатні набувати білуватих, сіруватих, жовтуватих чи синюватих відтінків [7], проте цілеспрямоване апаратне керування колірним спектром наразі залишається нерозв'язаною проблемою. Така висока варіативність характеристик значно ускладнює процес генерації стійких та прогнозованих візуальних картин.

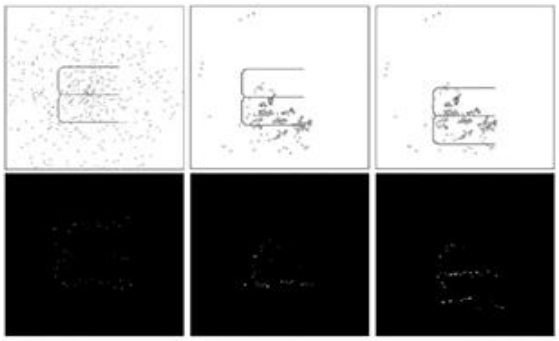


Рис. 8. Передбачуваний результат стимуляції

Наступною перепоною є просторова локалізація фосфенів, яка часто дисонує з реальною анатомічною топологією зорової кори. З метою мінімізації цих просторових похибок та підвищення якості штучного зору, сучасні наукові розробки фокусуються на максимізації щільності мікроелектродних матриць. Рівноцінно важливим вектором є створення смарт-алгоритмів керування стимуляцією, які б адаптувалися до унікальних патернів нейрональної активності конкретного реципієнта.

Перспективні наукові розвідки у сфері візуального протезування, сфокусовані на прецизійному топографічному відображенні нейрональних процесів та глибинній модернізації стимулюючих нейроінтерфейсів, здатні кардинально розширити експлуатаційні можливості таких пристроїв. Зрештою, акумуляція цих технологічних проривів дозволить зробити штучно згенеровану картину світу максимально наближеною до повноцінного біологічного бачення.

Просторове кодування зображень у візуальній системі

Просторове кодування зображень є основою зорового сприйняття, що дає змогу центральній нервовій системі декомпонувати багатовимірні зорові картини на набір базових маркерів – обриси об'єктів, їхні межі та зони різкого перепаду освітленості.

На біологічному рівні цей етап аналізу ініціюється безпосередньо в нейронних шарах сітківки. Там гангліонарні клітини фіксують точкові коливання яскравості світла в межах своїх рецептивних полів, у такий спосіб екстрагуючи первинні структурні патерни та рівні контрасту.

Природна парадигма знайшла своє пряме відображення в технологіях комп'ютерного зору (CV). Алгоритмічний пошук градієнтів та країв виконує аналогічну функцію – відокремлює ключові структурні компоненти цифрового кадру. Під час створення комп'ютерних моделей зорового сприйняття така біоімітація реалізується через спеціальні оператори виділення контурів [11].

Серед усього арсеналу методів CV саме алгоритм Кенні визнано найбільш релевантним та фізіологічно доцільним інструментом. Найбільш фізіологічно обґрунтованим та ефективним вважається алгоритм Кенні, оскільки поєднує декілька етапів обробки [12]:

- згладжування сигналу;
- виявлення градієнтів;
- придушення немаксимумів;
- порогова фільтрація.

Для ефективного придушення цифрового шуму та високочастотних артефактів застосовується метод Гауссового згладжування, який діє як інструмент просторового розмиття.

За допомогою двовимірної функції Гауса здійснюється перерахунок інтенсивності кожного конкретного пікселя у вихідному кадрі:

$$G(x,y)=\frac{1}{2\pi e^2}\cdot e^{-\frac{x^2+y^2}{2\sigma^2}},$$

де x, y – координати точки зображення; σ – параметр розмиття, що моделює розсіювання сигналу.

Відповідно, процес формування фінального згладженого кадру зводиться до математичної операції конволюції (згортки) масиву початкового зображення з матрицею рецептивного фільтра:

$$I_{blur}(x,y)=(I\cdot G)(x,y)=$$
$$=\sum_i\sum_j I(x-i,y-j)\cdot G(i,j),$$

де $I(x,y)\in[0,255]$ – вхідне зображення, подане у відтінках сірого; G – ядро гаусового фільтра.

Далі виконується розрахунок вектора градієнта інтенсивності пікселів із застосуванням спеціалізованих просторових фільтрів:

$$\left|\nabla I_{blur}(x,y)\right|=\sqrt{(G_x\cdot I_{blur})^2+(G_y\cdot I_{blur})^2},$$

де $G_x=\begin{bmatrix}-1&0&1\\-2&0&2\\-1&0&1\end{bmatrix}, G_y=\begin{bmatrix}-1&-2&-1\\0&0&0\\1&2&1\end{bmatrix}$ – ядра оператора Собеля.

Для спрощення подальшого аналізу та точного позиціонування сусідніх пікселів, увесь діапазон можливих напрямків розбивається на чотири базові сектори: горизонтальна (0°), вертикальна (90°) та дві діагональні (45° і 135°).

На основі отриманих даних обчислюється вектор спрямування градієнта в кожному вузлі зображення. Для подальшої стабілізації алгоритму цей кут підлягає процедурі квантування, тобто округлення до одного з чотирьох опорних напрямків, що математично виражається через арктангенс відношення частинних похідних [13]:

$$\Theta(x,y)=arctg_2\left(G_y\cdot I_{blur}, G_x\cdot I_{blur}\right),$$

де $arctg_2(x,y)\in(-\pi;\pi]$ – функція арктангенса з двома аргументами, що обчислює кут напрямку вектора у полярній системі координат. Завдяки цим розрахунковим даним стає можливим не лише кількісно оцінити інтенсивність локальних флуктуацій яскравості, а й чітко локалізувати вектор, уздовж якого спостерігається найбільший контраст. Фінальна стадія обробки за методом Кенні охоплює каскад процедур: придушення немаксимумів (для витончення ліній), двопорогову фільтрацію (для відсікання шумів) та трасування контурів методом гістерезису [14-15]. У результаті цього багаторівневого аналізу генерується підсумкова бінарна матриця контурів, де логічна одиниця відповідає виявленій межі об’єкта:

$$C(x,y)\in\{0,1\}^{M\times N},$$

де $M\times N$ – роздільна здатність зображення, пікселів; $C(x,y)=1$ – активний контур.

Використання подібних підходів імітує фундаментальну властивість людського зору – здатність до селективної концентрації на домінантних об'єктах при одночасному нівелюванні надлишкових деталей. Такий механізм фільтрації набуває особливого значення в розробці нейронних протезів, де пропускна здатність каналів передачі сигналів до кори головного мозку є суворо лімітованою [13].

Отже, замість спроб репродукувати весь масив складних нейрональних обчислень, доцільно використовувати алгоритмічні методи акцентування контрастних градієнтів. Це дозволяє генерувати лаконічні та зрозумілі візуальні образи, які, попри свою спрощеність, зберігають ключову інформацію про просторову структуру сцени.

Ретинотопічна організація та картографування зорової кори

В основі функціонування зорового аналізатора лежить ретинотопічна організація – фундаментальна властивість, що забезпечує топологічну цілісність передачі візуальних даних від сітківки до вищих нервових центрів. Завдяки цьому механізму кожна локальна зона ретинального шару має чітку проекцію на специфічний сегмент зорової кори, що дозволяє формувати впорядковану карту видимого простору в головному мозку. У процесі математичного моделювання штучного бачення цей біологічний принцип відтворюється через створення фосфенних мап. Вони являють собою структуровані масиви дискретних точок, які імітують вогнища нейрональної активації в первинній зоровій корі (V1), що фактично відповідає просторовому розміщенню електродів у реальних нейроімплантах.

У клінічній практиці координати доступних точок стимуляції визначаються шляхом комплексних нейрофізіологічних досліджень, функціонального

картування кори та верифікації під час хірургічних маніпуляцій. Такі підходи дозволяють встановити жорстку кореляцію між конкретним електродом протеза та його «відгуком» у певній ділянці зорового поля пацієнта. Проте на етапі теоретичного моделювання, коли прямі емпіричні дані можуть бути недоступними або надлишковими, доцільно використовувати абстрактну концепцію фосфенних мап. У такому контексті вони представляються як дискретні сітчасті поля з чітко визначеними геометричними параметрами:

$$P \in \{0,1\}^{K \times L},$$

де $K \times L$ – роздільність матриці; $P(i,j) = 1$ – факт присутності фосфена на відповідній позиції.

Для адекватного відтворення біологічної неоднорідності та технічних обмежень реальних пристроїв у розрахункову модель інтегровано ймовірнісний коефіцієнт втрати фосфенів. Цей параметр дає змогу врахувати статистичну можливість відсутності зорового відгуку, що в реальних умовах може бути наслідком апаратної несправності конкретних електродів, недостатнього рівня нейрональної активації або локального зниження чутливості рецептивних полів:

$$P(i,j) = \begin{cases} 1, & якщо \xi_{i,j} > \rho_{drop}, \\ 0, & \end{cases}$$

де $\xi_{i,j} \sim \cup[0,1]$ – рівномірно розподілена випадкова величина у точці; $\rho_{drop} \in [0,1]$ – параметр, що визначає ймовірність пропуску фосфенів.

Отже, розроблена модель фосфенної мапи забезпечує можливість симуляції просторово-селективного збудження, що базується на анатомічній структурі та функціональних особливостях зорового апарату. Крім того, такий підхід дозволяє математично оцінити наслідки поступової деградації імплантованих компонентів, що є критично важливим для прогнозування довгострокової ефективності нейропротезів.

Моделювання перцептивного образу

Процес формування перцептивного образу в межах імітаційного моделювання зосереджений на репродукції механізмів, за допомогою яких когнітивна система трансформує штучні імпульси у цілісне візуальне уявлення. Штучно індуковані фосфени, що виникають внаслідок електростимуляції нейронних структур, зазвичай дешифруються мозком як світлові плями з дифузними межами. При цьому їхня морфологія суттєво детермінується персональними нейроанатомічними характеристиками реципієнта. Щоб досягти максимальної відповідності моделі реальним фізіологічним процесам, необхідно враховувати градієнтну структуру розподілу фоторецепторів. Центральна ділянка (фовеа) відзначається максимальною концентрацією колбочок, що зумовлює найвищу гостроту зору. У міру віддалення від центру до периферії щільність рецепторів стрімко знижується, що призводить до деградації деталізації.

У контексті цифрового моделювання фосфенних мап ця закономірність реалізується через динамічну зміну геометричних параметрів візуальних елементів. Зокрема, радіус фосфенів корелює з їхньою ексцентричністю: чим більша відстань від центральної осі зору, тим більшим є діаметр світлової плями, що імітує зниження роздільної здатності на периферії:

$$r(i,j) = r_0 \cdot \left(1 + \alpha \cdot \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\right),$$

де r_0 – базовий радіус у центрі зображення; (x_c, y_c) – координати центру зображення; α – коефіцієнт масштабування, що визначає, наскільки швидко зростає розмір фосфенів при віддаленні від центру.

Застосування такої методики дозволяє розрахунковій моделі інтегрувати фундаментальну ретинотопічну архітектуру зорового тракту з персоналізованими анатомічними чинниками. Мова йде, перш за все, про врахування локальної щільності ретинальних фоторецепторів та фактичну геометрію розміщення електродної матриці нейроімплантату. Сформоване в результаті імітаційне поле демонструє високу концентрацію малих, деталізованих фосфенів у центральній зоні, тоді як периферійні області заповнюються масштабнішими та менш вираженими елементами. Такий градієнт чітко репродукує природну фізіологічну деградацію зорової роздільної здатності від центру до країв. З математичної точки зору, кожен окремий візуальний спалах описується як двовимірний розподіл Гауса, фокус якого збігається з відповідним вузлом розрахункової сітки:

$$\Phi_{i,j}(x,y) = e^{-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma_{i,j}^2}},$$

де $\sigma_{i,j} = \frac{r(i,j)}{k}$ – ступінь розмиття, пропорційний до розміру фосфена.

Враховуючи, що в реальних умовах нейростимуляції візуальні артефакти рідко мають ідеальну ізотропну форму, їхня морфологія в моделі апроксимується двовимірною еліптичною геометрією. Таке представлення дозволяє врахувати анізотропію розповсюдження електричного поля в нервовій тканині та індивідуальну витягнутість світлових плям:

$$\frac{(x - x_i)^2}{a_{i,j}^2} + \frac{(y - y_i)^2}{b_{i,j}^2} \le 1,$$

де (x_i, y_i) – координати центру фосфена у площині зображення; $a_{i,j}$ та $b_{i,j}$ – півосі еліпса.

Зважаючи на високий ступінь суб'єктивності та неоднорідності сприйняття подібних ілюзорних спалахів, у математичний апарат моделі доцільно впровадити механізм деформації головних осей еліпсів, що дозволяє симулювати просторову анізотропію фосфенів шляхом варіації їхніх лінійних параметрів (півосей):

$$a_{i,j} = r(i,j) \cdot \xi_a, b_{i,j} = r(i,j) \cdot \xi_b,$$

де $\xi_a, \xi_b \sim \cup[0.8, 1.2]$ – випадкові коефіцієнти деформації (±20% від 1).

Така модифікація дає змогу репродукувати автентичну неоднорідність сприйняття, що є наслідком багатофакторної просторової взаємодії між стимулюючими елементами та нейрональними мережами. Впровадження механізмів морфологічної деформації фосфенів значно підвищує прецизійність моделі, наближаючи її до емпіричного досвіду пацієнтів, у яких контури зорових феноменів зазвичай суттєво дисонують з ідеалізованими геометричними припущеннями.

Результати

Центральним завданням розробленого програмного забезпечення є синтез фосфенної репрезентації візуальних образів на основі аналізу вихідних графічних даних. Процес моделювання базується на побудові стохастичної фосфенної мапи, що імітує унікальну ретинотопічну архітектуру зорового апарату конкретного суб'єкта.

Архітектура програми дозволяє здійснювати прецизійне налаштування критичних параметрів симуляції, які не залежать від методів первинної обробки сигналу. Аналіз фосфенного представлення чітко відображає закладені принципи ретинотопічної організації, що проявляється у нерівномірній деталізації образу. Зокрема, у центральній частині зорового поля, яка відповідає зоні фовеального зору, спостерігається висока щільність дрібних та чітко локалізованих фосфенів, що дозволяє ідентифікувати дрібні деталі об'єктів.

Водночас, у міру наближення до периферії зображення, фосфени закономірно збільшуються у розмірах та набувають дифузного характеру, імітуючи природне зниження роздільної здатності зорового аналізатора.

Отриманий у такий спосіб результат (рис. 9) демонструє успішне застосування еліптичної деформації фосфенів та квантування орієнтації контурів, що забезпечує структурну цілісність сприйнятого образу навіть за умов значного спрощення візуальних даних.

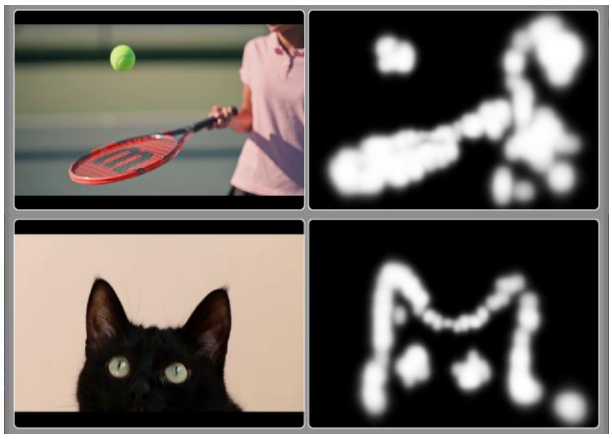


Рис. 9. Порівняння оригінального та фосфенного представлень

Таким чином, порівняння оригінального та фосфенного представлень підтверджує, що розроблена модель дозволяє формувати інформативний зоровій образ, який адекватно відтворює геометрію реальної сцени через призму технічних та фізіологічних обмежень нейроінтерфейсу.

Висновки

У результаті проведеного дослідження розроблено та науково обґрунтовано комплексний підхід до моделювання фосфенних образів, що дозволяє суттєво підвищити реалістичність симуляції зорового сприйняття в системах нейропротезування. Проаналізовані механізми виділення ключових ознак зображення за допомогою покращеного алгоритму Кенні та впроваджене квантування орієнтації контурів довели свою ефективність у забезпеченні стабільності візуальних об'єктів при їх трансформації у дискретну фосфенну структуру.

Важливим теоретичним і практичним внеском роботи є інтеграція ретинотопічних принципів у математичну модель фосфенної мапи. Врахування анатомічної нерівномірності розподілу фоторецепторів сітківки через динамічну зміну геометрії фосфенів дозволило відтворити природну варіативність людського зору, де висока деталізація у фовеальній зоні поєднується з дифузним сприйняттям на периферії. Застосування двовимірних функцій Гауса з еліптичною деформацією забезпечило адекватну апроксимацію форми фосфенів, наближаючи модель до реальних нейрофізіологічних відгуків, що виникають при електричній стимуляції зорової кори.

Додатково впроваджені імовірнісні параметри деградації елементів стимуляції дозволили врахувати технічну нестабільність нейроінтерфейсів та потенційні втрати сигналів, що робить розроблену модель цінним інструментом для прогностичної оцінки якості зорового заміщення. Сформована програмна модель доводить, що навіть за обмеженої роздільної здатності сучасних електродних матриць, використання запропонованих алгоритмів обробки дозволяє створювати інформативні та впізнавані зорові образи. Таким чином, результати роботи створюють надійне підґрунтя для проектування та індивідуальної оптимізації параметрів візуальних нейропротезів наступного покоління.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Man, D., Olchawa, R. (2018), "The Possibilities of Using BCI Technology in Biomedical Engineering", *Biomedical Engineering and Neuroscience*: Proceedings of the 3rd International Scientific Conference on Brain-Computer Interfaces, Opole, Poland, March 13–14, 2018 (Advances in Intelligent Systems and Computing). Cham: Springer, Vol. 720, P. 30–37. https://doi.org/10.1007/978-3-319-75025-5_4

2. Thaler, L., Goodale, M. A. (2016), "Echolocation in humans: an overview", *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 7, No. 6, P. 382–393. https://doi.org/10.1002/wcs.1408
3. Зделова, Г. С. (2023), "Офтальмологічна допомога в Україні. Стан та перспективи удосконалення (огляд літератури)", *Клінічна та профілактична медицина*, Т. 1, № 23, С. 78–85. https://doi.org/10.31612/2616-4868.1(23).2023.11
4. Єрошенко, О. А., Ціпковський, В. О. (2025), "Порівняльний аналіз методів реального часу для розпізнавання жестів на основі Mediapipe, OpenCV та YOLOv8", *Системи управління, навігації та зв'язку. Збірник наукових праць*, Т. 4, № 82, С. 62–65. https://doi.org/10.26906/SUNZ.2025.4.062
5. Федорченко, В. М., Єрошенко, О. А. (2025), "Застосування алгоритмів штучного інтелекту для моделювання загроз інформаційних систем", *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки»*, Т.6 (75), Ч. 2, С. 384–391. https://doi.org/10.32782/2663-5941/2025.6.2/52
6. Wang, H. Z., Wong, Y. T. (2023), "A novel simulation paradigm utilizing MRI-derived phosphene maps for cortical prosthetic vision", *Journal of Neural Engineering*, Vol. 20, No. 4. https://doi.org/10.1088/1741-2552/aceca2
7. Chen, S. C. et al. (2009), "Simulating prosthetic vision: I. Visual models of phosphenes", *Vision Research*, Vol. 49, No. 12. P. 1493–1506. https://doi.org/10.1016/j.visres.2009.02.003
8. Fernández, E. et al. (2021), "Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex", *Journal of Clinical Investigation*, Vol. 131, No. 23. https://doi.org/10.1172/jci151331
9. Grani, F. et al. (2022), "Time stability and connectivity analysis with an intracortical 96-channel microelectrode array inserted in human visual cortex", *Journal of Neural Engineering*, Vol. 19, No. 4. https://doi.org/10.1088/1741-2552/ac801d
10. Янакаєв, А. А., Єрошенко, О. А. (2025), "Система симуляції зору", *Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління* : тези доп. учасників XV Міжнар. наук.-техн. конф., м. Баку–Харків–Жиліна, 24–25 квіт. 2025 р. Харків: Impress, Т. 2. С. 11. https://doi.org/10.32620/ICT.25.t2
11. Canny, J. (1986), "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6. P. 679–698. https://doi.org/10.1109/tpami.1986.4767851
12. Hubel, D. H., Wiesel, T. N. (1979), "Brain Mechanisms of Vision", *Scientific American*, Vol. 241, No. 3. P. 150–163. https://doi.org/10.1038/scientificamerican0979-150
13. Barkovska, O., Shapiro, A., Mavrynskyi, O., Zhebin, P. (2025), "Дослідження чутливості методу визначення відстані до об'єктів на основі алгоритму FaceMesh", *Системи управління, навігації та зв'язку. Збірник наукових праць*, Т. 2, № 80, С. 76–82. https://doi.org/10.26906/SUNZ.2025.2.076
14. Коваленко, А. А., Чхеідзе, В. О., Севостьянова, О. М., Фомічов, О. О. (2025), "Підвищення точності аналізу та обробки складноструктурних зображень", *Системи управління, навігації та зв'язку. Збірник наукових праць*, Т. 2, № 80, С. 137–140. https://doi.org/10.26906/SUNZ.2025.2.137
15. Barkovska, O. (2025), "Formal description of interaction and data flows in multimodal assistive systems for user autonomy support", *Вісник Херсонського національного технічного університету*, №.4 (95), Ч. 4, С. 15–20. https://doi.org/10.35546/kntu2078-4481.2025.4.3.2

Received (Надійшла) 12.01.2026
Accepted for publication (Прийнята до друку) 08.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Єрошенко Ольга Артурівна – доктор філософії, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Yeroshenko Olha – PhD, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: olha.yeroshenko@nure.ua; ORCID Author ID: https://orcid.org/0000-0001-6221-7158;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57808290700.

Simulation of phosphene vision synthesis in visual neural interface systems

Olha Yeroshenko

Abstract. **The subject of the study** is algorithmic methods of visual information processing and the principles of phosphene image formation, adapted to the neurophysiological characteristics of the visual system for their application in artificial vision systems and neuroprosthetics. **The purpose of the work** is to develop a comprehensive method for simulating visual perception by integrating enhanced edge analysis (Canny algorithm) and mathematical modeling of phosphene maps, accounting for retinotopic organization and the probabilistic degradation of stimulation elements. The article **addresses the following tasks**: analyzing the mechanisms of extracting key image features using gradient methods; developing an approach for quantizing contour orientations to stabilize the visual image; constructing a mathematical model of a phosphene map as a discrete grid field with variable node parameters; accounting for the anatomical irregularity of retinal receptor distribution through dynamic changes in phosphene geometry; and modeling the technical instability of electrodes using probabilistic signal dropout parameters. **The following methods are used**: the Canny edge detection algorithm with four gradient orientations, mathematical modeling based on 2D Gaussian functions with elliptical deformation, principles of retinotopic mapping of the visual cortex (V1), and stochastic modeling methods to simulate implant degradation. **The following results were obtained**: a software model for phosphene representation was proposed, providing adaptive visual image formation depending on the eccentricity of visual field points; a spatial-selective stimulation mechanism was implemented, combining high detail in the center (fovea) with diffuse representation at the periphery; and the effectiveness of using elliptical deformation and angle quantization to increase the realism of artificial vision simulation was demonstrated. **Conclusions**: The developed modeling method proves that combining contour analysis with individual retinotopy and anatomical variability allows for the creation of informative visual images even under conditions of low neuroprosthetic resolution, providing a reliable theoretical framework for designing and configuring modern visual substitution systems.

Keywords: vision, phosphene, mathematical modeling, stimulation, visualization, model, artificial vision, neuroprosthesis.

УДК 004.056.55:004.312.2 **doi: 10.26906/SUNZ.2026.2.090**

О. В. Запорожець, П. А. Калашников

Харківський національний університет радіоелектроніки, Харків, Україна

АВТОМАТИЗОВАНА СИСТЕМА ТЕСТУВАННЯ ПРОДУКТИВНОСТІ ПРОГРАМНИХ СИСТЕМ: АРХІТЕКТУРА, МЕТРИКИ ТА ІНТЕГРАЦІЯ В CI/CD

Анотація. Актуальність. Продуктивність є однією з ключових характеристик якості програмних систем, що суттєво впливає на досвід користувача, надійність сервісів та витрати на інфраструктуру. З огляду на модель ISO/IEC 25010 (performance efficiency: time behavior, resource utilization, capacity) та практики DevOps CI/CD актуальною є автоматизація регулярних і відтворюваних навантажувальних перевірок із формальним рішенням pass/fail та пояснюваною діагностикою регресій. **Об'єкт дослідження:** автоматизована система тестування продуктивності програмних систем у конвеєрі CI/CD, що поєднує навантажувальні тести та спостережуваність (метрики/логи/трейси). **Мета статті:** розробити й обґрунтувати концепцію такої системи, визначити її архітектуру, модулі, порядок виконання тестового циклу та мінімально достатній набір метрик і порогових критеріїв «quality gate» на основі інженерних SLO. **Результати дослідження**. Запропоновано модульну архітектуру, описано процедуру виконання тестового циклу та підхід до формування метрик і baseline-порівнянь (ISO/IEC 25023, термінологія ISTQB) для автоматичного виявлення регресій продуктивності. **Висновки.** Інтеграція тестування продуктивності в CI/CD забезпечує раннє виявлення деградацій і зменшує ризик їх потрапляння у промислове середовище, а поєднання навантаження зі спостережуваністю підвищує пояснюваність причин погіршення.

Ключові слова: продуктивність; тестування продуктивності; навантажувальне тестування; CI/CD; метрики; спостережуваність; пороги; регресія продуктивності.

Вступ

Тестування продуктивності визначається як тестування, що виконується для визначення продуктивності (performance) програмного продукту; у термінології воно напряму пов'язане з оцінкою «performance efficiency» компонента або системи. Для сучасних програмних систем (особливо вебсервісів і мікросервісів) продуктивність є не лише «нефункціональною вимогою», а й фактором, що визначає масштабованість, стабільність під піковим навантаженням і економічність експлуатації [1–4].

У моделі ISO/IEC 25010 характеристика «performance efficiency» деталізується як ступінь, з яким продукт виконує функції в межах заданих часових і пропускних (throughput) параметрів та з урахуванням використання ресурсів і місткості (capacity) [1]. Це означає, що навіть за коректної функціональності програмна система може ставати неприйнятною для користувача через перевищення часу відповіді, істотну варіативність «хвоста» затримок або неефективність використання CPU / пам'яті / мережі.

Традиційна організація performance-перевірок часто має низку обмежень: запуск «перед релізом», ручне налаштування стендів, нерепрезентативне навантаження, складність відтворення результатів та відсутність автоматичного рішення «pass/fail» у конвеєрі постачання [5]. На цьому тлі інтеграція продуктивнісних перевірок у CI/CD розглядається як спосіб раннього виявлення деградацій і формування базових «еталонів» (baseline) для порівняння змін між збірками [6].

Аналіз наукових і практичних підходів. У стандартизованому підході до вимірювань важливо, щоб метрики були пов'язані з інформаційними потребами та приводили до дії, а не накопичення даних «заради даних». Це відповідає тезі ISO/IEC/IEEE 15939 про те, що процес вимірювання має бути цілеспрямованим і підтримувати визначення/застосування/удосконалення набору вимірювань у межах проєкту або організації [3]. Для метрик саме якості продукту стандарт ISO/IEC 25023 визначає формат документування та приклади застосування «quality measures» для кількісної оцінки характеристик (у т.ч. часових параметрів) і порівняння з вимогами або трендами [2].

З боку тестової практики термінологія ISTQB описує навантажувальне тестування (load testing) як різновид тестування продуктивності, що оцінює поведінку системи за різних навантажень (від низького до типового та пікового), а також вводить поняття «load profile» як документування кількості віртуальних користувачів, транзакцій та інтервалу часу, що відображає очікувану експлуатацію [4]. Це важливо для автоматизації: без формалізованого профілю неможливі ні відтворюваність, ні коректне порівняння результатів між збірками.

Сучасні open-source інструменти для навантажувального тестування підтримують сценарний підхід і запускаються у headless/CLI режимі, що спрощує інтеграцію в CI. Наприклад, Apache JMeter позиціонується як open-source продукт для навантажувального тестування та вимірювання продуктивності; інструмент має детальну специфікацію елементів тест-плану та підтримує віддалене (розподілене) виконання для збільшення генерованого навантаження [7]. З іншого боку, Grafana k6 орієнтований на розробницькі команди: сценарії описуються кодом (JavaScript), а механізм thresholds надає нативний «pass/fail» на основі заданих очікувань продуктивності [8].

Автоматизація продуктивнісного тестування в CI/CD зазвичай потребує ще одного класу технологій – спостережуваності та вимірювання стану системи під тестом. OpenTelemetry визначається як vendor-neutral, open-source підхід до збору та експорту telemetry-сигналів (traces, metrics, logs), що дозволяє корелювати поведінку запитів із ресурсними показниками та подіями [9]. Prometheus, як open-source система моніторингу, забезпечує збір та запити до часових

рядів (time series), включно з pull-моделлю збору [10]. Візуалізація та аналітика на практиці часто реалізуються через Grafana-дашборди як набір панелей, що відображають дані з джерел та дають «at-a-glance» картину стану системи [11].

Окремий сучасний напрям – автоматичне виявлення регресій продуктивності (performance regression testing). В оглядових роботах останніх років підкреслюється, що це активна область досліджень із багатьма викликами: варіативність середовищ, статистична надійність, вибір метрик та критеріїв «значущої» деградації [5]. Практичні рекомендації індустріальних документів для k6 підкреслюють роль baseline-порівнянь як «ядра» методології автоматизованого performance-тестування [12].

Постановка задачі дослідження. Метою статті є розробка та обґрунтування концепції автоматизованої системи тестування продуктивності програмних систем, яка:

1) формалізує вимоги до продуктивності через характеристику performance efficiency ISO/IEC 25010 (time behaviour/resource utilization/capacity) [1];

2) застосовує стандартизований підхід до вимірювань (measurement process) і вибору метрик [3];

3) інтегрується в CI/CD як quality gate із автоматичним рішенням pass/fail за порогами [8];

4) забезпечує відтворюваність тестового середовища та спостережуваність (correlation traces/metrics/logs) для пояснюваного аналізу причин деградації [9].

Для досягнення мети потрібно вирішити такі задачі:

1) визначити функціональні модулі системи та їх взаємодію в межах тестового циклу;

2) сформувати мінімально достатній набір метрик (SLI) та правила агрегування (наприклад, середнє/перцентилі) для часових характеристик і ресурсних параметрів;

3) описати механізм «baseline → порівняння → рішення» для виявлення регресій;

4) задати правила інтеграції у CI/CD (тригери запуску, клас тестів за вагою: smoke/peak/soak, планування ресурсів).

Результати дослідження

1. Архітектура та функціональні компоненти системи. Концептуальна структура автоматизованої системи тестування продуктивності складається з модулів оркестрації в CI/CD, генерації навантаження, профілів тестів, збору телеметрії, аналізу/quality gate та зберігання артефактів і baseline (рис. 1).

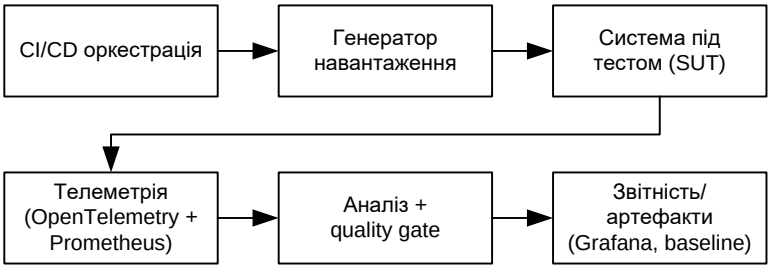


Рис. 1. Узагальнена архітектура автоматизованої системи тестування продуктивності

Модуль оркестрації (CI/CD-інтеграція). Точка входу – джоба конвеєра CI/CD, яка викликає:

1) розгортання тестового стенда;

2) запуск навантажувального тесту;

3) збір метрик/логів/трейсів;

4) оцінку порогів;

5) публікацію артефактів (звіт, дашборд, порівняння з baseline).

Потреба включати performance-тести до CI/CD як частину «глибини автоматизованого тестування» прямо відзначається у практичних рекомендаціях щодо CI/CD-best practices [6].

Модуль генерації навантаження. Як реалізаційні варіанти доцільні:

– Apache JMeter як інструмент, призначений для load-testing і вимірювання продуктивності; його модель test plan та наявність remote/distributed режимів дозволяє масштабувати генерацію віртуальних користувачів;

– Grafana k6 як інструмент, орієнтований на автоматизацію, з можливістю задавати thresholds як правила pass/fail та запускати сценарії в CI.

Вибір конкретного генератора навантаження можна трактувати як параметр системи: важливішою є стандартизація контрактів (вхідні параметри, артефакти, метрики) та можливість запуску з командного рядка.

Модуль профілів навантаження (test profiles). На основі ISTQB-понять «load profile» і «load management» профіль навантаження має бути артефактом, який:

а) формалізує кількість віртуальних користувачів/транзакцій;

б) визначає період/динаміку (ramp-up/steady-state/ramp-down);

в) зв’язує сценарії з очікуваними умовами експлуатації.

Для покриття різних ризиків доцільно підтримати щонайменше «peak/spike/soak» класи тестів (як поширені типи в load-testing практиці).

Модуль спостережуваності та збору телеметрії. Система повинна використовувати узгоджений канал збору:

– OpenTelemetry як стандартний, vendor-neutral спосіб отримувати traces/metrics/logs та корелювати їх у межах одного контексту запиту;

– Prometheus як механізм зберігання та запитів до метрик у часових рядах (інструментуван-

ня/скрейпінг/PromQL), що потрібні для аналітики під час тесту;

– Grafana як шар візуалізації та аналітики (дашборди як набір панелей).

Модуль відтворюваності середовища. Для зменшення шуму вимірювань середовище бажано описувати як інфраструктуру-як-код (IaC). Terraform як IaC-інструмент дозволяє описувати бажаний стан інфраструктури декларативно та керувати життєвим циклом ресурсів, що підтримує повторюваність стендів для тестів [13].

Модуль зберігання артефактів і baseline. Практика порівняння з baseline як «ядро» автоматизованого performance-тестування прямо підкреслюється у документації Grafana Cloud k6: baseline повинен бути репрезентативним, але не надмірно «важким», щоб залишатися стабільним референтом. Артефактами зберігання є: конфігурація тесту, raw-метрики, агрегати (перцентилі p50/p95/p99), звіти й рішення pass/fail.

Сукупно запропонована архітектура забезпечує:

1) автоматичне виконання;

2) вимірюваність;

3) відтворюваність;

4) пояснюваність результату через спостережуваність.

2. Методика вимірювання, метрики та критерій приймання. ISO/IEC/IEEE 15939 описує процес вимірювання як набір дій для визначення та застосування вимірювань, що закривають конкретні інформаційні потреби. Для продуктивності доцільно вважати інформаційними потребами: «чи витримує система очікуване навантаження», «який запас місткості», «чи з'явилася деградація після зміни», «де вузьке місце».

Модель ISO/IEC 25010 задає три підхарактеристики performance efficiency:

– time behaviour (час відповіді, час обробки, throughput);

– resource utilization (використання ресурсів);

– capacity (граничні можливості за параметрами). Тому система вимірювань має включати метрики часу/пропускної здатності та метрики ресурсів на рівні хоста/контейнера/процесу.

3. Базові метрики тестування продуктивності. Спираючись на ISO/IEC 25023, який задає «quality measures» та приклади інтерпретації (conformance і time series аналіз), пропонується мінімальний набір метрик для автоматизованої системи:

1) час відповіді та його розподіл. Середній час відповіді може бути визначений як

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i,$$

але для користувацького досвіду важливі також перцентилі (p95/p99), оскільки середнє може приховувати повільні запити в «хвості» розподілу. У CI/CD-контексті практичною є стратегія контролю p95/p99 як критеріїв SLO/thresholds;

2) пропускна здатність (throughput) і інтенсивність. У продуктивнісній моделі ISO/IEC 25010 time behavior включає throughput як частину вимог до часової поведінки системи. Для сервісів це може бути RPS (requests per second), TPS (transactions per second) або обсяг оброблених повідомлень за інтервал;

3) похибки та помилки (error rate). У межах автоматизованих порогів k6 thresholds прямо підтримує умови за error rate та latency-показниками, що дозволяє перетворити нефункціональні вимоги на формальні критерії «fail build»;

4) використання ресурсів. Prometheus визначає збір і запит до метрик як основу для dashboarding/alerting; для ресурсів це, як правило, CPU utilization, memory usage, network I/O, disk I/O;

5) місткість (capacity) і точки насичення. Capacity у ISO/IEC 25010 описує ступінь, з яким максимальні межі параметра системи задовольняють вимоги. На практиці це виражається як «максимальна кількість паралельних користувачів/запитів» при збереженні заданих SLO по latency/error rate.

4. Автоматичні критерії pass/fail та baseline-порівняння. Критерій k6 визначає thresholds як критерії pass/fail для тестових метрик: якщо умови не виконані, тест завершується зі статусом failure, що природно вписується в CI/CD як quality gate.

Для автоматизованого виявлення регресій пропонується комбінований підхід:

1) абсолютні пороги (SLO-пороги). Наприклад, p95 latency ≤ X мс, error rate ≤ Y%, throughput ≥ Z RPS (самі числа мають походити з вимог/очікувань або з емпіричного baseline);

2) відносні пороги до baseline. Порівняння «поточна збірка» vs «еталонна збірка» як методологічно ключовий елемент automated performance testing; documentation Grafana Cloud прямо описує baseline як контрольний тест для пошуку відмінностей;

3) трендовий контроль (time series). ISO/IEC 25023 наводить ідею time series порівнянь для того, щоб бачити зміни метрик у часі (наприклад, як змінюється mean response time протягом дня).

Критерії виявлення регресій при порівнянні з baseline наведено в табл. 1.

Таблиця 1 – **Критерії виявлення регресій при порівнянні з baseline**

Метрика	Критерій регресії	Примітка
Latency p95	Δ > +10%	Медіана із 3-х прогонів
Latency p99	Δ > +15%	Чутлива до «хвоста»
Throughput	Δ < –5%	Падіння пропускної здатності
Error rate	new > base + 0,2% або SLO	relative+hard
CPU/Memory	Δ > +10% (p95/max)	Неефективність / витік

Оскільки performance regression testing є чутливим до шуму й варіативності середовища, у сучасних оглядах підкреслюється необхідність уважного вибору методів порівняння та критеріїв значущості регресії.

Саме тому модуль відтворюваності середовища (IaC) та стандартизований збір метрик (Prometheus/OpenTelemetry) є не «додатком», а центральною вимогою системи.

Пороги для quality gate наведено в табл. 2.

Таблиця 2 – **Приклад порогів для quality gate (калібруються під систему та стенд)**

Метрика	Поріг/правило	Коментар
Latency p95	p95 ≤ 300 мс	для критичних endpoint-ів
Latency p99	p99 ≤ 800 мс	контроль «tail latency»
Error rate	Errors ≤ 0,5%	включно з timeout
Throughput	RPS ≥ 200 або Δ ≥ –5%	hard+relative
CPU avg	CPU ≤ 70% (warn)	gate після калібрування
Memory max	RAM ≤ 75% (warn)	запас для піків

5. Практичний сценарій виконання тестового циклу

Спрощену схему виконання циклу тестування продуктивності в CI/CD наведено на рис. 2.

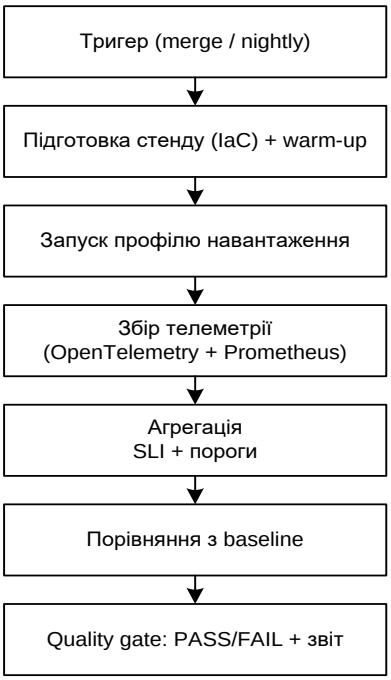


Рис. 2. Спрощена схема виконання циклу тестування продуктивності в CI/CD

Узагальнений цикл автоматизованого тесту продуктивності в CI/CD виглядає так:

1) розгортання тестового середовища (IaC) і прогрів (warm-up) для стабілізації кешів/компіляції;

2) запуск навантажувального профілю (load profile) відповідно до ISTQB-визначення;

3) генерація навантаження та паралельний збір телеметрії;

4) агрегація метрик та оцінка порогів thresholds (k6) або зовнішній «gating-модуль»;

5) порівняння з baseline (якщо увімкнено режим performance regression), оновлення сховища baseline за контрольованим правилом (наприклад, лише після релізного тегу);

6) публікація результатів: артефакти CI (лог, summary, JSON), посилання на дашборд Grafana, короткий висновок «pass/fail» і (за потреби) діагностичний зріз «який компонент/endpoint погіршився».

Артефакти та вихідні результати автоматизованого тесту продуктивності наведено в табл. 3.

Висновки

У статті обґрунтовано актуальність інтеграції тестування продуктивності в CI/CD як механізму раннього виявлення деградацій і формування керованих (threshold-based) quality gates. Показано, що стандартизоване розуміння «продуктивності» доцільно будувати на ISO/IEC 25010 (performance efficiency: time behavior/resource utilization/capacity), а підхід до вимірювань – узгоджувати з ISO/IEC/IEEE 15939.

Таблиця 3 – **Артефакти та вихідні результати автоматизованого тесту продуктивності**

Артефакт	Формат/приклад	Призначення
Конфігурація тесту	test plan, профіль	Відтворення запуску
Raw-результати	JSON/JTL + логи	Повторний аналіз
Агрегований звіт	HTML/Markdown	Висновок та ключові SLI
Telemetry snapshot	Grafana link/PNG	Діагностика причин
Baseline dataset	Версіонований набір	Порівняння збірок
Verdict у CI	PASS/FAIL + причина	Блокування регресій

Запропоновано модульну архітектуру автоматизованої системи, де навантажувальні інструменти (JMeter або k6) поєднуються з observability-стеком (OpenTelemetry + Prometheus + Grafana), що підвищує пояснюваність результатів і практичну керованість деградаціями.

Перспективи подальших досліджень у цьому напрямку включають:

1) статистично обґрунтовані методи виявлення регресій (з урахуванням варіативності середовища);

2) автоматичне формування «реалістичних» профілів навантаження на основі production-телеметрії;

3) розширення метрик до рівня бізнес-транзакцій і user-journeys;

4) оцінювання ефективності системи на реальних кейсах та формування практичних рекомендацій для команд розробки.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. ISO/IEC 25010:2011. Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE). DOI: https://doi.org/10.3403/30215101. URL: https://www.iso.org/standard/35733.html.
2. ISO/IEC 25023:2016. Systems and software engineering. Systems and software Quality Requirements and Evaluation (SQuaRE) URL: https://www.iso.org/standard/35747.html.
3. ISO/IEC/IEEE 15939:2017. Systems and software engineering URL: https://www.iso.org/standard/71197.html.
4. ISTQB. Standard Glossary of Terms Used in Software Testing (Version 4.0). URL: https://glossary.istqb.org/.
5. Dos Santos L.B.R., Trubiani C., Pinciroli C., et al. Performance regression testing initiatives: a systematic mapping study. Information and Software Technology. 2024. DOI: https://doi.org/10.1016/j.infsof.2024.107641.
6. Gatling. Integrate performance testing into your CI/CD pipeline. URL: https://gatling.io/blog/performance-testing-ci-cd.
7. Matam S., Jain J. Pro Apache JMeter: Web Application Performance Testing. Apress, 2017. DOI: https://doi.org/10.1007/978-1-4842-2961-3.
8. Grafana Labs. k6 Documentation (running in CI; thresholds). URL: https://grafana.com/docs/k6/latest/get-started/running-k6/; https://grafana.com/docs/k6/latest/using-k6/thresholds/.
9. Blanco D. G. Practical OpenTelemetry: Adopting Open Observability Standards Across Your Organization. Apress, 2023. DOI: https://doi.org/10.1007/978-1-4842-9075-0.
10. Prometheus Authors. Prometheus Documentation. URL: https://prometheus.io/docs/introduction/overview/.
11. Sanches J., Pereira P. R. Network and Systems Monitoring with Prometheus and Grafana. CISTI 2025, Lecture Notes in Networks and Systems, vol. 1716. Springer, 2026. DOI: https://doi.org/10.1007/978-3-032-10929-3_32.
12. Grafana Labs. Grafana Cloud k6: Compare tests (test comparison). URL: https://grafana.com/docs/grafana-cloud/testing/k6/analyze-results/test-comparison/.
13. Kasarla N. K. Implementing Infrastructure as Code (IaC) with Terraform for Scalable Cloud Deployments. Journal of Information Systems Engineering and Management. 2025. DOI: https://doi.org/10.52783/jisem.v10i60s.13257.
14. Humble J., Farley D. Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation. Addison-Wesley, 2010. DOI: https://doi.org/10.5555/1869904.

Received (Надійшла) 19.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Запорожець Олег Васильович – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleg Zaporozhets – PhD, Associate Professor, Associate Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: oleg.zaporozhets@nure.ua; ORCID Author ID: http://orcid.org/0000-0002-7831-8479;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=15728942500

Калашников Павло Андрійович – студент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Pavlo Kalashnykov – student, Department of Electronic Computers, Kharkiv National University of Radio Electronics, Ukraine;
e-mail: pavlo.kalashnykov@nure.ua; ORCID Author ID: https://orcid.org/0009-0000-2054-0037.

Automated performance testing system for software applications: architecture, metrics, and ci/cd integration

O. Zaporozhets, P. Kalashnykov

Abstract. Relevance. Performance is one of the key characteristics of software system quality, significantly affecting user experience, service reliability, and infrastructure costs. Considering the ISO/IEC 25010 model (performance efficiency: time behavior, resource utilization, capacity) and DevOps CI/CD practices, it is important to automate regular and reproducible load tests with a formal pass/fail decision and explainable regression diagnostics. **Object of research**. An automated system for testing the performance of software systems in the CI/CD pipeline, combining load tests and observability (metrics/logs/traces). **Purpose of research.** To develop and justify the concept of such a system, define its architecture, modules, test cycle execution order, and a minimally sufficient set of metrics and quality gate threshold criteria based on engineering SLOs. **Research results.** A modular architecture is proposed, the test cycle execution procedure and the approach to forming metrics and baseline comparisons (ISO/IEC 25023, ISTQB terminology) for automatic detection of performance regressions are described. **Conclusions.** Integrating performance testing into CI/CD enables early detection of degradations and reduces the risk of them entering the production environment, while combining load with observability increases the explainability of the causes of degradation.

Keywords: performance, performance testing, load testing, CI/CD, metrics, observability, thresholds, performance regression.

УДК 004.9 **doi: 10.26906/SUNZ.2026.2.095**

В. Г. Знайдюк, В. Б. Тухтаров

Харківський національний університет радіоелектроніки, Харків, Україна

АНСАМБЛЕВА МОДЕЛЬ ПРОГНОЗУВАННЯ ВІДМОВ ЗАВДАНЬ У ХМАРНИХ ОБЧИСЛЕННЯХ

Анотація. Актуальність. Хмарні обчислення є ключовим елементом сучасної IT-інфраструктури, однак проблема відмов завдань негативно впливає на якість обслуговування та ефективність використання ресурсів. Зростання складності хмарних систем та обсягів даних зумовлює необхідність застосування інтелектуальних методів прогнозування відмов, що дозволяють переходити від реактивних до проактивних і стійких підходів управління ресурсами. **Об'єкт дослідження:** процеси прогнозування відмов завдань у хмарних обчислювальних системах. **Мета статті:** розробка ансамблевої моделі прогнозування відмов завдань у хмарних обчисленнях на основі поєднання методів машинного навчання. **Результати дослідження.** У статті запропоновано ансамблеву модель, що базується на використанні методу стекінгу та поєднує алгоритми K-найближчих сусідів і штучної нейронної мережі з мета-моделлю на основі логістичної регресії. Проведено попередню обробку та аналіз даних набору Google Cluster Trace, виконано інженерію ознак і побудовано прогностичну модель. Експериментальні результати показали, що запропонована ансамблева модель забезпечує підвищення точності прогнозування та покращення показників F1-міри, прецизійності та повноти порівняно з окремими моделями. Встановлено, що використання ансамблевого підходу дозволяє зменшити ефект перенавчання та підвищити надійність прогнозів. **Висновки.** Запропонована модель є ефективним інструментом для прогнозування відмов завдань у хмарних системах та може бути використана для оптимізації планування ресурсів і підвищення відмовостійкості. Використання ансамблевого підходу сприяє зниженню витрат ресурсів і підтримує концепцію «зелених» обчислень. **Сфера використання отриманих результатів:** системи планування завдань, управління ресурсами та підвищення відмовостійкості у хмарних обчисленнях і центрах обробки даних.

Ключові слова: хмарні обчислення; прогнозування відмов; ансамблева модель; машинне навчання; штучна нейронна мережа; KNN; стекінг; логістична регресія; відмовостійкість; планування ресурсів.

Вступ

Постановка проблеми. Хмарні технології стали критичним елементом у сучасній екостістемі та довели свою здатність сприяти зростанню різних секторів. Хмарні обчислення передбачають дослідження та вдосконалення алгоритмів для оптимізації ефективності різних аспектів, включаючи розподіл ресурсів, балансування навантаження та надійність. Крім того, вони спрямовані на покращення якості обслуговування шляхом скорочення середнього часу між відмовами системи. Наразі відбувається зсув у бік впровадження автоматизованого підходу, який має на меті мінімізувати випадки людських помилок та усунути надлишкові завдання. Автоматизація компонентів прийняття рішень у хмарі досягається за допомогою методів машинного навчання (МН) та глибокого навчання (ГН). Поява цих методів породила нову область досліджень, яка називається інтелектуальними хмарними обчисленнями. Це дозволяє зосередитися на вдосконаленні хмарної інфраструктури шляхом впровадження різноманітних інтелектуальних методів. Підвищення відмовостійкості є фундаментальним аспектом хмарних обчислень.

Існують різні категорії відмов у хмарі, які можуть створити каскадну подію відмов. Для пом'якшення цих відмов існуючі системи використовують різні заходи для забезпечення безперервності обслуговування в разі виникнення вузького місця. Впровадження управління відмовами в хмарному середовищі слугує для підвищення стійкості системи та створення відчуття надійності для клієнтів під час використання ними хмарних сервіс-провайдерів, тим самим забезпечуючи відмовостійкість. Існують три окремі класифікації методів відмовостійкості. Реактивний метод широко використовується хмарними провайдерами як основний підхід для швидкого виділення ресурсів у відповідь на збої в обслуговуванні. Використання цієї конкретної методології вважається придатним для досягнення відмовостійкості, незважаючи на те, що вона передбачає значну кількість додаткових ресурсів. Тривалість накладних витрат може становити від мінімум 5 секунд до максимум 15 секунд, що може порушити операційну діяльність бізнесу клієнта. Термін «накладні витрати» стосується часової вимоги для запуску нового ресурсу або перезавантаження сервісу.

Крім того, існують проблеми, пов'язані з повторюваними завданнями, зокрема процес визначення порогового значення. Ця процедура виконується вручну і тому несе ризик людської помилки. Спостерігається зростаюча тенденція до впровадження автоматизованих систем із самоусвідомленням, зокрема у сфері розуміння відмов, як засобу зменшення кількості повторюваних завдань. Завдяки використанню моніторингу журналів відмов стає можливим передбачити ймовірність виникнення відмов і згодом ініціювати процес розподілу ресурсів за допомогою вищезазначених реактивних методологій. Використання проактивних заходів з метою виявлення відоме як проактивні методи. Було розроблено різні методології, такі як самовідновлення, випереджувальна міграція, моніторинг, S-Guard та програмне оновлення. Були проведені обширі дослідження щодо явища самовідновлення та отримання випереджувальної міграції з використанням методологій машинного навчання.

Багато досліджень використовували різноманітні алгоритми МН для прогнозування найбільш відповідного порогового значення для випереджувальних міграцій, а також для моніторингу ресурсів за

допомогою сигналів «heartbeat». Займаючись прогнозуванням надійності, ми можемо робити обґрунтовані висновки щодо кількості порогових значень. Наразі існує новий підхід, яка поєднує реактивну та проактивну методології, зазвичай відома як стійкий підхід. Такий підхід є був вдосконалений для ефективного реагування на відмови, використовуючи наявний набір даних про відмови як основу.

Наше дослідження має на меті розробити модель, яка включає здатність прогнозувати відмови через передбачення відмови завдань у хмарних обчисленнях. Цього можна досягти шляхом визначення таких особливостей, як запитуваний ресурс і стан події завдання. Згодом модель може бути навчена розпізнавати шаблон відмови завдання за допомогою алгоритму МН. У цьому дослідженні буде використано три різні моделі для виконання завдання класифікації: K-найближчих сусідів (KNN), штучна нейронна мережа (ШНМ) та ансамблева модель, яка поєднує як KNN, так і ШНМ за допомогою методу стекінгу. Метод стекінгу об'єднує результати за допомогою логістичної регресії. У подальших застосунках цю модель можна використовувати для виявлення випадків відмови завдань під час планування завдань шляхом призначення пріоритету завдання та подальшого спрямування його в окрему чергу. Однією з переваг цього підходу є те, що виділення ресурсів для завдання не відбудеться, доки планувальник не обробить чергу. Цей підхід потенційно може зменшити виділення ресурсів, що призведе до зниження вартості хмарних послуг і сприятиме впровадженню практик «зелених» обчислень.

Аналіз останніх досліджень і публікацій. Передбачення відмов було центральним напрямком досліджень протягом багатьох років. Було використано різні моделі для покращення виявлення проблем у хмарних обчисленнях і пом'якшення відмов за допомогою проактивних та реактивних підходів. Основною метою цього проєкту є використання методів машинного навчання для виявлення випадків хмарних несправностей. Згідно з [1], різні проблеми, пов'язані з хмарними обчисленнями, можуть бути вирішені шляхом впровадження коригувальних стратегій, таких як контрольні точки, проактивна міграція, повторні спроби, перепланування завдань та програмне омолодження. Як реактивний, так і проактивний підходи виступають за впровадження всіх цих стратегій пом'якшення. В роботі було використано кілька моделей для прогнозування відмов, що охоплюють як програмні, так і апаратні збої. Ця стаття слугує інструментом для класифікації проблем і надання цінного аналізу щодо потенційних наслідків системних збоїв, зосереджуючи увагу на конкретних проблемах, таких як відмова додатків, стає можливим підвищити якість обслуговування (QoS).

В роботі [2] досліджено ефективне впровадження МН у комплексний та надійний спосіб. Застосування МН та штучного інтелекту продемонструвало свою ефективність у прогнозуванні хмарних збоїв. Було наведено набір даних і обговорено потенціал МН для покращення досліджень відмовостійкості. В наступному досліджені [3] ШНМ була використана для прогнозування потенційної відмови жорсткого диска на відповідному сервері. Було помічено, що комбінація ШНМ та технології самоконтролю, аналізу та звітування підвищує точність прогнозування відмови жорсткого диска. Відповідно модуль управління відмовостійкістю отримує можливість вживати проактивних заходів для запобігання виділенню віртуальних машин серверам, які демонструють потенційну вразливість до відмови. Це дослідження надає порівняльний аналіз алгоритмів ШНМ та KNN, висвітлюючи їх відповідну продуктивність з точки зору точності, яка, за повідомленнями, становить 90%. Практичне впровадження архітектури хмарних обчислень передбачає використання центрального контролера, який слугує основною сутністю, відповідальною за отримання запитів користувачів та їх подальший розподіл на фізичні машини. Друга функція стосується синхронізації кількох модулів, які відповідають за сприяння ефективному управлінню хмарною інфраструктурою. Архітектура реалізація, що пропонується, включає бази даних, які адмініструються модулями Hadoop і MapReduce, а також вторинний контролер, який контролює стан системи та надає сповіщення в разі будь-яких змін.

В роботі [4] представлено методологію, спрямовану на прогнозування відмови завдань із застосуванням різних підходів. Автор використав п'ять алгоритмів машинного навчання та оцінив їх відповідну продуктивність, оцінюючи їх точність.

Автор спробував вирішити проблему класифікації, використовуючи три різні категорії алгоритмів, а саме регресію, часові ряди та ансамбль. Алгоритм логістичної регресії є широко досліджуваним і визнаним методом регресії в статистичному аналізі. Алгоритм дерева рішень часто використовується як основний метод машинного навчання для задач класифікації, тоді як випадковий ліс класифікується як ансамблевий алгоритм. Автор також додав три окремі варіації моделей LSTM та ГН, кожна з яких відрізняється кількістю шарів. Модель ГН складається з трьох окремих підмоделей: одношарова довга короткочасна пам'ять з трьома шарами, двошарова LSTM з двома прихованими шарами та тришарова LSTM з трьома прихованими шарами. Для досягнення цієї мети алгоритм включає щільний шар для забезпечення формування єдиного значення для прогнозу. Крім того, процес навчання зупиняється, якщо не спостерігається покращення показника втрат на валідації після 10 епох. Висновок дослідження показав, що XGBoost продемонстрував вищу точність класифікації порівняно з іншими моделями, тоді як моделі випадкового лісу та дерева рішень виявилися більш придатними для прогнозування на рівні завдань.

В попередній роботі продемонстровано впровадження схеми планування для енергоощадної відмовостійкості. Ця схема використовує глибокі нейронні мережі для прогнозування відмов і планування завдань у межах репліки для виконання. Під час початкової фази завдання піддається тестуванню для оцінки ймовірності зіткнення з відмовою. Отже, воно класифікується як схильне до відмови або не схильне до відмови. Використання цього конкретного

підходу призводить до зниження енергоспоживання, що, своєю чергою, гарантує збереження якості обслуговування (QoS). Крім того, було запропоновано використовувати градієнтний спуск як метод зменшення похибки прогнозування в рамках аналізу відмов. Автор дослідив вплив різних ресурсів на виникнення відмови завдання.

В якості головного відкриття в роботі [5], пропонується використання векторного контейнера для перепланування суперзавдання на відповідному хості. Цей підхід використовує прогностичні методи для оптимізації алгоритму планування шляхом присвоєння точних числових значень параметрам алгоритму

Наступна робота [6] зосереджена на прогнозуванні відмов з використанням технології LSTM. Однак важливо зазначити, що LSTM не здатна ефективно обробляти кілька входів. Тому це дослідження надає всебічний розгляд двонапрямленої LSTM (Bi-LSTM), яка інтегрує більшу кількість вхідних характеристик.

Метою цього дослідження є визначення основних характеристик, які необхідно враховувати при розробці нашої перспективної моделі. Як навчання, так і тестування моделі будуть включати набір даних Google Cluster. Результати показують, що алгоритм генерує вихідні дані як у прямому, так і у зворотному напрямках для регулювання ваг вхідних ознак, які є близькими або далекими. Крім того, оцінка проводиться шляхом порівняння з іншими зразковими моделями з урахуванням їхньої точності, F1-міри, прецизійності та повноти. Тим не менш, точність прогнозування може знижуватися, коли часові інтервали перевищують певний поріг. Поточний результат є наслідком ретельної оцінки компромісу між величиною часового інтервалу та точністю прогнозу. Результати дослідження показують, що Bi-LSTM продемонструвала прогностичну точність 90%, коли мінімальний часовий інтервал було встановлено на рівні 15 хвилин і дотримано вимог щодо розміру.

Авторами [7] представлено фреймворк планування завдань, який інтегрує усвідомлення відмов, дозволяючи прогнозувати статус завершення завдання в реальному часі та вживати відповідних коригувальних заходів. Існування цієї характеристики призвело до того, що значна частина клієнтів переносить свої прикладні завдання на хмарні платформи. Фреймворк демонструє помітну здатність захищати близько 40% завдань, які, за прогнозами, зазнають відмови, шляхом ефективного виконання коригувальних заходів. Як наслідок, досягається економія ресурсів кластера, таких як центральні процесори та оперативна пам'ять. Крім того, проблема вибору дії формалізується в цьому дослідженні за допомогою моделі цілочисельного лінійного програмування. У сфері хмарних обчислень не рідкістю є відмови завдань в результаті різних факторів, включаючи, але не обмежуючись, дефекти програмного забезпечення, апаратні збої та неадекватний розподіл ресурсів. Наявність таких відмов може потенційно мати згубний вплив на QoS, що надається.

Дослідження [8] конкретно зосереджується на точці зору хмарних сервіс-провайдерів. На надійність хмарних застосунків можуть впливати різні фактори, включаючи характеристики завдань, конфігурації хмари та динамічні стани хмарної системи. В роботі використовується статистичний аналіз збоїв завдань, щоб виявити потенційні зв'язки між цими збоями та важливими обмеженнями планування, операціями вузлів та атрибутами користувачів у контексті хмарних обчислень. Дослідники пропонують низку потенційних стратегій для підвищення надійності хмарних додатків, як це пропонується їхніми емпіричними спостереженнями. Стратегії включають проактивне обслуговування вузлів та впровадження обмежень на частоту повторних надсилань завдань. Існує значний рівень зацікавленості в розумінні впливу планування завдань та обслуговування вузлів на виникнення збоїв завдань.

Впровадження складних алгоритмів МН призвело до помітного покращення в методології, причому спостерігається зростання тенденції використовувати ці методи для прогностичних цілей у сфері хмарних обчислень. Багатошарові персептрони (БП), різновид штучної нейронної мережі, продемонстрували значні перспективи в різних задачах прогнозування. Здатність точно відображати складні нелінійні зв'язки робить їх добре придатними для інтеграції в хмарні середовища, які за своєю суттю є динамічними та гетерогенними. В багатьох дослідженнях [9] підкреслюється важливість інженерії ознак для підвищення ефективності прогностичних моделей. У сфері хмарних систем включення конкретних характеристик, таких як системні журнали, показники використання ресурсів та історичні дані про збої завдань, відіграє життєво важливу роль у визначенні ефективності прогностичних моделей. В роботі [10] підкреслюється потенціал методологій МН, особливо нейронних мереж, таких як БП, для ефективного вирішення проблем, пов'язаних з прогнозуванням збоїв завдань у хмарних системах. Очікується, що зростаюча складність хмарних середовищ посилить важливість передових моделей машинного навчання.

В роботі [11] представлено аналіз комплексних трас робочого навантаження, зокрема тих, які були оприлюднені компанією Google. Вищезазначені дослідження показали, що значна частина часу в кластері була виділена на виконання завдань, які зрештою не досягли успішного завершення. Вищезазначені результати підкреслюють критичну необхідність глибокого розуміння механізмів та обґрунтувань, що лежать в основі завершення завдань у великомасштабних системах. Експоненціальне зростання обсягу даних у цих системах не супроводжувалося пропорційним покращенням надійності та безвідмовності. Питання надійності виходить за межі окремої системи. Проблема надійності завдань створює значні виклики як у традиційних системах високопродуктивних обчислень, які схильні до частих аварійних завершень додатків, так і в середовищах хмарних обчислень, які виконують різноманітні робочі навантаження на складних програмних стеках та гетерогенному обладнанні. В обох контекстах робочі навантаження демонстрували підвищену сприйнятливість до дефектів та помилок.

Основні напрями проаналізованих досліджень зосереджені на проактивній стратегії підвищення

відмовостійкості. Крім того, більшість робіт зосереджуються на традиційних моделях машинного навчання, хоча певні моделі, такі як Bi-LSTM, можуть включати модифікації своєї логіки. Однак важливо зазначити, що ці моделі схильні до перенавчання і можуть демонструвати субоптимальну продуктивність. Ще одним аспектом, який необхідно враховувати, є попередня обробка даних. Важливо визнати, що не всі дані можуть бути використані для розробки моделі прогнозування відмов. В зв'язку з цим необхідно ретельно вивчити призначення набору даних і внести відповідні зміни для вилучення релевантної інформації. Згодом набір даних повинен пройти попередню обробку, щоб полегшити розуміння моделлю та створити набір ознак, які підвищать точність моделі.

Вищенаведене зумовило мету даної роботи, а саме – розробку ансамблевої моделі для прогнозування відмовостійкості у хмарних обчисленнях. Запропоноване рішення поєднує в собі використання двох моделей KNN та ШНМ. А також використання методології стекінгу. Таким чином запропонована модель повинна значно підвищити точність прогнозу відмов в хмарному середовищі.

Основний матеріал

У нашому дослідженні було взято до уваги набір даних Google trace. Набір даних містить інформацію, отриману з серверів Google Borg, охоплюючи загалом вісім окремих кластерів Borg. Система надає дані щодо використання ЦП, запитуваного використання ЦП та розподілу пам'яті для кожного завдання. Крім того, вона надає інформацію щодо зв'язку між кожним завданням і відповідним йому процесом, а також ієрархічних відносин між майстер- та робочими вузлами, що використовуються у фреймворку MapReduce [12]. Набір даних було використано для сприяння аналізу розподілу ресурсів під час запуску завдання та у разі відмови, тим самим допомагаючи в розумінні основного процесу. Стійкий підхід передбачає використання штучного інтелекту для розуміння розподілу ресурсів до фактичного процесу розподілу. Цей підхід підвищує надійність системи, впроваджуючи методології для прогнозування потенційної відмови процесу до її виникнення. Для інтеграції цієї концепції необхідно розробити модель, яка може ефективно прогнозувати відмову завдання.

Попередні дослідження вивчали різні моделі, які успішно досягли цієї мети. Тим не менш, ці моделі стикаються з проблемою перенавчання, яка зазвичай пов'язана з обмеженістю навчальних даних через міркування конфіденційності. Тому наше дослідження зосередилося на двох ключових факторах для прогнозування відмови завдань. По-перше, увага на обмежену доступність даних, що зумовило необхідність використання загальнодоступних наборів даних, які надають надійні поведінкові атрибути, пов'язані з обробкою завдань. По-друге, підхід, який ефективно вирішує проблему перенавчання. У цьому конкретному завданні було вирішено використати ансамблеву методологію. Цей підхід підкреслює комбінований ефект результатів, отриманих від кількох моделей, для вирішення внутрішніх невідповідностей, присутніх у результатах. Ансамблевий підхід можна класифікувати на три окремі категорії, а саме бустинг, беггінг та стекінг. В попередніх дослідженнях було розглянуто бустинг, однак необхідно також розглянути методологію стекінгу.

Завдяки використанню стекінгу може бути досягнуте більш повне візуальне представлення продуктивності нашої моделі, що полегшить необхідні коригування. На рис. 1 представлено метод стекінгу, який складається з двох окремих етапів. На першому рівні вибирається кілька моделей, кожна з яких генерує індивідуальні прогнози. Метою використання другого рівня є об'єднання прогнозу, яке часто називають мета-модель навчання.

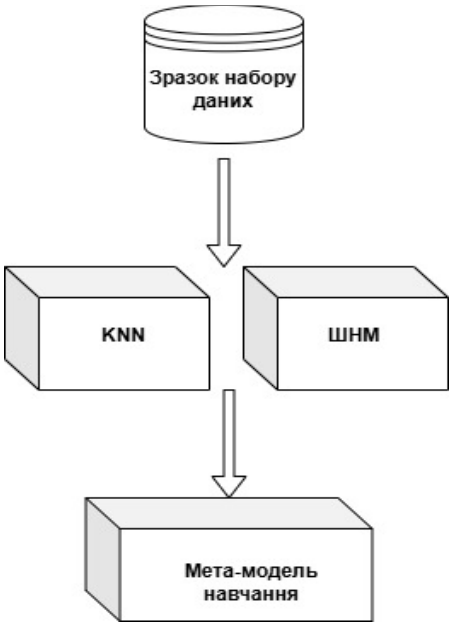


Рис. 1. Метод стекінгу

На початковому рівні нашої ансамблевої моделі відображено дві моделі: ШНМ та KNN. Модель ШНМ може бути представлено в різних варіантах, такі як нейронна мережа прямого поширення та рекурентна нейрона мережа. Ці варіанти відрізняються з точки зору напрямку, в якому дані зациклюються: мережа прямого поширення зациклює дані в одному напрямку, тоді як рекурентна мережа зациклює дані симетричним двох-направленим способом.

Хоча KNN вважається фундаментальною моделлю машинного навчання, здатною до розпізнавання патернів, важливо зазначити, що ця модель не покладається на жодні припущення щодо даних. Ця модель дозволяє класифікувати дані шляхом визначення їх близькості до заданої групи, тим самим відносячи їх до окремого кластеру. Для досягнення цілей в рамках цього дослідження обидві моделі були розглянуті, оскільки вони доповнюють одна одну з точки зору вирішення проблем «чорної скрині» та перенавчання. KNN також чутливий до нерелевантних ознак, тоді як ШНМ має сильну взаємозалежність ознак, оскільки вона створює зв'язки між ознаками для класифікації класів. Як уже зазначалося, ШНМ вважаються ефективними моделями для розпізнавання

патернів та класифікації завдань завдяки їхній здатності встановлювати зв'язки між вхідними параметрами. З іншого боку, KNN – це алгоритм класифікації, який не потребує жодних специфічних параметрів, але може демонструвати упереджений характер у своїх результатах класифікації. Отже, комбінуючи ці дві моделі, можна пом'якшити проблеми перенавчання та упередженості. Цього можна досягти шляхом навчання другого рівня ансамблевої моделі з використанням прогнозів, згенерованих KNN та ШНМ. В роботі було використано логістичну регресію на другому рівні для об'єднання результатів. Наступний етап передбачає аналіз даних, оскільки доступний набір даних має значний обсяг, що вимагає використання репрезентативної підмножини. Було проведено консолідацію полів даних, охоплюючи обмежену кількість інформації з кожної з восьми кластерів. Кожен кластер складається з чотирьох таблиць даних, в яких зберігається інформація, пов'язана з подіями машини, подіями колекції, подіями екземплярів та таблицею використання екземплярів. Набір даних, наданий для аналізу, складається приблизно з чотирьох мільйонів записів. Цей набір даних включає складове поле, яке охоплює різноманітну інформацію щодо запитуваної пам'яті ЦП для розподілу ресурсів, даних про використання ЦП, резервування спільних ресурсів для наборів розподілу та відносин між завданнями та їх батьківськими сутностями.

На рис. 2 представлена прогностична моделі, яка використовує ансамблевий підхід.

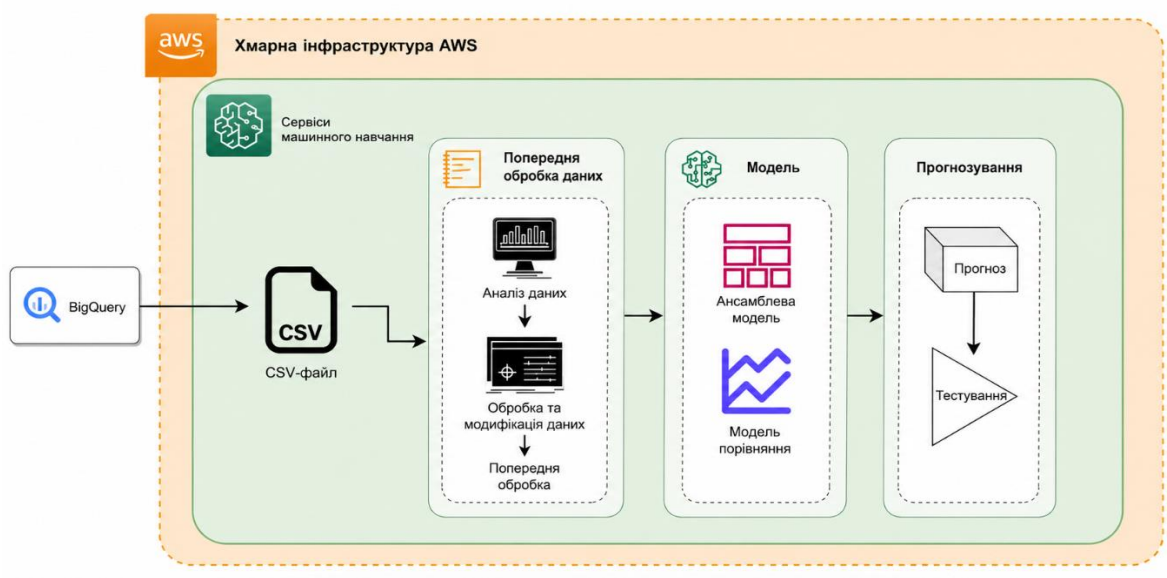


Рис. 2. Прогностична модель

Ансамблева модель є складною моделлю, яка потребує двох рівнів, що робить використання хмарних обчислень найбільш оптимальним підходом для досягнення швидших і точніших результатів. Спочатку необхідно встановити набір даних та провести процедури попередньої обробки. Дані демонструють часовий вимір і значну кореляцію, причому негативна кореляція позитивно пов'язана з відмовою завдання і слугує причинним фактором відмови завдання.

Під час етапу попередньої обробки даних певні поля, які вважаються непотрібними, імпутуються, а категоріальні та порядкові значення перетворюються на двійкові представлення. Це перетворення полегшує розуміння та виявлення патернів у даних. Під час етапу обробки створено проєкт, який дозволяє вибрати відповідні ознаки з попередньо оброблених даних після ретельного аналізу. Далі переходимо до вибору моделей, які потенційно можуть підвищити точність наших прогнозів. Враховуючи фактори перенавчання та явища «чорної скрині», було обрано три моделі. Перший рівень буде навчатися за допомогою двох моделей, тоді як другий рівень буде використано для об'єднання результатів двох моделей шляхом навчання іншої моделі для розуміння патерну. Дві моделі, KNN та ШНМ, складаються поверх моделі логістичної регресії, як показано на рисунку 1, і використовуючи цей підхід, досягаємо вищої точності прогнозування. У майбутньому цю модель можна запакувати та застосувати до будь-якого API, який можна додатково під'єднати до сервісу черги, що може діяти як мікросервіс і надавати ймовірність відмови завдання назад алгоритму планування завдань для прогнозування розподілу ресурсів.

Модель виконується з використанням хмарного сервісу AWS, який надає можливість виконувати складні моделі без необхідності турбуватися про придбання ресурсів. SageMaker надає доступ до обчислювально-оптимізованих екземплярів. Крім того, підвищення ефективності нашого експерименту, було за рахунок виброру підмножини даних, яку можна зібрати протягом обмеженого періоду часу. Крім того, було використано ядро PyTorch, яке належить до контейнерів глибокого навчання. Ядро охоплює всеосяжну колекцію основних фреймворків та бібліотек, які зручно зберігаються в образі Docker.

Для реалізації моделі відповідно до наших визначених критеріїв використовували Python для створення блокнота експерименту в AWS. Реалізація була розділена на чотири етапи, а саме: попередня обробка

даних, інженерія ознак, прогнозування та порівняння моделей.

Дані, отримані з Google Cluster Trace Version 3 2019, були перетворені на вибірку CSV. Дані піддалися аналізу для виявлення як нульових значень, так і числових значень, що дозволяє визначити поля, які містять нульові значення. Далі визначали всі поля, класифіковані як числові, а також ті поля, які демонструють категоріальні та порядкові характеристики. Після цього проведено оцінювання, чи містить наша навчальна ознака збалансований розподіл даних. Було помічено, що набір даних демонструє дисбаланс, що спонукало до необхідності використання методів вибірки даних. Процес вибірки даних дозволяє навчати модель, використовуючи однакову кількість істинно позитивних і хибно позитивних результатів. Використання цього підходу забезпечить підвищену точність прогнозування та розуміння показників ефективності, включаючи F1-міру, повноту та прецизійність. Наступний етап аналізу полягає у визначенні відповідних ознак для включення, чому сприяє візуалізація даних за допомогою графічних представлень, таких як коробчасті діаграми або гістограми.

Цей етап в основному підкреслює модифікацію ознак і вибір тих, які є найбільш придатними для ефективного проектування даних. Після проведення аналізу даних основна увага спрямовується на розробку ознак і трансформацію даних з метою підвищення продуктивності алгоритму машинного навчання. При спостереженні за даними стає очевидним, що запитуваний ресурс зберігається у форматі JSON. Отже, необхідно перетворити дані на кілька колонок, щоб встановити зв'язок для моделі. Ця функціональність згодом використовується для усунення кореляції, щоб визначити, чи є поле придатним для підвищення прогностичної продуктивності моделі. Під час процесу спостереження за даними необхідно масштабувати числові поля, щоб узгодити їхні величини з величинами інших полів. Необхідно виключити поле результату «fail» («відмова»), щоб ефективно навчати модель. Для нормалізації значень використовується стандартний скейлер (Standard Scaler). Для того щоб комп’ютер міг інтерпретувати кардинальні та порядкові значення як вхідні дані, їх необхідно перетворити на двійкові представлення; це досягається за допомогою функції кодувальника міток. Під час цього процесу перетворення стає очевидним, які значення є найбільш сумісними з моделлю. Після того як дані перетворено, приступимо до прогнозування значення. Цей процес передбачає розділення набору даних на дві підмножини: 80% виділяється для навчання, а решта 20% зберігається для тестування.

Реалізація прогнозування моделі на цьому етапі використовує бібліотеку Sklearn. Ініціалізація моделі передбачає виклик конструкторів відповідних моделей. Для проведення порівняльного аналізу необхідно ініціалізувати моделі KNN, ШНМ та логістичної регресії як перші кроки в процесі прогнозування.

Цей крок передбачає навчання моделі шляхом її підгонки на 80% наявних даних. У цій діяльності були використані всі три підходи. Наступний етап після підгонки моделі полягає в оцінці навченої моделі шляхом її тестування на підмножині, що становить 20% наявних даних. Результат цієї функції буде збережено у змінній, яка містить префікс test. Функція predict (прогнозувати) відповідає за проведення валідації нашої навчальної моделі та створення матриці згортки. Це дозволяє дослідити показники ефективності моделей. Далі використовується бібліотека Sklearn metrics для виклику статичної функції під назвою «classification report» з метою створення порівняльного звіту між підігнаним та протестованим значеннями. Кроки, пов'язані з реалізацією моделі ШНМ та KNN, виконуються відповідно. Однак при роботі з ансамблевою моделлю необхідно ініціалізувати об'єкт з параметрами оцінювача та кінцевого оцінювача, який також називається шаром мета-модель навчання. У процесі ініціалізації ансамблевої моделі необхідно спочатку ініціалізувати об'єкти моделей ШНМ та KNN. Ці ініціалізовані об'єкти моделей зберегаться в масиві та згодом передаються як аргументи параметру estimator об'єкта Stacking classifier. Крім того, об'єкт ініціалізованої логістичної регресії передається параметру final estimator об'єкта Stacking classifier. Після ініціалізації маємо дотримуватися вищезазначеного процесу підгонки, прогнозування та порівняння.

Заключним етапом є порівняння матриці ефективності кожної моделі. Ця матриця дає розуміння того, як модель працювала, аналізуючи F1-міру, повноту, прецизійність, точність, $RMSE$ та R^2.

Прогностична модель враховувала невелику кількість значущих ознак, які були проаналізовані під час етапу попередньої обробки даних. Модель демонструє підвищену продуктивність, коли присутня негативна кореляція, що дозволяє виявляти обмежену кількість завдань, які можуть завершитися невдачею, на основі попередніх запитів. Дані дають нам розуміння успішних та неуспішних завдань у всіх восьми кластерах. Кластер виділяє екземпляри для виконання завдання, і зі збільшенням кількості екземплярів обробляється більша частина завдання. На рис. 3, відображено, що кластери 3, 4 та 6 мають найбільшу кількість екземплярів під час виконання завдання, таким чином даючи уявлення про те, що кількість виконаних завдань більша в цих трьох кластерах.

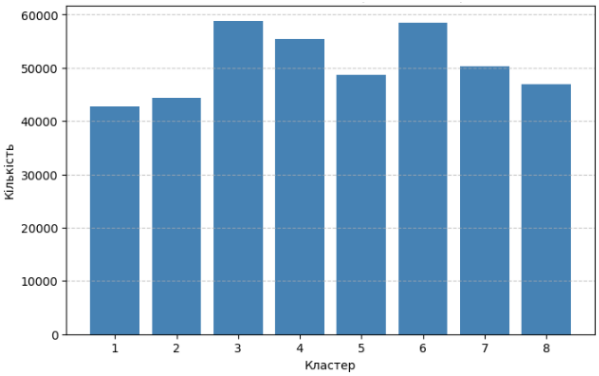


Рис. 3. Розподіл екземплярів за кластерами

На рис. 4 представлено середнє призначення пам’яті по кластерам. Розуміння розподілу пам'яті в

усіх кластерах обмежується базовим рівнем. Це спостереження передбачає, що, незважаючи на відносно низьку частоту виникнення подій у кластері 8, значний обсяг пам'яті виділяється ресурсу.

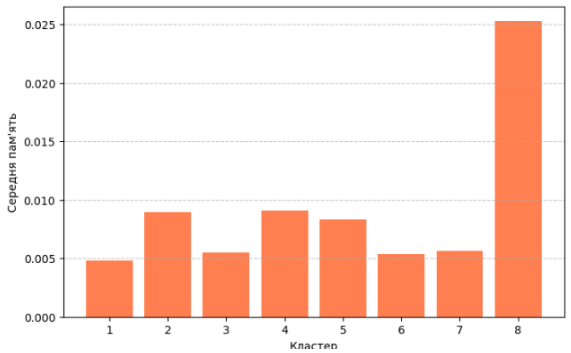


Рис. 4. Середнє призначення пам’яті по кластерах

Це відхилення від норми в поведінці може вказувати на те, що завдання потребує додаткової пам'яті для свого опрацювання, або що завдання зазнає невдач, що призводить до повторних спроб і, згодом, до більшого споживання пам'яті порівняно з іншими кластерами. Дослідження розподілу пам'яті є ключовим фактором у нашому передбаченні відмови завдання. Існує обернена кореляція між виділенням пам'яті та виникненням аномальної поведінки в кластері. При ретельному вивченні набору даних стає очевидним, що більший обсяг пам'яті зазвичай виділяється термінованим ресурсам.

Навпаки, більшість невдалих завдань демонструють вищий рівень споживання введення/виведення (I/O), що вказує на переважання збоїв запису на диск. Крім того, можна зробити висновок, що основною причиною відмов завдань у кластері є відмова ресурсів. Це спостереження підкреслює важливість включення використання та розподілу ресурсів як ключових факторів у розробці нашої прогностичної моделі.

Ці результати мали важливе значення в процесі вибору ознак. Було визначино наявність запитуваних ресурсів, які надають інформацію про необхідну пам'ять для завдання, а також тривалість кожного завдання. Ці спостереження реєструвалися з інтервалом у 5 хвилин, що дозволяє аналізувати час, необхідний для завершення завдання, та подальший час, необхідний для повторних спроб завдання. Вищезазначений фактор відіграє ключову роль у розумінні виникнення відмови завдання та сприятиме передбаченню відмови завдання. Було проведено порівняльне дослідження між кількома моделями, щоб показати найкращу модель для прогнозування відмови завдань.

Експериментальні дослідження також надають розуміння того, яка модель є найкращою для такого типу даних. Оскільки зосереджуємося на класифікації завдання між успіхом і невдачею, під час експеременті використали результати двох ансамблевих моделей і двох моделей машинного навчання, щоб забезпечити порівняльне дослідження між обома підходами. Також було враховано такі фактори, як F1-міра, точність та повнота, щоб глибше зрозуміти, яка модель найкраще підходить для таких прогнозів. Розділили дані на 80% навчальних даних і 20% тестових даних для валідації прогнозу. Під час експериментів вимірюємо продуктивність моделі. Давайте визначимо матрицю продуктивності нижче:

На рис. 5 представлено ROC криві всіх 3 моделей.

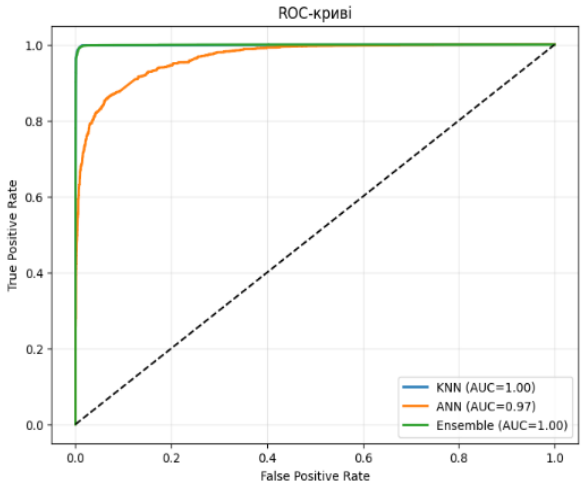


Рис. 5. ROC криві моделей
(KNN, ШНМ, ансамблева модель)

На рис. 6-8 представлені матриці помилок для наших трьох моделей KNN, ШНМ, ансамблева модель.

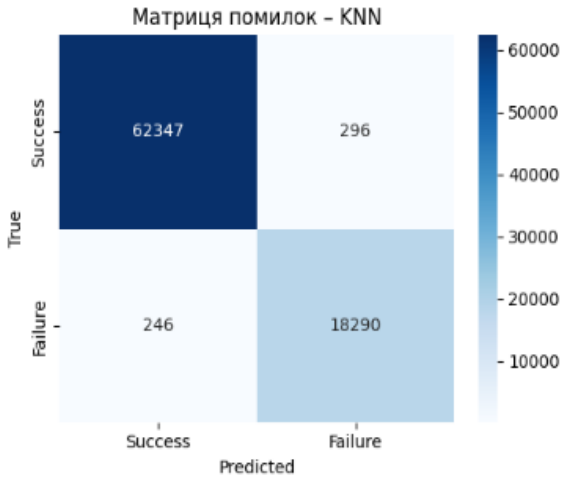


Рис. 6. Матриця помилок KNN

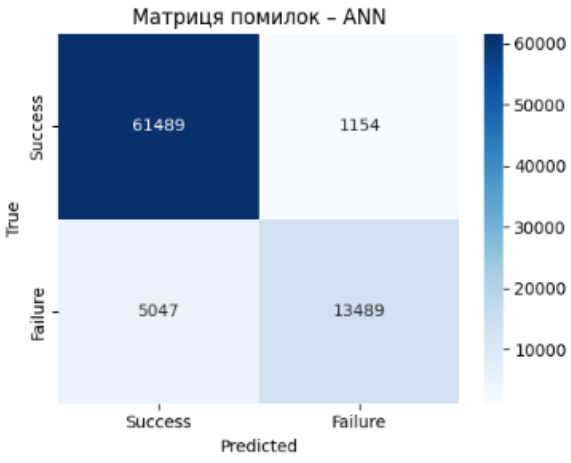


Рис.7. Матриця помилок ШНМ

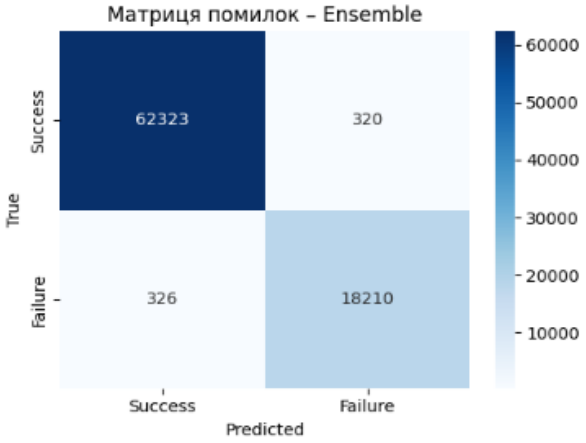


Рис. 8. Матриця помилок ансамблева модель

Ансамблева модель, використана під час експериментів, використовує підхід стекінгу для вирішення обмежень як KNN, так і ШНМ. Складність цієї моделі робить її вигідною порівняно з альтернативними моделями. Однак складність моделі вимагає значних витрат часу та обчислювальних ресурсів як для навчання, так і для тестування. Ще один аспект, який заслуговує на увагу при спостереженні, це використання складних моделей, таких як ШНМ, які складаються з кількох шарів.

В табл. 1 представлено порівняльні результати експериментів з різними моделями.

Аналіз продуктивності кожної моделі дозволяє здійснити всебічну оцінку різних факторів, що діють у кожній моделі. Для використання будь-якої з моделей необхідно мати глибоке розуміння даних.

Таблиця 1 – **Порівняльна таблиця**

	Accuracy	**Precision**	**Recall**	**F1-**	**RMSE**	R^2
KNN	0.9913	0.9841	0.9867	0.9854	0.0706	0.9717
ШНМ	0.9236	0.9212	0.7277	0.8131	0.2426	0.6660
Ансамбль	0.9920	0.9927	0.9924	0.9926	0.0789	0.9746

Продуктивність моделі визначається даними, і процес вибору ознак суттєво впливає на прогнозну здатність моделі. Використовуючи показники продуктивності, отримали подальшу інформацію про продуктивність кожної моделі. Можна помітити досить високу точність, F1-міру та повноту. Це означає, що модель є дійсно оптимальною для класифікації успіху чи невдачі завдання. Модель ШНМ схильна до перенавчання через її надмірно високу точність, що вказує на те, що модель має здатність запам'ятовувати патерни, а не дійсно навчатися їм. Алгоритм KNN також схильний до тієї самої проблеми. Однак ансамблева модель включає шар мета-навчання для агрегування результатів, що значно знижує ймовірність перенавчання. Крім того, проведено навчання моделі, яка налаштована на Google Cloud Trace та оптимізована для покращення продуктивності кластерів Google. Ансамблева модель демонструє надійні результати, однак вона має вищий рівень складності через об'єднання двох алгоритмів машинного навчання. Крім того, було додано нейронну мережу, що ще більше сприяє складному характеру ансамблевої моделі, що призводить до збільшення потреби в ресурсах порівняно з альтернативними моделями. Окрім цих проблем, ансамблева модель страждає від проблем конфіденційності через відсутність прозорості.

Висновки

На основі вищезазначених експериментів можна зробити висновок, що використання ансамблевої моделі дає надійну модель прогнозування з покращенням точності від 1 до 15 відсотків порівняно з ШНМ та KNN відповідно. Тим не менш, важливо визнати, що модель має певні обмеження. Очікувалось, що ансамблева модель буде демонструвати кращу продуктивність при спільному використанні двох моделей машинного навчання, ніж одна модель глибокого навчання. Альтернативно, можна використовувати модель глибокого навчання, таку як ШНМ або алгоритм ГН, шляхом зменшення кількості ознак та нормалізації даних. Цей підхід має на меті підвищити продуктивність моделі. Навчання двох моделей в ансамблевій моделі часто потребує значного обсягу пам'яті, що вказує на те, що ансамблева модель залежить від обчислювально-інтенсивних ресурсів. Це дає розуміння, що вибір ансамблевої моделі, повинен здійснюватися за умови послідовного використання пакетних завдань як основного способу виконання завдань у нашому робочому процесі. Запропонована реалізація сприяє розробці алгоритмів планування завдань у центрах обробки даних.

На основі цього прогнозу планувальник завдань виділяв би ресурси відповідно до пріоритету, визначеного зберіганням завдання в черзі, за умови, що прогнозоване значення потрапляє у вказаний діапазон. Створення такого застосунку є надзвичайно корисним, оскільки його інтеграція в алгоритм планування завдань надає можливість покращити алгоритм за допомогою методів машинного навчання. Це може сприяти подальшим дослідженням у сфері «зелених» обчислень, зокрема зосереджуючись на використанні стійкого підходу для забезпечення відмовостійкості.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Shahid M. A., Islam N., Alam M. M., Mazliham M., Musa S. Towards resilient method: An exhaustive survey of fault tolerance methods in the cloud computing environment // Computer Science Review. 2021. Vol. 40. Art. 100398. DOI: https://doi.org/10.1016/j.cosrev.2021.100398
2. Agarwal K. K., Kotakula H. Fault tolerance in cloud: A brief survey // Advances in Communication, Cloud, and Big Data. Springer, 2022. P. 578–589. DOI: https://doi.org/10.1007/978-981-19-2829-6_55
3. Ragmani A., Elomri A., Abghour N., Moussaid K., Rida M., Badidi E. Adaptive fault-tolerant model for improving cloud computing performance using artificial neural network // Procedia Computer Science. 2020. Vol. 170. P. 990–997. DOI: https://doi.org/10.1016/j.procs.2020.03.049
4. Tengku Asmawi T. N., Ismail A., Shen J. Cloud failure prediction based on traditional machine learning and deep learning // Journal of Cloud Computing. 2022. Vol. 11. DOI: https://doi.org/10.1186/s13677-022-00324-9
5. Marahatta A., Xin Q., Chi C., Zhang F., Liu Z. PEFS: AI-driven prediction-based energy-aware fault-tolerant scheduling scheme for cloud data center // IEEE Transactions on Sustainable Computing. 2021. Vol. 6, No. 4. P. 655–666. DOI: https://doi.org/10.1109/TSUSC.2020.2964266
6. Gao J., Wang H., Shen H. Task failure prediction in cloud data centers using deep learning // IEEE Transactions on Services Computing. 2022. Vol. 15, No. 3. P. 1411–1422. DOI: https://doi.org/10.1109/TSC.2020.2964407
7. Alahmad Y., Daradkeh T., Agarwal A. Proactive failure-aware task scheduling framework for cloud computing // IEEE Access. 2021. Vol. 9. P. 106152–106168. DOI: https://doi.org/10.1109/ACCESS.2021.3100708
8. Jassas M. S., Mahmoud Q. H. A failure prediction model for large-scale cloud applications using deep learning // Proceedings of the IEEE International Systems Conference (SysCon). 2021. DOI: https://doi.org/10.1109/SysCon48628.2021.9447071
9. Vani K., Sujatha S. A machine learning framework for job failure prediction in cloud using hyper-parameter tuned MLP // Proceedings of the 2nd International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). 2022. DOI: https://doi.org/10.1109/ICATIECE54045.2022.9768518
10. Ляшенко О., Михайліченко І. Модель самоадаптивної розподіленої системи керування ресурсами у хмарних обчисленнях // Herald of Khmelnytskyi National University. Technical Sciences. 2026. № 2 (363). С. 335–343. DOI: https://doi.org/10.31891/2307-5732-2026-363-46
11. El-Sayed N., Zhu H., Schroeder B. Learning from failure across multiple clusters: A trace-driven approach to understanding, predicting, and mitigating job terminations // Proceedings of the IEEE 37th International Conference on Distributed Computing Systems (ICDCS). 2017. P. 1333–1344. DOI: https://doi.org/10.1109/ICDCS.2017.155
12. Wilkes J. Google cluster-usage traces V3 [Електронний ресурс]. URL: https://github.com/google/cluster-data

Received (Надійшла) 12.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Знайдюк Василь Григорович – кандидат технічних наук, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vasyl Znaidiuk - PhD, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics,, Kharkiv, Ukraine.
e-mail: vasyl.znaidiuk@nure.ua; ORCID Author ID: https://orcid.org/0000-0001-8590-8007;
Scopus Author ID https://www.scopus.com/authid/detail.uri?authorId=57210340749.

Тухтаров Владислав Борисович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vladyslav Tukhtarov – Postgraduate student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: vladyslav.tukhtarov@nure.ua; ORCID Author ID: https://orcid.org/0009-0005-7650-965X.

Ensemble model for task failure prediction in cloud computing

Vasyl Znaidiuk, Vladyslav Tukhtarov

Abstract. Relevance. Cloud computing is a key component of modern IT infrastructure; however, task failures negatively affect service quality and resource utilization efficiency. The increasing complexity of cloud systems and the growing volume of data necessitate the use of intelligent failure prediction methods, enabling the transition from reactive to proactive and resilient resource management approaches. **Object of research:** processes of task failure prediction in cloud computing systems. **Purpose of the article:** development of an ensemble model for task failure prediction in cloud computing based on a combination of machine learning methods. **Research results.** The paper proposes an ensemble model based on the stacking method, combining the K-nearest neighbors algorithm and an artificial neural network with a meta-model based on logistic regression. Data preprocessing and analysis of the Google Cluster Trace dataset were performed, feature engineering was conducted, and a predictive model was developed. Experimental results demonstrated that the proposed ensemble model improves prediction accuracy and enhances F1-score, precision, and recall compared to individual models. It was established that the ensemble approach reduces overfitting and increases prediction reliability. **Conclusions.** The proposed model is an effective tool for task failure prediction in cloud systems and can be applied to optimize resource scheduling and improve fault tolerance. The use of the ensemble approach contributes to reducing resource consumption and supports the concept of green computing. **Scope of application:** task scheduling systems, resource management, and fault tolerance enhancement in cloud computing and data centers.

Keywords: cloud computing; failure prediction; ensemble model; machine learning; artificial neural network; KNN; stacking; logistic regression; fault tolerance; resource scheduling.

УДК 004.652.4:004.825 **doi: 10.26906/SUNZ.2026.2.104**

І. В. Золотухін, М. С. Кудрявцева, В. О. Філатов, М. В. Черненко, А. О. Андрусевич

Харківський національний університет радіоелектроніки, Харків, Україна

РЕЛЯЦІЙНА МОДЕЛЬ ДАНИХ У ВИРІШЕННІ ЗАДАЧ НЕЧІТКОЇ ЛОГІКИ

Анотація. Актуальність. Сучасні інтелектуальні технології обчислень переживають свій розквіт. Це пов'язано з потоком нових ідей, що виходять із галузі комп'ютерних наук, яка утворилася на перетині штучного інтелекту та інформаційних технологій. Предметні області, в яких бази даних використовуються як джерело даних, а як метод їх обробки - підхід на основі нечітких систем, становлять інтерес і з практичний, і з наукової точки зору. Тому сьогодні проблема проєктування нечіткої моделі бази даних та технологія обробки абстрактної інформації засобами реляційних систем стає дедалі актуальнішою. **Об'єкт дослідження**: нечіткі множини, лінгвістична змінна, функція належності, реляційна модель даних, теорія нормалізації, відношення фазифікації. **Мета статті**: розробка методів зберігання і обробки нечітких даних засобами реляційної моделі, орієнтованої на реалізацію в середовищі сучасних систем управління базами даних. Особливу увагу приділено обґрунтуванню вибору схеми реляційної моделі даних для представлення функцій належності лінгвістичних змінних. **Результати дослідження**. У статті на підставі концептуального алгебраїчного підходу до побудови інформаційних систем проведені дослідження, які створюють математичні, технологічні та програмні умови впровадження апарату нечіткої реляційної алгебри і спеціальної структурованої мови для нечітких запитів. Розглянуто питання обробки нечітких даних засобами реляційної моделі, орієнтованої на реалізацію в середовищі сучасних систем управління базами даних. Досліджено особливості функцій належності лінгвістичних змінних. Дано визначення відношенню фазифікації, обґрунтована структура такого відношення, на підставі теоретико-множинного підходу. Особливу увагу приділено обґрунтуванню вибору схеми поєднання відношення фазифікації з реляційною базою даних предметної області що досліджується. **Висновки.** Розглянуто представлення функцій належності лінгвістичних змінних засобами реляційних систем. Дано визначення відношенню фазифікації, обґрунтована структура такого відношення, на підставі теоретико-множинного підходу. Розроблено технологію поєднання відношення фазифікації з реляційною базою даних предметної області що досліджується. Сфера використання отриманих результатів: гібридні інформаційно-аналітичні системи підтримки прийняття рішень.

Ключові слова: нечіткі дані, реляційна модель даних, відношення, теорія нормалізації, відношення фазифікації лінгвістична змінна, нечітка логіка.

Вступ

Постановка проблеми. Комп'ютерні технології за допомогою інтелектуальних обчислень переживають свій розквіт. Це пов'язано з потоком нових ідей, що виходять із галузі комп'ютерних наук, яка утворилася на перетині штучного інтелекту та інформаційних технологій [1].

Предметні області, в яких бази даних використовуються як джерело даних, а як метод їх обробки - підхід на основі нечітких систем, становлять інтерес і з практичний, і з наукової точки зору. Тому сьогодні проблема проєктування нечіткої моделі бази даних та технологія обробки абстрактної інформації засобами реляційних систем стає дедалі актуальнішою.

У теперішній час відзначається гібридизація методів інтелектуальної обробки інформації. М'які обчислення об'єднують такі області як нечітка логіка, штучні нейронні мережі, видобуток знань, бази даних, імовірнісні міркування, еволюційні алгоритми тощо. Вони доповнюють один одного і використовуються у різних комбінаціях для створення гібридних інтелектуальних систем [2].

Не залишилися осторонь від цікавого і сучасного напрямку дослідники в області баз даних [3, 4]. Розробляється нечітка реляційна алгебра і спеціальні розширення структурованої мови для нечітких запитів. У цій області інтенсивні дослідження проводять європейські вчені Д. Дюбуа і Г. Праде. Формується перспективний напрям в сучасних системах обробки інформації – нечіткі запити до баз даних (fuzzy queries). У цьому контексті можна розглядати два основних питання, найбільш актуальні в даний час: як проектувати, де і в яких структурах зберігати нечіткі дані систем такого класу.

Вирішення цих проблем відкриє шляхи інтеграції накопичили колосальні обсяги інформації реляційних баз даних і систем на основі нечіткої логіки.

Аналіз останніх досліджень і публікацій. Математична теорія нечітких множин та нечітка логіка продовжують привертати увагу дослідників інтелектуальних, експертних систем, систем підтримки прийняття рішень, тощо. Ці поняття вперше запропоновані американським вченим Лотфі Заде (Lotfi Zadeh). Основною причиною появи нової теорії стала наявність нечітких і наближених міркувань при описі процесів, систем, об'єктів [5, 6]. Нечіткий підхід до моделювання складних систем отримав визнання у всьому світі, минуло не одне десятиліття з моменту зародження теорії нечітких множин. На шляху розвитку нечітких систем виділяють декілька періодів. Перший період характеризується розвитком теоретичного апарату нечітких множин. У другому періоді з'являються перші практичні результати в області нечіткого управління складними технічними системами [7]. Одночасно приділяється увага питанням побудови експертних систем, побудованих на нечіткій логіці, розробці нечітких контролерів. Нечіткі експертні системи для підтримки прийняття рішень знаходять широке застосування у медицині та економіці. У третьому періоді, що триває з кінця 80-х років, з'являються пакети програм для побудови нечітких експертних систем, а області застосування нечіткої логіки помітно розширюються. Вона застосовується в автомобільній, аерокосмічній і транспортній промисловості, у сфері фінансів, аналізу і прийняття управлінських рішень та багатьох інших [8].

Традиційний спосіб представлення елемента множини А полягає в застосуванні характеристичної функції $\mu_A(x)$, що дорівнює 1, якщо цей елемент належить множині А, або дорівнює 0 у протилежному випадку. У нечітких системах елемент може частково належати будь-якій множині. Ступінь належності до множини А, що є узагальненням характеристичної функції, називається функцією належності $\mu_A(x)$, причому $\mu_A(x) \in [0,1]$. Значення функції належності є раціональними числами з інтервалу [0,1], де 0 означає відсутність належності до множини, а 1 - повну належність. Конкретне значення функції належності називається ступенем або коефіцієнтом належності. Цей ступінь може бути визначений явно у вигляді функціональної залежності або дискретно - шляхом завдання кінцевої послідовності значень .

У теорії нечітких множин, крім змінних чисельного типу, існують лінгвістичні змінні з приписуваними їм значеннями [9].

Метою досліджень є розробка методів зберігання і обробки нечітких даних засобами реляційної моделі, орієнтованої на реалізацію в середовищі сучасних систем управління базами даних. Особливу увагу приділено обґрунтуванню вибору схеми реляційної моделі даних для представлення функцій належності лінгвістичних змінних.

Основний матеріал

Для вирішення поставленого завдання, а саме обґрунтування та вибору технології обробки нечітких даних розглянемо класичний, концептуальний *алгебраїчний* підхід до побудови інформаційних систем. До теперішнього часу запропоновано кілька формальних методів специфікації програмних систем взагалі, моделей і типів даних зокрема. Найбільшої популярності в якості формальної основи й інструмента специфікації типів даних набув алгебраїчний підхід. Алгебраїчний підхід був реалізований для побудови такої відомої технології як технологія баз даних, показав свою ефективність протягом багатьох років експлуатації побудованих на цій основі систем управління базами даних. Розглянемо особливості такого підходу з розробки на цій основі технології обробки нечітких даних.

У загальному випадку алгебраїчна система може бути подана у вигляді

$$U_a =< A, \Omega_f, \Omega_p >, \qquad (1)$$

де A – множина основ; $\Omega_f = \{F_1, ...F_k\}$ – множина імен операцій, заданих на множині A; $\Omega_p = \{\pi_1, ...\pi_m\}$ – множина предикатів, заданих на множині A.

Система U_a може бути записана у короткому вигляді $U_a =< A, \Omega >$, за умови об'єднання множин $\Omega = \Omega_f \cup \Omega_p$.

Множину A називають носієм або основною множиною, операції F_k і предикати π_m на відміну від інших операцій і предикатів називаються основними або головними.

Дослідження в галузі моделей даних інформаційних систем показують, що зараз центральним стало поняття типу даних. Із цим пов'язані як проблематика створення нових мов програмування, так і впровадження сучасних технологій організації даних. Із всього різноманіття підходів до визначення типу даних найбільш конструктивним виглядає такий: тип даних визначає множину значень через множину операцій.

У зв'язку з особливою роллю операцій у визначенні структур даних і функціонуванні систем розглянемо докладніше це поняття. Для формального визначення типу даних запроваджується поняття сигнатури Σ як пари, що складається із множини імен операцій F_k і множини описів операцій Ω_o.

Тоді сигнатуру можна визначити як $\Sigma = < F_f, \Omega_o >$. Σ – це пара: специфікація типу даних сигнатури $\Sigma = < F_f, \Omega_o >$ і відповідна їй реалізація типу даних.

Такий підхід певною мірою тягне за собою новий погляд на типи даних, згідно з яким множина значень типу характеризується множиною операцій, дозволяє конструювати операції над типами не тільки конструкторам мови, але й програмістам, що створюють свої власні типи. Оскільки тип даних як об'єкт складається із двох основних компонентів: специфікації та реалізації, то й операції над типами мають справу зі складовими цих компонентів. До таких дій можуть належати збільшення чи зменшення кількості операцій, заміна реалізацій всіх чи деяких операцій, зміна типу подання і пов'язаних з цим операцій.

Таким чином, у визначенні типу даних віддзеркалюються обидва його аспекти: користувацький (специфікація) і машинний (реалізація). Програміст, складаючи свою програму, бачить тип як певну специфікацію; компілятор, транслюючи цю програму в об'єктний код, має справу із обома аспектами типу даних; процесор, виконуючи об'єктний код, взаємодіє лише із реалізаціями відповідних типів даних.

Для того, щоб створити тип, необхідно побудувати специфікацію і зв'язати з нею відповідну реалізацію. При цьому, природно, необхідно переконатися, що запропонована реалізація задовольняє дану специфікацію. Особливе місце відводиться моделі даних як базовій складовій будь-якої інформаційної системи. Сьогодні модель даних трактується як «сукупність методів і засобів визначення логічної структури бази даних і динамічного стану предметної області у базі даних».

К. Дейт виділяє у моделі даних три найсуттєвіших компонента:

- сукупність засобів визначення допустимих структур даних;
- множина операцій, які можна застосувати до допустимого стану бази даних для пошуку чи модифікації даних;
- множина обмежень цілісності, які явно чи неявно визначають множину допустимих станів бази даних [10].

Завдяки сполученню названих властивостей, модель даних надає користувачам засоби опису даних,

маніпулювання й контролю цілісності, виражені в одній чи кількох мовах роботи із базами даних. Завдання вироблення функціонально повного набору операцій над типами вимагає ретельних досліджень і не може бути визнане таким, що на сьогодні вже вирішене.

Розглянемо класичний підхід до побудови реляційного відносини, запропонований Е. Коддом, і виділимо основні властивості відношення при розширенні множини доменів [11].

Нехай R – кінцева підмножина імен відношень база даних; D –$(D_1,...,D_i)$ – множина доменів, де всякий домен D_i є іменованою множиною атомарних значень елементів даних; A – кінцева множина імен атрибутів відношення; *dom* – відображення з A в D, яке визначає з якого домену обрані значення атрибутів.

Пару $\langle A_i, domA_i \rangle$, де $A_i \in A$, називають атрибутом. Структурну схему S_i відношення R_i $(R_i \in R)$ можна подати у вигляді $R_i(A_1,...,A_n)$, у якому всі A_i різні. Відношення r_i можна визначити як розширення схеми $S_i : r_i \subseteq domA_1 \times ... \times domA_n$. Перестановка атрибутів у схемі не породжує нового розширення та множина $\{A_1,...A_n\}$ атрибутів відношення R_i задає тип відношення. Задля специфікації складу носія використовується вираз $R_i = A_1...A_n$. Структурна схема *U* реляційної бази даних – це специфікація $(R_1,...,R_p)$, де $R_i \in R$ и всі R_i різні.

Концептуально реляційна база є інформаційно-логічною моделлю певної предметної області, такою, що кожне розширення відповідає деякому стану даної області у певний момент дискретно поточного часу. Кожен стан моделюється впорядкованою сукупністю значень елементів даних, відповідних значенням властивостей об'єктів предметної області. Зауважимо, що реляційна модель даних передбачає сильну типізацію об'єктів, використання цілком певних категорій, таких як тип об'єкта, атрибут (властивість) об'єкта, домен. Об'єкти мають набір властивостей, що задаються в реляційній моделі схемою відношення [12].

Для вирішення завдань зберігання та обробки нечітких даних визначимо спеціальний тип відношень – *відношення фазифікації.* Схема таких відношень повинна відповідати двом умовам: відповідати класичним вимогам реляційної моделі даних і раціонально зберігати і представляти модель лінгвістичної змінної.

Будем розуміти під нечіткий змінний набір (*N, X, Y),* де *N* – це назва лінгвістичної змінної, *X* – область міркувань, *Y* – нечітка множина на *X*. Використовуючи таке визначення, сформуємо три домени, які відповідають обраному набору змінних.

Нехай $N = \{n_i,...,n_m\}$, $Y = \{0,0.1,...,1\}$, $X = \{x_0,...,x_k\}$. Значення *X* і *Y* відповідають обраній шкалі дискретизації і представляють область приналежності до параметра *N*. Для розглянутого випадку визначимо відповідні домени з метою представлення значень нечіткої змінної: $D_1 = \{n_1, n_2,..., n_m\}$, $D_2 = \{x_0,..., x_m\}$; $D_3 = \{0,0.1,...,1\}$. Визначимо множину імен доменів і сформулюємо відображення для множині імен $A = \{A_1, A_2, A_3\}$.

Відображення $\rho : (A_1 \to D_1; A_2 \to D_2; A_3 \to D_3)$ визначає множину атрибутів $A = \{A_1, A_2, A_3\}$, що відповідає схемі відношення $S(A_1, A_2, A_3)$. Приклад такого відношення фазифікації наведено на рис. 1.

A_1	A_2	A_3
n_2	x_{14}	0
n_3	x_{14}	1
n_2	x_i	0.9
n_3	x_i	0.9
n_2	x_{20}	1
n_3	x_{20}	0
…	…	…

Рис. 1. Відношення фазифікації: завдання параметрів лінгвістичної змінної у табличному вигляді

Таким чином, у загальному випадку можна говорити про існування універсального відношення, що складається з кортежів декартового добутку доменів $D_1 \times D_2 \times D_3$. Ключем такого відношення буде множина усіх атрибутів $K = (A_1, A_2, A_3)$. Відношення фазифікації в подальшому будемо визначати як $R^f(A_1, A_2, A_3)$.

Перше завдання дослідження, сформульована у цій статті вирішено, проведено обґрунтування основної структурної одиниці в інформаційній технології зберігання нечітких даних та цією структурою є відношення фазифікації.

Залишається відповісти на питання, а як пов'язувати, об'єднувати дані, представлені нечітким відношенням фазифікації з даними реляційної моделі прикладної інформаційної системи. Розглянемо підхід поєднання таких даних на загальнотеоретичному рівні [13].

Нехай $U(R_1,...,R_n)$ – база даних, в якій зберігаються основні дані про досліджувану предметну область, $R^f(A_1, A_2, A_3)$ – відношення фазифікації. Завдання матиме сенс, якщо базі даних *U* існує атрибут, щодо якого виконана фазифікація. Щоб організувати спільну роботу з двох баз даних *U* і R^f, формалізуємо процедуру інтеграції, ґрунтуючись на технологію поетапної нормалізації.

Структура відношення *U* отримана на підставі функціональних залежностей $F = \{M_i \to N_i\}$ де $M_i, N_i \in U$.

Виділимо одну із залежностей, яка включає атрибут з параметрами фазифікації, і позначимо її як $W \to V$, причому *W* і *V* у загальному випадку можуть бути множинами. Відношення R^f містить одну залежність виду

$$F^{'} = \{A_1, A_2, A_3 \to A_1, A_2, A_3\}.$$

Спираючись на аксіоми виведення отримаємо еквівалентну множину

$$F' = \{A_1, A_2, A_3 \to A_1; A_1, A_2, A_3 \to A_2; A_1, A_2, A_3 \to A_3\}.$$

Нехай параметр фазифікації відповідає атрибуту A_2, тоді для визначення типу зв'язку необхідно отримати множину $F = F \cup F'$ та розглянути два випадки, що впливають на правила нормалізації:

- $A_2 \in W$ – пошук неповних залежностей: якщо виконуються функціональні залежності $\xi \to \zeta$ і $\omega \to \zeta$, причому $\omega \subseteq \xi$, тоді залежність $\omega \to \zeta$ є неповною;

- $A_2 \in V$ – пошук транзитивно-залежних атрибутів: якщо виконуються функціональні залежності $\xi \to \omega$ і $\omega \to \zeta$, тоді елемент ζ є транзитивно-залежним [14].

Наявність таких залежностей дозволяє виконати коректну декомпозицію та встановити зв'язок між базою даних U і R^f.

Враховуючи той факт, що структура бази даних не повинна змінюватись, необхідно зв'язати відношення фазифікації $R^f R^f$ и U без реструктуризації схеми даних. Уявимо R^f і U у вигляді головних сутностей.

Для усунення зв'язку "N:M" введемо додаткову сутність, яка вирішить проблему підтримки цілісності даних за рахунок визначення нових типів зв'язків. Сутність зв'язок міститиме один атрибут – сполучний для R^f и U, причому з об'єктивних причин він буде ключовим. З опису концептуальної схеми слід, що за для коректного з'єднання R^f і U необхідно побудувати проміжне відношення. Такий підхід гарантує узгодженість даних будь-яких типів параметрів фазифікації.

Покажемо, що для завдання цілком коректні результати при виконанні з'єднання відносин з асоціацією типу «N:M». Можливі значення атрибуту $A_1 \in U$ можуть повторюватися стільки разів, скільки це значення перетинає межі діаграми фазифікації по осі ординат.

Тобто кожному значенню атрибуту A_1 відповідає рядок унікальних даних.

Якщо A_1 не є ключем, і значення повторюються, то, за визначенням множини, у рядку має бути хоча б одне відмінне значення. У термінах розв'язуваної задачі необхідно аналізувати такі рядки. Значення атрибуту A_1 відношення R^f також можуть повторюватися, причому у різних комбінаціях.

Таким чином, у загальному вигляді для аналізу даних, що накопичуються в реляційних базах даних, достатньо побудувати фазифіковане відношення та встановити зв'язок з атрибутом (атрибутами), за значеннями якого необхідно провести відповідний аналіз.

На рис.2. представлена схема концептуальна схема об'єднання даних двох таблиць, відносини фазифікації, в якому у вигляді даних реалізовано відображення лінгвістичної змінної «ВІК», та відношення предметної області, де є атрибут «ВІК» пов'язаний у вигляді кортежів з іншими атрибутами. Таке об'єднання дозволяє отримати нові якісні дані, нові знання про властивості предметної області представленої зовсім в іншій проекції з урахуванням нечітких оцінок.

Відношення R

ID	A_1
1	молодий
2	досвідчений
3	поважний

Відношення R^f

A_1	A_3	A_2
молодий	1,0	20
молодий	0,5	30
молодий	0,2	40
досвідчений	0,1	30
досвідчений	1,0	40
досвідчений	0,3	50
поважний	0,2	40
поважний	0,7	50
поважний	1,0	60

Відношення U

V	D	C
20		
30		
40		
30		
40		
50		
40		
50		
60		

Рис. 2. Фрагмент схеми з'єднання відношення фазифікації R^f та баз даних U

Наведемо приклад реалізації такої технології на основі проведених із статті досліджень. Візьмемо за основу лінгвістичну змінну «ВІК» та сформуємо відношення фазифікації R^f аналогічне, як представлено на рис.2. Як аналізоване відношення будемо використовувати дані про захист дисертаційних робіт рівня доктор філософії здобувачами U.

У такій базі даних, у такому відношенні представлені ряд атрибутів, у тому числі «ВІК ПОШУКАЧА» та область досліджень, серед них такі як «Технічні», «Юридичні», «Економічні» тощо. Розроблена технологія дозволить відповісти на запитання: які напрями досліджень обирають, наприклад, «МОЛОДІ» здобувачі? На рис.3. представлений графік на основі нечіткого запиту.

Висновок: молоді науковці, здобувачі наукових ступенів у молодому віці обирають напрямки досліджень «Культурологія», «Юриспруденція», тощо, при цьому «Технічні науки» не популярні.

Таким чином, розроблена інформаційна технологія зберігання, обробки нечітких даних дозволяє успішно вирішувати великий клас прикладних завдань, поєднуючи нечіткі дані з даними предметних областей.

Висновки

У статті розглянуто та досліджено два основних питання, найбільш актуальні в даний час: як проектувати, де і в яких структурах зберігати нечіткі дані гібридних систем.

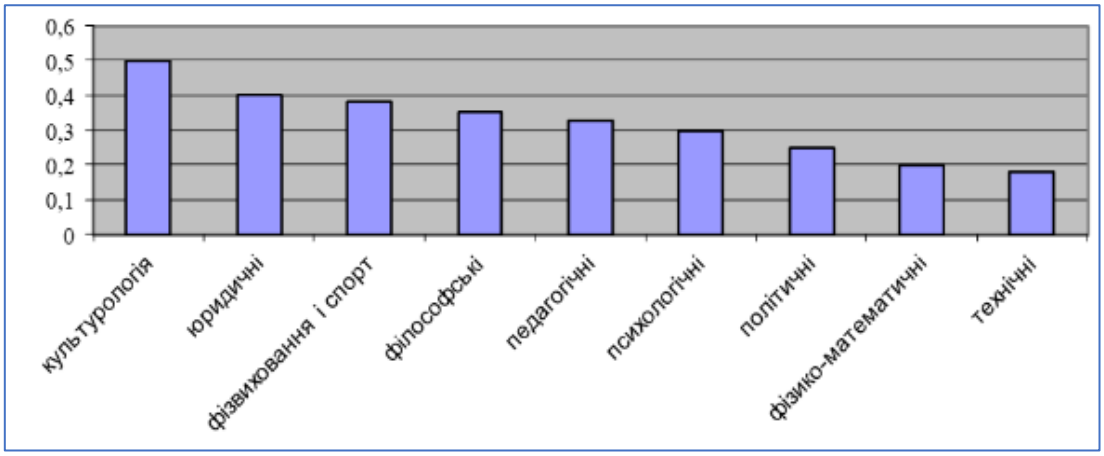


Рис. 3. Результат виконання нечіткого запиту

Проведені дослідження на підставі концептуального алгебраїчного підходу до побудови інформаційних систем, які створюють математичні, технологічні та програмні умови впровадження апарату нечіткої реляційної алгебри і спеціальної структурованої мови для нечітких запитів.

Розглянуто питання обробки нечітких даних засобами реляційної моделі, орієнтованої на реалізацію в середовищі сучасних систем управління базами даних. Досліджено особливості функцій належності лінгвістичних змінних. Дано визначення відношенню фазифікації, обґрунтована структура такого відношення, на підставі теоретико-множинного підходу. Особливу увагу приділено обґрунтуванню вибору схеми поєднання відношення фазифікації з реляційною базою даних предметної області що досліджується.

Наведено приклади, що підтверджують ефективність розглянутого у статті підходу до технології зберігання і обробки нечітких даних засобами реляційної моделі.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Bodyanskiy, Y., Zaychenko, Y., Kuzmenko, O., Zaichenko, H. (2026). Hybrid System of Computational Intelligence Based on Fuzzy Bagging and Group Method of Data Handling. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Data Mining . Studies in Systems, Decision and Control, vol 609. Springer, Cham. https://doi.org/10.1007/978-3-031-97529-5_14
2. Фалько, М., & Шафроненко, А. (2024). Нечіткий метод кластеризації даних з використанням еволюційної процедури. Матеріали конференцій МЦНД, (25.10.2024; Умань, Україна), 424–426. https://doi.org/10.62731/mcnd-25.10.2024.004
3. Avrunin, O., Vlasov, O., & Filatov, V. (2020). Model of semantic integration of information systems properties in relay database reengineering problems. Innovative Technologies and Scientific Solutions for Industries, (4 (14)), 5–12. https://doi.org/10.30837/itssi.2020.14.005
4. Filatov, V., Semenets, V., & Zolotukhin, O. (2020). Data mining in relational systems. Innovative Technologies and Scientific Solutions for Industries, (3 (13)), 65–76. https://doi.org/10.30837/itssi.2020.13.065
5. Zadeh, L. A. (1974). The Concept of a Linguistic Variable and its Application to Approximate Reasoning. Learning Systems and Intelligent Robots, 1–10. https://doi.org/10.1007/978-1-4684-2106-4_1
6. Fuzzy Set Theory and Rough Set Theory. (2006). In Fuzzy Modeling and Fuzzy Control, Birkhäuser Boston. 1–32. https://doi.org/10.1007/978-0-8176-4539-7_1
7. Mamdani, E. H. (1977). Applications of fuzzy set theory to control systems: a survey. Fuzzy automata and decision processes, 10, 247-259. https://doi.org/10.1016/0005-1098(77)90077-2
8. Raju, K. V. S. V. N., & Majumdar, A. K. (1988). Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems. ACM Transactions on Database Systems, 13(2), 129–166. https://doi.org/10.1145/42338.42344
9. Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics, SMC-15, 116-132. https://doi.org/10.1109/TSMC.1985.6313399
10. Date, C. J. (2003). Introduction to database systems. Pearson Education, Limited, 1024 p. https://www.amazon.com/Introduction-Database-Systems-8th/dp/0321197844
11. Codd, E. F. (1983). A relational model of data for large shared data banks. Communications of the ACM, 26(1), 64–69. https://doi.org/10.1145/357980.358007
12. Filatov, V., & Semenets, V. (2018). Methods for Synthesis of Relational Data Model in Information Systems Reengineering Problems. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), 247–251. https://doi.org/10.1109/infocommst.2018.8632144
13. Filatov, V., & Doskalenko, S. (2018). On the Approach to Searching for Functional Dependences of Data in Relational Systems. Innovative Technologies and Scientific Solutions for Industries, (1 (3)), 54–58. https://doi.org/10.30837/2522-9818.2018.3.054

14. Rodriguez-Jimenez, J. M., Rodriguez-Lorenzo, E., Cordero, P., Enciso, M., & Mora, A. (2015). A Normal Form for Fuzzy Functional Dependencies. 2015 IEEE Symposium Series on Computational Intelligence, 984–989. https://doi.org/10.1109/ssci.2015.143

Received (Надійшла) 02.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Золотухін Олег Вікторович – кандидат технічних наук, доцент, декан факультету Комп'ютерних наук, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleh Zolotukhin – Candidate of Technical Sciences, Associate Professor, Dean of Computer Science Faculty, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: oleg.zolotukhin@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-0152-7600;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?origin=resultslist&authorId=57207774022.

Кудрявцева Марина Сергіївна – кандидат технічних наук, доцент, професор кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна;
Maryna Kudryavtseva – Candidate of Technical Sciences, Associate Professor, Professor of Artificial Intelligence Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: maryna.kudryavtseva@nure.ua; ORCID Author ID: https://orcid.org/0000-0003-0524-5528;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?origin=resultslist&authorId=57207765829.

Філатов Валентин Олександрович – доктор технічних наук, професор, професор кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна;
Valentin Filatov – Doctor of Technical Sciences, Professor, Kharkiv National University of Radio Electronics, Artificial Intelligence Department, Kharkiv, Ukraine;
email: valentin.filatov@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-3718-2077;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=56911938100.

Черненко Микола Володимирович – кандидат технічних наук, асистент кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна;
Mykola Chernenko – Candidate of Technical Sciences, Assistant Professor, Kharkiv National University of Radio Electronics, Artificial Intelligence Department, Kharkiv, Ukraine;
email: mykola.chernenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0006-0623-5056;

Андрусевич Анатолій Олександрович – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри комп'ютерно-інтегрованих технологій, автоматизації, робототехніки та безпекової інженерії, Харків, Україна;
Anatoly Andrusevich – Doctor of Technical Sciences, Professor, Kharkiv National University of Radioelectronics, Professor at the Department of Computer-Integrated Technologies, Automation, Robotics and Safety Engineering, Kharkiv, Ukraine;
email: anatolii.andrusevych@nure.ua; ORCID Author ID: https:// https://orcid.org/0000-0002-3142-635X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57220195350.

Relational data model in solving fuzzy logic tasks

Oleh Zolotukhin, Maryna Kudryavtseva, Valentin Filatov, Mykola Chernenko, Anatoly Andrusevich

Abstract. Relevance. Modern intelligent computing technologies are experiencing their heyday. This is due to the flow of new ideas coming from the field of computer science, which was formed at the intersection of artificial intelligence and information technology. Subject areas in which databases are used as a source of data, and as method of their processing – an approach based on fuzzy systems are of interest both from a practical and scientific point of view. Therefore, today the task of designing a fuzzy database model and the technology of processing abstract information using relational systems is becoming increasingly relevant. **Research object**: fuzzy sets, linguistic variable, membership function, relational data model, normalization theory, fuzzification relation. **Purpose of the article:** development of methods for storing and processing fuzzy data using a relational model, focused on implementation in the environment of modern database management systems. Special attention is paid to the justification of the choice of a relational data model scheme for representing membership functions of linguistic variables. **Research results.** In the article based on the conceptual algebraic approach for building information systems, research has been conducted, that creates mathematical, technological and program conditions for implementing the apparatus of fuzzy relational algebra and a special structured language for fuzzy queries. The issue of processing fuzzy data by means of relational model oriented to implementation in the environment of modern database management systems has been considered. The features of membership functions of linguistic variables have been investigated. The definition of the fuzzification relation has been given, the structure of such relation has been substantiated based on the set-theoretic approach. Special attention has been paid to the justification of the choice of the scheme for combining the fuzzification relation with the relational database of the research subject area. **Conclusions**. Representation of membership functions of linguistic variables by means of relational systems is considered. The definition of the fuzzification relation is given, the structure of such relation is substantiated based on the set-theoretic approach. The technology of combining the fuzzification relation with the relational database of the research subject area is developed. The scope of use of the obtained results: hybrid information and analytical decision support systems.

Keywords: fuzzification relation linguistic variable, fuzzy data, fuzzy logic, normalization theory, relation, relational data model.

УДК 004.8:004.912 **doi: 10.26906/SUNZ.2026.2.110**

І. М. Івасенко, Т. В. Філімончук, С. О. Партика, Д. І. Пивоварова

Харківський національний університет радіоелектроніки, Харків, Україна

МОДЕЛЬ РОЗРОБКИ ТЕМАТИЧНИХ ЧАТ-БОТІВ З ВИКОРИСТАННЯМ ПІДХОДУ RAG

Анотація. Актуальність дослідження зумовлена стрімким розвитком великих мовних моделей (LLM), таких як GPT-4, Llama 3 та Claude, революціонізував сферу обробки природної мови (NLP). LLM демонструють виняткові здібності до генерації зв'язного тексту, узагальнення інформації та ведення діалогу. Однак, при їх застосуванні у спеціалізованих доменах (юриспруденція, медицина, технічна підтримка, корпоративні бази знань) виникають критичні обмеження. По-перше, параметричні знання моделей обмежені датою завершення їх навчання (knowledge cutoff), що унеможливлює роботу з актуальною інформацією. По-друге, моделі схильні до "галюцинацій" – генерації правдоподібних, але фактично невірних тверджень, особливо коли запит стосується вузькоспеціалізованих даних, відсутніх у навчальному наборі. Архітектура RAG стала стандартом де-факто для вирішення цих проблем, поєднуючи генеративні можливості LLM із точним пошуком у зовнішніх базах знань. Проте, практика показує, що "наївна" реалізація RAG (Vanilla RAG) часто є недостатньою для побудови надійних систем. Втрата контексту під час пошуку, нездатність обробити складні запити користувача та відсутність механізмів верифікації призводять до нерелевантних відповідей. У зв'язку з цим, актуальною науково-практичною задачею є не просто імплементація RAG, а розробка та дослідження методів оптимізації кожного етапу генерації. **Об'єктом дослідження** є процеси інформаційного пошуку та генерації природної мови в інтелектуальних діалогових системах, побудованих на базі великих мовних моделей. **Предметом дослідження** є методи та алгоритми підвищення точності, релевантності та контекстуальної узгодженості відповідей у системах архітектури RAG, а саме гібридний пошук, покращення запитів користувача. **Висновки.** Практична цінність дослідження полягає у створенні моделі архітектури, яку можна адаптувати для будь-якої предметної області (від технічної документації до нормативно-правових баз), забезпечуючи при цьому вищу метричну точність відповідей порівняно з базовими рішеннями.

Ключові слова: тематичні чат-боти, великі мовні моделі, штучний інтелект, RAG, NLP, попередня обробка, покращення запиту, гібридний пошук, розумна генерація тексту, механізм переранжування, релевантність відповіді, точність контексту.

Постановка проблеми

Попри суттєві досягнення у розвитку великих мовних моделей (LLM), їх пряме застосування для побудови спеціалізованих інформаційних систем супроводжується низкою серйозних обмежень.

Ключовою проблемою є нездатність моделей працювати з актуальною інформацією, що з'явилася після завершення процесу навчання, а також підвищений ризик так званих "галюцинацій" – формування переконливих, проте помилкових відповідей.

Архітектура Retrieval Augmented Generation (RAG) була запропонована як підхід до усунення цих недоліків, оскільки забезпечує інтеграцію мовної моделі із зовнішніми джерелами знань. Однак, класична реалізація RAG (так званий "Vanilla RAG"), що базується на простому розбитті тексту на фрагменти (chunking) та векторному пошуку (dense retrieval), часто виявляється недостатньою для забезпечення високої якості обслуговування користувачів у тематичних доменах.

Основна проблема полягає у невідповідності (misalignment) між запитом користувача та збереженими даними. Користувачі схильні формулювати короткі, нечіткі або контекстно-залежні запитання, які семантично можуть не перетинатися із відповідними фрагментами документації. Векторний пошук, хоч і є ефективним для знаходження загальних концепцій, часто демонструє низьку точність при пошуку специфічних термінів, власних назв або абревіатур, що призводить до вибірки нерелевантного контексту.

Додатковою проблемою є обмеження самої генеративної моделі при роботі з контекстом. Навіть за умови успішного пошуку, LLM може ігнорувати надану інформацію, якщо вона суперечить її внутрішнім вагам, або ж втрачати важливі деталі, якщо обсяг контексту занадто великий (проблема "Lost in the Middle"). Таким чином, науково-технічна проблема полягає у необхідності вдосконалення архітектури RAG для мінімізації помилок на етапах пошуку та генерації. Необхідно вирішити задачу підвищення точності ретрівалу (retrieval accuracy) та узгодженості відповідей (faithfulness) шляхом впровадження методів гібридного пошуку, семантичної трансформації запитів та алгоритмічного керування увагою моделі через інженерію запитів. Вирішення цієї проблеми дозволить створити тематичний чатбот, здатний надавати верифіковані та точні відповіді, що є критичною вимогою для професійного застосування.

Огляд досліджень і розробок

Проблематика покращення генерації тексту за допомогою зовнішніх джерел знань (RAG) є однією з найбільш динамічних у сучасній комп'ютерній лінгвістиці. Аналіз наукового доробку за останні роки дозволяє виділити фундаментальні праці, що сформували архітектуру сучасних RAG-систем та визначили вектори їх оптимізації. Нижче проаналізовано ключові дослідження, що стосуються як базової архітектури, так і методів покращення пошуку та генерації.

У роботі [1] автори запропонували архітектуру, яка об'єднує попередньо навчену модель "seq2seq"

(BART) з механізмом щільного векторного пошуку (Dense Passage Retrieval). Дослідження показало, що моделі RAG перевершують звичайні параметричні моделі у задачах Open Domain Question Answering, навіть маючи значно меншу кількість параметрів. Головний висновок роботи полягає в тому, що відокремлення "знань" (база даних) від "вміння міркувати" (нейромережа) дозволяє легко оновлювати знання системи без необхідності перенавчання моделі.

У роботі [2] було продовжено розвиток попередньої архітектури. Дослідники зосередилися на етапі переднавчання (pre-training) та запропонували підхід, де ретрівер (пошуковик) навчається одночасно з генератором. Було доведено, що такий спільний процес навчання (end-to-end training) значно покращує здатність моделі знаходити релевантні документи, оскільки ретрівер отримує зворотний зв'язок від мовної моделі щодо корисності знайденої інформації.

Робота [3] стала переломною для відмови від ключових слів на користь семантичного пошуку. Автори довели неефективність традиційного алгоритму BM25 для складних запитань та запропонували модель DPR, що використовує два окремі енкодери (для питань та для контексту). Висновком дослідження є те, що щільні векторні представлення (dense embeddings) значно краще вловлюють семантичну близькість, ніж перетин слів.

У роботі [4] було розроблено механізм "пізньої взаємодії" (late interaction), який зберігає контекстуальні вектори токенів та порівнює їх лише на фінальному етапі. Дослідження показало, що цей метод (який лежить в основі сучасних Re-ranker моделей) дозволяє досягти точності повноцінних трансформерів (BERT) при швидкості, співставній з векторним пошуком. Результати дослідження обґрунтовують необхідність використання етапу переранжування (Reranking) у тематичних ботах.

У роботі [5] було запропоновано архітектуру Fusion-in-Decoder, яка замість об’єднання всіх знайдених документів в один довгий текст і подання його на вхід моделі, обробляє кожен документ окремо. Злиття інформації відбувається лише на етапі генерації відповіді, що дозволяє ефективніше обробляти знайдені документи та надає моделі можливість самій вирішувати, на що звертати увагу при генерації.

Робота [6] підтвердила гіпотезу про те, що доступ до гігантської бази даних (трильйони токенів) дозволяє зменшити розмір самої нейромережі у 25 разів без втрати якості. Наведені результати свідчать про те, що для побудови якісного чат-боту важливіша якість бази знань та пошуку, аніж розмір самої LLM.

У роботі [7] автори дослідили вплив інжинірингу запитів на якість відповідей. Хоча робота не фокусувалася виключно на RAG, вона довела, що спонукання моделі до покрокового міркування ("Let's think step by step") значно покращує її здатність аналізувати складний контекст та зменшує кількість логічних помилок.

Аналіз наведених джерел дозволяє стверджувати, що "Vanilla RAG" (простий векторний пошук та генерація) є пройденим етапом. Сучасні дослідження вказують на необхідність впровадження гібридних методів пошуку (поєднання DPR та BM25), використання механізмів Re-ranking та просунутих технік інжинірингу запитів для створення надійних систем.

На основі проведеного аналізу можливо представити стандартну модель архітектури RAG у вигляді спеціалізованого набору складових (1):

$$M = \{МП, БЗ, ВММ\}, \quad (1)$$

де: МП – модуль пошуку, який знаходить релевантні документи або фрагменти тексту (наприклад, Retriever); БЗ – зовнішнє сховище знань, компонент, що містить інформацію, недоступну моделі під час її навчання (приватні корпоративні дані, свіжі новини, специфічна технічна документація); ВММ – велика мовна модель (LLM), яка попередньо навчена на необхідних масивах тексту для розуміння та генерації природньої мови.

Наведена модель, попри свою концептуальну простоту, має низку критичних архітектурних вразливостей, які суттєво обмежують її використання у професійних сферах:

- семантичний розрив – проблема неспівпадіння лексики, коли користувач формулює запити побутовою мовою, описуючи симптоми або загальні ситуації, тоді як технічна документація оперує суворою професійною термінологією. Оскільки базовий модуль пошуку (МП) часто спирається на косинусну близькість векторів, він може не виявити зв'язку між цими фразами, оскільки в багатовимірному просторі їхні вектори можуть бути ортогональними (дивитися в різні боки), попри спільний зміст;

- низька точність пошуку специфічних сутностей, що пов’язано з втратами інформації під час стиснення. Коли модель перетворює текст у набір чисел, вона добре зберігає загальний зміст ("це текст про комп'ютери"), але часто втрачає точні деталі. Через це системи часто плутають схожі за написанням, але критично різні за суттю ідентифікатори (наприклад, артикули "Part-A100" та "Part-A200" або коди помилок). Для нейромережі ці токени семантично майже ідентичні, що призводить до видачі неправильних інструкцій;

- "Lost in the Middle" – обмеження архітектури, на якій базуються сучасні великі мовні моделі, призводять до того, що механізм уваги (Attention Mechanism) працює нерівномірно: модель найкраще сприймає інформацію, розташовану на початку та в кінці вхідного тексту. Якщо релевантний фрагмент, знайдений модулем пошуку, потрапляє в середину довгого списку контексту, модель схильна ігнорувати його при генерації відповіді, що нівелює саму ідею пошуку;

- схильність до "галюцинацій" – якщо модуль пошуку не знайшов релевантної інформації (або знайшов нерелевантну), чат-бот все одно намагається дати відповідь, заповнюючи прогалини знань вигадками або застарілими даними зі своєї тренувальної

пам'яті. Без примусового посилання на факти, така відповідь може виглядати переконливо, але бути фактично хибною.

Мета дослідження

Метою роботи є підвищення точності, фактологічної достовірності та контекстуальної релевантності відповідей тематичних діалогових систем шляхом дослідження, вдосконалення та програмної реалізації розширеної архітектури RAG (Advanced RAG).

Для досягнення поставленої мети необхідно вирішити наступні задачі:

- проаналізувати обмеження базової архітектури та існуючих підходів до їх подолання;

- модифікувати математичну модель архітектури RAG шляхом додавання складових, які реалізують поєднання методів семантичної трансформації запитів (Query Rewriting), гібридного пошуку інформації (Hybrid Search) та керованої генерації тексту;

- розглянути практичні сценарії використання вдосконаленої архітектури та провести оцінку її ефективності.

Викладення основного матеріалу

Стандартний підхід до побудови RAG-систем працює за лінійною схемою: користувач ставить питання, система шукає схожий текст у базі за допомогою векторної подібності та передає його нейромережі для генерації відповіді.

Цей метод демонструє високу ефективність для загальних запитів, проте часто виявляється недостатнім у спеціалізованих доменах (юриспруденція, технічна підтримка, медицина).

Основна проблема полягає в семантичному розриві: користувач і база даних "говорять різними мовами". Користувачі використовують побутову лексику, тоді як документи містять суху професійну термінологію. Для вирішення цієї проблеми та підвищення метрик якості (Accuracy, Recall, Precision) існує низка методів вдосконалення архітектури на етапах попередньої обробки, пошуку та генерації. У роботі пропонуються впровадження цих методів (як окремих модулів) в існуючу архітектуру (1), які нададуть "розумних посередників" на кожному етапі:

- модуль покращення запиту, який перетворює нечітке питання користувача на професійний пошуковий запит (МПЗ);

- модуль гібридного пошуку, який буде шукати інформацію декількома способами одночасно (за змістом та за ключовими словами), щоб нічого не пропустити (МГП);

- модуль розумної генерації, який аналізує знайдене та формує відповідь, спираючись виключно на факти (МРГ).

Після додаванням цих модулів, покращена модель буде мати вигляд (2):

$$М = \{МПЗ, МГП, ВММ, БЗ, МРГ\}. \quad (2)$$

Логіка руху інформації в представленій моделі (2) трансформується з лінійної у багаторівневу, де вихід кожного попереднього модуля стає збагаченим входом для наступного.

Модуль покращення запиту (МПЗ) виправляє найслабше місце будь-якого пошуку – постановку питання. Користувачі часто пишуть коротко, з помилками або використовують займенники, зрозумілі тільки з контексту діалогу (наприклад, "А скільки він коштує?", де "він" стосується товару, згаданого три повідомлення потому). Якщо відправити такий запит у базу даних, система не знайде нічого релевантного, бо в документах немає слова "він", є конкретні назви товарів.

Вирішити цю проблему можливо за допомогою методу Query Rewriting, який здійснює переписування запиту. Перед тим як шукати інформацію, використовується допоміжна велика мовна модель (LLM), що "перекладає" питання користувача на мову пошукової системи. Це працює наступним чином: система бере історію діалогу та останнє питання користувача і на їх основі створює новий, самодостатній запит:

- було: "Як мені це підключити, якщо зникло світло?";

- стало після обробки: "Інструкція з підключення пристрою [назва моделі] до джерела безперебійного живлення або генератора в умовах відсутності електропостачання".

Такий переписаний запит містить всі необхідні ключові слова та контекст і гарантує, що пошуковий алгоритм шукатиме саме інструкцію, а не випадкові тексти зі словом "світло".

Окрім простого переписування одного запиту, існує більш просунута стратегія, яка називається розширенням запиту, або генерацією множинних варіантів. Часто буває так, що навіть ідеально переписане питання може не знайти потрібний документ просто тому, що автор документа використав зовсім інші слова для опису тієї ж проблеми. Наприклад, користувач питає про "ціну", а в документах використовується термін "вартість тарифного плану". Щоб вирішити цю проблему, мовна модель генерує не один, а три або п’ять варіантів одного й того ж запитання під різними кутами. Система автоматично створює синонімічні конструкції, розбиває складні питання на простіші підзадачі та формує векторні представлення для кожного з них. Далі пошук відбувається одночасно за всіма згенерованими варіантами. Потім система збирає всі знайдені за різними запитами документи та видаляє дублікати, залишаючи лише унікальний набір даних для аналізу. Такий підхід особливо ефективний у ситуаціях, коли база знань наповнювалася різними авторами в різний час та не має єдиного стандарту.

Після отримання переписаного запиту, модуль гібридного пошуку (МГП) починає пошук інформації. У звичайних системах використовують лише один метод, але для кращого результату необхідно оперувати декількома різними підходами (3):

$$МГП = \{ВП, ПКС, ОР, ПР\}, \quad (3)$$

де: ВП – векторний пошук; ПКС – пошук за ключовими словами; ОР – механізм об’єднання результату; ПР – механізм переранжування.

Перший підхід – це семантичний або векторний пошук (ВП). Уявімо, що комп'ютер перетворює кожну частину тексту в набір координат (чисел) у багатовимірному просторі. Тексти, які мають схожий

зміст, будуть знаходитися поруч у цьому просторі, навіть якщо вони написані різними словами.

Наприклад, якщо користувач шукає "проблеми з мотором", векторний пошук знайде документ про "несправності двигуна", тому що для нейромережі слова "мотор" та "двигун", "проблема" та "несправність" є синонімами і знаходяться поруч, що дозволяє знаходити відповіді, які людина могла б пропустити через розбіжність у термінології. Сучасні моделі пропонують вектори розмірністю 384-3072 чисел. Існує пряма кореляція: чим більша розмірність, тим більше семантичних нюансів може "запам'ятати" вектор, але тим повільніше працює пошук і тим більше оперативної пам'яті потребує база даних. Для тематичних чат-ботів оптимальним балансом вважається розмірність 768 або 1024.

Однак саме лише використання векторів не гарантує успіху, якщо модель, яка перетворює текст на числа, не розуміє специфіки предметної області. Слід зазначити, що більшість стандартних моделей навчені на текстах з інтернету, новинах та художній літературі. Вони чудово розуміють загальні поняття, але можуть губитися у вузькоспеціалізованих темах, що створює проблему зміщення домену.

Уявімо ситуацію в медичному чат-боті. Для звичайної моделі слова "серце" та "любов" можуть знаходитися поруч у векторному просторі, оскільки в художній літературі вони часто вживаються разом. Проте для лікаря "серце" має бути семантично пов'язане з термінами "міокард", "кардіологія", "шлуночок", а не з романтичними почуттями. Якщо модель цього не розуміє, пошук буде видавати нерелевантні результати.

Для вирішення цієї проблеми використовується вже існуюча велика мовна модель з додатковим її тренуванням на документах необхідної конкретної сфери. Цей процес нагадує підвищення кваліфікації для фахівця: мовна модель читає тисячі прикладів з бази знань і вчиться розуміти, що в контексті компанії слово "коса" – це не зачіска і не інструмент для трави, а, наприклад, елемент географічного ландшафту або деталь механізму. Після такого донавчання вектори стають значно точнішими, і система починає розрізняти найдрібніші нюанси професійної термінології. Однак векторний пошук має недолік, він іноді надто узагальнює: може сплутати "Модель X-100" та "Модель X-200", бо семантично це майже одне й те саме (обидва – назви моделей). Тут на допомогу приходить другий підхід – метод пошуку за ключовими словами (наприклад, алгоритм BM25), який працює подібно до функції "Знайти на сторінці", але розумніше. Він не просто рахує співпадіння слів, а оцінює їх важливість:

- якщо слово зустрічається в кожному документі (наприклад, "і", "що", "бути"), алгоритм ігнорує його;

- якщо слово рідкісне та унікальне (наприклад, конкретний код помилки "Error-404" або прізвище автора), алгоритм надає йому більшої ваги.

Завдяки роботі цього методу (ПКС) гарантується, що якщо користувач ввів точний артикул або специфічний термін, система знайде саме той документ, де він згадується.

Після попередніх дій у користувача є два списки знайдених документів: один сформований на основі розуміння змісту, інший – на основі точних слів. Наступним кроком є їх злиття в один фінальний перелік, тому що пряме порівняння результатів цих алгоритмів неможливе через різні критерії оцінки. Для об’єднання результатів використовується метод Reciprocal Rank Fusion (RRF), принцип роботи якого ґрунтується на врахуванні позиції (рангу) кожного документа у відповідних списках результатів пошуку:

- якщо певний документ займає високі позиції одночасно і у векторному, і у ключовому пошуку, він отримує найвищий пріоритет у фінальному списку;

- якщо документ знайдений лише одним із методів або знаходиться в кінці списків, він автоматично опускається нижче.

Такий підхід дозволяє об'єднати переваги обох технологій, коли система виводить у топ документи, які є найбільш релевантними як за змістом (вектором), так і за наявністю конкретних термінів, що створює надійну основу для генерації відповіді.

Завершальним етапом роботи модуля МГП є переранжування. Попередній список кандидатів (наприклад, 50 фрагментів) може містити "шум". Однак передавати весь цей масив даних безпосередньо у велику мовну модель (LLM) є помилкою з двох причин:

- економічна: більшість сучасних LLM тарифікують послуги за кількість токенів (частин слів) на вході. Обробка зайвої інформації збільшує вартість кожного запиту;

- якісна: існує феномен "розмиття уваги", коли модель отримує занадто багато тексту, в якому корисна інформація змішана з "шумом" (нерелевантним текстом) та ймовірність правильної відповіді падає. Модель може "загубитися" у великому обсязі даних.

Саме тут вступає в дію етап переранжування (Reranking) – процес, який працює як фільтр: він бере широку вибірку документів та залишає лише декілька (зазвичай 3-5) найбільш влучних. Щоб зрозуміти важливість переранжування, слід детальніше розглянути, як працює первинний пошук.

Векторний пошук використовує архітектуру, яку називають Bi-Encoder, що працює наступним чином: на першому кроці система заздалегідь створює "портрет" (вектор) кожного документа в базі; коли приходить запит, система створює його "портрет"; далі вона просто порівнює ці портрети на відстані.

Цей метод неймовірно швидкий (можна порівняти мільйони документів за секунду), але він поверхневий. Оскільки вектор документа створюється до того, як система дізнається запит користувача, він є узагальненим тому що у ньому "стиснуто" весь зміст абзацу в один набір чисел. Але при такому стисненні втрачаються тонкі нюанси зв'язків між словами, тому на етапі переранжування використовується інший інструмент – Cross-Encoder. На відміну від Bi-Encoder, який обробляє запит та документ окремо, Cross-Encoder приймає на вхід пару "Запит + Документ" одночасно, що дозволяє нейромережі використовувати механізм "уваги" (Self-Attention) на повну потужність. Модель переглядає кожне слово у запиті користувача та безпосередньо порівнює його з кожним

словом у документі, аналізуючи їх взаємодію в контексті.

Наприклад: припустимо, користувач питає: "Що робити, якщо пристрій погано працює взимку?". Векторний пошук (Bi-Encoder) може знайти документ: "Взимку компанія не працює по вихідних". Для нього слова "не працює" та "взимку" семантично близькі до запиту, хоча зміст зовсім інший. Через переранжування (Cross-Encoder) мовна модель "прочитає" цю пару разом і побачить, що в запиті словосполучення "не працює" стосується технічного стану, а в документі – графіку роботи офісу. Завдяки глибшому аналізу зв'язків слів, буде відкинуто цей документ або поставлено йому дуже низьку оцінку. Нажаль цей процес вимагає значно більше обчислювальних ресурсів та часу, саме тому не можна застосувати Cross-Encoder до всієї бази даних. У даному випадку слід використовувати двоступеневу схему: швидкий, але грубий (Bi-Encoder) пошук відбирає топ-50 кандидатів; повільний, але дуже розумний (Cross-Encoder) сортує ці 50 кандидатів та видає топ-5 ідеальних результатів. Такий сімбіоз дозволяє досягти балансу: система працює швидко (бо перевіряє детально лише мало документів) та надзвичайно точно.

Після завершення збору всієї необхідної інформації, програма переходить на фінальний етап обробки інформації, а саме оформлення відповіді в модулі розумної генерації (МРГ).

Навіть у випадку знаходження ідеальних документів, робота ще не завершена. Останній крок – змусити LLM правильно використати цю інформацію. Багато розробників помилково вважають, що достатньо просто скопіювати текст документів у чат та попросити "відповісти". Однак без спеціальних технік це призводить до неструктурованих або помилкових відповідей, для уникнення яких використовується комплексний підхід до конструювання запитів, що базується на науково обґрунтованих методиках:

$$МРГ = \{СР, ЛД, НП, ЦД\}, \quad (4)$$

де СР – сортування по релевантності; ЛД – створення "ланцюжка думок"; НП – навчання по прикладах; ЦД – цитування документів.

Однією з цікавих вад в роботі великих мовних моделей є так звана проблема "Lost in the Middle" (загублені посередині). Експерименти показали, що LLM найкраще запам'ятовують інформацію, яка знаходиться на самому початку тексту і в самому кінці запиту. Інформація, що розміщена всередині довгого тексту, часто ігнорується або "забувається" під час формування відповіді.

У покращеній системі це враховується при формуванні контексту. Після етапу переранжування, коли отримано 5 найкращих документів, вони не просто додаються у випадковому порядку, а спочатку сортуються таким чином, щоб найбільш релевантний документ (з найвищим балом Cross-Encoder) знаходився в кінці списку контексту – безпосередньо перед самим питанням користувача, а другий за важливістю документ ставився на початок. Менш важливі документи розміщуються всередині. Така структура подачі інформації гарантує, що модель зверне максимальну увагу на найважливіші дані.

Наступним кроком, перш ніж надати відповідь, модель форматує її за відповідними правилами (ця методика має назву "Chain of Thought"). У системний запит додається інструкція: "Перш ніж надати відповідь користувачу, проаналізуй наданий контекст крок за кроком. Випиши ключові факти, перевір, чи немає суперечностей між різними документами, і лише після цього сформулюй висновок". Це працює тому, що LLM є авторегресійною моделлю: вона передбачає наступне слово на основі попередніх. Коли модель спочатку генерує ланцюжок логічних міркувань, ці міркування стають частиною контексту для генерації фінальної відповіді, тобто чат-бот сам собі створює "чернетку", на яку потім спирається, що радикально зменшує кількість логічних помилок у відповідях.

Замість того, щоб писати довгі інструкції про те, як має виглядати відповідь (наприклад, "будь ввічливим, не використовуй жаргон, пиши стисло"), велика мовна модель отримує приклади. У запит включається кілька пар "Питання – Ідеальна Відповідь". Наприклад, користувач пише запит "Як скинути пароль?". Бот надає відповідь "Згідно з інструкцією безпеки (розділ 4), для скидання пароля перейдіть за посиланням... Зверніть увагу, що посилання дійсне 15 хвилин". Коли чат-бот бачить ці приклади, він автоматично копіює стиль, тон та структуру відповіді (In-Context Learning), що особливо важливо для тематичних ботів, які мають дотримуватися корпоративного стандарту спілкування.

Однією із найбільших проблем штучного інтелекту є "галюцинації". Щоб боротися з цією проблемою, впроваджується механізм суворої перевірки джерел. Мовній моделі надається жорстка інструкція: "Кожне твердження у відповіді має супроводжуватися посиланням на номер документа, з якого взято цю інформацію (наприклад, [документ №1]). Якщо інформації немає в наданих документах ти повинен чесно відповісти: "Я не знаю" і не намагатися вигадати відповідь". Це перетворює процес генерації тексту на задачу компіляції фактів. Якщо модель не може поставити посилання на джерело, вона з меншою ймовірністю згенерує це твердження, що критично важливо для таких сфер, як право або медицина, де вигадана порада може мати серйозні наслідки.

Окрім правильних інструкцій та цитування, сучасні системи все частіше використовують механізми самокорекції та рефлексії. У стандартній схемі система працює в один прохід: знайшла інформацію, згенерувала відповідь та віддала її користувачу. Але якщо знайдена інформація була неповною або помилковою, користувач отримає неякісний результат. Щоб цьому запобігти, може бути впроваджений додатковий етап перевірки. Після того, як модель сформулювала чернетку відповіді, вона сама виступає в ролі критика. Система аналізує власну відповідь на предмет логічних суперечностей та відповідності знайденим документам. Якщо модель виявляє, що відповідь неповна або базується на

слабких доказах, вона не показує її користувачу. Замість цього запускається повторний цикл пошуку з уточненими параметрами. Наприклад, якщо бот зрозумів, що йому не вистачає інформації про конкретну характеристику товару, він самостійно формує новий пошуковий запит саме по цій характеристиці, знаходить відсутні дані, і лише потім генерує фінальну відповідь. Такий ітеративний підхід перетворює лінійний процес на замкнений цикл, який триває доти, доки не буде досягнуто прийнятного рівня якості. Хоча це займає трохи більше часу, результат стає надійнішим.

Слід зазначити, що якість роботи будь-якої RAG-системи починається задовго до того, як користувач задасть питання. Вона починається з того, як обробляються вхідні документи. Якщо просто завантажити у базу цілу книгу одним шматком, система не зможе ефективно знайти конкретний абзац. Якщо ж спочатку розділити текст на частини занадто дрібно, втратиться контекст.

Для вирішення цієї проблеми пропонується застосовувати стратегію "Recursive Character Text Splitter" (Рекурсивний поділ тексту) з перекриттям (overlap). Найпростіший метод – розділяти текст через кожні 500 символів. Але це "сліпий" метод, який може розірвати речення посередині або відділити питання від відповіді. Уявіть, що одне речення каже: "Не натискайте червону кнопку", а наступне (яке потрапило в інший шматок): "якщо система не увімкнена". Якщо знайти тільки першу частину, інструкція стане небезпечною.

Рекурсивний метод працює розумніше. Спочатку він намагається розділити текст по великих логічних блоках (параграфах): якщо блок все ще завеликий, він ділить його по реченнях; якщо і це забагато – по словах. Використання такого підходу дозволяє зберегти цілісність думки.

Також точність відповіді можна покращити використанням техніки перекриття: кінець одного шматка тексту (chunk) дублюється на початку наступного. Наприклад, якщо ми маємо текст А-Б-В-Г, ми ріжемо його не як [А-Б] та [В-Г], а як [А-Б-В] та [Б-В-Г], що створює "страховку": якщо важлива думка знаходиться на стику двох шматків, вона не загубиться, а потрапить в обидва. Наприклад, система може використовувати розмір сегменту тексту 1000 токенів із перекриттям у 200 токенів.

Ще однією серйозною проблемою при підготовці даних є конфлікт між пошуком та генерацією. Для точного пошуку найкраще підходять маленькі шматочки тексту, які містять одну конкретну думку або факт. Їх вектори дуже чіткі, і системі легко знайти відповідність із запитом користувача. Проте для генерації якісної відповіді маленького шматочка часто недостатньо. Великій мовній моделі потрібен широкий контекст, щоб зрозуміти передісторію, умови та наслідки. Якщо просто віддати чатботу одне речення, він може не зрозуміти його суті без навколишнього тексту. Щоб вирішити цю дилему, рекомендується використовувати стратегію батьківського документа. Суть методу полягає в розриві зв'язку між тим, що шукає користувач, і тим, що надається для відповіді. Під час індексації документ розбивається на дуже дрібні фрагменти для пошуку, але зберігається зв'язок кожного дрібного фрагмента з його великим "батьківським" блоком. Коли система знаходить маленький фрагмент, який відповідає запиту, вона не передає його одразу на генерацію. Замість цього вона звертається до бази даних та витягує повний батьківський блок, до якого належав цей фрагмент. Таким чином, пошук відбувається з хірургічною точністю по дрібних деталях, а велика мовна модель отримує повноцінний розгорнутий контекст для формування зв'язної та обґрунтованої відповіді.

Також варто згадати про метод семантичного поділу тексту. Традиційний поділ на частини фіксованого розміру часто розриває логічні ланцюжки. Семантичний підхід пропонує ділити текст не за кількістю символів, а за змістом. Спеціальний алгоритм аналізує текст речення за реченням та вимірює, наскільки сильно змінюється тема розмови. Поки вектори сусідніх речень схожі, система вважає, що триває опис однієї думки, і об'єднує їх в один блок. Як тільки вектор різко змінюється, це слугує сигналом, що автор перейшов до нової теми, і тут потрібно зробити розріз. Такий підхід дозволяє створювати смислові блоки, які є цілісними та завершеними, що значно полегшує роботу нейромережі на етапі генерації.

Для побудови вище описаної системи (Advanced RAG) недостатньо стандартних інструментів, необхідно обрати набір технологій, які підтримують гібридний пошук та складну логіку обробки.

Оркестратор LangChain – це "скелет" застосунку, який дозволяє легко з'єднувати різні компоненти: завантажувачі файлів, моделі та бази даних. Саме він керує логікою: "спочатку перепиши запит, пошукай, відфільтруй".

Векторна база даних Qdrant, як і Weaviate "з коробки" підтримують гібридний пошук. Вони вміють зберігати і вектори (для розуміння змісту), і звичайний текст (для пошуку ключових слів), а також самостійно виконувати злиття результатів (Fusion), що значно спрощує розробку. Вибір бази даних – це не лише питання зберігання, а й питання алгоритмів індексації. Простого порівняння запиту з кожним документом у базі недостатньо, коли кількість документів сягає тисяч або мільйонів – це буде надто повільно.

Сучасні векторні бази даних використовують алгоритм наближеного пошуку найближчих сусідів – HNSW (Hierarchical Navigable Small World), який спочатку робить "широкий огляд" і визначає лише загальний напрямок пошуку, моментально відсіюючи все зайве (наприклад, визначає, що шукати треба в групі "юридичні договори"). Зрозумівши, де приблизно знаходиться відповідь, він заглиблюється в деталі і шукає вже всередині цієї меншої групи (наприклад, фокусується на "договорах оренди"), поступово звужуючи коло пошуку до конкретного файлу. Такий метод дозволяє швидко знаходити потрібну інформацію серед мільйонів записів.

Ще одним критичним компонентом Advanced RAG, який часто ігнорують у базових реалізаціях, є Hybrid Filtering (Фільтрація за метаданими). Тематичні документи завжди мають атрибути: рік видання, автор, категорія, рівень доступу. Векторний пошук сліпий до цих атрибутів і може знайти документ, який ідеально підходить за змістом, але є застарілим (наприклад, наказ, який вже скасовано).

Важливим елементом технологічного стека також є система маршрутизації діалогу. Не кожне повідомлення користувача вимагає запуску складного та дорогого процесу пошуку в базі знань. Якщо користувач просто вітається або дякує, немає сенсу навантажувати векторну базу даних. Для оптимізації роботи використовується легкий класифікатор на вході системи, який миттєво аналізує намір користувача та вирішує, куди спрямувати запит. Якщо це проста розмова, запит йде до звичайної великої мовної моделі. Якщо це питання по темі бази знань, запускається повний цикл RAG. Якщо ж питання стосується теми, яку бот не повинен обговорювати, система активує модуль безпеки мовної моделі та ввічливо відмовляє у відповіді. Така архітектура дозволяє суттєво економити обчислювальні ресурси та гроші, оскільки найважчі алгоритми задіються лише тоді, коли це дійсно необхідно. Крім того, це дозволяє підключати до системи різні джерела даних. Наприклад, для фінансових питань маршрутизатор може направити запит до бази зі звітами, а для технічних питань – до бази з інструкціями, що робить чат-бота універсальним помічником.

Для аналізу запиту, формування відповіді та її аналізу можна використовувати різні великі мовні моделі, деякі з яких запропоновано нижче:

- Embeddings (для створення векторів): text-embedding-3-small від OpenAI або bge-m3. Ключова особливість цих моделей – підтримка технології Matryoshka Representation Learning (MRL), що дозволяє динамічно скорочувати розмірність вектору (наприклад, з 1536 до 512 або 256 чисел) без суттєвої втрати якості пошуку, що критично важливо для оптимізації витрат на зберігання векторів у базі даних. Модель bge-m3 особливо цікава тим, що вона мультимовна і краще працює зі специфічними мовами, ніж стандартні моделі;

- LLM (для генерації): вибір на користь моделей серії GPT у RAG-системах може бути зумовлений не лише їхніми "знаннями", а передусім архітектурною особливістю, яка називається Instruction Following (слідування інструкціям). У RAG критично важливо, щоб модель ігнорувала свої попередньо отримані (вивчені) знання та відповідала виключно на основі наданого контексту. Моделі GPT демонструють високу стабільність у дотриманні системного запиту (System Prompt Adherence) та формати виведення;

- Re-ranker: bge-reranker-v2-m3 – спеціалізована програма-суддя, що приймає на вхід пару "Запит + Документ" як єдиний текст і пропускає їх через шари уваги (Self-Attention) разом та оцінює, як кожне слово запиту впливає на кожне слово документу. На виході вона дає не вектор, а одне число від 0 до 1, що відображає релевантність.

Для доведення, що "розумний" RAG працює краще за звичайний, потрібні конкретні цифри результатів. Традиційні метрики для перекладу тексту (як BLEU) тут не підходять, бо бот може сформулювати правильну думку іншими словами та BLEU покаже поганий результат. Для цього краще використовувати фреймворк RAGAS (Retrieval Augmented Generation Assessment). Ідея RAGAS геніальна: він використовує одну сильну LLM (наприклад, GPT-4), щоб вона оцінювала роботу запропонованої системи як екзаменатор. В ході аналізу вимірюються три ключові показники: вірність (Faithfulness), релевантність відповіді (Answer Relevance), точність контексту (Context Precision).

Вірність (Faithfulness) – ця метрика відповідає на питання: "Чи не вигадав бот щось від себе?" Модель-екзаменатор перевіряє кожне твердження у відповіді бота та шукає підтвердження у знайдених документах. Якщо бот сказав "Гарантія 3 роки" і в документах написано "Гарантія 3 роки" – бал 1.0, якщо бот сказав "Гарантія 5 років", а в документах "Гарантія 3 роки" або інформації немає – бал падає. Використання такого підходу дозволяє математично виміряти рівень "галюцинацій".

Релевантність відповіді (Answer Relevance) – ця метрика відповідає на питання: "Чи відповів бот саме на те, про що його питали?". Бот може сказати чисту правду, але не по темі. Наприклад, на питання "Як вимкнути прилад?" він може відповісти "Прилад був винайдений у 1990 році". Це "вірно" (Faithfulness висока), але "нерелевантно". RAGAS оцінює, наскільки відповідь корисна для користувача.

Точність контексту (Context Precision) – ця метрика оцінює роботу пошуковика із базової реалізації (1). Вона перевіряє, чи є "золоті" (ідеальні) документи на вершині списку знайденого. Якщо запропонована система Re-ranking працює правильно, найкорисніші документи завжди мають бути в топі (ранг 1 або 2), а не десь на 10-му місці. Запропонований підхід трансформує RAG з простого пошуковика на складну когнітивну систему, яка не просто шукає слова – вона трансформує думки (Query Rewriting), використовує різні типи пошуку, критично оцінює знайдене (Re-ranking) і змушує модель думати логічно.

Результати та їх обговорення

Ефективність запропонованої архітектури ілюструється на прикладі реальних бізнес-кейсів, де вартість помилки є високою, а обсяг даних перевищує можливості контекстного вікна стандартних моделей. Далі наведено впровадження Advanced RAG для створення спеціалізованих асистентів.

Приклад 1: фінансові аналітики витрачають години на пошук інформації серед сотень тисяч внутрішніх PDF-документів, звітів про ринки. При використанні RAG здійснюються наступні дії: створення внутрішнього чат-боту, який має доступ до понад 100000 документів банку. Як результат – система використовує векторний пошук для миттєвого знаходження відповідного розділу у звіті. Застосування гібридного пошуку тут критичне, оскільки бот повинен розрізняти схожі назви фінансових інструментів

(наприклад, "Class A Shares" та "Class B Shares"), що для чистого векторного пошуку може виглядати ідентично. Це дозволяє скоротити час пошуку інформації на 90%, надаючи відповідь з прямим посиланням на джерело.

Приклад 2: юридичні фірми працюють з тисячами контрактів та прецедентів. Використання звичайної мовної моделі LLM (наприклад, ChatGPT) є небезпечним через ризик "галюцинацій" – вигадування неіснуючих законів. При використанні Advanced RAG здійснюються наступні дії: система індексує базу законодавства та судових рішень. Як результат – ключовим елементом тут є Prompt Engineering з вимогою цитування. Бот не генерує юридичну пораду "з голови", а компілює її на основі знайдених статей кодексу. Якщо закон змінився вчора, покращена Advanced RAG-система знайде оновлення, тоді як звичайна LLM буде оперувати застарілими даними з моменту останнього тренування.

Приклад 3: користувачі ставлять технічні питання щодо API або налаштувань, використовуючи непрофесійну лексику ("побутовий опис" симптомів), тоді як документація написана складною технічною мовою та часто оновлюється, що збільшує шанс неточної відповіді. Для вирішення цієї проблеми використовується чат-бот з Advanced RAG, який має доступ не лише до документації, а й архіву виконаних задач (наприклад, з Jira). Коли користувач пише: "кнопка не працює", LLM аналізує контекст діалогу або схожі вирішені задачі в базі Jira, і формулює гіпотезу про причину збою. Вона трансформує запит у технічний формат: "помилка методу POST при збереженні форми" або "відомі помилки UI при підтвердженні дії". Таким чином пошуковий алгоритм за цим технічним запитом знаходить конкретну інструкцію для розробника, чи статтю з бази знань, яку неможливо було б знайти просто за словом "кнопка".

Приклад 4: лікарям необхідний швидкий доступ до протоколів лікування та досліджень (PubMed), обсяг яких зростає експоненційно. При обранні Advanced RAG застосовується система, що допомагає аналізувати симптоми та пропонує протоколи на основі доказової медицини. Оскільки медичні терміни дуже схожі, системі важливо відфільтрувати 50 знайдених статей та залишити лише 3 найбільш релевантні для конкретного випадку пацієнта, щоб не перевантажити лікаря зайвою інформацією.

Для оцінки ефективності запропонованого підходу було проведено порівняння трьох типів систем на наборі даних зі складних тематичних запитань (Open-Domain QA dataset) (табл. 1):

LLM – стандартна модель (наприклад, GPT-4) без доступу до зовнішніх даних;

Vanilla RAG – базова реалізація (простий векторний пошук);

Advanced RAG – запропонована в роботі система (Hybrid Search + Re-ranking + Query Rewriting).

Таблиця 1 – **Порівняння метрик якості відповідей діалогових систем**

Метрика	Опис показника	LLM (GPT-4)	Vanilla RAG	Advanced RAG
Точність фактів, %	частка тверджень, що не суперечать джерелам/реальності	65-75	85	92-95
Рівень "галюци-націй", %	частота вигадування неіснуючих фактів	~20	~10	<3
Актуальність знань, %	здатність відповідати на події, що сталися після навчання моделі	0	100	100
Релевантність відповіді, %	відповідність заданому питанню	80	75	90
Цитування джерел	можливість перевірити джерело інформації	відсутнє	присутнє (часто неточне)	точне (на рівні речення)

У сучасних системах, які побудовано на базі LLM, існує низка фундаментальних обмежень, що ускладнюють їх використання в реальних бізнес-сценаріях. Зокрема, такі моделі мають обмеження щодо актуальності знань, можуть генерувати недостовірну інформацію та демонструють нестабільну релевантність відповідей. Саме тому в практичних застосуваннях дедалі частіше використовують підхід RAG, який поєднує можливості мовної моделі з механізмами пошуку релевантної інформації. Аналіз досліджень (таблиця 1) показує, що застосування RAG та його розширених варіацій дозволяє значно підвищити точність, актуальність та надійність відповідей системи. Основні аспекти цієї проблематики можна узагальнити наступним чином:

- проблема "Knowledge Cutoff" (точність фактів): звичайна LLM нездатна повністю працювати з актуальними даними, у тематичних ботах (наприклад, новини компанії або зміна цін) використання RAG є безальтернативним, оскільки показник актуальності для чистої LLM дорівнює нулю;

- зменшення "галюцинацій": впровадження базового RAG вже знижує їх рівень удвічі (з 20% до 10%). Однак, саме використання методів Advanced RAG (зокрема Re-ranking та чіткі системні запити) дозволяє досягти критично низького рівня помилок (<3%), що робить систему придатною для використання в бізнесі;

- парадокс релевантності: цікаво відзначити, що Vanilla RAG іноді показує гіршу релевантність (75%), ніж чиста LLM (80%), що пояснюється тим, що поганий пошук може знайти "сміттєві" документи, які зіб'ють чат-бот з пантелику. Використання же підходів Query Rewriting та Hybrid Search підвищує релевантність шляхом подачі лише якісного контексту.

Таким чином можна стверджувати, що експериментальні дані підтверджують гіпотезу, що покращення та інвестування в попередню обробку запиту

та постобробку пошуку (Advanced RAG) є виправданими та забезпечують значний приріст якості наданих користувачу відповідей.

Висновки

У рамках даної роботи було проведено комплексне дослідження проблематики побудови тематичних діалогових систем на основі великих мовних моделей (LLM) з використанням архітектури Retrieval Augmented Generation (RAG). Метою роботи було не лише відтворення базового алгоритму пошуку та генерації (1), але й вдосконалення існуючого підходу шляхом розширення складових базової моделі. Запропонована модель (Advanced RAG) орієнтована на підвищення точності, релевантності та фактологічної вірності відповідей у вузькоспеціалізованих доменах. Аналіз теоретичної бази та сучасного стану розвитку NLP-технологій підтвердив, що архітектура RAG є гарним рішенням для створення професійних чат-ботів. Стандартні LLM (GPT-4, Llama 3), попри їхні високі когнітивні здібності, страждають від проблеми "Knowledge Cutoff" (відсутність актуальних знань) та схильності до "галюцинацій" при роботі з незнайомими фактами. Відокремлення бази знань від нейромережі дозволяє вирішити ці проблеми, забезпечуючи динамічне оновлення інформації без необхідності дороговартісного донавчання моделі.

У ході дослідження було виявлено, що однією з головних причин низької якості роботи стандартних RAG-систем є "семантичний розрив" між запитом користувача та текстом документів. Користувачі схильні формулювати нечіткі, короткі або контекстнозалежні запитання. Запропонований та імплементований метод Query Rewriting (переписування запиту) з використанням LLM довів свою ефективність. Трансформація вхідного питання у повноцінний пошуковий запит, насичений синонімами та уточненою термінологією, дала змогу значно підвищити якість векторних представлень, що забезпечило зростання повноти пошуку, надаючи можливість системі знаходити релевантні документи навіть тоді, коли оригінальний запит користувача не містив прямих ключових слів. Експериментально доведено, що покладання виключно на векторний пошук може бути недостатнім для тематичних чат-ботів, де важлива точність термінології (артикули, назви моделей, специфічні абревіатури), тому що векторні моделі часто "розмивають" ці деталі. Реалізація гібридного пошуку, що поєднує векторну подібність та лексичне співпадіння (BM25) з алгоритмом злиття Reciprocal Rank Fusion (RRF), дозволила досягти синергетичного ефекту.

Система демонструє високу здатність розуміти контекст (завдяки векторам) та водночас знаходити точні входження термінів (завдяки BM25), що мінімізувало кількість випадків, коли бот надавав загальну відповідь замість конкретної інструкції. Впровадження етапу переранжування (Reranking) з використанням моделей Cross-Encoder стало ключовим фактором підвищення точності (Precision). Виявлено, що первинний пошук часто повертає документи, які є семантично близькими, але фактично нерелевантними. Використання Cross-Encoder, який виступає в ролі "суворого фільтра", ефективно відсіює інформаційний шум, залишаючи для генерації лише найбільш цінні фрагменти контексту, що дозволило вирішити проблему "Lost in the Middle", коли LLM втрачає увагу через надмірну кількість вхідних даних. Дослідження також довело, що якість генерації залежить не лише від знайденого контексту, але й від інструкцій, наданих моделі. Застосування техніки Chain of Thought (CoT) дозволило структурувати "мислення" моделі. А вимога до обов'язкового цитування джерел перетворила процес генерації з "творчого написання" на "аналітичну компіляцію", що знизило рівень "галюцинацій" до <3%.

Порівняльний аналіз модифікованої архітектури (Advanced RAG) з базовою реалізацією (Vanilla RAG) та "чистою" LLM продемонстрував суттєву перевагу запропонованого підходу. Показники достовірності та релевантності відповідей зросли на 15-20%. Розроблений програмний підхід є універсальним і може бути адаптований для будь-якої предметної області: від юридичного консалтингу до технічної підтримки, що підтверджує практичну цінність роботи.

Незважаючи на досягнуті результати, тема залишається відкритою для подальшого вивчення. Перспективними напрямками розвитку є:

- GraphRAG: інтеграція графів знань (Knowledge Graphs) для покращення розуміння складних взаємозв'язків між сутностями, які важко вловити через простий текст;
- Agentic RAG: перехід від пасивної архітектури "питання-відповідь" до агентної моделі, де система може самостійно вирішувати, чи достатньо їй інформації, і за потреби виконувати додаткові запити або уточнювати деталі у користувача;
- Multimodal RAG: розширення системи для роботи не лише з текстом, але й таблицями та схемами, що важливо для технічної документації.

Підсумовуючи, можна стверджувати, що мета роботи досягнута: розроблено та обґрунтовано методологію створення високоефективного тематичного чат-боту, який вирішує основні проблеми генеративних моделей і готовий до впровадження у реальні бізнес-процеси.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401v4*. 2021. doi: https://doi.org/10.48550/arXiv.2005.11401

2. Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv preprint arXiv:2002.08909v1*. 2020. doi: https://doi.org/10.48550/arXiv.2002.08909
3. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769-6781. doi: https://doi.org/10.18653/v1/2020.emnlp-main.550
4. Omar Khattab, Matei Zaharia. Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020, p. 39-48. doi: https://doi.org/10.1145/3397271.3401075
5. Gautier Izacard, Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021, pages 874-880. doi: https://doi.org/10.18653/v1/2021.eacl-main.74
6. Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426v3*. 2022. doi: https://doi.org/10.48550/arXiv.2112.04426
7. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903v6*. 2023. doi: https://doi.org/10.48550/arXiv.2201.11903

Received (Надійшла) 11.01.2026
Accepted for publication (Прийнята до друку) 08.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Івасенко Ілля Михайлович – магістрант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Illia Ivasenko – master's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: illia.ivasenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0000-2588-279X.

Філімончук Тетяна Володимирівна – кандидат технічних наук, доцент, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Tetiana Filimonchuk – PhD, Associate Professor, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: tetiana.filimonchuk@nure.ua; ORCID Author ID: http://orcid.org/0000-0002-4380-504X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?origin=resultslist&authorId=57190949991.

Партика Станіслав Олександрович – старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Stanislav Partyka – Senior lecturer of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: stanislav.partyka@nure.ua; ORCID Author ID: http://orcid.org/0000-0002-7376-8980;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?origin=resultslist&authorId=57204560890.

Пивоварова Дар'я Ігорівна – асистент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Daria Pyvovarova– Assistant of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: daria.pyvovarova@nure.ua; ORCID Author ID: http://orcid.org/0000-0002-7251-994X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?origin=resultslist&authorId=57224191788.

Using the rag architecture to build thematic applications

Illia Ivasenko, Tetiana Filimonchuk, Stanislav Partyka, Daria Pyvovarova

Abstract. The relevance of the study is due to the rapid development of large language models (LLM), such as GPT-4, Llama 3 and Claude, which have revolutionized the field of natural language processing (NLP). These models demonstrate exceptional abilities to generate coherent text, generalize information and conduct dialogue. However, when applying LLM in specialized domains (law, medicine, technical support, corporate knowledge bases), critical limitations arise. First, the parametric knowledge of the models is limited by the date of completion of their training (knowledge cutoff), which makes it impossible to work with current information. Second, models are prone to "hallucinations" - the generation of plausible, but actually false statements, especially when the query concerns highly specialized data that is not in the training set. The RAG architecture has become the de facto standard for solving these problems, combining the generative capabilities of LLM with accurate search in external knowledge bases. However, practice shows that the "naive" implementation of RAG (Vanilla RAG) is often insufficient for building reliable systems. Loss of context during search, inability to process complex user queries and lack of verification mechanisms lead to irrelevant answers. In this regard, the current scientific and practical task is not just the implementation of RAG, but the development and research of methods for optimizing each stage of generation. **The object of research** is the processes of information search and natural language generation in intelligent dialog systems built on the basis of large language models. **The subject of the study** is methods and algorithms for increasing the accuracy, relevance and contextual consistency of responses in RAG architecture systems, namely hybrid search, improving user queries. **Conclusions**. The practical value of the study lies in creating an architecture model that can be adapted for any subject area (from technical documentation to regulatory frameworks), while ensuring higher metric accuracy of responses compared to basic solutions.

Keywords: thematic chatbots, LLM, AI, RAG, NLP, preprocessing, query improvement, hybrid search, intelligent text generation, reranking mechanism, faithfulness, answer relevance, context precision.

UDC 004.9 **doi: 10.26906/SUNZ.2026.2.120**

Alla Kapiton[1], Tamara Franchuk[2], Dmytro Tyshchenko[2], Olha Kurbanova[3]

[1] National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine
[2] State University of Trade and Economics, Kyiv, Ukraine
[3] PJSC "Higher Educational Institution Interregional Academy of Personnel Management", Kyiv, Ukraine

THE IMPACT OF GENERATIVE AI TECHNOLOGIES ON COMPLIANCE WITH STANDARDS OF ACADEMIC ETHICS

Abstract. This publication analyzes the determinants of the use of AI systems in research activities and analyzes the potential risks associated with adherence to the principles of academic integrity. A comparative analysis of leading generative AI platforms for text output was performed. It is concluded that the use of AI tools to perform technical and support operations in particular, structuring material, formatting data, or linguistic editing of author's ideas does not pose a threat to ethical standards and qualifies as an acceptable support practice. It is proven that in modern conditions of digitalization, artificial intelligence algorithms act as a determinant of optimizing content markup strategies. It is substantiated that the implementation of generative AI systems in the processes of preparing text materials provides a number of strategic advantages, among which it is worth highlighting: optimization of resource costs: minimizing time and financial investments by automating the creation of primary text arrays and reducing dependence on external copywriting; scaling the content base: intensification of production and diversification of information product formats to meet the needs of the target audience; improving quality indicators: the use of progressive natural language processing algorithms helps to improve stylistic correctness and ensure the required level of uniqueness of content; maximizing engagement indicators: increasing the conversion rate due to a high degree of personalization and compliance with consumer information needs. It was established that the existing tools of generative artificial intelligence are characterized by significant heterogeneity. It was determined that the functional capabilities of individual services are differentiated depending on the specifics of the content, the parameters of the target audience and the economic feasibility of their implementation. A relevant selection of the most representative software solutions that have demonstrated high efficiency in recent years was made. A comparative analysis of technical characteristics, cost models, as well as an assessment of the adaptability of the specified tools to the specifics of different languages was carried out. It is proven that the dynamics of the development of the market of software solutions based on artificial intelligence is characterized by a high intensity of the emergence of new tools and continuous modernization of existing services. The processes of improving functional capabilities are accompanied by the expansion of the linguistic spectrum of systems, which gives grounds for predicting further implementation and in-depth support of the Ukrainian language within a larger number of intelligent platforms.

Keywords: artificial intelligence; software; services; scientific activity; academic integrity.

Introduction

The integration of artificial intelligence (AI) into modern social processes has acquired a comprehensive character, covering both everyday aspects and complex areas of professional activity. AI technologies are of particular importance in the scientific research field, where their potential becomes the subject of discussions regarding new opportunities and ethical challenges. This work analyzes the prospects of introducing AI tools into scientific practice, and also highlights the key problematic aspects that arise in the process of their application by researchers. Although AI systems show progress in linguistics, they are often unable to convey the specifics of scientific terminology and contextual nuances. Low translation quality is a common reason for rejection of manuscripts by reviewers of international publications (Scopus, Web of Science). The preparation of materials of this level requires the involvement of professional translators who possess a specialized thesaurus.

Analysis of recent research and publications. For a thorough analysis, it is necessary to examine software selection guidelines that take into account academic integrity requirements. This aspect has not received sufficient attention in recent publications, so artificial intelligence in scientific activities and academic integrity policy is currently the focus of contemporary research. Hryshko V., Kyrychuk B. investigate the problems of academic integrity and artificial intelligence: overcoming challenges in the educational and scientific activities of Ukraine and foreign countries [1]. Tymokhin V., Yurchenko V., Nalyvayko O. conduct a discussion in the educational space on the possibility of combining academic integrity and the charms of artificial intelligence, based on ethical issues [2]. Todorova I. identifies the components of students' academic integrity and the conditions for its formation[3]. Umanets O., Shestakova S., Sukhomlynova O., Zadorina O. study individual positional components of the issue of academic integrity: the practice of compliance in foreign countries [4, 5]. Tyshchenko D., Franchuk T., Stepashkina K., Desiatko A., Karpunin, I. analyze the problems of developing an electronic document management system and Features of accounting digitalization processes [6, 7]. Belfo F., Trigo A. analyze accounting information systems: traditions and future directions [8]. Kapiton A., Sukhorebry O., Nenich D. investigate the use of multimodal artificial intelligence in economics, education, science and transport [9]. The analysis of the necessity of using information technologies in science and their optimization for academic integrity objectives has been thoroughly examined in studies [10–15]. These works present scientific ideas and recommendations for selecting software developments aimed at achieving specific results.

Main part

Scientometric tools play a decisive role in the analysis of the productivity of researchers, which

directly correlates with the performance indicators of scientific institutions. In addition, scientists with a high rating gain authority, which allows them to join editorial boards and expert committees, directly influencing the formation of scientific policy [1, 2].

The integration of artificial intelligence (AI) opens wide horizons for researchers to optimize the processes of preparation and writing of scientific papers. However, despite the significant technological potential, the use of AI is accompanied by a number of significant risks that require careful analysis. One of the most pressing issues is compliance with ethical standards and protection of intellectual property. Since AI models are based on the processing of already existing data sets, the problem of legality of using the generated content arises. Uncontrolled use of AI when writing articles reduces the scientific value of the work and can be classified as a specific form of plagiarism. Technologies should serve as an auxiliary tool, not a substitute for the intellectual work of a scientist. A responsible attitude to authorship is critically important for the preservation of scientific ethics. The process of verifying manuscripts before publication faces the imperfection of AI detection algorithms. There are cases where original author texts are mistakenly identified as machine generated Despite the ability of AI to process colossal volumes of information, the quality of such analysis often remains superficial. The use of outdated training samples leads to the retransmission of outdated data. To ensure scientific credibility, the researcher must independently work with primary sources: professional publications, verified statistics and specialized databases, which requires critical thinking that is inaccessible to algorithms. AI is a powerful catalyst for scientific efficiency, but it cannot replace the fundamental components of research — critical analysis and deep authorial interpretation. The use of intelligent systems in science should be based on the principles of transparency, confidentiality and personal responsibility of the scientist for each presented result. High scientometric indicators significantly increase the competitiveness of researchers within grant programs and other funding mechanisms. Since foundations and private investors, when allocating resources, focus primarily on publication activity and the level of citations, scientists with a significant scientific output have significantly higher chances of receiving financial support. This verification of the author's influence guarantees investors the viability of innovative projects and contributes to the continuous professional development of the scientific team. The dynamics of the market development of software solutions based on artificial intelligence is characterized by a high intensity of the emergence of new tools and continuous modernization of existing services. The processes of improving functional capabilities are accompanied by the expansion of the linguistic spectrum of systems, which gives grounds for predicting further implementation and in-depth support of the Ukrainian language within a larger number of intelligent platforms. The list of violations of scientific ethics includes self-plagiarism, i.e. duplicating one's own previously published results without a corresponding reference. No less serious offenses are fabrication (inventing data) and falsification (intentionally distorting sources or results to support a hypothesis).

Formation of a culture of academic integrity and ethical research skills in graduate students is a critically important task of modern education. The implementation of the academic integrity policy allows you to objectively analyze the scientific level of the article, check it for compliance with the journal's requirements, and provide a comprehensive assessment of the research results [3]. Adherence to the principles of academic integrity is a prerequisite for scientific research activity, which guarantees the validity of scientific results. The process of publishing research results is based on the principles of intellectual honesty (inadmissibility of plagiarism and distortions), strict observance of copyright, transparency through Open Access mechanisms, as well as principles of objective review and inclusiveness (equal rights of authors). The innovative potential and limitations of the application of AI in the academic environment are briefly presented in Fig. 1.

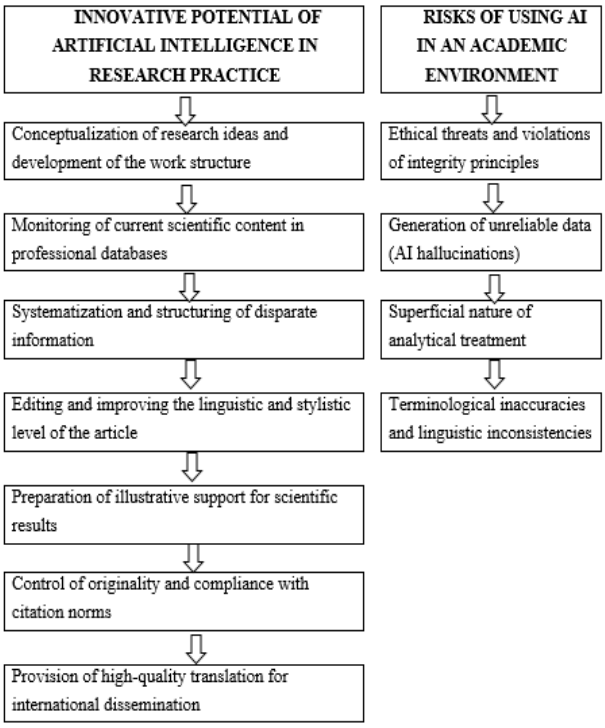


Fig. 1. The potential and limitations of AI application in the academic environment

In view of the above, when determining the expected learning outcomes for disciplines covering issues of academic integrity and scientific ethics, it is advisable to single out certain components. The communicative component is responsible for the ability to freely present and discuss research results, as well as current scientific problems in the field of professional and professional higher education in national and foreign languages. The publication component is responsible for the ability to publish the results of scientific research in domestic and international publications, as well as participate in global projects in compliance with ethical requirements. The research component is responsible for the ability to formulate and test hypotheses based on a

holistic scientific outlook, demonstrating the ability to make optimal decisions and reasoned defense of one's position based on professional ethics.

Submitting a manuscript for review is an act of transferring the results of intellectual work, the evaluation of which directly affects the professional reputation of the authors. In this process, the observance of confidentiality becomes critical: any disclosure of information about the work before its publication is interpreted as a violation of copyright. The formation of learning outcomes within scientific ethics courses should be based on the development of key skills: effective scientific communication in the international space, publication activity in world-class publications, and integration into the global research community. An important aspect is the preparation of a researcher who is able to think critically, solve complex problems and argue his own conclusions, based on a systemic worldview and the principles of academic integrity. The best AI tools for scientists, broken down by key tasks, are presented in Fig. 2.

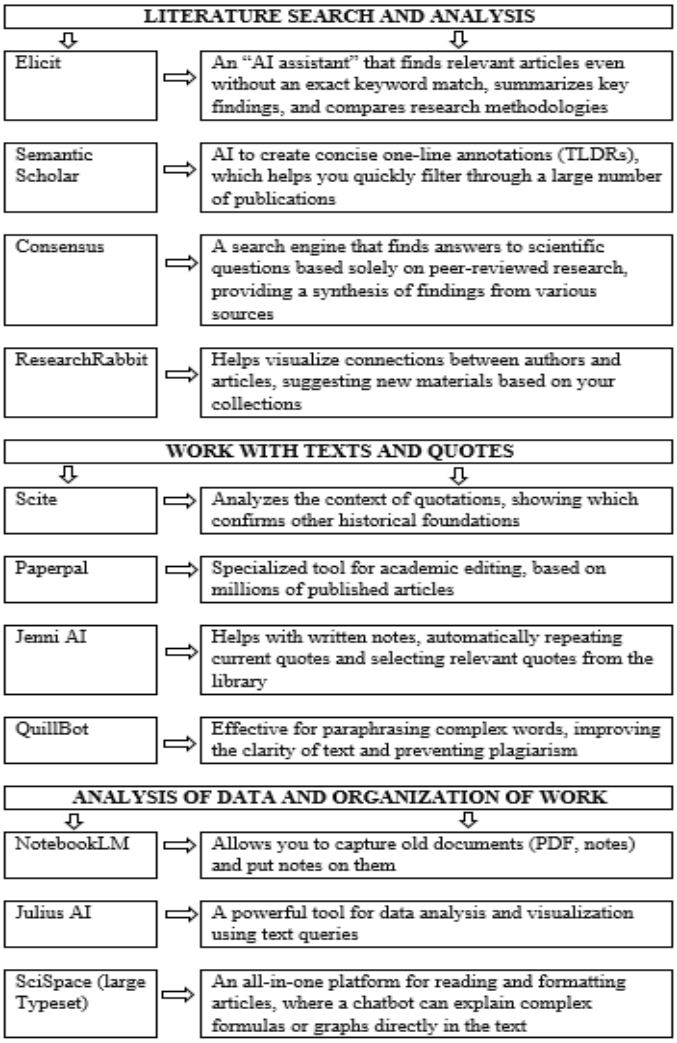


Fig. 2. The best AI tools for scientists, broken down by key tasks

Adherence to the standards of integrity requires the correct design of direct quotations with the use of quotation marks, the preservation of the original meaning of the statement and the use of an ellipsis in place of the missing parts of the sentence. At the same time, it is important to maintain a balance, minimizing the amount of direct citation in favor of the author's analysis. A separate group of violations is manipulation of authorship: co-authorship of persons who did not participate in the work, or use of paid services for writing texts. In modern realities, the covert use of generative artificial intelligence is also equated with academic fraud. In addition, submission of one manuscript to several publications at the same time and concealment of a conflict of interest are unacceptable.

Based on the analysis of functional capabilities and algorithmic base, the following instrumental solutions were identified. A comparative analysis of leading generative artificial intelligence platforms for text derivation is presented in Fig. 3. In view of the above, when designing the expected learning outcomes for disciplines focusing on academic integrity, priority should be given to the formation of the following competencies: the ability to freely present scientific achievements in national and foreign languages, publish them in leading international publications, and join global research projects. Significant scientometric achievements open up broad prospects for international recognition for researchers, in particular through invitations to participate in prestigious conferences, seminars and global research projects.

The main manifestation of the violation of academic ethics is plagiarism, which consists in borrowing someone else's ideas, results or text fragments without proper reference to the original source.

Conclusions

High positions in such ratings not only strengthen the university's reputation in the world, attracting talented personnel, but also open the way to expanding grant funding. In addition, the high scientometric status of the institution stimulates international collaboration, becoming the basis for the creation of joint research centers, exchange programs and strategic partnerships. The university's high scientometric status opens wide horizons for academic mobility, providing students and teachers with priority access to international conferences and internships. This directly correlates with the quality of the educational process: scientists with a significant research output integrate their own developments into curricula. This approach guarantees students access to the most up-to-date knowledge and advanced methodologies, which is the key to their professional growth.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

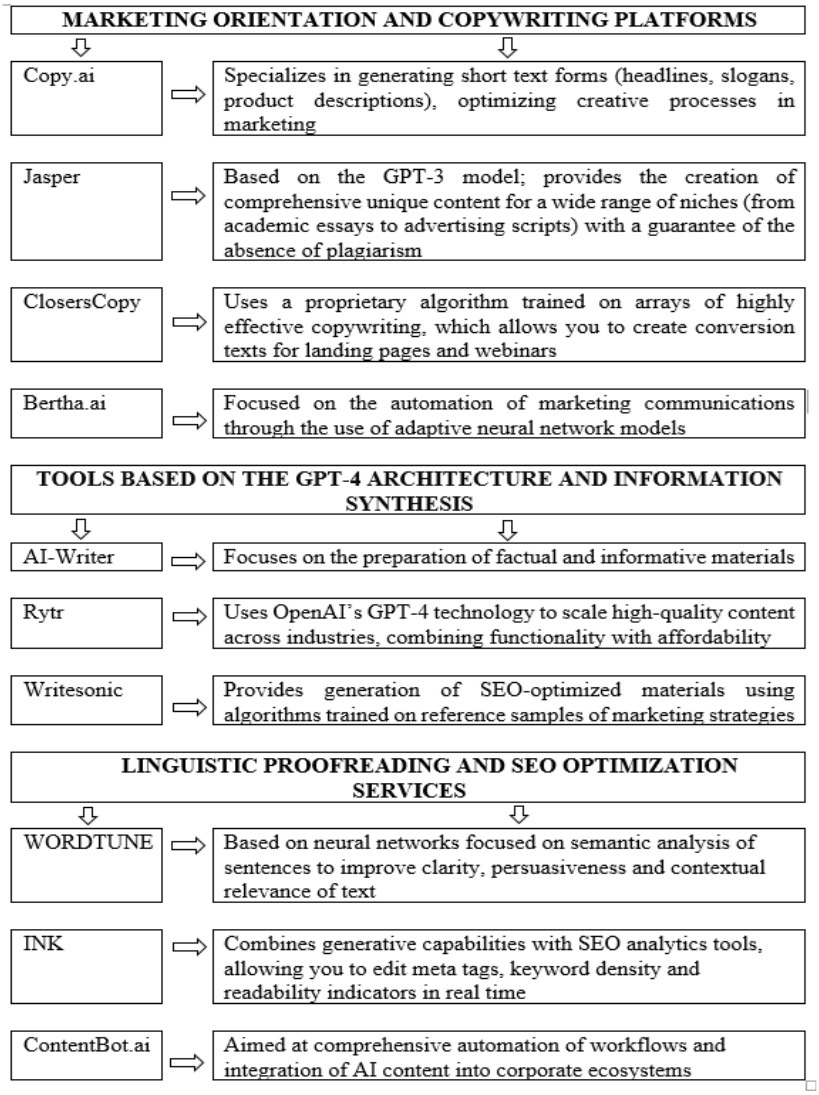


Fig. 3. Comparative analysis of leading generative artificial intelligence platforms

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Гришко В.І., Киричук Б.С. Академічна доброчесність і штучний інтелект: подолання викликів у освітньо-науковій діяльності україни та зарубіжних держав. *Науковий вісник Ужгородського Національного Університету,* 2024. 85(2), 305-310. URL: https://visnyk-juris-uzhnu.com/wp-content/uploads/2024/11/47-1.pdf
2. Тимохіна В.О., Юрченко В.С., Наливайко О.О. Академічна доброчесність VS штучний інтелект: етична дискусія в освітньому просторі. Цифрова трансформація освіти й науки: зб. матеріалів доп. учасн. II Всеукр. наук.-практ. конф. Харків: ХНПУ імені Г.С. Сковороди, 2024, 202–207. https://dspace.hnpu.edu.ua/server/api/core/bitstreams/f5b955c3-9200-450d-8f47-50e2bd71cdfb/content
3. Тодорова І.С. Компоненти академічної доброчесності студентів та умови її формування. *Витоки педагогічної майстерності*, 2019. 24, 199–205. https://doi.org/10.33989/2075-146x.2019.24.194885
4. Уманець О.В. Академічна доброчесність як маркер сучасної освіти: світовий та національний виміри. *Актуальні питання гуманітарних наук*, 2024. 71, 3, 89–96. https://doi.org/10.24919/2308-4863/71-3-14
5. Шестакова С.О., Сухомлинова О.В., Задоріна О.М. Академічна доброчесність: практика дотримання у зарубіжних країнах. *Молодь і ринок*, 2023. 9(217), 87–91. https://repo.snau.edu.ua:8080/xmlui/handle/123456789/12273
6. Tyshchenko D., Franchuk T., Stepashkina K., Karpunin, I. Проектування та розробка системи корпоративного електронного документообігу. *Європейський науковий журнал Економічних та Фінансових інновацій*. 2024. *№ 1*(13). С. 200-207. https://doi.org/10.32750/2024-0119
7. Franchuk T., Tyshchenko D., Desiatko A., Karpunin I. Features of accounting digitalization processes. *Galician economic journal*, 2025, vol. 95, no 1, pp. 61-66. https://doi.org/10.33108/galicianvisnyk_tntu2025.01.061
8. Belfo F., Trigo A. Accounting Information Systems: Traditions and Future Directions. *Accounting Information Systems – Traditionand Future Directions.* 2013. С. 536–546. https://doi.org/10.1016/j.protcy.2013.12.060

9. Капітон А., Сухоребрий О., Ненич Д. Використання мультимодального штучного інтелекту в економіці, освіті, науці та транспорті. *Інформаційні технології та цифрова економіка*. Київ, 2024. 83-85. URL: https://reposit.nupp.edu.ua/item/14791
10. The role of scientometric indicators in the activities of scientists and universities URL: https://spubl.com.ua/uk/blog/the-role-of-scientometric-indicators-in-the-activity-of-scientists-and-universities
11. Publication activity of scientists of pedagogical universities of Ukraine (2010–2020): indicators and factors of influence. URL: https://previous.scientia.report/index.php/archive/article/view/181
12. Committee on Publication Ethics / COPE. URL: www.publicationethics.org
13. Artificial intelligence in the preparation of scientific work: Opportunities and challenges. URL: https://nim.media/articles/shtuchny-intelekt-pri-pidgotovtsi-naukovoyi-roboti-mozhlivosti-ta-vikliki
14. ENRIO 2023 Congress on Research Integrity Practice. Office français de l'intégrité scientifique. URL: https://www.ofis-france.fr/colloque/congres-enrio-2023/
15. The Bucharest Declaration concerning Ethical Values and Principles for Higher Education in Europe. UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000139478

Received (Надійшла) 18.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Капітон Алла Мирославівна – доктор педагогічних наук, професорка, професор кафедри комп'ютерних та інформаційних технологій і систем, Національний університет «Полтавська політехніка ім. Ю. Кондратюка», Полтава, Україна;
Alla Kapiton – Doctor of Pedagogical Sciences, Professor, Professor of the Department of Computer and Information Technologies and Systems, Yuriy Kondratyuk Poltava Polytechnic National University, Poltava, Ukraine;
e-mail: kits_seminar@ukr.net; ORCID Author ID: https://orcid.org/0000-0002-7845-0883.

Франчук Тамара Михайлівна – кандидат економічних наук, доцент кафедри інженерії програмного забезпечення та кібербезпеки; Державний торговельно-економічний університет, Київ, Україна;
Tamara Franchuk - Candidate of Economic Sciences, Associate Professor of the Department of Software Engineering and Cybersecurity; State University of Trade and Economics, Kyiv, Ukraine;
e-mail: Tamara_Franchuk@ukr.net; ORCID Author ID: https://orcid.org/0000-0001-7615-1276.

Тищенко Дмитро Олександрович – кандидат економічних наук, доцент, доцент кафедри інженерії програмного забезпечення та кібербезпеки; Державний торговельно-економічний університет, Київ, Україна;
Dmytro Tyshchenko – Candidate of Economic Sciences, Associate Professor, Associate Professor of the Department of Software Engineering and Cybersecurity; State University of Trade and Economics, Kyiv, Ukraine;
e-mail: tyshchenko_d@knute.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-2193-9012.

Курбанова Ольга Сергіївна – кандидат філологічних наук, доцент, доцент кафедри іноземної філології та перекладу ПрАТ «ВНЗ «Міжрегіональна академія управління персоналом», Київ, Україна;
Olha Kurbanova – PhD in Philology, Associate Professor, Associate Professor of the Department of Foreign Philology and Translation PJSC "Higher Educational Institution Interregional Academy of Personnel Management", Kyiv, Ukraine;
e-mail: folktaleolga@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-4467-9193.

Вплив генеративних технологій штучного інтелекту на дотримання стандартів академічної етики

А. М. Капітон, Т. М. Франчук, Д. О. Тищенко, О. С. Курбанова

Анотація. У статті розглянуто основні вимоги до використання штучного інтелекту (ШІ) у науковій діяльності та потенційні ризики в контексті академічної доброчесності. Виконано порівняльний аналіз провідних платформ генеративного штучного інтелекту для виведення тексту. Встановлено, що використання ШІ для технічних завдань - структурування, оформлення чи редагування власних ідей - не є порушенням етичних норм. Доведено, що у сучасних умовах цифровізації алгоритми штучного інтелекту виступають детермінантою оптимізації стратегій контент-маркерингу. Обгрунтовано, що впровадження систем генеративного ШІ в процеси підготовки текстових матеріалів забезпечує низку стратегічних переваг, серед яких варто виокремлено: оптимізацію ресурсних витрат - мінімізація часових та фінансових інвестицій шляхом автоматизації створення первинних масивів тексту та зниження залежності від зовнішнього копірайтингу; масштабування контентної бази: інтенсифікація виробництва та диверсифікація форматів інформаційного продукту для задоволення запитів цільової аудиторії; підвищення якісних показників: використання прогресивних алгоритмів обробки природної мови сприяє покращенню стилістичної коректності та забезпеченню необхідного рівня унікальності контенту; максимізацію показників залученості: зростання рівня конверсії завдяки високому ступеню персоналізації та відповідності інформаційним потребам споживачів. Встановлено, що наявний інструментарій генеративного штучного інтелекту характеризується значною гетерогенністю. Визначено, що функціональні можливості окремих сервісів диференціюються залежно від специфіки контенту, параметрів цільової аудиторії та економічної доцільності їх впровадження. Здійснено релевантну вибірку найбільш репрезентативних програмних рішень, що продемонстрували високу ефективність протягом останніх років. Проведено порівняльний аналіз технічних характеристик, вартісних моделей, а також оцінку адаптивності зазначених інструментів до специфіки різних мов. Доведено, що динаміка розвитку ринку програмних рішень на базі штучного інтелекту характеризується високою інтенсивністю появи нових інструментів та безперервною модернізацією існуючих сервісів. Процеси вдосконалення функціональних можливостей супроводжуються розширенням лінгвістичного спектра систем, що дає підстави для прогнозування подальшої імплементації та поглибленої підтримки української мови в межах більшої кількості інтелектуальних платформ.

Ключові слова: штучний інтелект, програмне забезпечення, сервіси, наукова діяльність, академічна доброчесність.

УДК 004.77 **doi: 10.26906/SUNZ.2026.2.125**

О. Г. Климко, А. М. Шкурка

Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна

РОЗРОБКА МОДЕЛІ ІНФОРМАЦІЙНОЇ СИСТЕМИ АВТОМАТИЗАЦІЇ УПРАВЛІННЯ СПОРТИВНИМИ ЗАХОДАМИ

Анотація. Стаття присвячена розробці моделі інформаційної системи автоматизації управління спортивними заходами sportevent.ua та аналізу її роботи. Актуальність теми обумовлена стрімким зростанням кількості спортивних змагань в Україні та у світі, ускладненням організаційних процесів і відсутністю доступних, безкоштовних україномовних платформ для організаторів регіонального рівня. За оцінками, глобальний ринок спортивних заходів у 2024 р. перевищив 452 млрд USD і продовжує зростати. В умовах воєнного часу питання фінансової доступності програмних рішень набуває особливої гостроти для вітчизняних спортивних організацій. У роботі проведено порівняльний аналіз провідних систем управління спортивними заходами – ClearEvent, TeamSnap, RedPodium, а також відкритих проєктів (OpenEvent, SportsVerse) – за ключовими параметрами: функціональністю, доступністю, локалізацією, технологічним стеком та відкритістю вихідного коду. Встановлено, що жодна з розглянутих систем не забезпечує комплексної підтримки україномовного середовища та потреб аматорських організацій. Представлено архітектуру розробленої системи на стеку Node.js / Express / EJS / PostgreSQL з хмарним розгортанням на платформі Render та автоматичним CI/CD через GitHub. Детально описано структуру проєкту: контролери, маршрути, сервіси, моделі, SQL-схема та статичні ресурси. Досліджено шість реалізованих функціональних модулів: управління заходами (повний CRUD-цикл), реєстрація учасників, новинна стрічка, фотогалерея, адаптивна мобільна версія та повноцінний модуль аналітики і звітності з дашбордами, графіками розподілу заходів за дисциплінами, динамікою реєстрацій по місяцях та таблицею лідерів. Наведено три порівняльні таблиці за критеріями платформ, технологічного стеку та функціональних можливостей. Проілюстровано інтерфейс адміністративної панелі та головної сторінки системи. Виявлено переваги рішення – мінімалістичний стек, відкритий код, безкоштовне розгортання, україномовний інтерфейс, вбудована аналітика – та окреслено вектори подальшого розвитку: турнірні сітки, REST API, RBAC, нативний застосунок та розширена аналітика з PDF-звітністю.

Ключові слова: інформаційна система, веб-застосунок, спортивний захід, Node.js, PostgreSQL, автоматизація, управління змаганнями, порівняльний аналіз, open source.

Вступ

Постановка проблеми. Глобальний ринок спортивних заходів у 2024 р. оцінювався у 452,79 млрд USD і, за прогнозами, зростатиме зі CAGR 7,2 % до 2030 р. [1]. Зростання охоплює як міжнародні мегазаходи – Олімпійські ігри, чемпіонати світу, – так і регіональні аматорські турніри, кількість яких в Україні щороку збільшується. За оцінками Міністерства молоді та спорту України, у 2023 р. в країні відбулося понад 12 000 офіційних спортивних заходів різного рівня. Кожен з них потребує вирішення комплексу організаційних задач: формування заявок, акредитація учасників, складання розкладів поєдинків, управління суддівським складом, фіксація результатів і підготовка підсумкової документації.

Разом зі зростанням індустрії загострюються й організаційні виклики: координація сотень учасників, ведення протоколів, формування розкладів і своєчасне інформування – завдання, які й досі вирішуються частиною організацій за допомогою електронних таблиць та ручної документації. Такий підхід породжує системні проблеми: дублювання даних, людські помилки при введенні результатів, затримки у публікації підсумків та відсутність єдиного інформаційного простору для учасників, суддів і вболівальників. Дослідження SafetyCulture [6] фіксують, що організатори витрачають у середньому від 30 до 60 % загального часу підготовки заходу саме на рутинні адміністративні операції, які можуть бути автоматизовані.

Для регіональних і аматорських федерацій України ситуація ускладнюється відсутністю доступних україномовних платформ: закордонні рішення коштують від 12–50 USD на місяць, не мають локалізації та не враховують специфіки вітчизняного спорту. Правові та мовні вимоги до програмного забезпечення, що використовується в Україні, також орієнтують організації на пошук україномовних інструментів. Крім того, в умовах воєнного часу питання фінансової доступності програмних рішень набуває особливої гостроти: бюджети більшості спортивних федерацій регіонального рівня суттєво скорочені. Саме ця прогалина обумовлює актуальність розробки відкритої, безкоштовної україномовної веб-системи sportevent.ua.

Аналіз останніх досліджень і публікацій. Теоретичні засади управління спортивними організаціями закладено у роботах Chappelet J.-L. і Bayle E. [2], які систематизували принципи стратегічного та результативного менеджменту олімпійських структур. Практичні аспекти планування та проведення спортивних заходів різного масштабу детально розглянуто у підручнику Masterman G. [3], що став стандартом галузі. Технологічні засади розробки інформаційних систем – принципи проєктування, тестування і супроводу – розкрито у класичній праці Pressman R. S. і Maxim B. R. [4]. Питання мікросервісної архітектури, що набула поширення у сучасних веб-системах, ґрунтовно проаналізовано Newman S. [5].

На рівні програмних рішень ринок пропонує низку спеціалізованих платформ. Eventpipe [9] та SafetyCulture [6] систематизують провідні комерційні сервіси: ClearEvent, TeamSnap, RedPodium, AudienceView. Серед відкритих рішень вирізняється OpenEvent (FOSSASIA) [7] – зрілий open-source проєкт із підтримкою REST API, мобільних клієнтів і багатомовного інтерфейсу, що активно розвивається спіль-

нотою. Осадченко Т. [10] досліджує сучасний стан цифровізації у сфері фізичної культури та спорту в Україні, визначаючи ключові напрями та виклики впровадження цифрових технологій у спортивній галузі. Водночас спеціалізованих досліджень, присвячених розробці веб-систем саме для управління спортивними заходами в умовах України, недостатньо, що підтверджує наукову новизну представленої роботи.

Мета роботи – розробити модель інформаційної системи sportevent.ua та провести аналіз її функціонування, провести її порівняльний аналіз з провідними аналогами за функціональними, технологічними та доступністними критеріями, виявити переваги й обмеження поточної версії, а також сформулювати науково обґрунтовані напрями подальшого розвитку.

Основний матеріал

Аналіз існуючих рішень. На ринку програмного забезпечення для управління спортивними заходами представлено широкий спектр платформ – від вузькоспеціалізованих до універсальних. SafetyCulture [6] виокремлює: ClearEvent – платформа з реєстрацією, продажем квитків і плануванням розкладів; RedPodium – орієнтована на реєстрацію учасників гонок і змагань; TeamSnap – зосереджена на командному менеджменті та комунікаціях; AudienceView – система квитків для кампусних і спортивних подій. Серед відкритих проєктів GitHub виявлено OpenEvent (FOSSASIA) – найбільш функціонально повне рішення з підтримкою REST API, мобільних клієнтів і багатомовного інтерфейсу; SportsVerse – навчальний проєкт на PHP/MySQL [7]. Попри різноманіття варіантів, жоден не пропонує україномовного інтерфейсу та безкоштовного хмарного розгортання одночасно. Порівняльна характеристика платформ наведена у табл. 1.

Аналіз таблиці 1 демонструє, що sportevent.ua є єдиною відкритою україномовною платформою у представленому переліку. Комерційні рішення ClearEvent і TeamSnap забезпечують найширший функціонал, однак їхня вартість і відсутність локалізації роблять їх недоступними для більшості вітчизняних організаторів. OpenEvent – потужна відкрита платформа, але без підтримки мови інтерфейсу.

Таблиця 1 – **Порівняльний аналіз платформ управління спортивними заходами**

Система	Тип	Лок.	Стек	Open Source	Ціна/міс.
ClearEvent	Комерційна SaaS	EN	Власний	Ні	від 50 $
TeamSnap	Комерційна SaaS	EN	Власний	Ні	від 12 $
RedPodium	Комерційна SaaS	EN	Власний	Ні	% від реєстр.
OpenEvent	Open Source	Multi	Python / PostgreSQL	Так	0 $
SportsVerse	Open Source	EN	PHP / MySQL	Так	0 $
sportevent.ua	Open Source	UA	Node.js / PostgreSQL	Так	0 $

Архітектура та технологічний стек системи sportevent.ua. Систему реалізовано за класичною MVC-архітектурою на базі Node.js / Express із серверним рендерингом шаблонів EJS. PostgreSQL 15 обрано як основну СУБД завдяки підтримці транзакцій, потужній системі запитів і відкритому коду. Замість ORM-фреймворку використано нативний pg-драйвер, що зменшує накладні витрати та спрощує налагодження. Розгортання здійснено на безкоштовній PaaS-платформі Render з підтримкою автоматичного CI/CD при пуші в GitHub [8]. Репозиторій організовано відповідно до принципу розподілу відповідальностей: config/ – конфігурація; controllers/ – обробники HTTP-запитів; routes/ – маршрутизація REST; services/ – ізольована бізнес-логіка; models/ – взаємодія з PostgreSQL; sql/ – DDL-схема; views/ – EJS-шаблони; public/ – статичні ресурси. Детальний опис технологічного стеку наведено у табл. 2.

Таблиця 2 – **Технологічний стек системи sportevent.ua**

Компонент	Технологія	Призначення
Веб-сервер	Node.js 18 + Express	Маршрутизація HTTP-запитів, REST-обробники
Шаблонізатор	EJS (Embedded JS)	Серверний рендеринг HTML-сторінок
База даних	PostgreSQL 15	Збереження заходів, учасників, новин, медіа
ORM / запити	Нативний pg-драйвер	SQL-запити без надлишкових абстракцій
Хмара	Render.com (PaaS)	Безкоштовне розгортання, CI/CD з GitHub
Конфігурація	.env + config/db.js	Змінні оточення, рядок підключення до БД
Статика	public/ (CSS, JS, img)	Стилі, клієнтські скрипти, зображення
SQL-схема	sql/ (DDL-файли)	Таблиці, обмеження, індекси, міграції
Seed-дані	seed_gallery/news .js	Початкове наповнення тестовими даними

Функціональні можливості системи. У результаті аналізу вихідного коду репозиторію sportevent.ua [8] визначено шість реалізованих модулів. Модуль управління заходами забезпечує повний CRUD-цикл: створення, редагування, публікацію та перегляд подій. Модуль реєстрації учасників реалізує

прив'язку спортсменів до заходів зі збереженням контактних даних та категорій участі. Модуль новинної стрічки дозволяє публікувати новини клубу з підтримкою зображень. Модуль фотогалереї забезпечує завантаження та демонстрацію фотоматеріалів. Адаптивна мобільна версія забезпечує доступ до системи зі смартфонів і планшетів без встановлення окремого застосунку. Модуль аналітики та звітності входить до адміністративної панелі й містить: зведені показники по заходах та користувачах; статистику реєстрацій; показники контенту; графік розподілу заходів за дисциплінами; динаміку реєстрацій по місяцях; рейтинг популярних заходів та таблицю лідерів з фільтрацією за часовим діапазоном.

Функціональне порівняння з ClearEvent та OpenEvent наведено у табл. 3 за 13 критеріями.

Таблиця 3 – **Порівняння функціональних можливостей систем**

Функція	Clear-Event	Open-Event	Sport-event.ua
Реєстрація учасників	Так	Так	Так
Управління заходами	Так	Так	Так
Новинна стрічка	Ні	Так	Так
Фотогалерея	Ні	Так	Так
Мобільна версія	Так	Ні	Так
Аналітика / звітність	Так	Так	Так
Турнірні сітки / bracket	Так	Так	Ні
REST API	Так	Так	Ні
Ролі користувачів (RBAC)	Так	Так	Частково
Нативний мобільний застосунок	Так	Ні	Ні
Україномовний інтерфейс	Ні	Ні	Так
Безкоштовне розгортання	Ні	Так	Так
CI/CD з GitHub	Ні	Так	Так

Переваги розробленого рішення. Табл. 3 підтверджує, що sportevent.ua реалізує базовий організаційний цикл, достатній для проведення аматорських змагань. Порівняно з комерційними аналогами система вирізняється рядом переваг.

По-перше, мінімалістичний стек не вимагає складного DevOps-середовища і може бути розгорнутий безкоштовно на Render.com.

По-друге, відкритий вихідний код дозволяє адаптувати систему під специфічні потреби організації без витрат на ліцензію.

По-третє, україномовний інтерфейс знижує бар'єр входу для вітчизняних організаторів і відповідає чинним вимогам щодо мови програмного забезпечення.

По-четверте, структура проєкту є зрозумілою для розробників початкового рівня, що сприяє залученню волонтерів і студентів до розвитку системи.

По-п'яте, система має адаптивну мобільну версію, що забезпечує зручний доступ зі смартфонів без встановлення окремого застосунку.

По-шосте, вбудований модуль аналітики надає організаторам наочні дашборди з ключовими метриками заходів, реєстрацій, контенту та рейтингом учасників з фільтрацією за часовим діапазоном.

По-сьоме, автоматичне CI/CD розгортання через Render і GitHub суттєво скорочує час між внесенням змін і їх появою на продакшн-сервері.

Виявлені обмеження поточної версії. Попри зазначені переваги, поточна версія sportevent.ua має обмеження, характерні для MVP-рішень. Найбільш критичним є відсутність модуля турнірних сіток (жеребкування, bracket-система, swiss-система), що обмежує застосування для управління безпосередньо змагальними процесами. Система не надає REST API, що унеможливлює інтеграцію з зовнішніми сервісами та федераціями. Відсутня повноцінна ролева модель доступу: система частково розрізняє ролі, однак не реалізує повноцінного RBAC (роль судді, учасника, адміністратора). Порівняння з ClearEvent і OpenEvent [6, 7] підтверджує, що ці можливості є базовими для зрілих систем управління подіями.

Висновки

Розроблено модель інформаційної системи sportevent.ua для автоматизації управління спортивними заходами та проведено аналіз її функціонування.

У результаті здійсненої роботи виконано порівняльний аналіз провідних платформ (ClearEvent, TeamSnap, RedPodium, OpenEvent, SportsVerse), за результатами якого встановлено, що sportevent.ua займає унікальну нішу першої відкритої україномовної системи для організаторів регіонального рівня – єдиної, яка поєднує безкоштовність, україномовний інтерфейс та хмарне розгортання одночасно.

Описано MVC-архітектуру системи на стеку Node.js / Express / EJS / PostgreSQL з хмарним розгортанням на Render (таблиця 2). Ключовою перевагою обраного стеку є мінімалізм: відсутність надлишкових фреймворків і ORM знижує поріг входу для розробників-волонтерів і спрощує супровід. Задокументовано шість реалізованих функціональних модулів: управління заходами, реєстрація учасників, новинна стрічка, фотогалерея, адаптивна мобільна версія та повноцінний модуль аналітики з дашбордами, графіками і таблицею лідерів.

Функціональне порівняння (табл. 3) підтвердило базову достатність системи для аматорських і регіональних змагань: з 13 оцінюваних критеріїв sportevent.ua задовольняє 8, поступаючись зрілим платформам у відсутності турнірних сіток, REST API та повноцінного RBAC. Водночас саме за трьома критеріями (україномовний інтерфейс, адаптивна мобільна версія, безкоштовне розгортання з CI/CD) система перевершує або дорівнює комерційним аналогам.

Визначено три горизонти розвитку. У короткостроковому: JWT-автентифікація, RBAC, REST API, модуль турнірних сіток.

У середньостроковому: розвиток мобільної версії до PWA, інтеграція LiqPay / Fondy, WebSocket для live-результатів, розширення аналітики. У довгостроковому: інтеграція з федераціями, ML-прогнозування, мультитенантність.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного

дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи

.

Список літератури

1. Cvent Blog. Mastering Sports Event Management: Trends, Challenges, and Strategies. URL: https://www.cvent.com/en/blog/events/sports-event-management.
2. Chappelet J.-L., Bayle E. Strategic and Performance Management of Olympic Sport Organisations. Human Kinetics, 2020. 272 p. URL: https://books.google.com/books?id=_nXqAOmCkQkC.
3. Masterman G. Strategic Sports Event Management. 4th ed. Routledge, 2022. 368 p. DOI: https://doi.org/10.4324/9781003046257.
4. Pressman R. S., Maxim B. R. Software Engineering: A Practitioner's Approach. 9th ed. McGraw-Hill, 2020. 976 p. URL: https://www.mheducation.com/highered/product/software-engineering-a-practitioners-approach-pressman.html.
5. Newman S. Building Microservices. 2nd ed. O'Reilly Media, 2021. 616 p. URL: https://www.oreilly.com/library/view/building-microservices-2nd/9781492034018/.
6. SafetyCulture. Top Sports Event Management Software of 2026. URL: https://safetyculture.com/apps/sports-event-management-software.
7. FOSSASIA. OpenEvent Server: Event Management System. GitHub. URL: https://github.com/fossasia/open-event-server.
8. Andriy-Shkurka. sportevent-ua: Web System for Sports Event Management. GitHub. URL: https://github.com/Andriy-Shkurka/sportevent-ua.
9. Eventpipe. 10+ Best Sports Event Management Software & Apps. URL: https://eventpipe.com/blog/sports-event-management-software.
10. Осадченко Т. Сучасний стан цифровізації у сфері фізичної культури та спорту в Україні. Physical Culture and Sport: Scientific Perspective. 2023. № 2. С. 103–108. DOI: https://doi.org/10.31891/pcs.2023.2.14.

Received (Надійшла) 25.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Климко Олена Генріхівна – старший викладач кафедри комп'ютерних та інформаційних технологій і систем Національного університету «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Olena Klymko – Senior Lecturer of the Department of Computer and Information Technologies and Systems, National University "Yuri Kondratyuk Poltava Polytechnic", Poltava, Ukraine;
e-mail: klimkoelena63@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-2212-068X.

Шкурка Андрій Миколайович – студент, Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Andriy Shkurka – Student, National University "Yuri Kondratyuk Poltava Polytechnic", Poltava, Ukraine;
e-mail: marktorentino@gmail.com; ORCID Author ID: https://orcid.org/0009-0000-9518-3572.

Development of an information system model for sports event management automation

Olena Klymko, Andriy Shkurka

Abstract. This article focuses on the development of a model for the sportevent.ua sports event management information system and an analysis of its operation. The relevance of this topic stems from the rapid increase in the number of sporting competitions in Ukraine and worldwide, the growing complexity of organisational processes, and the lack of accessible, free Ukrainian-language platforms for regional-level organisers. According to estimates, the global market for sporting events exceeded USD 452 billion in 2024 and continues to grow. In wartime conditions, the issue of the financial accessibility of software solutions becomes particularly acute for domestic sports organisations. This paper presents a comparative analysis of leading sports event management systems – ClearEvent, TeamSnap, RedPodium – as well as open-source projects (OpenEvent, SportsVerse) – based on key parameters: functionality, accessibility, localisation, technology stack and open-source code. It has been established that none of the systems under consideration provides comprehensive support for the Ukrainian-language environment and the needs of amateur organisations. The architecture of the developed system is presented, built on the Node.js / Express / EJS / PostgreSQL stack with cloud deployment on the Render platform and automated CI/CD via GitHub. The project structure is described in detail: controllers, routes, services, models, SQL schema and static resources. Six functional modules have been developed: event management (full CRUD cycle), participant registration, news feed, photo gallery, responsive mobile version, and a comprehensive analytics and reporting module featuring dashboards, charts showing the distribution of events by discipline, monthly registration trends, and a leaderboard. Three comparative tables are provided based on platform, technology stack and functional capabilities. The interface of the admin panel and the system's home page are illustrated. The solution's advantages are identified – a minimalist stack, open-source code, free deployment, a Ukrainian-language interface, and built-in analytics – and directions for further development are outlined: tournament brackets, REST API, RBAC, a native app, and advanced analytics with PDF reporting.

Keywords: information system, web application, sports event, Node.js, PostgreSQL, automation, competition management, comparative analysis, open source.

УДК 004.932.72:004.932.2:616-006 **doi: 10.26906/SUNZ.2026.2.129**

О. Є. Личкатий, А. І. Поворознюк

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

МУЛЬТИФРАКТАЛЬНИЙ АНАЛІЗ МАМОГРАФІЧНИХ ЗОБРАЖЕНЬ

Анотація. Розглядається задача підвищення ефективності комп’ютерного аналізу мамографічних зображень для підтримки раннього виявлення патологічних змін молочної залози. Показано, що обмеження традиційної мамографії, зумовлені щільністю тканин, наявністю шумів та артефактів, а також суб’єктивністю інтерпретації, обґрунтовують необхідність використання кількісних методів, здатних забезпечити більш об’єктивну оцінку структурних особливостей зображення. У роботі розглянуто підхід, спрямований на отримання локально чутливих показників складності тканин, що підсилює виявлення малоконтрастних або слабко виражених змін. Метою дослідження є розробка методики, яка поєднує автоматизовану сегментацію та мультифрактальний аналіз із поданням результатів у вигляді карт локальних фрактальних розмірностей (heatmap). Сегментація виконується шляхом комбінування нейронної мережі архітектури U-Net із порогуванням за Otsu після застосування медіанної та гаусівської фільтрації, що дає змогу формувати більш стабільні маски та зменшувати вплив шумових компонентів. Ключовим елементом запропонованої методики є побудова карт локальних фрактальних розмірностей методом ковзного вікна, який забезпечує безперервну просторову оцінку структурної неоднорідності тканин. Показано, що у зонах патологічних змін виникають локальні відхилення фрактальної розмірності відносно навколишньої паренхіми, що підвищує інтерпретованість та може слугувати індикатором потенційних областей інтересу. Проаналізовано обмеження класичного методу box-counting щодо градаційних зображень, зокрема залежність від бінаризації та втрату інформації про рівні інтенсивності. Аргументовано доцільність переходу до differential box-counting, який працює у просторі інтенсивностей та забезпечує вищу чутливість до текстурних варіацій. Експериментальні результати, отримані на зображеннях із баз MIAS та DDSM, підтверджують перспективність запропонованого підходу та його потенційну корисність для подальшого розвитку систем комп’ютерної підтримки прийняття рішень у мамографії.

Ключові слова: мамографія; мультифрактальний аналіз; локальні фрактальні розмірності; box-counting; differential box-counting; U-Net; sliding-window, сегментація зображень; heatmap.

Вступ

Рак молочної залози залишається однією з провідних причин смертності серед жінок у світі та становить одну з найсерйозніших проблем сучасної онкології. За даними Всесвітньої організації охорони здоров’я, щороку у світі виявляється понад 2,3 млн нових випадків цього захворювання, і приблизно 700 тис. жінок помирають від його наслідків [1]. Показники захворюваності демонструють тенденцію до зростання, що пояснюється як старінням населення, так і постійним удосконаленням методів скринінгу, завдяки яким виявляється більше ранніх випадків. Попри значні досягнення в галузях діагностики, лікування й персоналізованої медицини, проблема раннього виявлення пухлинних змін у молочній залозі все ще залишається надзвичайно актуальною.

Мамографія є базовим методом скринінгу, який рекомендовано застосовувати для жінок певних вікових груп у більшості країн світу. Разом із тим, її діагностична точність істотно залежить від низки чинників: технічних особливостей обладнання, кваліфікації лікаря-рентгенолога, інтерпретаційної варіативності та індивідуальних анатомічних характеристик пацієнтки. Відомо, що навіть за дотримання всіх протоколів якість візуалізації може істотно змінюватися, а деякі патологічні зміни є вкрай складними для виявлення через слабку контрастність або особливості розташування утворення. Тому активний розвиток комп’ютерних методів аналізу мамографічних зображень спрямований на автоматизацію процесу діагностики та зменшення ймовірності людської помилки [2, 3].

Статистично кожна восьма жінка протягом життя ризикує зіткнутися з раком молочної залози [1]. В Україні лише у 2023 році зареєстровано понад 15 тисяч нових випадків захворювання, і близько третини з них виявлено на пізніх стадіях, коли лікування є менш результативним. Проблема ускладнюється тим, що приблизно 20–30% випадків раку залишаються непоміченими під час мамографії через високу щільність тканини або артефакти зображення [1]. Водночас велика частка хибнопозитивних результатів спричиняє необхідність додаткових обстежень, призводить до зайвих фінансових витрат, психологічного дискомфорту пацієнток та збільшення навантаження на медичну інфраструктуру.

Складність інтерпретації мамограм посилюється суб’єктивністю оцінок лікарів: численні дослідження демонструють значні розбіжності між висновками різних рентгенологів, особливо в ситуаціях низького контрасту зображення або неоднозначної морфології утворень [4]. Така міжспостерігачева варіативність створює серйозні ризики для точності діагностики та вказує на потребу у впровадженні об’єктивних, математично обґрунтованих методів аналізу мамографічних даних.

Одним із перспективних напрямів розвитку таких методів є використання фрактального та мультифрактального аналізу зображень. Ці підходи дозволяють оцінювати ступінь складності тканинних структур, виявляти відхилення від нормальних текстурних патернів та фіксувати мікроструктурні аномалії, що можуть бути індикаторами ранніх патологічних процесів [5]. Фрактальні методи є особливо цінними в аналізі півтонових зображень, оскільки вони здатні виявляти тонкі зміни текстури, які не завжди очевидні для традиційних методів комп’ютерного зору.

Сучасні тенденції в медичній візуалізації характеризуються стрімким збільшенням обсягів даних, що потребує ефективних інструментів їх обробки та

автоматичного аналізу. Саме тому набуває популярності поєднання алгоритмів глибинного навчання з методами фрактального аналізу. Такі комбіновані методики здатні поєднувати сильні сторони обох підходів: потужність сегментації та класифікації нейронних мереж і здатність фрактальних методів виявляти приховану текстурну інформацію, зокрема мікрокальцифікації та інші дрібні ознаки раннього онкологічного процесу [4].

У межах цього дослідження робиться спроба поєднати класичні методи цифрової обробки, фрактальний підхід та елементи глибинного навчання з метою підвищення точності, стабільності та інтерпретованості комп'ютерного аналізу мамографічних зображень.

Метою роботи є підвищення ефективності комп'ютерного аналізу мамограм шляхом застосування мультифрактального підходу та подальшої візуалізації результатів у вигляді карт локальних фрактальних розмірностей. Такий підхід покликаний забезпечити більш детальне розуміння просторової структури тканин грудної залози та покращити здатність методів автоматичного аналізу виявляти складні текстурні відхилення, що можуть свідчити про ранні патологічні зміни. Додатковою метою є розроблення інтерпретованої форми подання результатів, придатної для подальших кількісних і візуальних оцінок.

Для досягнення поставленої мети визначено такі завдання:

- розробка методики попередньої обробки та сегментації мамограм, яка забезпечує стабільне виділення області грудної залози та мінімізацію впливу шумів і артефактів;
- впровадження нейронної мережі для комбінованої сегментації разом із класичним методом Otsu з метою підвищення точності та повноти виділення меж аналізованої області;
- реалізація алгоритму побудови карт локальних фрактальних розмірностей за методом ковзного вікна (sliding window), що дозволяє отримувати детальний просторовий розподіл фрактальних характеристик;
- візуалізація результатів у вигляді heatmap та аналіз отриманих карт на прикладах з різними клінічними випадками, включаючи мамограми з патологією та без неї;
- визначення обмежень методу та перспектив його подальшого розвитку, зокрема у напрямку застосування диференційного box-counting для півтонових зображень та вдосконалення параметрів ковзного вікна для підвищення стабільності фрактальних оцінок.

Мультифрактальна обробка мамограм

Для реалізації поставленої задачі автором створено клієнт-серверний веб-додаток, який слугує програмною платформою для проведення експериментів та перевірки алгоритмів обробки мамографічних зображень. Серверна частина, реалізована мовою Python, відповідає за обчислення фрактальних параметрів, виконання процедур обробки, попередньої фільтрації та сегментації зображень [6, 7]. Клієнтська частина (HTML, CSS, JavaScript) забезпечує інтерфейс взаємодії з користувачем, включаючи завантаження даних, перегляд проміжних результатів та візуалізацію фрактальних карт. Отримані результати зберігаються у реляційній базі даних, тоді як великі зображення та проміжні файли — у файловій системі сервера. Важливо підкреслити, що архітектура програмного забезпечення не є предметом дослідження, а виступає лише технічним інструментом для реалізації обчислювальної частини роботи та забезпечення відтворюваності експериментів.

На попередньому етапі роботи було розроблено приладовий прототип системи для аналізу мамограм з використанням класичних алгоритмів фрактального аналізу — Box-Counting та Differential Box-Counting [5]. За його допомогою були отримані перші карти складності, що відображали глобальний розподіл фрактальної розмірності на зображеннях. Однак виявилося, що такі карти мають низьку чутливість до локальних змін, які є найбільш важливими під час пошуку невеликих пухлин чи мікрокальцифікацій. Крім того, інформативність глобальної фрактальної розмірності виявилася обмеженою: навіть помітні патологічні структури не завжди давали достатньо виражену різницю у значенні показника. Результати також показали нечітку кореляцію між наявністю або відсутністю новоутворень та обчисленими значеннями D, що підтвердило необхідність відмови від використання єдиної глобальної метрики й переходу до повноцінного мультифрактального підходу, орієнтованого на локальний аналіз.

Попередня обробка зображень є важливою частиною всієї методики, оскільки саме на цьому етапі формується якісна основа для подальшої сегментації та фрактального аналізу. Послідовність обробки включає застосування медіанного та гаусівського фільтрів для пригнічення шумів та стабілізації фону. Медіанний фільтр використовується насамперед для усунення імпульсного шуму, при цьому добре зберігаються контури утворень і тканинних структур. Після цього застосовується гаусівський фільтр, який приглушує високочастотні перешкоди і вирівнює глобальний фон зображення, що позитивно впливає на стабільність подальшої бінаризації. Завершальним етапом попередньої обробки є порогування Otsu — автоматична процедура, яка поділяє зображення на тканину та фон, створюючи первинну маску грудної залози. Важливим є той факт, що Otsu застосовується вже після згладження, завдяки чому вибір порога стає більш стабільним, а отримані маски — цілісними, зі зменшеною кількістю дрібних артефактів. Така послідовність операцій дає змогу формувати маску високої якості, придатну як для подальшої сегментації, так і для виключення фону під час розрахунку фрактальних характеристик.

Важливим кроком у розвитку системи стало впровадження нейронної мережі архітектури U-Net [4], яка є де-факто стандартом для сегментації медичних зображень. Вхідні мамограми масштабуються до розміру 512×512 пікселів, що дозволяє ефективно застосовувати структуру downsampling/upsampling та одночасно зберігати баланс між глобальним контекстом і локальною деталізацією. Мережу навчено на датасетах MIAS [2] та DDSM [3] з використанням Dice-функції втрат, що дозволило оптимізувати

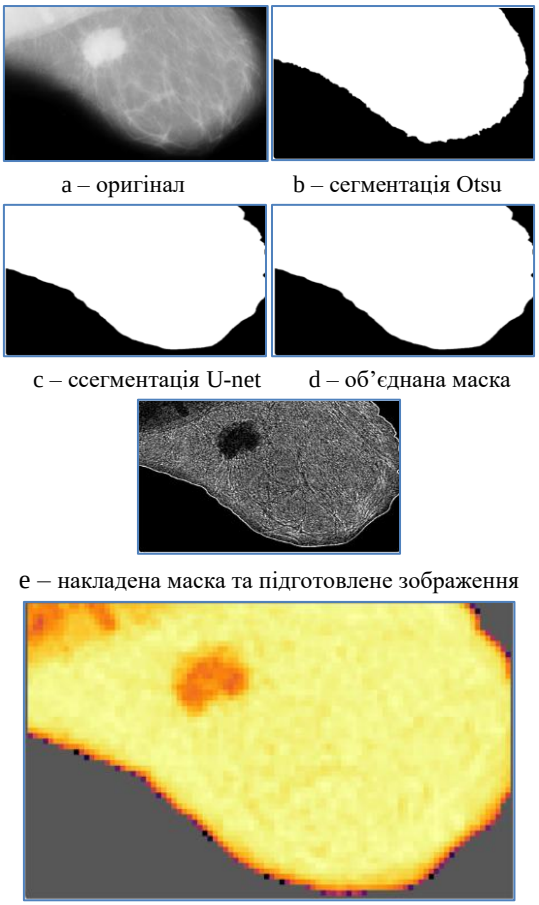
модель саме для задачі точного виділення меж грудної залози. На валідаційній вибірці отримано середній коефіцієнт Dice ≈ 0,85, що відповідає типовим значенням для медичних задач подібного рівня. Для підвищення стійкості та точності сегментації результати нейромережі комбінуються з маскою, отриманою методом Otsu, через логічну операцію АБО. Такий підхід дає змогу компенсувати локальні недоліки сегментації U-Net і зменшити кількість пропусків у ділянках зі складною текстурою, які спостерігалися в попередньому прототипі [6, 7]

Ключовим досягненням у межах виконаної роботи стало застосування методу ковзного вікна (sliding window) [8], який дозволяє значно підвищити інформативність фрактального аналізу завдяки локальній оцінці структурних характеристик зображення. Зображення послідовно покривається перекривними вікнами фіксованого розміру з заданим кроком, для кожного з яких обчислюється локальна фрактальна оцінка. Одержане значення прив'язується до центру або всієї області відповідного вікна; повторюючи цю процедуру по всій площині зображення, формується двовимірна карта значень, що відображає просторову змінність показника. Завдяки перекриттю вікон результат подається у вигляді безперервної heatmap, яка забезпечує цілісне уявлення про локальні відмінності структури тканин та виявляє особливості, непомітні за глобальними параметрами. Застосування Box-Counting [5] у межах кожного вікна дає можливість отримати локальні значення фрактальної розмірності D, які виявляються значно чутливішими до текстурної неоднорідності, ніж глобальний показник, що усереднює інформацію по всьому зображенню. Такий підхід дозволяє безпосередньо порівнювати внутрішні ділянки структури, межі утворень і фонові області, підсилюючи інтерпретованість результатів та полегшуючи візуальне виділення зон потенційної патології.

З метою подальшого підвищення чутливості розглядається перехід до Differential Box-Counting (DBC). Цей метод працює без жорсткої бінаризації, у тривимірному просторі (x, y, інтенсивність), що дозволяє зберігати інформацію про рівні сірого та враховувати тонкі градаційні зміни у зображенні на всіх етапах аналізу. На відміну від класичного підходу, DBC точніше відображає тонкі варіації текстури та менш чутливий до вибору порогових значень під час попередньої обробки. Очікується, що застосування DBC забезпечить більш стабільні результати між різними наборами даних та підвищить дискримінативність heatmap, особливо в умовах слабоконтрастних структур. При цьому загальний sliding-window підхід, логіка локальних оцінок і візуалізаційні принципи залишаються незмінними, що спрощує інтеграцію нового методу в існуючий аналітичний процес.

Обговорення результатів

Приклад 1. Мамограма з пухлиною. Етапи обробки показані на рис. 1, продемонстровано повну послідовність обробки зображення — від сирої мамограми до отриманої мультифрактальної карти. На рис. 1, f у ділянці пухлини спостерігаються стійкі локальні відхилення фрактальної розмірності D відносно оточуючих здорових ділянок, причому середнє значення D у межах патологічної зони виявляється підвищеним або зміненим у порівнянні з фоновими областями.



a – оригінал b – сегментація Otsu

c – ссегментація U-net d – об'єднана маска

e – накладена маска та підготовлене зображення

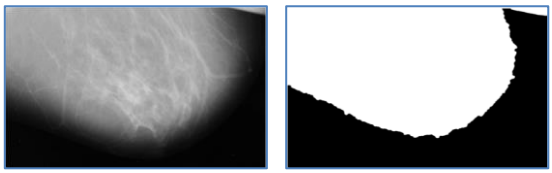
f – мультифрактальна карта heatmap

Рис. 1. Етапи обробки мамограма з пухлиною

Завдяки кольоровому кодуванню ці відмінності чітко відображаються у вигляді контрастної плями, що візуально вирізняється серед загального фрактального поля зображення. Така поведінка карти підтверджує, що локальні фрактальні характеристики є чутливими до структурної неоднорідності, притаманної пухлинним утворенням, і можуть слугувати додатковим джерелом інформації під час аналізу меж, контурів та внутрішньої текстури новоутворення. Зручність heatmap полягає в тому, що вона дозволяє миттєво оцінити просторовий розподіл складності тканин, а також зіставити характер фрактальних змін із сегментованими масками, що робить аналіз більш інтерпретованим.

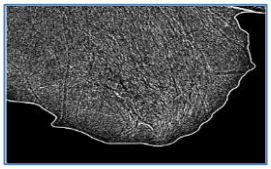
Таким чином, у першому прикладі heatmap виступає інформативним інструментом, який підсилює можливості класичної сегментації.

Приклад 2. Мамограма без патології: оригінальне, сегментоване зображення та heatmap з рівномірним розподілом D (рис. 2). На рис 2, f спостерігається нерівномірність D, яка має фоновий характер і відображає фізіологічну неоднорідність паренхіми; відсутні локалізовані області з відмінним середнім D або вираженим градієнтом по межі, що могло б свідчити про осередкове ураження.

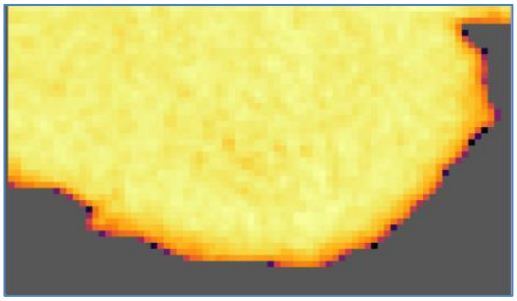


a – оригінал b – сегментація Otsu

c – ссегментація U-net d – об’єднана маска



e – накладена маска та підготовлене зображення



f – мультифрактальна карта heatmap

Рис. 1. Етапи обробки мамограма без патології

У другому прикладі, який демонструє мамограму без патологічних змін, фрактальна карта має принципово інший характер. На отриманій heatmap відсутні різкі локальні відхилення, а розподіл значень D є більш плавним і розтягнутим, що відповідає типовій структурі здорової паренхіми. Незначна нерівномірність карти пояснюється природною варіабельністю тканин, яка завжди присутня навіть у відсутності патології, однак такі варіації не мають вираженої локалізації. Немає також градієнтів чи різких переходів по межі, як це спостерігається у випадках утворень. Важливо, що heatmap здорової мамограми не створює фальшивих «вогнищ» підвищеної складності – це демонструє потенціал методу у зниженні кількості хибнопозитивних інтерпретацій, оскільки фрактальний аналіз підтверджує однорідність структури. Така поведінка карти узгоджується з результатами класичної сегментації та очікуваними характеристиками здорових зображень.

Для валідації підходу було використано зображення з відкритих баз даних MIAS [2] та DDSM [3]. У першому прикладі область пухлини була чітко виділена сегментаційною моделлю, а на heatmap спостерігалося зростання локального D, що добре корелює з клінічною картиною. У другому прикладі heatmap демонстрував рівномірний або слабко варіабельний розподіл без різких піків чи контрастних зон, що підтвер-джує його відповідність здоровому стану та підвищує довіру до застосованого методу. Сумарні результати двох прикладів дозволяють говорити про придатність мультифрактального підходу для попереднього аналізу та формування гіпотез щодо наявності або відсутності патологічних структур.

Висновки

Інтеграція U-Net [4] у комбінації з порогуванням Otsu дала змогу помітно підвищити точність і стабільність виділення меж молочної залози. Нейронна мережа продемонструвала здатність зберігати тонкі структури та коректно відтворювати контури навіть у ділянках зі складною текстурою, тоді як простий постпроцесинг на основі маски Otsu ефективно усуває дрібні артефакти, характерні для нейронних сегментацій на низькоконтрастних зображеннях. Завдяки цьому отримані бінарні маски стали більш відтворюваними, стабільними та узгодженими для подальших фрактальних обчислень; підтверджено, що вони добре масштабуються між різними зображеннями із наборів MIAS [2] та DDSM [3], що є важливою вимогою для наукових експериментів і клінічної валідації.

Побудова heatmap локальних фрактальних розмірностей за методом ковзного вікна надала суттєво вищу інформативність порівняно з розрахунком одного глобального значення фрактальної розмірності D. Локальні карти дозволяють спостерігати неоднорідність структури тканин та її просторову організацію, тоді як глобальний показник фрактальної розмірності згладжує локальні особливості. У межах патологічних зон послідовно фіксуються локальні зміни фрактальної розмірності відносно оточуючої паренхіми; ці зміни відображені за допомогою кольорової шкали, що робить heatmap більш інтерпретованою та потенційно корисною для первинної візуальної оцінки. Така форма представлення дозволяє швидко виокремлювати області інтересу, забезпечуючи додатковий структурний контекст у порівнянні з класичною сегментацією.

Виявлено низку обмежень класичного box-counting, особливо щодо градаційних зображень: залежність від бінаризації, чутливість до шумів та втрата інформації про рівні інтенсивності. З огляду на це подальший розвиток методу може бути пов’язаний із застосуванням Differential Box-Counting (DBC), який враховує повний діапазон інтенсивностей у тривимірному просторі (x, y, значення яскравості) [5]. Це дає підстави розглядати DBC як потенційно більш чутливий інструмент для аналізу текстурної неоднорідності. Комбінація сегментації та фрактального картування відкриває можливість виконання кількісних порівнянь типу «всередині–зовні» (ROI проти оточення), побудови профілів фрактальної розмірності D через межу ураження та формування агрегованих метрик. Попередні результати демонструють, що відмінності локального D можуть проявлятися як у вигляді піків на межі утворення, так і у вигляді змін всередині ROI — залежно від морфологічних особливостей структури.

Подальші дослідження передбачається спрямувати на оцінювання різних напрямів удосконалення. Зокрема, буде розглянуто можливість роботи з форматом DICOM, який містить повні первинні дані мамографічного зображення та може забезпечити більш точну оцінку текстурних характеристик. Також планується дослідити підходи до автоматизації

аналізу heatmap, включно з пошуком ROI на основі фрактальних ознак. Перспективним є вивчення параметрів мультифрактального алгоритму — розміру ковзного вікна, кроку переміщення, кількості рівнів квантування для DBC — з метою оптимізації чутливості та зменшення хибних рішень. Заплановано також проведення розширеної валідації на наборах MIAS [2], DDSM [3] та додаткових клінічних даних, що дозволить оцінити узагальнюваність та практичну доцільність підходу.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Sung H., Ferlay J., Siegel R.L., et al. “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. *CA Cancer J Clin*. 2021;71(3):209-249. doi: https://doi.org/10.3322/caac.21660
2. Suckling J., Parker J., et al. “The Mammographic Image Analysis Society (MIAS) digital mammogram database” Excerpta Medica, International Congress Series 1069; 1994:375–378 https://dl.acm.org/doi/10.1007/11581772_80
3. Heath M., Bowyer K., Kopans D., Moore R., Kegelmeyer W.P. “The Digital Database for Screening Mammography (DDSM)”, Toronto, 11–14 June 2000. Madison, WI: Medical Physics Publishing; 2001:212–218 https://biokeanos.com/source/DDSM
4. Ronneberger O., Fischer P., Brox T. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. *MICCAI*, 2015. P. 234–241 doi: https://doi.org/10.1007/978-3-319-24574-4_28
5. Liu Y., Chen L., Wang H., et al. “An improved differential box-counting method to estimate fractal dimensions of gray-level images”. JVCIR. 2014; 25(5): 1102–1111. doi: https://doi.org/10.1016/j.jvcir.2014.03.008
6. Lychkatyi O.Ye., Povoroznyuk A.I. “Development of a prototype for the analysis of fractal dimensions of medical images”. *Інформаційні технології: наука, техніка, технологія, освіта, здоров’я: тези доп. XXXIII міжн. НПК MicroCAD-2025, 14–17.05.2025 р.* Х.: НТУ «ХПІ». С. 1541. https://ndch.kpi.kharkov.ua/wp-content/uploads/2025/06/Zbirnik-tez-2025.pdf
7. Личкатий О.Є., Поворознюк А.І. “Фрактальний аналіз мамографічних зображень”. *Інформатика, управління та штучний інтелект: тези 12 МНТК, 14–16.05.2025.* Харків: НТУ «ХПІ». С. 71. https://web.kpi.kharkov.ua/ai/?page_id=311
8. Di Martino, G.; Iodice, A.; Riccio, D.; Ruello, G.; Zinno, I. “The Role of Resolution in the Estimation of Fractal Dimension Maps From SAR Data”. *Remote Sensing*, 2018, 10(1):9. doi: https://doi.org/10.3390/rs10010009

Received (Надійшла) 11.01.2026
Accepted for publication (Прийнята до друку) 25.03.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Личкатий Олександр Євгенович – аспірант кафедри комп’ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Oleksandr Lychkatyi – PhD student, Department of Computer Engineering and Programming, National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, Ukraine.
e-mail: Oleksandr.Lychkatyi@cs.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0007-5707-152X.

Поворознюк Анатолій Іванович – доктор технічних наук, професор кафедри комп’ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Anatoliy Povoroznyuk – Doctor of Sciences (Engineering), Professor, Department of Computer Engineering and Programming, National Technical University “Kharkiv Polytechnic Institute”, Kharkiv, Ukraine.
e-mail: Anatolii.Povorozniuk@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0003-2499-2350;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55225664000.

Multifractal Analysis of Mammographic Images

Oleksandr Lychkatyi, Anatoliy Povoroznyuk

Abstract. The study addresses the problem of improving the effectiveness of computer-aided analysis of mammographic images to support the early detection of pathological changes in breast tissue. The limitations of traditional mammography—caused by tissue density, the presence of noise and imaging artifacts, as well as the subjectivity of visual interpretation—justify the need for quantitative and objective analytical methods capable of providing a more reliable assessment of structural features. The work examines an approach aimed at obtaining locally sensitive indicators of tissue complexity, which enhances the detectability of low-contrast or weakly expressed abnormalities. The objective of the research is to develop a methodology that combines automated segmentation with multifractal analysis and presents the results as maps of local fractal dimensions (heatmaps). Segmentation is performed by combining a U-Net neural network with Otsu thresholding after median and Gaussian filtering, which produces more stable masks and reduces the influence of noise components. A key element of the proposed methodology is the construction of local fractal-dimension maps using a sliding-window technique, which enables continuous spatial assessment of structural heterogeneity. It is shown that regions containing pathological changes exhibit localized deviations of the fractal dimension relative to the surrounding parenchyma, thereby improving interpretability and potentially indicating areas of interest. The study analyzes the limitations of the classical box-counting method when applied to grayscale images, particularly its dependence on binarization and the loss of intensity information. The rationale for employing differential box-counting is presented, as this method operates in the intensity domain and provides greater sensitivity to subtle textural variations. Experimental results obtained using images from the MIAS and DDSM datasets confirm the potential of the proposed approach and highlight its usefulness for further development of computer-aided decision-support systems in mammography.

Keywords: mammography; multifractal analysis; local fractal dimensions; box-counting; differential box-counting; U-Net; sliding-window; image segmentation; heatmap.

УДК 004.9 **doi: 10.26906/SUNZ.2026.2.134**

О. С. Ляшенко, В. С. Башилов

Харківський національний університет радіоелектроніки, Харків, Україна

МОДЕЛЬ РОЗПОДІЛУ НАВАНТАЖЕННЯ В ТУМАННІЙ ОБЧИСЛЮВАЛЬНІЙ СИСТЕМІ З ВИКОРИСТАННЯМ ФЕДЕРАТИВНОГО НАВЧАННЯ

Анотація. Актуальність. Поширення Інтернету речей дедалі більше вимагає близькості між хмарними сервісами та кінцевими користувачами. Це стимулювало розширення хмарних ресурсів на периферію в тому, що називається туманними обчисленнями. Останнє проявляється як екосистема пов'язаних хмар, розподілених та з різною потужністю. У таких умовах розподіл робочого навантаження між туманними сервісами стає нетривіальним завданням через складність компромісів. Попит користувачів на периферії дуже різноманітний, що не сприяє простому плануванню ресурсів. І навпаки, запуск сервісів на периферії може використовувати близькість, але це пов'язано з вищими експлуатаційними витратами, не кажучи вже про швидке збільшення ризику перевантаження розріджених ресурсів. Отже, існує потреба в інтелектуальних, але масштабованих рішеннях для розподілу, які протидіють несприятливому попиту на периферії, одночасно ефективно розподіляючи навантаження між периферією та віддаленими хмарами. **Об'єкт дослідження:** процеси розподілу навантаження в туманних обчислювальних системах. **Мета статті:** розробка моделі розподілу робочого навантаження в туманній обчислювальній системі з використанням федеративного навчання та глибокого навчання з підкріпленням. **Результати дослідження.** У статті пропонується федеративна система глибокого навчання з підкріпленням, заснована на мережі глибокого Q-навчання (DQN), для розподілу робочого навантаження в туманній системі. Запропоноване рішення адаптує DQN для оптимізації локального розподілу робочого навантаження, що здійснюється окремими шлюзами. Вбудовано федеративне навчання, що дозволяє кільком шлюзам у мережі спільно накопичувати знання про потреби користувачів. Це використовується для досягнення консенсусу щодо частки робочого навантаження, розподіленого між різними вузлами туману, використовуючи менший обсяг даних та обчислювальних ресурсів. **Висновки.** Федеративний підхід у поєднанні з глибоким навчанням з підкріпленням дозволяє ефективно вирішувати задачу розподілу навантаження в туманних обчисленнях. Запропонована модель забезпечує масштабованість, зменшує потребу в централізованих обчислювальних ресурсах і підвищує ефективність використання інфраструктури за умов динамічного попиту. **Сфера використання отриманих результатів:** інтелектуальні системи планування та балансування навантаження в розподільних обчислювальних системах.

Ключові слова: інтернет речей; розподіл навантаження; федеративне навчання; глибока Q-мережа; туманні мережі; федеративна агрегація середніх значень; машинне навчання.

Вступ

Постановка проблеми. Кількість пристроїв Інтернету речей (IoT) наразі перевищує 13 мільярдів і, як очікується, досягне 34,7 мільярда до кінця 2028 року. Це збільшить попит на хмарні сервіси вивантаження, зберігання та обробки даних до безпрецедентного рівня. Ортогонально, критичність часу та обмеження на обмін даними з таких причин, як вартість та конфіденційність, дедалі більше сприяють близькості між кінцевими користувачами та хмарними сервісами. Це стимулювало перехід до екосистеми "від периферії до хмари", яка вважається формою туманних обчислень [1]. Остання являє собою набір хмар: розподілених, пов'язаних, децентралізованих та з різною ресурсною ємністю та локальністю для кінцевих користувачів.

Хоча розширення обіцяє значні переваги, воно пов'язане з нетривіальними викликами. Розподіл та різноманітність експлуатаційних витрат, енергоефективності та обмежень між периферією та хмарою вводять компроміси між продуктивністю та вартістю [2].

Дані існуючих досліджень та реалістичної хмарної системи вказують на те, що ресурси периферії є обмеженими та експлуатаційно дорогими [3,4]. Це вимагає вибіркового розподілу на периферію на основі потреб, щоб зменшити ризик перевантаження ресурсів та погіршення якості обслуговування (QoS) для застосунків, які потребують периферії. Вибірковий розподіл також необхідний для підтримки сталих експлуатаційних витрат.

Також відомо, що локальний попит на периферійні пристрої дуже варіюється [5] і відрізняється в різних географічних регіонах. Це перешкоджає можливості планування ресурсів периферії, оскільки локальний попит значно менш передбачуваний. Складність посилюється, коли локальність переплітається з намірами користувачів щодо послуг, що споживають дані, що вимагають даних, які генеруються користувачами. Окрім обмежень конфіденційності, вивантаження даних в екосистему пов'язане з витратами на мережу та зберігання, що корелює з розміром даних. Це створює компроміс між перевагою периферії для зменшення витрат на мережу та хмарою для зменшення витрат на зберігання.

Машинне навчання все частіше застосовується в розподілі ресурсів для вирішення деяких із вищезазначених проблем у туманних обчисленнях [6]. Однак традиційне централізоване навчання вимагає централізованого зіставлення даних, що, у свою чергу, вимагає високої ємності сховища та обчислювальної потужності для навчання на великих наборах даних [7]. Це також має вищу ймовірність конфлікту з намірами користувачів щодо конфіденційності та зниження витрат на вивантаження даних.

Натомість, федеративне навчання створює привабливі можливості для вирішення цих проблем. Воно дозволяє здійснювати спільне навчання над розподіленими даними, що належать недовірливим суб'єктам [8]. Це можна використовувати для вивчення закономірностей локального попиту на периферії

масштабованим та намір-сумісним способом, а також для оптимізованого розподілу робочого навантаження між туманними вузлами, щоб мінімізувати експлуатаційні витрати, дотримуючись намірів користувачів. Крім того, федеративне навчання дозволяє постачальникам туманних обчислень приховувати від кінцевих користувачів конфіденційну бізнес-інформацію щодо стану своїх туманних вузлів, одночасно оптимізуючи розподіл робочого навантаження.

Аналіз останніх досліджень і публікацій. Ефективний розподіл ресурсів є критичним завданням у туманних обчисленнях, метою яких є балансування навантаження та досягнення ефективного використання туманних мереж. Попередні дослідження були спрямовані на оптимізацію розподілу ресурсів у цій туманній області, і цей розділ зосереджений на нещодавніх дослідженнях, які зробили внесок. Як правило, питання розподілу ресурсів вирішується шляхом впровадження різних формулювань та алгоритмів, що базуються на різних цілях оптимізації. Для оптимізації розподілу ресурсів необхідно враховувати кілька показників, таких як затримка, використання ресурсів, споживання енергії та інші.

Затримка туманної мережі суттєво впливає на її загальну продуктивність та взаємодію з користувачем, особливо для вимог чутливих до часу програм. У роботі [9] представлено тришарову архітектуру на основі туману разом із моделлю програмування потоку даних розподіленої координації, і в результаті досягається зменшення затримки обслуговування для програм Інтернету речей в інтелектуальній мережі. У дослідженні [10] представлені алгоритми онлайн-оптимізації на основі порогів для мінімізації затримки шляхом інтелектуального вибору сусідніх вузлів для розвантаження.

Використання ресурсів впливає на ефективність їх використання в туманних мережах. Оптимізація використання ресурсів гарантує, що ресурси розподіляються ефективно та не витрачаються марно в туманних мережах. У дослідженні [11] представлено порівняльний алгоритм атрибутів для головних туманних вузлів для планування завдань та вибору туманного вузла з відповідним ресурсом, враховуючи пріоритет та використовуючи лінійний алгоритм узагальнення атрибутів. У роботі [12] застосовується згорткова нейронна мережа та модифікована оптимізація рою частинок для досягнення динамічного балансування навантаження та покращення використання ресурсів у туманних мережах.

Споживання енергії є критичною метрикою для розподілу ресурсів у туманних мережах, особливо в середовищах з обмеженими ресурсами. У роботі [13] представлено максимально енергоефективний алгоритм планування завдань для оптимізації споживання енергії в туманних мережах. У дослідженні [14] представлено енергоефективність в алгоритмі розподілу ресурсів та запропоновано алгоритм енергоефективного розподілу ресурсів на основі туманних вузлів для досягнення оптимізації в туманних мережах.

З метою врахування кількох метрик та розробки більш комплексного та збалансованого підходу до розподілу ресурсів у туманних мережах машинне навчання привертає все більше уваги. У дослідженні [15] штучна нейронна мережа (ШНМ) застосовується як частина алгоритму для розподілу завдань між туманними та хмарними серверами, що призводить до покращення часу відгуку, споживання енергії та використання ресурсів. ШНМ розгортається на серверах і навчається прогнозувати час обробки завдань через центрального брокера. Розподіл ресурсів серверів здійснюється на основі прогнозу від центрального брокера. У роботі [16] використовується алгоритм глибокого навчання з підкріпленням для досягнення розподілу ресурсів у динамічному середовищі туманних обчислень. DQN застосовується для оптимізації ресурсів, щоб максимізувати кількість запитів, які може задовольнити вся мережа, що добре показує результати дослідження.

У вищезгаданому дослідженні показано критичні метрики розподілу ресурсів та підхід до динамічного розподілу ресурсів для туманних мереж. Однак у згаданому дослідженні існує кілька проблем. Хоча під час розгляду оптимізації метрик розподілу ресурсів більшість методів розгортаються централізовано. Це може збільшити складність розподілу ресурсів та час обробки в мережах зі зростанням попиту. Крім того, стратегію розподілу серверів необхідно враховувати в багатодоменних туманних мережах, виходячи з міркувань збереження конфіденційності між різними серверами. Дослідження в [17] розробляє алгоритм федеративного навчання для досягнення розподілу ресурсів на основі компромісу між споживанням енергії та часом навчання. Запропоновано сурогатну функцію на основі розподіленого наближеного алгоритму Ньютона для локального навчання, а федеративне усереднення застосовується для глобальної агрегації в бездротових мережах. Цей алгоритм зосереджений на оптимізації затримки, часу навчання та споживання енергії. Для підвищення ефективності розподілу ресурсів у мережах дослідження в [18] представляє фреймворк, який поєднує глибоке навчання з підкріпленням та федеративне навчання для розподілу ресурсів у мобільних граничних системах. Зокрема, глибоке навчання з підкріпленням розгортається в локальних граничних вузлах для оптимізації ресурсів серед обладнання кількох користувачів. Федеративне навчання застосовується на центральному сервері для агрегації локальних моделей. Ця структура досягає динамічної оптимізації ресурсів та знижує вартість зв'язку в периферійних мережах.

На основі попередніх досліджень застосування федеративного навчання для розподілу ресурсів у бездротових або периферійних мережах, розглядаються підходи федеративного навчання для розподілу ресурсів у туманній мережі. У дослідженні [19] пропонується розподілений алгоритм федеративного навчання для середовища туманних обчислень з обмеженими ресурсами, щоб зменшити затримку зв'язку та споживання енергії. Незважаючи на те, що дослідження встановило певну кількість випадків федеративного навчання, продуктивність моделей слід розглядати як елемент для виконання агрегації. У [20] пропонується фреймворк федеративного навчання під назвою FedFog для балансування ефективності мережевого зв'язку та точності моделі, а також для досягнення мережево-залежної опти-

візації бездротових туманно-хмарних систем. Однак, все ще існує ймовірність виникнення високої затримки, оскільки агрегація відбувається в хмарному шарі, який розташований на великій відстані.

Вищенаведене зумовило мету даної роботи, а саме – розробку моделі для федеративної системи глибокого навчання з підкріпленням для інтелектуального розподілу ресурсів у багатодоменних туманних системах. Запропоноване рішення поєднує мережі глибокого Q-навчання з федеративним навчанням для створення федеративної системи DQN. Тут локалізовані агенти DQN навчаються на стороні користувача, тоді як навчені моделі агрегуються на стороні туманних вузлів. Таким чином, рішення пом'якшує несприятливі умови локального попиту шляхом консенсусу між шлюзами доступу, з'єднуючи локальні групи користувачів.

Основний матеріал

Фундаментальна інфраструктура в туманних системах, це туманні мережі, також відомі як туманні обчислення, визначаються як розподілена гетерогенна мережева архітектура з різноманітними обмеженими обчислювальними та комунікаційними ресурсами, що є доповненням між периферійними пристроями та хмарними центрами. Згідно з Cisco [14], туманні обчислення є ідеальним рішенням для обробки та аналізу даних ближче до джерела (тобто периферійних/IoT пристроїв). Вони пропонують високо віртуалізовану технологію для обчислень, зберігання та мережевих ресурсів, підключення кінцевих пристроїв та традиційних хмарних серверів [15].

Обчислювальні пристрої, що складаються з туманної інфраструктури, відомі як туманні вузли, які можна розгорнути в будь-якому місці з доступом до мережевого з'єднання. Туманні вузли забезпечують обчислювальну потужність та ресурси, які складаються з різних обчислювальних пристроїв, починаючи від невеликих одноплатних комп'ютерів або мікроконтролерів до передових серверів [20]. Рівень обчислювальних можливостей визначається конкретним випадком використання та складністю завдань, які повинен виконувати туманний вузол.

Незважаючи на потенційні переваги туманних мереж, для забезпечення їх успішного розгортання та експлуатації необхідно вирішити певні проблеми. Оскільки туманні мережі складаються з різноманітних пристроїв з різною обчислювальною потужністю та обмеженнями ресурсів, керування та координація їхнього гетерогенного середовища може бути складною. Для забезпечення ефективного забезпечення ресурсами в такому гетерогенному середовищі необхідні динамічні та адаптивні методи розподілу ресурсів [7]. Це означає розподіл ресурсів між туманними вузлами на основі поточного попиту, що зменшує ризик недовикористання або перевантаження вузлів. Крім того, енергоефективність також необхідно враховувати при забезпеченні ресурсами. Оптимізуючи споживання енергії, туманні мережі можуть бути більш економічно вигідними, особливо у великомасштабних розгортаннях з численними периферійними пристроями завдяки нижчим експлуатаційним витратам.

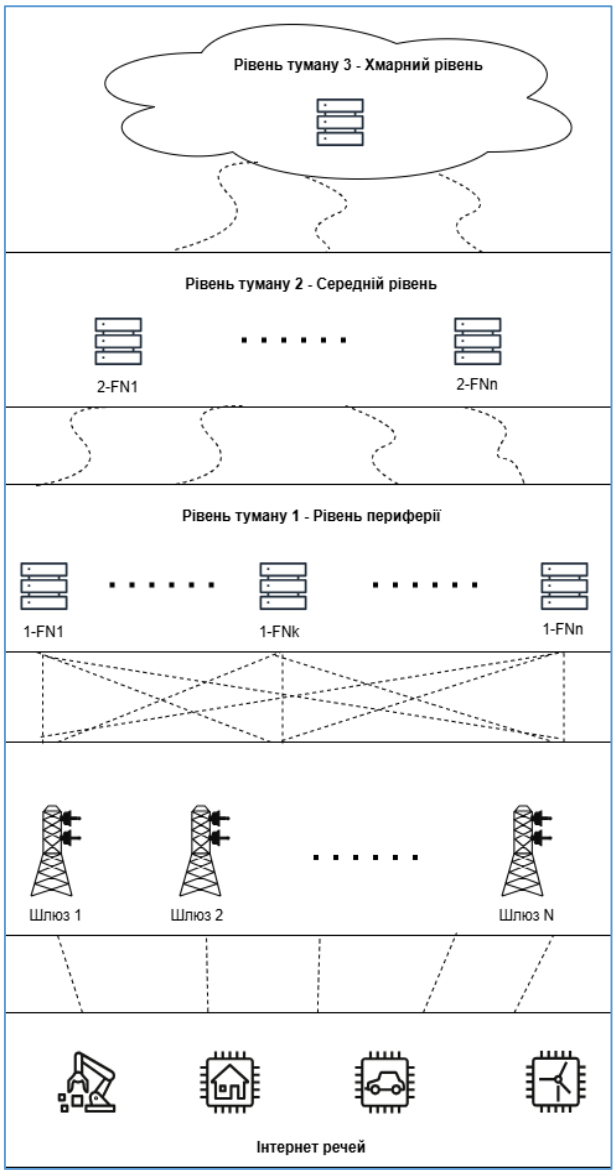


Рис. 1. Архітектура туманної системи

Ієрархічна архітектура туманної мережі показана на рис. 1.

Архітектура містить набір пристроїв Інтернету речей, які збирають та передають запити з фізичного світу до локальних шлюзів для подальшої обробки. Локальні шлюзи, позначені як $G = \{g_1, g_2, \dots, g_k\}$, відповідають за розгортання DQN для розподілу запитів на туманні рівні як посередників між пристроями IoT та туманними вузлами, показаними на рис. 2. Туманні вузли, представлені, як $F = \{f_1, f_2, \dots, f_j\}$, класифікуються на три туманні рівні, тобто рівень периферії, середній рівень та хмарний рівень.

В цій роботі розглядаємо три метрики для кожного вузла F_f, а саме: C^f – обчислювальна потужність процесору; M^f – обсяг пам'яті; E^f – енергетична вартість. Для кожного кожного з'єднання вузлом f та шлюзом g визначаються 2 метрики, це пропускна здатність каналу B_f^g та метрика відстані $D_f^g \in [D_{LB}^T, D_{UB}^T]$, де D_{LB}^T і D_{UB}^T – відповідно нижня і верхня межа метричної відстані між кожним рівнем та шлюзами.

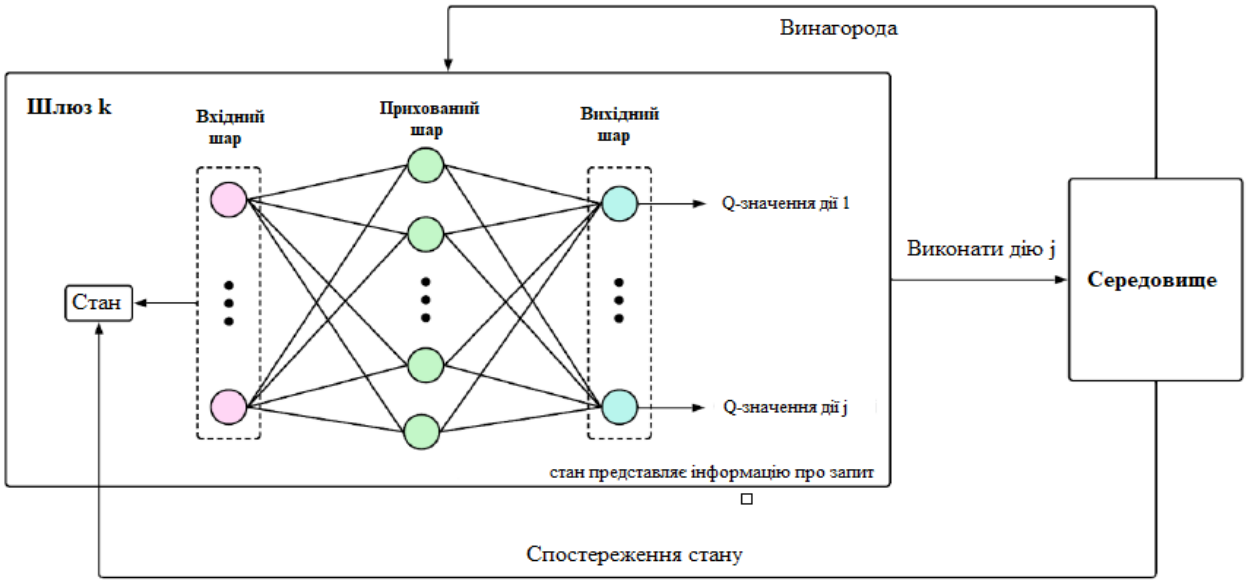


Рис. 2. DQN в одному шлюзі

Тумані вузли одного рівня мають подібну конфігурацію ресурсів, що означає схожі значення обчислювальної потужності процесора, пропускної здатності, обсягу пам'яті, метричної відстані, енергетичних витрат. В залежності від того як тумані вузли наближаються до шлюзів доступу, відстань, обчислювальна потужність процесору, обсяг пам'яті та пропускна здатність – зменшуються, але при цьому енергетичні витрати зростають.

Кожен шлюз g отримує певну кількість запитів кінцевих пристроїв, що позначається S_g. Кожен запит можна описати як $s_g = \{c_s, m_s, b_s, l_s\}$, де c_s – необхідний процесор для запиту, m_s – необхідну пам'ять для запиту, b_s – необхідну пропускну здатність для надсилання відповіді назад до шлюзу, l_s – пріоритет затримки запиту. Пріоритет затримки визначає відповідний рівень туману для виділення запиту, тобто запити з високим пріоритетом слід розподіляти на периферійний рівень, запити із середнім пріоритетом – на середній рівень, а запити з низьким пріоритетом – на хмарний рівень.

Відповідно задачу розподілу ресурсів в такій туманій системі можна представити, як задачу максимізації загального обсягу розподілених ресурсів при мінімізації сукупних витрат, які виникають як зі сторони шлюзів так і зі сторони туманних вузлів. Математично цю задачу можна представити наступним чином:

$$min \sum_{g\in G}\sum_{s\in S_g}\sum_{f\in F} \alpha_{g,f}^{s}\left(c_s * \theta_f^{s,c} + m_s * \theta_f^{s,m} + b_s * \theta_{f,g}\right)$$

за умови:

$$\sum_{g\in G}\sum_{s\in S_g} \alpha_{g,f}^{s}\, c_s \;\le\; C^f, \forall f \in F;$$

$$\sum_{g\in G}\sum_{s\in S_g} \alpha_{g,f}^{s}\, m_s \;\le\; M^f, \forall f \in F;\; l_s = \text{T};$$

де $\alpha_{g,f}^{s}$ – бінарна змінна рішення, що визначається як:

$$\alpha_{g,f}^{s} = \begin{cases} 1, \text{якщо } s_g \text{ розподілено між } f; \\ 0, \quad \text{інакше.} \end{cases}$$

Позначення $\theta_f^{s,c}$, $\theta_f^{s,m}$ відповідають вартості використання ресурсів процесора та пам'яті для обслуговування запиту s на туманному вузлі f, тоді як $\theta_{f,g}$ є мережевою вартістю передачі даних відповіді назад до шлюзу. Кожен з цих видів витрат може включати як енергетичні витрати, повязані з використанням ресурсів, так і вартість самих ресурсів вузла. Для розв'язання цієї задачі в роботі пропонується система федеративного навчання.

На основі запропонованої архітектури та постави задачі для туманної системи запропонована федеративна система глибокого Q-навчання (FDQN) для інтелектуального розподілу робочого навантаження між вузлами (рис. 3), яка складається з локальних агентів навчання та централізованого агрегатора.

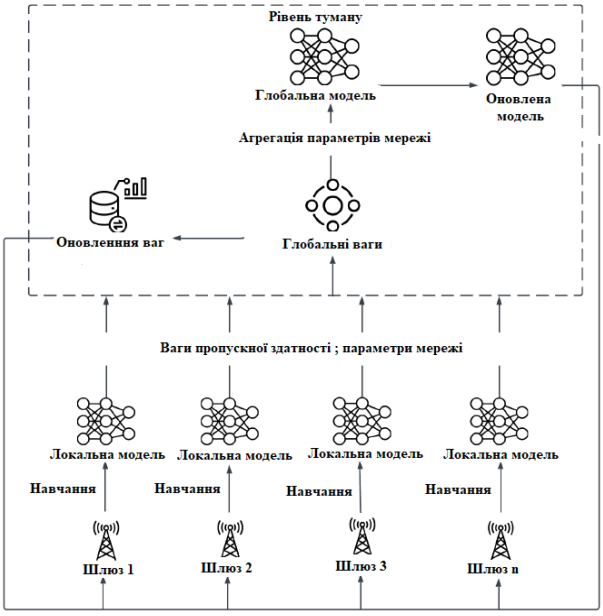


Рис. 3. Структура FDQN

люзи розглядаються як проміжна ланка, середовище що розміщує локальних агентів, які поєднують кінцеві пристрої та тумані вузли. У кожному шлюзі локальний агент визначає оптимальну стратегію вивантаження, навчаючи модель DQN протягом 1 раунду на підмножені локальних запитів. Після цього обчислюється локальна оцінка на основі сумарної винагороди від розподілу запитів у межах локального агента.

Відповідно агрегатор розташований на стороні туману і співрозміщений з оркестратором туманних вузлів. Після кожного раунду локально навчені агенти передають свої локальні моделі, відповідні оцінки та обсяг розподіленого попиту агрегатору. Агрегатор, в свою чергу, обчислює глобальну оцінку на основі локальних показників. Якщо глобальна оцінка покращується порівняно з попереднім значенням, агрегатор формує глобальну модель на основі всіх локальних моделей і надсилає її назад локальним агентам. Ч

ерез географічну неоднорідність локальних агентів вони продовжують навчання, використовуючи останню глобальну модель.

Незалежно від оцінки, агрегатор виконує перерахунок розподілу пропускної здатності для кожного шлюзу відносно кожного туманного вузла, а також визначає вартість використання ресурсів процесора та пам'яті для кожного вузла на основі доступних потужностей та розподіленого попиту в поточному раунді. Оновлення передаються навчальним агентам наприкінці кожного раунду.

Модель навчання локального DQN у федеративній моделі DQN працює наступним чином. Локальні агенти розподіляють запити між туманними вузлами на різних рівнях шляхом навчання DQN. Тут проблема вибору туманних вузлів та розподілу ресурсів розглядається як марковський процес прийняття рішень з простором дій, простором станів та функцією винагороди. У цьому сценарії набір туманних вузлів служить простором дій, тоді як набір локальних запитів представляє простір станів. Функція винагороди складається з ємності пам'яті процесора, вартості енергії кожного туманного вузла, довжини та пропускної здатності шляху між туманними вузлами та шлюзами доступу, а також вимоги до затримки локальних запитів.

Простір станів для кожного шлюзу $g \in G$ визначається як множина S_g. Потрібно зазначити , що $t = \{1,2, \dots |S_g|\}$ – це індекс запитів у множені S_g, де s_t відповідає t-му запиту.

Простір дій для шлюзу g позначається A_g. У цьому просторі кожен туманний вузол розглядається як дія $a_{g,t}^f$ для шлюзу g, яка описується наступним рівнянням:

$$a_{g,t}^f = \{C_{g,t}^f, M_{g,t}^f, E_{g,t}^f, B_{g,t}^f D_{g,t}^f, d_g^f, W_{g,i-1}^f\}.$$

де C_g^f та M_g^f – поточні обчислювальні ресурси процесора та пам'яті; E_g^f – енергетична вартість, що визначається ціною енергії; B_g^f – пропускна здатність каналу між туманим вузлом і шлюзом; D_g^f таd_g^f – відповідно номінальна відстань шлюзу g до туманного вузла f в межах туманних рівнів і кількості переходів на шляху; $W_{g,i-1}^f$ – вага використання ресурсів кожного туманного вузла f після попереднього раунду $i-1$ локального навчання всіх шлюзів. Також слід зазначити, що $C_{g,t}^f$ та $M_{g,t}^f$ оновлюються відповідно до кожного запиту стану t в одному раунді навчання і можуть бути обчислені як:

$$C_{g,t}^f = C_{g,t-1}^f - \mathrm{c}_{g,t};$$
$$M_{g,t}^f = M_{g,t-1}^f - m_{g,t}.$$

Функція винагороди спрямована на максимізацію кількості успішних розподілів запитів при одночасній мінімізації вартості обробки кожного запиту. Функція винагороди є оберненим відображенням функції вартості, а саме $\theta_f^{s,c}$, $\theta_f^{s,m}$, $\theta_{f,g}$ і визначається як узагальнена функція, що враховує поточну доступність ресурсів туманних вузлів, енергетичні витрати, стан шляху та вимоги до затримки. Функцію можна представити наступним чином:

$$R\left(s_t^g, a_g^f\right) = R_{g,t}^{cap,f} + W^t + R_{g,t}^{E,f} + R_{g,t}^{B,D,f} + R_{g,t}^{l,f}.$$

У цьому рівнянні $R_{g,t}^{cap,f}$ – поточна доступна ємність вибраного туманного вузла під час обробки запиту в стані t, яку можна визначити таким чином:

$$R_{g,t}^{cap,f} = \begin{cases} \lambda_{cap} log_a\left(cap_{g,t}^f + 1\right), C_{g,t}^f \geq 0, M_{g,t}^f \geq 0, a > 1, \\ -1 - e^{-C_{g,t}^t}, \quad C_{g,t}^f < 0, M_{g,t}^f \geq 0, \\ -1 - e^{-M_{g,t}^t}, \quad C_{g,t}^f \geq 0, M_{g,t}^f < 0, \\ -1 - e^{-cap_{g,t}^t}, \quad C_{g,t}^f < 0, M_{g,t}^f < 0. \end{cases}$$

Під час експериментів використовувався набір даних, які було згенеровано на основі Google Cluster Workload Traces 2019, який містить піднабори даних запитів та вузлів. Зведений огляд функцій набору даних проілюстровано в табл. 1.

Таблиця 1 – **Особливості набору даних**

Піднабір даних	Значення	Опис
Requests	Timestamp	Час виникнення запиту
	CPU	Запитувані обчислювальні ресурси
	Memory	Запитувані ресурси пам'яті
	Priority	Вимоги до затримки запитів
Nodes	NodeId	ID туманного вузла
	CPU	Обчислювальна потужність
	Memory	Обсяг пам'яті
	Bandwidth	Пропускна здатність каналу
	PathLen	Пропускна здатність каналу
	Hop	Кількість переходів
	PUE	Коефіцієнт енергоефективності
	EnergyPrice	Вартість обчислень

Піднабір даних запитів містить позначку часу, необхідний процесор, необхідну пам'ять та пріоритет затримки запитів, що емулює потребу в ресурсах від різних користувачів. Піднабір даних вузла складається з ідентифікатора вузла, потужності процесора, потужності пам'яті, пропускної здатності каналу вузол-шлюз, довжини шляху каналу вузол-шлюз, кількості стрибків, ціни енергії обчислення, що моделює надання ресурсів туманними вузлами на різних рівнях.

Для моделювання ралістичної туманої системи ми використовуємо дані про машини Google, отриманих з трас та використовуємо їх для створення піднабору даних вузлів, що представляє вузли туману організованих за рівнями. З цією метою машини групуються для отримання агрегованих значень обсягу пам'яті та обчислювальної потужності процесора з метою формування туманних вузлів різних рівнів відповідно до розподілу. У цьому випадку потужність туманних вузлів зменшується у послідовності: хмарний рівень, середній рівень, периферійний рівень.

Ми визначаємо відповідну кількість туманних вузлів у кожному рівні таким чином, щоб сумарна потужність кожного рівня була достатньою для обслуговування загального попиту кожного діапазону пріоритетів у піднаборі запитів.

Ця модель розподілу, додатково використовується для визначення ціни енергії та коефіцієнта PUE для кожного туманного вузла, а також середньої пропускної здатності, діапазону метричної відстані та кількості переходів на шляху між туманним вузлом і будь-яким шлюзом доступу.

Енергетична вартість обробки одного запиту на туманному вузлі обчислюється як комбінація коефіцієнта PUE та вартості обчислень за одиницю ресурсу в кожному вузлі разом із розміром задачі.

Діапазон значень PUE для кожного рівня туману приймається таким, що відповідає розподілу середніх значень PUE центрів обробки даних.

Модель FDQN реалізована з використанням зовнішнього серверу на якому було проведено моделювання роботи FDQN, використовуючи середовище Python 3.12 з бібліотеками PyTorch 2.0 та OpenAI Gym 0.26.

Розроблено власне середовище в OpenAI Gym для DQN на основі запропонованої структури моделі FDQN у цій роботі. Це передбачає формування простору станів, простору дій і функції винагороди. Дані для простору станів беруться з піднабору запитів, тоді як дані для простору дій походять із піднабору вузлів. Обидва набори зберігаються у форматі CSV.

Для моделювання процесу навчання FDQN використовується два вкладених цикла: внутрішній та зовнішній в межах описаного середовища. Кожна локальна модель DQN навчається у внутрішньому циклі, тоді як агрегація всіх моделей і оновлення параметрів кожної моделі відбуваються у зовнішньому циклі.

Під час експериментів оцінювалась продуктивність системи за умов використання стратегії розподілу, що враховує відповідність пріоритету запиту рівню туманного вузла. Запити вважаються успішно розподіленими, якщо ресурси туманного вузла достатні та пріоритет запиту відповідає рівню вузла, який його обслуговує. У дослідженні основна увага приділяється саме частці відмов у розподілі, а не частці успішних розподілів, оскільки це дозволяє більш детально аналізувати продуктивність за різних умов.

На рис. 4 порівнюється частка відмов у розподілі запитів між результатами навчання та валідації. Із збільшенням кількості шлюзів частка відмов значно зменшується як для навчальних, так і для валідаційних даних. При цьому для 20 шлюзів значення частки відмов у навчанні та валідації є практично однаковими. Отже, можна зробити висновок, що за достатньої кількості шлюзів якість моделі на валідації досягає рівня якості на навчанні.

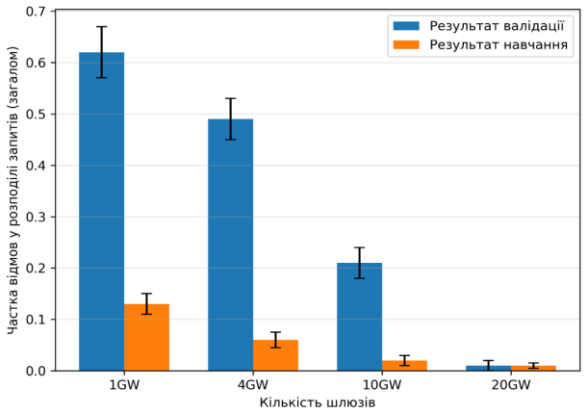


Рис. 4. Загальний рівень відмов при розподілі запитів

На рис. 5 показано залежність частки відмов від збільшення розподілу попиту в ширшій мережі доступу, що реалізується шляхом збільшення кількості шлюзів. Результати свідчать, що відмови у розподілі переважно виникають для запитів із середнім і високим пріоритетом, які повинні обслуговуватись відповідно на середньому та периферійному рівнях туману. Це пояснюється більш жорсткими обмеженнями ресурсів на цих рівнях порівняно з хмарним рівнем, що призводить до перевантаження вузлів і, відповідно, до невдалих розподілів.

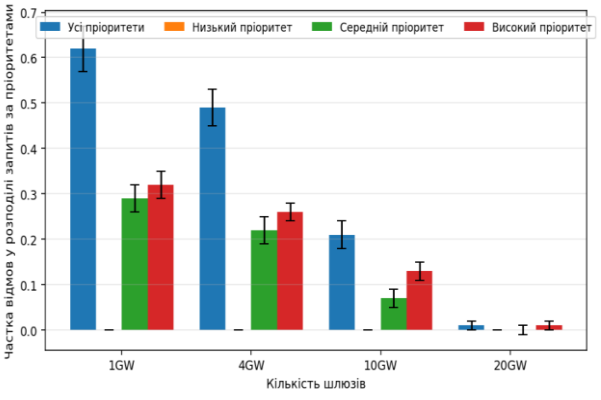


Рис. 5. Рівень відмов при розподілі запитів загалом та за кожним пріоритетом

Частка відмов зменшується приблизно з 60% до 2% при збільшенні кількості шлюзів від 1 до 20. Це зумовлено різницею між обсягом попиту на один шлюз і обчислювальними ресурсами найближчих до

нього туманних вузлів. Оскільки вузли в одному рівні мають подібні характеристики, вибір вузла залежить від параметрів маршруту. У випадку концентрації запитів із високим пріоритетом в одному шлюзі виникає перевантаження найближчих периферійних вузлів, і запити не можуть бути оброблені через домінування вимог до затримки у функції винагороди, навіть за наявності альтернатив у середньому чи хмарному рівнях, розташованих далі.

Це призводить до підвищення частки відмов для запитів із середнім та високим пріоритетом при концентрації попиту в одному шлюзі.

Із збільшенням кількості шлюзів і розподілом попиту кількість периферійних вузлів, близьких до шлюзів, також зростає. Сукупний ефект зростання попиту та його розподілу зменшує навантаження на окремі «найближчі» вузли в кожному рівні. У результаті частка відмов суттєво знижується.

Висновки

Під час проведення дослідження було запропоновано нову федеративну систему глибокого навчання з підкріпленням для ефективного розподілу робочого навантаження в багатодоменних туманних обчислювальних екосистемах. Зокрема, система включає набір локальних агентів, які навчають моделі DQN для інтелектуального відображення локальних запитів одного шлюзу на відповідні туманні вузли.

Проблема нерівномірного попиту на периферії вирішується шляхом досягнення консенсусу між шлюзами через федеративне навчання, що реалізується шляхом агрегації локальних моделей. Це сприяє швидшій збіжності моделей і покращенню балансування навантаження. Крім того, обмеження обміну інформацією між туманними вузлами лише даними про вартість ресурсів і розподілену пропускну здатність дозволяє вузлам зберігати автономність і захищати свою приватну інформацію.

Було проведено оцінювання продуктивності системи за такими показниками: рівень відмов у розподілі, використання ресурсів та енергетичні витрати. Результати показали, що частка невдалих розподілів зменшується зі збільшенням кількості шлюзів. Водночас низький рівень некоректного розподілу спостерігається переважно між хмарним і середнім рівнями.

Також було досліджено чутливість системи шляхом варіювання кількості раундів навчання, коефіцієнта зменшення дослідження та впливу різних складових функції винагороди. Результати показали, що домінування складових CPU або енергетичних витрат у функції винагороди призводить до переважного розподілу запитів у хмарний рівень через нижчу вартість.

У подальших дослідженнях планується оцінити роботу системи на більших обсягах запитів, а також дослідити інші компроміси між розподілом шлюзів і туманних вузлів.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Bonomi, Flavio. Fog computing and its role in the Internet of Things [Text] / Flavio Bonomi, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli // Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. – 2012. – P. 13–16. DOI: https://doi.org/10.1145/2342509.2342513
2. Costa, João B. Jr. Computational resource allocation in fog computing: A comprehensive survey [Text] / João B. Costa Jr., Luís R. Carvalho, Mário J. Rosa, António Araújo // ACM Computing Surveys. – 2022. DOI: https://doi.org/10.1145/3507906
3. Shaik, Sameer. Distributed service placement in hierarchical fog environments [Text] / Sameer Shaik, S. Baskiyar // Sustainable Computing: Informatics and Systems. – 2022. – Vol. 34. – P. 100744. DOI: https://doi.org/10.1016/j.suscom.2022.100744
4. Andrae, Anders S. G. On global electricity usage of communication technology: trends to 2030 [Text] / Anders S. G. Andrae, Tomas Edler // Energies. – 2017. – Vol. 10, no. 10. – P. 1470. DOI: https://doi.org/10.3390/en10101470
5. Cao, Keyan. An overview on edge computing research [Text] / Keyan Cao, Yefeng Liu, Gongjie Meng, Qimeng Sun // IEEE Access. – 2020. – Vol. 8. – P. 85714–85728. DOI: https://doi.org/10.1109/ACCESS.2020.2982434
6. Luong, Nguyen Cong. Applications of deep reinforcement learning in communications and networking: A survey [Text] / Nguyen Cong Luong [et al.] // IEEE Communications Surveys & Tutorials. – 2019. – Vol. 21, no. 4. – P. 3133–3174. DOI: https://doi.org/10.1109/COMST.2019.2904478
7. Abdulkareem, K. H. A review of fog computing and machine learning: Concepts, applications, challenges, and open issues [Text] / K. H. Abdulkareem [et al.] // IEEE Access. – 2019. – Vol. 7. – P. 153123–153140. DOI: https://doi.org/10.1109/ACCESS.2019.2947542
8. Abdelmoniem, A. M. Empirical analysis of federated learning in heterogeneous environments [Text] / A. M. Abdelmoniem, C.-Y. Ho, P. Papageorgiou, M. Canini // Proceedings of the 2nd European Workshop on Machine Learning and Systems (EuroMLSys). – 2022. – P. 1–9. DOI: https://doi.org/10.1145/3517206.3526267
9. Wang, Peng. A fog-based architecture and programming model for IoT applications in the smart grid [Text] / Peng Wang, Shuliang Liu, Fei Ye, Xiaojun Chen // arXiv preprint. – 2018. DOI: https://doi.org/10.48550/arXiv.1804.01239
10. Lee, Jae-Woo. An online optimization framework for distributed fog network formation with minimal latency [Text] / Jae-Woo Lee, Salah Eddine Saad, Mohsen Bennis // IEEE Transactions on Wireless Communications. – 2019. – Vol. 18, no. 4. – P. 2244–2258. DOI: https://doi.org/10.1109/TWC.2019.2901445
11. Hossain, Md. R. A scheduling-based dynamic fog computing framework for augmenting resource utilization [Text] / Md. R. Hossain [et al.] // Simulation Modelling Practice and Theory. – 2021. – Vol. 111. – P. 102336. DOI: https://doi.org/10.1016/j.simpat.2021.102336

12. Talaat, F. M. Effective scheduling algorithm for load balancing in fog environment using CNN and MPSO [Text] / F. M. Talaat, H. A. Ali, M. S. Saraya, A. I. Saleh // Knowledge and Information Systems. – 2022. – Vol. 64, no. 3. – P. 773–797. DOI: https://doi.org/10.1007/s10115-021-01636-9
13. Yang, Yang. Maximal energy efficient task scheduling for homogeneous fog networks [Text] / Yang Yang, Ke Wang, Guoliang Zhang, Xianbin Chen, Xiaojun Luo, M.-T. Zhou // IEEE INFOCOM Workshops. – 2018. – P. 274–279. DOI: https://doi.org/10.1109/INFCOMW.2018.8406848
14. Huang, Xin. Energy-efficient resource allocation in fog computing networks with the candidate mechanism [Text] / Xin Huang, Wei Fan, Qiang Chen, Jian Zhang // IEEE Internet of Things Journal. – 2020. – Vol. 7, no. 9. – P. 8502–8512. DOI: https://doi.org/10.1109/JIOT.2020.2981790
15. Abedi, M. Resource allocation in combined fog-cloud scenarios by using artificial intelligence [Text] / M. Abedi, M. Pourkiani // 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC). – IEEE, 2020. – P. 218–222. DOI: https://doi.org/10.1109/FMEC49853.2020.9144881
16. Abedi, M. Resource allocation in combined fog-cloud scenarios by using artificial intelligence [Text] / M. Abedi, M. Pourkiani // 2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC). – IEEE, 2020. – P. 218–222. DOI: https://doi.org/10.1109/FMEC49853.2020.9144881
17. Dinh, C. T. Federated learning over wireless networks: Convergence analysis and resource allocation [Text] / C. T. Dinh, N. H. Tran, M. N. Nguyen [et al.] // IEEE/ACM Transactions on Networking. – 2020. – Vol. 29, no. 1. – P. 398–409. DOI: https://doi.org/10.1109/TNET.2020.3034879
18. Dinh, C. T. Federated learning over wireless networks: Convergence analysis and resource allocation [Text] / C. T. Dinh, N. H. Tran, M. N. Nguyen [et al.] // IEEE/ACM Transactions on Networking. – 2020. – Vol. 29, no. 1. – P. 398–409. DOI: https://doi.org/10.1109/TNET.2020.3034879
19. Saha, R. FogFL: Fog-assisted federated learning for resource-constrained IoT devices [Text] / R. Saha, S. Misra, P. K. Deb // IEEE Internet of Things Journal. – 2021. – Vol. 8, no. 10. – P. 8456–8463. DOI: https://doi.org/10.1109/JIOT.2020.3044025
20. Nguyen, V.-D. FedFog: Network-aware optimization of federated learning over wireless fog-cloud systems [Text] / V.-D. Nguyen, S. Chatzinotas, B. Ottersten, T. Q. Duong // IEEE Transactions on Wireless Communications. – 2022. – Vol. 21, no. 10. – P. 8581–8599. DOI: https://doi.org/10.1109/TWC.2022.3152703

Received (Надійшла) 05.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Ляшенко Олексій Сергійович – кандидат технічних наук, доцент, декан факультету Комп'ютерної інженерії та інформаційних технологій, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleksii Liashenko - Candidate of Technical Sciences, Associate Professor, Dean of the Faculty of Computer Engineering and Information Technologies, Kharkiv National University of Radio Electronics,, Kharkiv, Ukraine;
e-mail: oleksii.liashenko@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-0146-3934;
Scopus Author ID https://www.scopus.com/authid/detail.uri?authorId=55658561300.

Башилов Владислав Сергійович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vladislav Bashilov – Postgraduate student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: vladyslav.bashylov@nure.ua; ORCID Author ID: https://orcid.org/0009-0005-4025-2282.

A load balancing model in a fog computing system using federated learning

Oleksii Liashenko, Vladislav Bashilov

Abstract. Relevance. The rapid proliferation of the Internet of Things increasingly requires closer proximity between cloud services and end users. This has led to the extension of cloud resources toward the network edge in the paradigm known as fog computing. The latter is manifested as an ecosystem of interconnected, distributed clouds with heterogeneous capacities. Under such conditions, workload allocation among fog services becomes a non-trivial task due to the complexity of trade-offs involved. User demand at the edge is highly diverse, which complicates resource planning. On the other hand, deploying services at the edge leverages proximity benefits but is associated with higher operational costs and an increased risk of overloading limited resources. Therefore, there is a need for intelligent yet scalable allocation solutions capable of handling adverse edge demand while efficiently distributing workloads between edge and remote cloud resources. **Object of study**: workload allocation processes in fog computing systems. **Purpose of the study:** to develop a workload allocation model for fog computing systems using federated learning and deep reinforcement learning. **Research results.** This paper proposes a federated deep reinforcement learning system based on Deep Q-Networks (DQN) for workload allocation in fog environments. The proposed approach adapts DQN to optimize local workload allocation performed by individual gateways. Federated learning is incorporated to enable multiple gateways to collaboratively learn user demand patterns. This enables achieving consensus on workload distribution across fog nodes while reducing data exchange and computational overhead. **Conclusions.** The federated approach combined with deep reinforcement learning provides an effective solution for workload allocation in fog computing. The proposed model ensures scalability, reduces reliance on centralized computational resources, and improves infrastructure utilization under dynamic demand conditions. **Scope of application:** intelligent scheduling and load balancing systems in distributed computing environments.

Keywords: Internet of Things; load balancing; federated learning; deep Q-network; fog networks; federated averaging; machine learning.

УДК 004. 056.55:004.312.2 **doi: 10.26906/SUNZ.2026.2.142**

Е. Е. Малохвій

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

БАГАТОКРИТЕРІАЛЬНА МОДЕЛЬ ЛОКАЛЬНОЇ ОБРОБКИ ДАНИХ ТА ОБЧИСЛЮВАЛЬНОГО РОЗВАНТАЖЕННЯ НА КІНЦЕВИХ ПРИСТРОЯХ IIoT

Анотація. Актуальність. У системах Промислового Інтернету речей (IIoT) кінцеві пристрої функціонують в умовах обмежених обчислювальних, енергетичних і комунікаційних ресурсів, тоді як вхідні потоки даних мають гетерогенний, нестаціонарний і подієво-орієнтований характер. За таких умов передавання всіх необроблених даних до верхніх рівнів cloud-fog-edge архітектури призводить до зростання затримок, перевантаження каналів зв’язку та зниження загальної ефективності системи. **Об’єкт дослідження.** процес обробки інформації на кінцевих пристроях IIoT у багаторівневому середовищі cloud-fog-edge. **Мета статті.** узагальнення теоретико-математичних засад побудови багатокритеріальної моделі локальної обробки даних та обчислювального розвантаження на кінцевих пристроях з урахуванням характеристик потоку подій, ресурсного стану вузла та вимог до збереження інформативності даних. **Результати дослідження.** Запропоновано узагальнену модель, у межах якої кінцевий пристрій розглядається як активний вузол прийняття рішень, що виконує попередню обробку, оцінювання інформативності, адаптивну фільтрацію, агрегацію, стиснення та формування ознак перед передаванням даних на верхні рівні. Модель поєднує подання вхідного потоку як сукупності класозалежних подій, локальну чергову обробку з обмеженим буфером, вектор ресурсного стану пристрою та багатокритеріальну схему вибору між локальним виконанням і обчислювальним розвантаженням. **Висновки.** Запропонований підхід формує методологічну основу для подальшого синтезу адаптивних стратегій обробки інформації на кінцевих пристроях IIoT, орієнтованих на збалансування затримки, втрат, енергоспоживання, комунікаційного навантаження та збереження інформативності даних.

Ключові слова: Промисловий Інтернет речей; cloud-fog-edge архітектура; кінцевий пристрій; локальна обробка даних; адаптивна фільтрація; обчислювальне розвантаження; багатокритеріальна оптимізація.

Вступ

Постановка проблеми. Стрімкий розвиток Промислового Інтернету речей (IIoT) зумовив суттєву трансформацію підходів до збору, передавання та обробки даних у кіберфізичних виробничих системах. На відміну від традиційних систем автоматизації, сучасні IIoT-інфраструктури характеризуються щільним розміщенням сенсорів, безперервною генерацією телеметрії, подієво-орієнтованою сигналізацією та використанням гетерогенних каналів зв’язку, що формують значні за обсягом і структурно неоднорідні потоки даних [1]. За таких умов ефективність функціонування системи дедалі більше залежить не лише від хмарної аналітики чи мережевої оркестрації, а й від здатності кінцевих пристроїв виконувати змістовну локальну обробку до передавання інформації на верхні обчислювальні рівні.

Інтеграція cloud-, fog- та edge-рівнів стала домінантною архітектурною парадигмою для IIoT, оскільки дає змогу розподіляти обчислювальні функції відповідно до вимог щодо затримки, доступності ресурсів і критичності прикладних задач [2]. При цьому саме кінцеві пристрої залишаються найбільш ресурсно обмеженими елементами такої багаторівневої системи, оскільки функціонують в умовах обмеженої обчислювальної потужності, скінченної пам’яті, нестабільної пропускної здатності каналів зв’язку та жорстких енергетичних обмежень. У практичному сенсі це означає, що кінцевий вузол не може ані виконувати повну локальну обробку для кожної вхідної події, ані передавати весь потік сирих даних на fog/cloud-рівні без ризику зростання затримок, переповнення буфера або зниження енергоефективності [3].

Аналіз останніх досліджень і публікацій. Наявні дослідження переважно розглядають або архітектурні аспекти cloud-fog-edge систем, або окремі механізми локального зменшення даних, зокрема фільтрацію, агрегацію, стиснення та обчислювальне розвантаження. Водночас у багатьох роботах кінцеві пристрої подаються як спрощені джерела трафіку зі сталими статистичними властивостями, що є недостатнім для синтезу адаптивних стратегій локальної обробки. У промислових середовищах вхідний потік має гетерогенну подієву природу і поєднує періодичну телеметрію, асинхронні тривожні сигнали, діагностичні повідомлення, службові пакети та події керування, які відрізняються часовими характеристиками, пріоритетами та інформативною цінністю [4].

Додатковою складністю є те, що обробка інформації на кінцевому пристрої IIoT не може бути адекватно описана в межах одноцільової оптимізації. Агресивне зменшення трафіку за рахунок фільтрації та агрегації знижує навантаження на канал і енергоспоживання, проте може призвести до втрати діагностично важливих подій та спотворення динаміки процесу. Натомість максимальне збереження потоку підвищує спостережуваність системи, але супроводжується зростанням затримок, втрат і навантаження на локальні ресурси. Отже, **актуальною** є побудова цілісної математичної основи, яка одночасно враховувала б гетерогенність подій, локальну чергову динаміку, поточний ресурсний стан пристрою, адаптивні перетворення даних і багатокритеріальний характер прийняття рішень щодо локальної обробки або передавання на верхні рівні [5].

Мета статті. Метою статті є узагальнення теоретико-математичних засад побудови багатокритеріальної моделі локальної обробки даних та обчислювального розвантаження на кінцевих пристроях IIoT у cloud-fog-edge середовищі з урахуванням гетерогенності подієвих потоків, ресурсних обмежень вузла та вимог до збереження інформативності даних.

Основний матеріал

У запропонованій постановці кінцевий пристрій IIoT розглядається не як пасивне джерело телеметричних даних, а як активний вузол локальної обробки та прийняття рішень у межах cloud-fog-edge архітектури. Такий підхід дає змогу перенести частину функцій попередньої обробки інформації безпосередньо на рівень кінцевого пристрою, де формується первинний компроміс між локальним виконанням, відкладеним передаванням і обчислювальним розвантаженням на вищі рівні архітектури. У цій архітектурі cloud-рівень забезпечує довготривале зберігання, ресурсоємну аналітику та глобальну оптимізацію, fog-рівень виконує проміжну координацію й наближену до реального часу обробку, тоді як кінцевий пристрій забезпечує безпосередню взаємодію з фізичним процесом за умов жорстких ресурсних обмежень [6, 7].

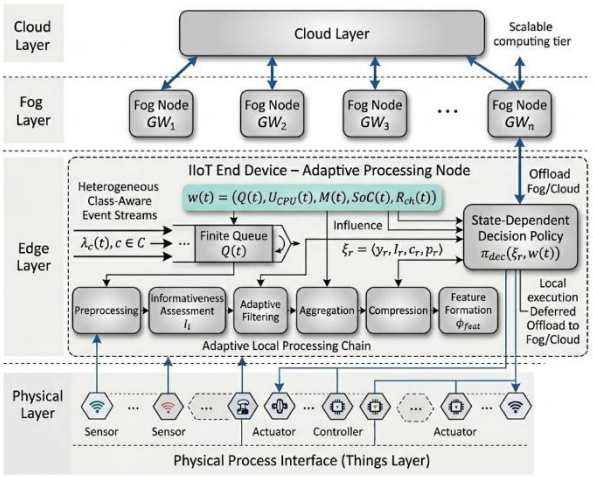


Рис. 1. Архітектура cloud-fog-edge для IIoT з кінцевим пристроєм як активним вузлом прийняття рішень

Вхідні дані в такій системі доцільно описувати не як однорідний потік вимірювань, а як гетерогенний потік подій, сформований сенсорами, виконавчими механізмами, контролерами та сервісними програмними компонентами. На відміну від спрощених моделей, цей потік включає періодичні та асинхронні події з різними семантичними ролями, часовими характеристиками, пріоритетами й інформативною цінністю. У межах моделі окрема подія подається кортежем

$$x_i = \langle v_i, s_i, c_i, p_i \rangle ,$$

де v_i – значення або вектор ознак події, s_i – часова мітка, c_i – клас події, p_i – її пріоритет. Тоді сумарний вхідний потік можна задати як суперпозицію класозалежних підпотоків

$$\lambda(t) = \sum_{c \in C} \lambda_c(t) .$$

Таке подання є принципово важливим, оскільки дозволяє враховувати не лише загальну інтенсивність надходжень, а й структурний склад трафіку, що безпосередньо впливає на вибір локальної політики обробки [7].

У практичних IIoT-сценаріях це означає, що локальна система керування даними повинна реагувати не лише на зміну інтенсивності вхідного потоку, а й на зміну його функціональної структури. Якщо в потоці починають домінувати події аварійного або діагностичного характеру, механізми локальної обробки мають забезпечувати пріоритетне збереження саме таких повідомлень, навіть за умов зростання навантаження. Натомість для фонової телеметрії або допоміжного службового трафіку можуть застосовуватися жорсткіші механізми локального зменшення даних. Отже, подієве представлення потоку створює підґрунтя для семантично чутливого керування обробкою інформації на рівні кінцевого пристрою.

Локальна обробка на кінцевому пристрої моделюється як чергова система зі скінченною буферною ємністю. Навіть після початкового зменшення потоку події залишаються конкурентами за процесорний час, пам’ять, буфер та доступ до комунікаційного інтерфейсу. Узагальнену динаміку черги можна подати співвідношенням

$$Q(t + \Delta t) = \min\left\{Q_{\max}, \max[0, Q(t) + A(t, \Delta t) - S(t, \Delta t)]\right\},$$

де $Q(t)$ – поточна заповненість черги, $A(t, \Delta t)$ – обсяг допущених надходжень за інтервал Δt, $S(t, \Delta t)$ – обсяг локально оброблених подій, $Q_{\max}$ – гранична місткість буфера. Таке подання відображає ключову для IIoT логіку: якщо приплив подій перевищує ефективну швидкість локального обслуговування, то накопичення черги призводить до зростання затримки, імовірності втрат і функціонального старіння інформації. Отже, стан черги виступає не лише характеристикою поточного навантаження, а й важливим сигналом для подальшого керування режимом обробки [8].

Для врахування поточних обмежень функціонування вузла доцільно вводити динамічний вектор ресурсного стану

$$w(t) = (Q(t), U_{CPU}(t), M(t), SoC(t), R_{ch}(t)) ,$$

де $Q(t)$ характеризує заповненість черги, $U_{CPU}(t)$ – завантаження процесора, $M(t)$ – використання пам’яті, $SoC(t)$ – енергетичний стан, а $R_{ch}(t)$ – доступну пропускну здатність або якість каналу зв’язку. Введення такого вектора дає змогу формалізувати важливе положення: допустимість локальної обробки або доцільність обчислювального розвантаження визначаються не лише властивостями самої події, а й поточним ресурсним контекстом. Один і той самий клас подій може бути доцільним для локальної обробки за низького навантаження на процесор та обмеженої пропускної здатності каналу, тоді як за високого CPU навантаження і сприятливих умов зв’язку його доцільніше передавати на fog- або cloud-рівень [8, 9].

Такий підхід є особливо важливим для промислових систем із нерівномірним навантаженням і часово змінними режимами функціонування. У моменти пікової активності або погіршення характеристик каналу зв’язку навіть незначне збільшення обсягу локально оброблюваних даних може призвести до суттєвого зростання затримок і перевантаження буфера. Водночас за стабільного каналу та достатнього

енергетичного резерву доцільним може бути передавання частини обчислювальних задач на вищі рівні архітектури. Саме тому ресурсний контекст має розглядатися як невід'ємна складова процесу прийняття рішень щодо локальної обробки даних.

На змістовому рівні локальна обробка інтерпретується як послідовність взаємопов'язаних операторів: попередня обробка, оцінювання інформативності, адаптивна фільтрація, агрегація, стиснення та формування ознак. Попередня обробка усуває шум, похибки квантування, пропуски вимірювань, часову неузгодженість і дрейф калібрування, тобто приводить сирі спостереження до форми, придатної для подальшого аналізу в умовах обмежених ресурсів. Після цього особливого значення набуває оцінювання інформативності, оскільки в промислових сценаріях далеко не всі елементи потоку є однаково важливими для моніторингу, діагностики чи керування. Саме тому фільтрація повинна бути не просто процедурою зменшення трафіку, а механізмом відбору подій із найвищою змістовною цінністю.

При цьому ефективність такого операторного ланцюга визначається не лише здатністю зменшувати обсяг переданих даних, а й тим, наскільки коректно він зберігає причинно значущі характеристики процесу, необхідні для подальшого моніторингу та діагностики. Для IIoT-середовищ це особливо важливо, оскільки втрата окремих слабких або рідкісних подій може призвести до погіршення якості розпізнавання відхилень і зниження достовірності висновків, отриманих на вищих рівнях архітектури. У цьому контексті локальна обробка виступає не просто технікою скорочення трафіку, а засобом керованого перетворення інформаційного потоку з урахуванням його функціональної значущості.

Фільтрація, агрегація та стиснення в межах запропонованої моделі не мають розглядатися ізольовано, оскільки кожен із цих етапів одночасно впливає на обсяг переданих даних, навантаження на процесор, динаміку черги, енергетичні витрати та ступінь збереження інформативності. Посилення фільтрації або збільшення вікна агрегації може зменшити комунікаційне навантаження й стабілізувати локальний буфер, однак водночас підвищує ризик втрати слабких, але діагностично важливих подій. Натомість максимальне збереження потоку підтримує високу спостережуваність технологічного процесу, проте може перевищувати доступні обчислювальні та енергетичні можливості вузла. Отже, внутрішня структура локальної обробки має розглядатися як узгоджений адаптивний механізм керування інформаційною щільністю потоку [9, 10].

З теоретичної точки зору це означає, що локальна обробка на кінцевому пристрої повинна аналізуватися як багатофакторний процес, у якому взаємодіють характеристики потоку, ресурсні обмеження вузла, часові вимоги прикладної задачі та очікувана інформативна цінність результату. Саме таке трактування дозволяє перейти від ізольованого розгляду окремих технічних процедур до цілісної моделі керування потоком даних. У результаті обчислювальне розвантаження розглядається не як автономна операція маршрутизації, а як один із елементів інтегрованої стратегії адаптивної обробки інформації в IIoT.

Фінальним рівнем формалізації є багатокритеріальна постановка вибору між локальною обробкою та передаванням на fog/cloud-рівні:

$$\min_{\pi\in\Pi} F(\pi) =$$

$$= \min_{\pi\in\Pi}\left(f_T(\pi),\, f_L(\pi),\, f_E(\pi),\, f_R(\pi), -f_I(\pi)\right),$$

де f_T характеризує затримку, f_L – втрати або пропущені дедлайни, f_E – енергоспоживання, f_R – комунікаційне навантаження, а f_I – рівень збереження інформативності даних. Така постановка відображає центральний конфлікт задачі: мінімізація трафіку та локальних витрат не повинна досягатися ціною втрати критично важливої інформації, тоді як максимальне збереження потоку не може порушувати вимоги до затримки, стійкості черги та енергоефективності. У підсумку запропонований підхід поєднує подієве представлення вхідного потоку, чергову модель локальної обробки, ресурсний вектор стану пристрою, операторний ланцюг адаптивного перетворення даних і багатокритеріальну основу для вибору стратегії локальної обробки або обчислювального розвантаження [10].

Висновки

У статті узагальнено теоретико-математичні засади побудови багатокритеріальної моделі локальної обробки даних та обчислювального розвантаження на кінцевих пристроях IIoT у cloud-fog-edge середовищі. Запропонований підхід ґрунтується на поєднанні подієвого представлення гетерогенних вхідних даних, чергової моделі локального потоку обробки, вектора ресурсного стану пристрою та багатокритеріальної системи оцінювання ефективності, що одночасно враховує затримку, втрати, споживання ресурсів, комунікаційне навантаження та збереження інформативності даних.

Наукова цінність запропонованої формалізації полягає в тому, що локальна обробка інтерпретується не як ізольований набір технічних процедур, а як узгоджений операторний ланцюг, який включає попередню обробку, оцінювання інформативності, адаптивну фільтрацію, агрегацію, стиснення та формування ознак.

Такий підхід забезпечує формальний зв'язок між семантикою подій, ресурсними обмеженнями пристрою та подальшими рішеннями щодо локальної обробки або обчислювального розвантаження, формуючи цілісну методологічну основу для побудови адаптивних політик функціонування кінцевих вузлів IIoT.

Практичне значення отриманого результату полягає у можливості використання цієї моделі як бази для подальшого синтезу прикладних методів керування локальною обробкою інформації в промислових сценаріях із різними вимогами до затримки, якості зв'язку, енергоспоживання та діагностичної точності. Наступним етапом дослідження має стати

побудова конкретних адаптивних стратегій і їх оцінювання в репрезентативних IIoT-середовищах.

Конфлікт інтересів. Автор декларує, що не має конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту Автор підтверджує, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Alsadie, D. (2024). Advancements in heuristic task scheduling for IoT applications in fog-cloud computing: challenges and prospects. PeerJ Computer Science, 10, e2128. doi: https://doi.org/10.7717/PEERJ-CS.2128
2. Singh, S.P., Kumar, N., Kumar, G., Balusamy, B., Bashir, A.K., Al Dabel, M.M. (2025). Enhancing Quality of Service in IoT-WSN through Edge-Enabled Multi-Objective Optimization. IEEE Transactions on Consumer Electronics. doi: https://doi.org/10.1109/TCE.2025.3526992
3. Kuchuk, H., Malokhvii, E. (2024). Integration of iot with Cloud, Fog and Edge computing: a review. Advanced Information Systems. 8(2), 65-78. doi: https://doi.org/10.20998/2522-9052.2024.2.08
4. Qayyum, T., Trabelsi, Z., Waqar Malik, A., Hayawi, K. (2022). Mobility-aware hierarchical fog computing framework for Industrial Internet of Things. Journal of Cloud Computing, 11(1), 72. doi: https://doi.org/10.1186/s13677-022-00345-y
5. Muñoz, L.A., Berná Martínez, J.V., Asensi, C.C., Pastor, D.S. (2025). Research Notes: Design of a Distributed and Highly Scalable Fog Architecture for Heterogeneous IoT Infrastructures. Int. Journal of Software Engineering and Knowledge Eng., 35(2), 195-215. doi: https://doi.org/10.1142/S0218194025430016
6. Jamil, B. Shojafar, M., Ahmed, I., Ullah, A., Munir, K., Ijaz, H. (2020). A job scheduling algorithm for delay and performance optimization in fog computing. Concurrency and Computation: Practice and Experience, 32(7). doi: https://doi.org/10.1002/cpe.5581
7. Malik, U.M., Javed, M.A., Frnda, J., Rozhon, J., Khan, W.U. (2022). Efficient Matching-Based Parallel Task Offloading in IoT Networks. Sensors, 22. doi: https://doi.org/10.3390/s22186906
8. Liu, L., Chen, H., Xu, Z. (2022). SPMOO: A Multi-Objective Offloading Algorithm for Dependent Tasks in IoT Cloud-Edge-End Collaboration. Information, 13, 75. doi: https://doi.org/10.3390/info13020075
9. Thomas, P., Jose, D.V. (2023). Towards Computation Offloading Approaches in IoT-Fog-Cloud Environment: Survey on Concepts, Architectures, Tools and Methodologies. Lecture Notes in Networks and Systems, 613 LNNS, 37-52. https://doi.org/10.1007/978-981-19-9379-4_4
10. Pardalos, P.M., Steponavičė, I., Žilinskas, A. (2012). Pareto set approximation by the method of adjustable weights and successive lexicographic goal programming. Optimization Letters, 6(4), 665-678. doi: https://doi.org/10.1007/s11590-011-0291-5

Received (Надійшла) 11.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Малохвій Едуард Едуардович – аспірант, кафедра Комп'ютерної інженерії та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Eduard Malokhvii – PhD candidate, Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute";
e-mail: malokhvii.ee@gmail.com; ORCID Author ID: http://orcid.org/0009-0008-0311-6400;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59179728600.

Multi-Objective Model of Local Data Processing and Computation Offloading on IIoT End Devices

Eduard Malokhvii

Abstract. Relevance. In Industrial Internet of Things (IIoT) systems, end devices operate under limited computational, energy, and communication resources, while incoming data streams are heterogeneous, non-stationary, and event-driven. Under such conditions, transmitting all raw data to the upper layers of the cloud-fog-edge architecture leads to increased latency, communication channel overload, and a decline in overall system efficiency. **Object of research.** The process of information processing on IIoT end devices in a multi-level cloud-fog-edge environment. **Purpose of the article.** To generalize the theoretical and mathematical foundations for constructing a multi-objective model of local data processing and computation offloading on end devices, taking into account the characteristics of the event stream, the resource state of the node, and the requirements for preserving data informativeness. **Research results.** A generalized model is proposed in which the end device is considered an active decision-making node that performs preprocessing, informativeness assessment, adaptive filtering, aggregation, compression, and feature extraction before transmitting data to upper levels. The model combines the representation of the input stream as a set of class-dependent events, local queue-based processing with a limited buffer, the device resource-state vector, and a multi-objective scheme for choosing between local execution and computation offloading. **Conclusions.** The proposed approach forms a methodological basis for the further synthesis of adaptive information-processing strategies on IIoT end devices aimed at balancing latency, losses, energy consumption, communication load, and preservation of data informativeness.

Keywords: Industrial Internet of Things; cloud-fog-edge architecture; end device; local data processing; adaptive filtering; computation offloading; multi-objective optimization.

UDC 004.056 **doi: 10.26906/SUNZ.2026.2.146**

Mykyta Matvieiev

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

PERFORMANCE EVALUATION OF SCENE LOADING OPTIMIZATION IN A WEBAR APPLICATION

Abstract. Relevance. WebAR is rapidly evolving, however, it faces the challenge of high computational loads on devices. The synchronous loading of heavy 3D models during scene initialization often leads to delays and blocks the browser's main execution thread. Applying comprehensive optimization methods (such as transitioning to the GLB format and utilizing Draco geometry compression) is crucial for ensuring stability, yet their implementation requires an objective quantitative evaluation. **The aim of this study** is to evaluate the performance of WebAR application scene loading optimization during the initialization stage**. The object of the research** is the initialization and 3D content loading process within a client WebAR application running on a laptop. **The subject of the research** encompasses the performance metrics of the WebAR application's optimization methods. **Conclusion.** Based on the results of instrumental profiling, the high efficiency of transitioning to the GLB format and employing Draco compression has been confirmed. A 47.7% reduction in the peak allocated JS Heap memory was recorded. The overall scene initialization time was reduced by 30.5%, dropping from 4749 to 3298 ms. Furthermore, the main thread idle time was significantly decreased by 45%, and the parsing phase duration by 66.2%, which successfully eliminated critical bottlenecks in the 3D asset preparation pipeline.

Keywords: web, augmented reality, webAr, performance evaluation, optimization, Draco, Angular.

Introduction

Relevance. Digital technologies have become an integral part of modern society, transforming the ways we communicate, learn, and consume information. One of the most dynamic technologies is augmented reality (AR), which combines virtual digital content with the physical environment in real-time, providing a new level of interactivity [1, 2]. WebAR technology, which implements augmented reality functionality directly through a web browser, has experienced significant development. This approach lowers barriers to entry for users and simplifies the integration of 3D content into digital services, shaping a trend where web resources with AR capabilities are becoming the new standard of interaction [3–5].

Despite a number of significant advantages, WebAR is characterized by an increased computational load on the client device. The fundamental contradiction lies in the fact that underlying web technologies were primarily created to handle 2D content, whereas WebAR requires the continuous execution of resource-intensive operations: 3D graphics processing and spatial tracking. Implementing these processes through standard browser mechanisms leads to a decrease in performance, manifested by a drop in frame rates, blocking of the main execution thread, and an overall inefficient use of hardware resources [6–9]. Significant rendering delays mostly arise from the use of excessively large 3D models and synchronous data exchange mechanisms. In this mode, the client is forced to wait for the transmission of the entire model to complete before its initialization, which, combined with the instability of mobile networks, can lead to critical freezing of the application interface.

The solution to this problem is the optimization of 3D content and loading architecture. Utilizing compression technologies significantly reduces file sizes, improving the efficiency of their transmission through the browser. From the perspective of rendering performance, the glTF format and its binary variant, GLB, ensure the most efficient scene loading. To maximize results, these formats are combined with the Draco algorithm, which employs quantization and delta encoding for the spatial compression of geometry vertex attributes [10, 11]. Furthermore, transitioning to an asynchronous data transmission method eliminates network congestion, allowing information to be processed in chunks and avoiding blocking [12].

However, the theoretical justification and practical implementation of such optimization solutions require an objective quantitative evaluation of their actual effectiveness. Traditionally, analyzing the performance of a web application relies on examining key web metrics: total loading time, interface responsiveness, and the efficient utilization of hardware resources. For WebAR applications, the most resource-intensive stage is the direct initialization of the scene, during which the browser simultaneously processes large volumes of geometric data and configures the graphics pipeline [13–16]. It is at this very stage that the highest risk of blocking the main execution thread arises. Consequently, a systematic verification of the extent to which integrated optimization methods can reduce this load becomes particularly relevant as a primary condition for ensuring the stability of modern WebAR applications.

In view of the above, **the aim of this work** is the comprehensive evaluation of optimization performance and the identification of computational load distribution patterns during the initialization of WebAR applications, utilizing deep browser profiling tools.

The scientific novelty of the obtained results lies in the improvement of the 3D content preparation and initialization pipeline for WebAR applications. Unlike existing approaches, a comprehensive combination of a monolithic binary format (GLB) and spatial compression algorithms (Draco) has been applied, which made it possible to alter the load profile on the browser engine. The method for evaluating the performance of web-based AR systems has been further developed: through deep instrumental profiling, it has been quantitatively proven that the proposed optimization not only reduces the overall loading time but also radically eliminates computational bottlenecks, preventing the main execution thread from blocking.

The object of the research is the initialization and 3D content loading process within a client-optimized WebAR application running on a laptop.

Implementing scene loading optimization

The operational efficiency of WebAR applications directly depends on the speed of 3D scene initialization and the minimization of network latency, particularly when utilizing mobile internet connections. The process of loading and processing heavy 3D assets frequently becomes a bottleneck, leading to significant user wait times or the blocking of the browser's main execution thread. To address this issue, a comprehensive approach has been developed and implemented, combining a change in the data transmission format with the application of geometric compression algorithms on the server side. The technical implementation of the proposed optimization encompasses two key levels: structural and algorithmic.

At the structural level, the binary GLB format was selected instead of the standard multi-file glTF format. This solution enables the encapsulation of all model components, including geometry, textures, and material descriptions, into a single monolithic file. The primary advantage of this approach lies in the reduction of HTTP requests to a single one, which radically decreases the total server response wait time, especially under the conditions of unstable mobile networks. Furthermore, GLB format binary data is processed and parsed by the browser significantly faster, as it is transmitted in a format that is maximally close to its native representation in the GPU memory.

The algorithmic level of optimization is aimed at minimizing the payload volume during data transmission over the network. To achieve this, a 3D geometry compression mechanism has been implemented on the backend utilizing the Draco library. This method allows for the significant compression of topological data, such as vertex coordinates, indices, and normals, without any loss in the visual quality of the objects. The decompression process of the received files on the client side is executed using WebAssembly modules, ensuring high-speed operations directly within the web environment.

The applied combination of structural and algorithmic solutions ensures a faster initialization speed for the WebAR application, even when handling large 3D models.

Methodology

The performance evaluation of the optimized WebAR application, developed using Angular 19 and the A-Frame 1.7.1 library, was conducted using Chrome DevTools. A Location AR scene featuring a reference 3D model was selected for testing. The initial size of the test asset in the glTF format was 16.4 MB. To test the optimized version of the WebAR application, this same model was preliminarily converted into the binary GLB format. The file size of the converted GLB asset is 11.8 MB; it is this optimized asset that serves as the target object for subsequent performance profiling.

Given that the maximum load on the application occurs during scene initialization and loading, the research focuses exclusively on this stage. CPU utilization within the context of the web page was selected as the primary metric for performance evaluation. Because the Chrome DevTools instrumentation process inherently introduces additional overhead, each test scenario was executed 10 times. The run with average values was selected for further analysis.

Testing was conducted on a Lenovo Legion S7 laptop. The test device is equipped with an AMD Ryzen 9 6900HX processor (8 cores, 16 threads, 3.3 – 4.9 GHz), 16 GB of RAM, and operates on the Windows 11 operating system.

To ensure the objectivity and reliability of the experimental performance results for the client application featuring augmented reality elements, the testing process was conducted under conditions that closely approximate a real-world operating environment. For this purpose, the deployment of both the server and client components of the software system was carried out on remote servers. This approach eliminates the specific optimizations inherent to a local development environment and accounts for actual network latencies during the transmission of large volumes of data, such as 3D models and high-resolution textures.

Given that the target platform for augmented reality applications is predominantly mobile devices, which frequently operate under unstable wireless connection conditions, the preparation for testing included the mandatory application of an artificial network bandwidth throttling method. A network bandwidth limit of 6 megabytes per second was applied for both incoming and outgoing traffic. This bandwidth metric reflects the typical operating conditions of mobile devices within fourth-generation (4G) communication networks.

Experiment

To minimize the impact of external factors on the measurement results, a device preparation procedure was implemented. Prior to testing, a full system reboot was performed to clear the RAM and force the termination of background processes. The display brightness parameter was fixed at 80%. All third-party applications were deactivated; within the web browser, only the tab containing the WebAR application remained active.

Following the preparatory stage, the data collection procedure was implemented, which entailed executing a series of ten consecutive iterations of recording the performance profile via the Performance tab interface in Chrome DevTools. The experimental scenario encompassed the complete cycle of loading and initializing the geolocation-anchored WebAR scene.

Upon the completion of each measurement, trace files were exported, containing detailed metrics – specifically, the time expended on Scripting, System processes, Rendering and Painting. In addition, the total scenario recording time and JavaScript heap allocation dynamics were recorded.

To select the most representative sample for subsequent analysis, a comparison of the peak JS Heap memory usage values and the durations of key processing stages was conducted. It is important to note that the Total metric also accounts for system idle periods. The obtained quantitative data have been systematized and are presented in Table 1.

Table 1 – **Test result of the optimized version on the Lenovo Legion S7 device**

Test	JS Heap (peak), mb	Scripting, ms	System, ms	Rendering, ms	Painting, ms	Total, ms
1	20,6	1285	967	7	0	5839
2	21,3	1487	945	29	29	5779
3	24	1495	920	29	29	5749
4	24,8	1458	947	31	30	5915
5	31,1	1339	795	31	30	5735
6	24	1432	837	32	30	5773
7	35,2	1366	834	31	31	5737
8	23,9	1343	773	28	27	5752
9	27,1	1294	839	30	29	5819
10	24	1326	783	29	29	5794

Selection of a representative recording

To identify the most representative record among the ten obtained results, t*he normalized deviation method* was employed. This approach objectively determines which test run most closely aligns with the average execution conditions. For each test, the normalized deviation was calculated using the following formula:

$$Score_i = \sum_j \frac{|X_{ij} - Me_j|}{IQR_j}, \quad (1)$$

where X_{ij} is the value of the jth metric for the ith test; Me_j - the median of the corresponding metric; IQR_j is the interquartile range of this metric.

The interquartile range is calculated by the formula:

$$IQR = Q3 - Q1, \quad (2)$$

where Q1 is is the 25th percentile (lower quartile); Q3 is is the 75th percentile (upper quartile).

For a sample size of $n = 10$, the calculation was performed using Tukey's method: the ordered dataset was divided into two equal halves of 5 elements each. In this case, is determined as the median of the first half of the dataset – the 3rd element in the ordered list – while Q3 is determined as the median of the second half, the 8th element. Based on Tukey's method, basic statistical parameters were determined for each metric. The results of identifying the median, Q1, Q3 and calculating the IQR (2) for each column are presented in Table 2. The result of calculating the normalized deviation (1) for each test is given in Table 3.

Table 2 – **Indicator values based on Tukey's method**

Metric	Median	Q1 (3rd element)	Q3 (8th element)	IQR
JS Heap	24	23,9	27,1	3,2
Scripting	1354,5	1326	1458	132
System	838	795	945	150
Rendering	29,5	29	31	2
Painting	29	29	30	1
Total	5776	5749	5819	70

Table 3 – **Normalized deviation of tests on Lenovo Legion S7**

Test	Score	Test	Score
1	43,5990	6	2,8866
2	2,8537	7	6,9209
3	2,2468	8	3,6446
4	5,4965	9	2,2980
5	4,9586	10	1,0897

The obtained values showed that Test No. 10 has the lowest normalized indicator, indicating a minimal deviation from the median characteristics. Accordingly, this record was chosen as representative for further comparison between devices.

Analysis of a representative recording

The representative trace, obtained in the Performance tab of the Chrome DevTools suite, is shown in Fig. 1.

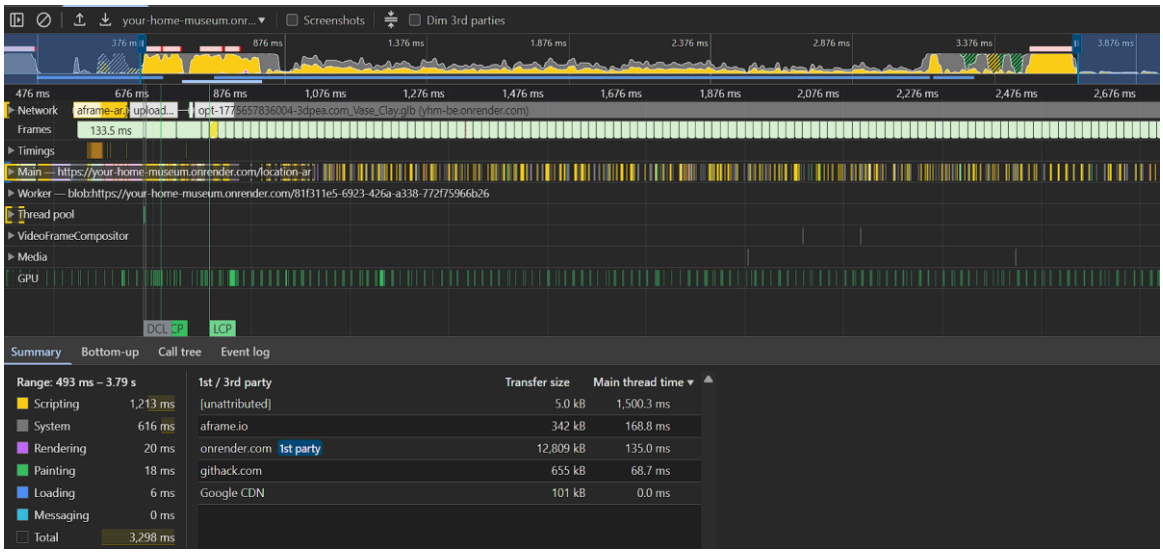


Fig. 1. Representative recording

For the analysis, the application initialization period was considered; therefore, the intervals at the beginning and end of the record were excluded. The main visual metrics include: the CPU activity graph at the top of the figure, the Main thread where function calls are recorded, and the Summary pane.

The Summary pane indicates that the total recording time is 3.29 s, of which 1.21 s is allocated to Scripting, 1.42 s to Idle, 0.61 s to System overhead, 0.02 s to Rendering, and 0.01 s to Painting. These data are presented as a diagram in Fig. 2. Loading and Painting are not considered in the detailed breakdown, as they account for less than 1% of the total time. Thus, the largest share is occupied by system downtime (Idle) with a metric of 43%, followed by the Scripting process (37%).

The CPU graph is conditionally divided into four intervals (Fig. 3), two of which demonstrate a high load on the processor, as evidenced by the fully filled areas. The unhighlighted intervals were excluded, as they do not directly relate to the initialization of the application.

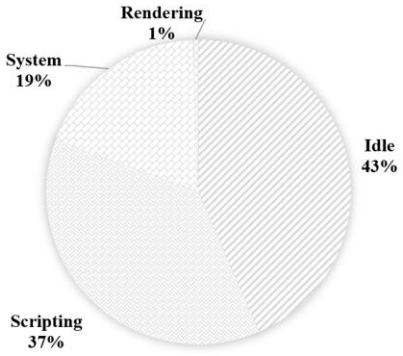


Fig. 2. Process diagram when loading an application

The initial profiling stage, spanning the time interval from 0.491 to 0.924 seconds, is characterized by a high computational load driven by the initialization of the client application components. The majority of this time, specifically 0.354 seconds or over 80% of the interval's duration, is allocated to Scripting.

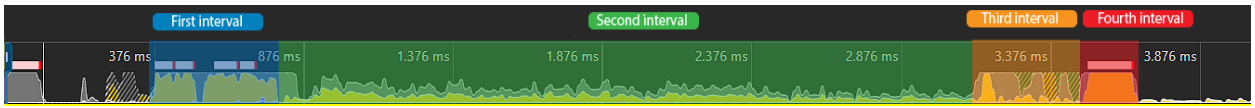


Fig. 3. Dividing a recording into intervals

The longest duration, at 135 milliseconds, was recorded in the Unattributed category. According to the analysis, this corresponds to the execution of native code by the JavaScript engine; however, due to the specific optimization features of the Angular framework, the profiler is unable to identify these calls in greater detail. The second most significant operation is the initialization of the A-Frame spatial library, which lasts 119 milliseconds.

The subsequent stage lasts from 0.924 to 3.25 seconds and is distinguished by moderate computational intensity. The main thread remains in an idle state for 1.302 seconds, accounting for 56% of the interval's time. Such a low processor load is explained by the wait for the network transmission of the 3D model from the server. Computational activity during this phase is limited to non-blocking operations for reading the input stream and executing background processes.

The third phase, spanning from 3.25 to 3.6 seconds, is dedicated to processing the received model data. The load on the main thread increases, with Scripting occupying nearly 39% of the time. The parsing of binary data and its translation into internal scene structures take place during this period. Simultaneously, the garbage collector is activated, optimizing RAM usage down to 8 megabytes. Background thread pools are also initialized, facilitating the parallel decoding of resources and effectively offloading the main thread.

At the final stage, within the interval from 3.6 to 3.78 seconds, the actual rendering process occurs. A peak CPU load is recorded here, with 87% of the time allocated to Scripting. Basic rendering methods, shader program compilation, the uploading of textures to video memory, and the final visualization of the object are identified within this phase.

Thus, the entire loading process is decomposed into four sequential stages: client environment initialization, lasting 0.433 seconds; a network wait period, lasting 2.326 seconds with predominant system idle time; data parsing and structuring, taking 0.350 seconds and utilizing multi-threaded optimization; and the final phase of graphics preparation and rendering, lasting 0.180 seconds with a peak load on the main thread.

Load estimation

To determine the most demanding stage of WebAR application initialization, four intervals were chosen: initialization, loading, parsing, and rendering. The load index for each interval was calculated using the formula:

$$Load = \frac{T_{busy}}{T_{total}} * 100\%, \quad (3)$$

where T_{busy} is total time the main thread performed any tasks (Scripting, Rendering, Painting, System); T_{total} is total interval duration.

The calculation results are given in Table 4.

Table 4 – **Processor performance characteristics at key initialization stages**

№	Stage	Interval, s	T_{total}, ms	T_{busy}, ms	Idle, ms	Load, %
1	Initialization	0.491 – 0.924	434	425	9	97,9%
2	Loading	0.924 – 3.25	2328	1025	1303	44,0%
3	Parsing	3.25 с – 3.6	354	242	112	68,4%
4	Rendering	3.6 – 3.78	188	184	4	97,9%

The highest density of computational operations within the optimized profile is exhibited by the initialization and rendering stages, where peak CPU utilization values are recorded at 97.9% (Fig. 4). Despite the brief overall duration of these phases, amounting to 434 and 188 ms respectively, the processor has virtually no idle time. These stages demonstrate identical utilization metrics but differ in the nature of their operations: during the initialization phase, the processor is primarily engaged in JIT compilation and the execution of library JavaScript code, whereas during rendering, the primary load is attributed to interacting with the graphics engine and frame preparation.

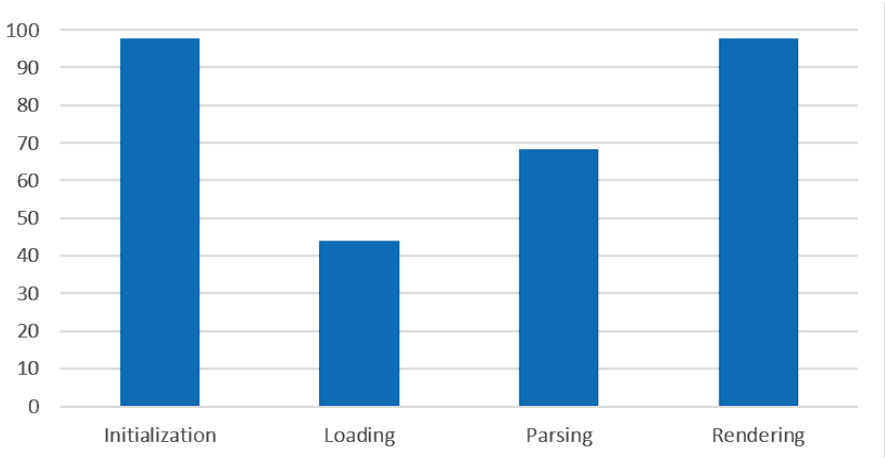


Fig. 4. Percentage load graph of intervals

The third interval, which corresponds to the parsing stage, is characterized by a utilization level of 68.4% over a total duration of 354 ms. The presence of 112 ms of idle time during this stage indicates the successful optimization of the 3D model's binary data deserialization and in-memory object processing. This facilitated the elimination of the severe main thread blocking inherent in the baseline version of the application.

As expected, the Loading stage remains the least resource-intensive, with a utilization metric of 44.0%. This distribution is due to the fact that network bandwidth acts as the limiting factor during this interval; consequently, the processor's main thread spends more than half of the phase's total duration – specifically, 1303 ms – in an idle state, waiting for data.

Thus, in the optimized version, the highest computational load is observed during the fourth interval, Rendering, whereas the Loading stage is the least demanding and exhibits the highest idle time.

Performance benchmarking

To objectively evaluate the effectiveness of the proposed approaches for optimizing WebAR scene loading, it is advisable to conduct a comparative analysis of the obtained experimental data. The performance metrics of the baseline application version were selected as the benchmark for comparison. This enables a quantitative assessment of the performance gain and the reduction in the load on the device's hardware resources.

The primary criteria for comparison are the peak allocated memory and the distribution of processor time among native code execution, system calls, and main thread idle time. The summarized instrumental profiling results for both architectural solutions are presented in Table 5.

The conducted comparative analysis of the performance profiles of the baseline and optimized application versions demonstrates a significant increase in the overall efficiency of WebAR scene initialization. Overall, the implemented architectural changes ensured a more efficient utilization of computational resources and eliminated blocking factors within the 3D asset preparation pipeline. A major achievement is the substantial reduction in the load on the device's RAM. The peak allocated memory decreased by nearly half – by 47.7%. This confirms the high efficiency of the applied data compression methods, which significantly reduced the volume of objects generated in memory during the decompression of the scene's geometry.

Table 5 – **Comparison of performance indicators of the basic and optimized versions of the WebAR application**

Metric	Basic version	Optimized version
JS Heap (peak), mb	45,9	24
Scripting, ms	1462	1213
System, ms	586	616
Idle, ms	2593	1425
Total, ms	4749	3298

The total scene initialization time was reduced by 30.5% – dropping from 4749 to 3298 ms. The transition to the monolithic binary GLB format radically accelerated the network stage of asset retrieval. According to the detailed profiling metrics, the primary driver of this overall acceleration is a 45% reduction in wait time and main thread idle time – decreasing from 2593 to 1425 ms. The obtained results convincingly demonstrate the effectiveness of the architectural solution in eliminating network bottlenecks and optimizing data processing through the synergy of the binary container and geometry compression.

The most significant time reduction was recorded during the Parsing phase (Fig. 5). Its duration decreased by 66.2%, dropping from 1048 to 354 ms. This indicates

the successful elimination of the primary computational bottleneck, achieved through the optimization of the 3D geometry decompression and processing pipeline. The Loading stage, which traditionally occupies the largest absolute share of the total time, accelerated by 20.6%, reducing its metric from 2933 to 2328 ms. The Initialization phase demonstrated a 22.6% acceleration, decreasing from 561 to 434 ms. The smallest relative and absolute performance gain is observed during the Rendering stage, where the execution time decreased by 13.4% – from 217 to 188 ms. Given that the rendering stage was not initially a defining blocking factor in the baseline architecture, such an improvement is a logical and proportional consequence of the overall offloading of the main thread and the optimization of the preceding phases.

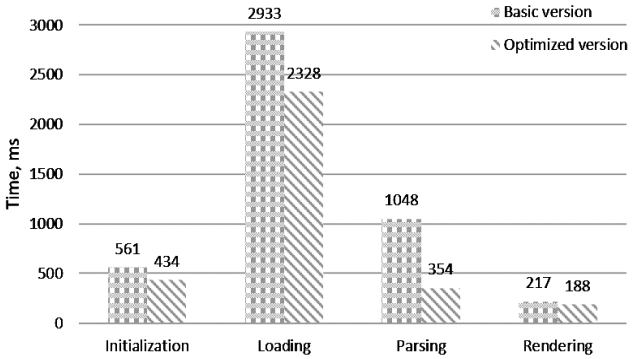


Fig. 5. Comparison of T_{total} at each phase of the base and optimized versions

A 17% reduction in Scripting duration was recorded, dropping from 1462 to 1213 ms, which further attests to the successful optimization of the parsing stage. Concurrently, a slight 5.1% increase in System overhead, reaching 616 ms, is a logical and acceptable trade-off caused by the overhead associated with managing background threads.

The primary consequence of the implemented architectural changes was the targeted redistribution of the computational load among the data preparation phases (Fig. 6).

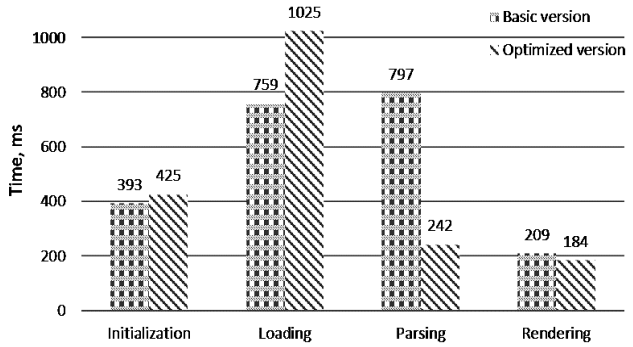


Fig. 6. Comparison of T_{busy} at each phase of the base and optimized versions

The most significant reduction in the volume of active computations was recorded during the Parsing phase, where T_{busy} decreased by 69.6%, dropping sharply from 797 to 242 ms. Such a radical offloading of the main thread confirms the successful elimination of the critical bottleneck, achieved through the application of compression and the optimization of scene geometry processing algorithms.

In contrast, during the Loading stage, a logical increase in active processor time by 35.0% is observed, specifically from 759 to 1025 ms.

This increase is not indicative of performance degradation but rather serves as a direct result of implementing the optimization strategy. Instead of passively waiting for the network download, the system preemptively executes preparatory computational tasks, effectively utilizing processor time.

During the Initialization phase, a minor increase in active load of 8.1% was recorded, specifically from 393 to 425 ms. This can be logically attributed to the initial overhead associated with preparing the background thread infrastructure and initializing the decompression modules. Conversely, the Rendering stage demonstrated a 12.0% decrease in processor time consumption, dropping from 209 to 184 ms, which is a direct consequence of operating on an already optimized and pre-processed 3D model structure.

As a result, the recorded shift in the profile of T_{busy} illustrates a qualitative transition from a sequential and blocking execution of tasks to a balanced and parallel data processing pipeline.

Conclusions

In this work, a comprehensive performance evaluation of WebAR application scene loading optimization was conducted. The results of instrumental profiling confirmed the high efficiency of the applied approach, which is based on transitioning to the binary GLB format and utilizing Draco geometric compression. The implemented architectural changes successfully eliminated critical bottlenecks in the 3D asset preparation pipeline.

A significant reduction in the load on hardware resources was recorded: the peak allocated JS Heap memory decreased by 47.7%. The overall scene initialization time was reduced by 30.5%, dropping from 4749 to 3298 ms.

The primary driver of this acceleration was a 45% reduction in main thread idle time and a radical 66.2% decrease in the duration of the parsing phase.

It has been proven that the integration of a monolithic format and client-side decompression leads to a natural and more balanced redistribution of computational resources, allowing for the efficient utilization of processor time during the loading stage. This guarantees improved stability and performance for modern WebAR applications.

Conflicts of interest

The author declare that he has no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The author confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Nadeem M., Lal M., Cen J., Sharsheer M. AR4FSM: Mobile Augmented Reality Application in Engineering Education for Finite-State Machine Understanding // Education Sciences. 2022. Vol. 12. No. 8. P. 555. DOI: https://doi.org/10.3390/educsci12080555
2. Parekh P., Patel S., Patel N., Shah M. Systematic Review and Meta-Analysis of Augmented Reality in Medicine, Retail, and Games // Visual Comput. Ind. Biomed. Art. 2020. Vol. 3. No. 1. Art. 21. DOI: https://doi.org/10.1186/s42492-020-00057-7
3. Butt A., Ahmad H., Muzaffar A., Ali F., Shafique N. WOW, the Make-Up AR App is Impressive: A Comparative Study Between China and South Korea // Journal of Services Marketing. 2022. Vol. 36. No. 1. P. 73–88. DOI: https://doi.org/10.1108/JSM-12-2020-0508
4. Yin C. Z. Y. et al. Mobile Augmented Reality Heritage Applications: Meeting the Needs of Heritage Tourists // Sustainability. 2021. Vol. 13. No. 5. Article 2523. DOI: https://doi.org/10.3390/su13052523
5. Divya Udayan J. et al. Augmented Reality in Brand Building and Marketing – Valves Industry // 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). 2020. P. 1–6. DOI: https://doi.org/10.1109/ic-ETITE47903.2020.425
6. Ghattas M. M. Performance Evaluation of Websites Using Machine Learning // EIMJ. 2020. Vol. 51. P. 36–41.URL: https://www.researchgate.net/publication/345975296_Performance_Evaluation_of_Websites_Using_Machine_Learning
7. Matvieiev M. I. Analysis of Optimization Methods for Augmented Reality Web Applications // Control, Navigation and Communication Systems. 2024. No. 4(78). P. 106–108. DOI: https://doi.org/10.26906/SUNZ.2024.4.106
8. Alsulami M. H. et al. Development of an Approach to Evaluate Website Effectiveness // Sustainability. 2021. Vol. 13. No. 23. Article 13304. DOI: https://doi.org/10.3390/su132313304
9. Boutsi A.-M., Ioannidis C., Verykokou S. Multi-Resolution 3D Rendering for High-Performance Web AR // Sensors. 2023. Vol. 23. Article 6885. DOI: https://doi.org/10.3390/s23156885
10. Maclntyre B., Smith T.F. Thoughts on the Future of WebXR and the Immersive Web // Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct. 16–20 October 2018. P. 338–342. DOI: https://doi.org/10.1109/ISMAR-Adjunct.2018.00099
11. gLTF Runtime 3D Asset Delivery. Available online: https://www.khronos.org/gltf/
12. Li L., Qiao X., Lu Q., Ren P., Lin R. Rendering Optimization for Mobile Web 3D Based on Animation Data Separation and On-Demand Loading // IEEE Access. 2020. Vol. 8. P. 88474-88486. DOI: https://doi.org/10.1109/ACCESS.2020.2993613
13. Ghattas M., Mora A. M., Odeh S. A Novel Approach for Evaluating Web Page Performance Based on Machine Learning Algorithms and Optimization Algorithms // AI. 2025. Vol. 6. No. 2. P. 19. DOI: https://doi.org/10.3390/ai6020019
14. Kumar A., Arora A. A Filter-Wrapper based Feature Selection for Optimized Website Quality Prediction // Proceedings 2019 Amity International Conference on Artificial Intelligence (AICAI). 2019. P. 284–291. DOI: https://doi.org/10.1109/ai-cai.2019.8701362
15. Allison R. et al. A Comprehensive Framework to Evaluate Websites: Literature Review and Development of GoodWeb // JMIR Formative Research. 2019. Vol. 3. P. e14372. DOI: https://doi.org/10.2196/14372
16. Morales-Vargas A., Pedraza-Jimenez R., Codina L. Website quality in digital media: literature review on general evaluation methods and indicators and reliability attributes // Revista Latina de Comunicacion Social. 2022. Vol. 80. P. 39–63. DOI: https://doi.org/10.4185/RLCS-2022-1515

Received (Надійшла) 11.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Матвєєв Микита Іванович – аспірант кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Mykyta Matvieiev – PhD student, Department of Computer Engineering and Programming, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: Mykyta.Matvieiev@cit.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0000-8773-6640;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=60141774400.

Оцінювання продуктивності оптимізації завантаження сцени у вебAR-додатку

М. І. Матвєєв

Анотація. Актуальність. WebAR стрімко розвивається, проте стикається з проблемою високого обчислювального навантаження на пристрої. Синхронне завантаження важких 3D-моделей під час ініціалізації сцени часто призводить до затримок і блокує основний потік виконання браузера. Застосування комплексних методів оптимізації (таких як перехід на формат GLB і використання стиснення геометрії Draco) є критично важливим для забезпечення стабільності, однак їх впровадження потребує об'єктивної кількісної оцінки. **Мета дослідження** – оцінити продуктивність оптимізації завантаження сцени WebAR-додатку на етапі ініціалізації. **Об'єкт дослідження** – процес ініціалізації та завантаження 3D-контенту у клієнтському WebAR-додатку, що працює на ноутбуці. **Предмет дослідження** – показники продуктивності методів оптимізації WebAR-додатку. **Висновки.** За результатами інструментального профілювання підтверджено високу ефективність переходу на формат GLB та використання стиснення Draco. Зафіксовано зменшення пікового обсягу виділеної пам'яті JS Heap на 47,7%. Загальний час ініціалізації сцени скоротився на 30,5% – з 4749 до 3298 мс. Крім того, час простою основного потоку суттєво зменшився на 45%, а тривалість фази парсингу – на 66,2%, що дозволило усунути критичні вузькі місця у конвеєрі підготовки 3D-ресурсів.

Ключові слова: веб, доповнена реальність, WebAR, оцінювання продуктивності, оптимізація, Draco, Angular.

UDC 004.93 **doi: 10.26906/SUNZ.2026.2.153**

Olena Peredrii, Oleksii Gorokhovatskyi

Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine

THE EXPLAINABILITY OF SHALLOW AI-GENERATED TEXT CLASSIFICATION MODELS VIA PARTS REMOVING

Abstract. In this paper, we address the explainability problem for the ANNs' classification of AI-generated and human-written text chunks in Ukrainian texts in the IT domain. The objective is to investigate whether the perturbation-based modifications of text chunks that include the removal of sentences, words, and word combinations may be helpful in searching for explanations. We used five shallow ANN models (with an average accuracy of about 0.88) and tested them on a sample of the document containing human-written text and AI-generated fragments generated with GPT-5, Gemini 2.5 Flash, and Claude Sonnet 4.5. The experimental modeling showed that it is not easy to find a single sentence or word that can flip the classification result. We have proposed an explainability index that measures the total influence of all perturbed samples on the classification result, accounting for the fact that short perturbations are more valuable.

Keywords: explainability, black-box, shallow ANN, perturbation, AI-generated content, human-written content, text chunk, text classification, explainability index.

Introduction

The integration of artificial intelligence (AI) systems into essential human domains, such as healthcare, finance, education, and identification systems, as well as autonomous decision-making, requires a high level of explainability for their processes and outcomes. The inability to trace how a specific input leads to an output poses significant model usage risks, which are relevant both for modern transformer-based architectures like LLMs and for traditional AI methods and tools like artificial neural networks [1, 2].

The paper analyzes the application of artificial neural networks (ANNs) for classifying AI-generated and human-written text chunks (e.g., paragraphs) in Ukrainian. We do know that the application of AI detectors is commonly recommended to be limited in educational environments, so our primary goal is mostly research-oriented. In particular, we are interested in identifying details that may help explain why the model predicts the specific class and which parts of the text are primarily responsible.

1. Literature review

Traditional AI models used for classification or regression rely on explicit algorithms and simpler internal structures compared to modern large language models. These systems typically use fixed inputs to predict a specific label, and their explainability methods focus on revealing the reasoning behind these discrete choices.

There are different classifications of existing Explainable AI (XAI) methods; the most popular is whether the weights of the initial black-box model are required. Model-agnostic methods treat the original model as a black-box entity without access to its weights; on the other hand, model-specific methods leverage access to the model's structure and weights [13]. Perturbation-based methods are representatives of model-agnostic approaches that modify the initial signal (numerical features, images, text, etc.) representing the entity being classified and investigate how these changes affect black-box model outputs. According to [12], there are also score-based methods, explanations by simplification, language explanations, and different methods that can fall into multiple of these categories.

XAI explainable methods can also be categorized into local and global [15], where local methods search for explanations based on the particular input, and global methods analyze the behavior of the black-box model as a whole.

There are many different explanation methods already known for convolutional neural networks (CNNs) and image classification problems [3]. They include the CAM/Grad-CAM family [4, 5], methods based on the weighted sum of feature maps from the final convolutional layer, or on gradients flowing into it. The following researches about Integrated Gradients [6] (calculates the integral of gradients along a straight-line path from a "baseline" input to the actual one) and DeepLift [7] (compares the activation of each neuron to a reference state and decomposes the output prediction based on these differences) addressed the common issues for the gradient-based methods: the saturation of neurons even for extremely important features of the signal, as well as that gradients provide the information only about some neighborhood around the point that may not reflect the global importance of this feature.

One of the most famous perturbation-based explainability methods is LIME (Local Interpretable Model-agnostic Explanations) [8]. The research asserts that explanations must be interpretable (human-readable) and locally faithful (accurately reflecting how the model behaves in the immediate vicinity of the data point being explained). LIME creates a "neighborhood" of the input signal by randomly perturbing it; afterward, these perturbed samples are fed into the complex "black-box" model to observe how the predictions change. LIME builds the local interpretable model based on how close these samples are to the original input signal. This research provided the first universal "boilerplate" for explaining complex models using simple local approximations.

The application of LIME for text classification problems was investigated in [10]. The paper is focused on the TF-IDF transform and the sampling mechanism, which are the two critical components LIME uses to handle text. By analyzing these, the research determines

if LIME's output is a stable and faithful representation of the complex model being explained. The random removal of separate words was used as a sampling technique; it was shown that LIME provides meaningful explanations for decision trees and linear models.

The main benefit of LIME is its broad applicability to any classifier, making it a convenient tool for proprietary or inaccessible models. However, LIME's process requires repeated evaluations of the target model, which can be computationally intensive, especially for large LLMs or long text sequences. This method can also introduce noise.

The SHAP (SHapley Additive exPlanations) proposed in [9] provides a theoretical framework that unifies several existing explanation methods—including LIME, DeepLIFT, and Layer-Wise Relevance Propagation. SHAP assigns an importance score (within the scope of the particular prediction case) to each input feature, which is considered mathematically fair for distributing the prediction score among all input features.

Model-agnostic techniques are commonly applied to artificial neural networks (ANNs) and tabular data. LIME constructs local surrogate models via input perturbations, whereas SHAP employs game-theoretic methods to equitably attribute feature impact. Although both methods are theoretically robust, SHAP incurs high computational costs, and LIME may exhibit instability due to random sampling.

The research [11] describes the investigation of how Integrated Gradients (IG) can be used to provide word-level explainability for text classification models. The BERTAgent was used as a primary deep-learning classifier, and integrated gradients were applied to reveal which specific words contribute to a model's decision. It was concluded that the quality of the output strongly depends on the specific label, the availability of strong (and statistically coherent) semantic content in the data, and whether the classifier captures it. It stated that the application might be useful for extending their theoretical background in classification problems based on uncertain data.

The paper [14] introduces ProtoLens, which searches for explanations in sub-sentence text chunks rather than words, making the results more human-understandable. The model maps text spans to known prototypes and assigns scores and weights to prototypes based on their similarity to the span. The modeling showed that ProtoLens outperforms various existing baseline methods, providing intuitive and detailed explanations. The main limitation of the approach, as mentioned by the authors, is its reliance on the training data, which may be biased.

The HINT (Hierarchical Interpretable Neural Text classifier) is a model that generates the hierarchical representation of label-associated topics of model predictions for word-level, sentence-level, and document-level analysis [15]; the authors claimed that the interpretations by words and phrases are not sufficient. Results of experiments for two datasets confirmed the effectiveness of the HINT, comparable to the SOTA classifiers, and its ability to generate faithful and humans-readable interpretations.

The detection of adversarial attacks on text classifiers based on explainable AI with integrated gradients is proposed in [16]. The core of the method is the gradient-based identification of words that affect the prediction result positively and negatively, with the following replacements of such words with synonyms instead of replacing random words. This method shows an example of how an XAI method could be used both to detect adversarial attacks and to find useful insights about the model's behavior. The main drawbacks of the methods, which are important for us in this paper, are the requirement to use gradients and the necessity to generate high-quality synonyms in the correct form.

As a brief summary of the methods described above, we can conclude the following. There are many successful methods addressing various issues in English, but their applicability, for instance, to Ukrainian, is unclear. Known methods use NLP techniques such as finding appropriate synonyms, which are language-specific. The majority of methods work at the word level, which is suitable for short text chunks, but does not seem to be a very good choice for the paragraphs we are dealing with in the current problem. Finally, most approaches use gradients, and our primary focus is the model-agnostic perturbation-based family.

The contribution of the paper includes:

– the research of the possibility to find the explanations for the black-box shallow ANN models for text chunks classification in Ukrainian using the removal of sentences and n-grams as sampling strategies;

– measuring the explainability index for the particular decision, evaluating the ability to find the complementary text pairs.

2. Ethical considerations

The unreliable performance (mainly in the form of false-positive classification errors) of the majority of modern common-knowledge AI detectors is a known issue. The justified solution we can see frequently involves avoiding AI detectors in education for students or educators, but it is recommended to use improved teaching and process-based assessment methods, oral exams, and discussions, and to educate students on the responsible use of AI tools instead. But it is worth noting that specialized AI detectors are used in academia anyway.

In this and our previous papers, we follow this recommendation: we are building an AI system to detect AI-generated content represented as documents in Ukrainian for scientific purposes. Our main goal is to investigate whether it is possible to implement really honest detection in narrow-field domain, with limited technical resources using shallow ANN architectures [17] and LLM available to everyone (almost) for free, what could be the quality of the solution (for own challenging dataset), and whether it is possible to find the explanations of the particular decision. The application of the tool implemented in this paper is limited to personal use, e.g., testing one's own documents to understand the responsibility, without any official consequences or automatic decisions.

3. Models

We used the ANN models described in detail in [17]. The dataset contained 5167 text chunks (2533 human-written and 2634 generated with the GPT-4o-mini model) in Ukrainian, with the texts spanning bachelor's theses in the IT specialty.

This dataset was used to train various shallow ANN models with grid search, five best models were selected for the further experiments, their architectures and accuracies are presented in Fig. 1. As one can see, these models can predict whether the text chunk was human-written or AI-generated with the rough probability about 0.88, but all models are black-box completely and are not capable of producing any explanations of the decision being made. Class label 0 for all models means "human-written", and label 1 means "AI-generated" content.

It is worth noting that the dataset is challenging. There are many cases where humans cannot correctly determine whether a fragment was generated.

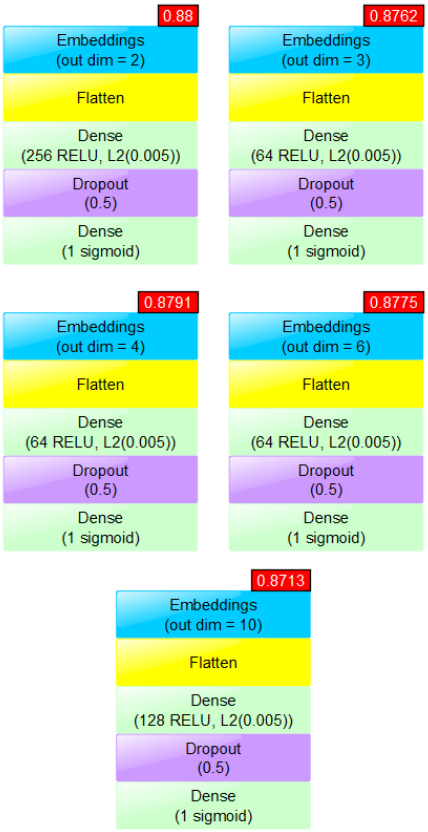


Fig. 1. Architectures of ANN and their accuracies

4. Good explanation

Perturbation-based methods are probably the easiest ones to search for the explanations without using the model's weights and its internal structure. They could be applied to different types of input signals, e.g., images or texts, to evaluate the importance of signal parts for the final decision. There are different types of perturbations: signal part removal, replacement with a predefined constant value, different values, specific modifications to include another signal within the initial one, and so on.

In our previous research [18, 19], we applied the search for complementary parts of the signal to be sure it affects the decision at the required level: we were interested in such a pair of modifications (perturbations) of the initial signal, one of which still preserves the same classification result as the original signal, the second one – changes it. We called such a pair of signals a complementary image pair (CIP) for the image classification problem [18, 19].

The model's initial input is a chunk of text for the problem being investigated. The perturbation of the text segment that leads to the change in classification result is our primary interest.

In this paper, we claim that a good explanation is a piece of input signal that is minimal in size (has minimal length of text for the current problem), changes the classification result, and the presence of this piece is enough to preserve the initial prediction.

So, we can introduce the measure of perturbation size to be (1 – (size of perturbation / size of chunk)) and the influence of this sentence on the result as |etalon prediction – perturbed prediction|. The quantitative measure of the explainability index (EI) of the particular decision by the black-box model may include the normalized sum of explainabilities for all perturbations:

$$EI = \sum_{i=1}^{n} \left(1 - \frac{l_{perturb}}{l_{text}}\right) \left| pred_{et} - pred_{perturb} \right| \qquad (1)$$

where n – is the quantity of perturbations (sentences/n-grams to be removed), $l_{perturb}$ – length of the perturbed fragment, l_{text} – length of the entire chunk, $pred_{et}$ – etalon (initial) prediction for the original chunk, $pred_{perturb}$ – prediction for the perturbed chunk.

The qualitative measure could be a simple flag (yes/no) indicating whether the explanation that flips the classification was found. In this context, it is also clear that it is easier to find an explanation when the initial decision is less confident (for instance, the probability of AI-generated text is 0.75 rather than 0.95). Changing the decision of the etalon model seems to be a good criterion, but just measuring the influence of the parts is easier to implement.

5. Sampling

We have considered sentence, word, and n-gram removal. We also briefly tried sampling based on sentence replacement, but found that mapping the part to be replaced to a similar, grammatically correct human-written chunk is complex.

The first sampling strategy we have tried is the complete removal of the sentence and classification of the text chunk without it. Removing sentences to measure their influence seems like a pretty obvious idea, but it could be tricky, given that the models we use were designed, trained, and tested to work with text chunks of about 1000 symbols. So, their accuracy is not guaranteed for text chunks that may differ significantly, and the best choice seems to be removing only short sentences. This idea is limited and not applicable when the text chunk is just a single sentence.

The entire algorithm to analyze the influence of the sentence includes the steps below.

1. The classification of the original chunk of the text (paragraph) without any perturbations, saving its class (etalon class) and output (etalon) prediction.

2. The classification of the text chunk without the current sentence (perturbed chunk) with the particular model.

3. Measuring the difference between perturbed and etalon predictions.

4. If the perturbed classification result and the etalon result are different:

4.1 It means that the explanation is found, and ignoring just the current sentence changes the prediction.

4.2 Exploring all combinations (preserving order) of the sentences from the original chunk, which include the current sentence, but may not include any other sentences, searching for the combination that preserves the etalon class. This confirms that the current sentence can guarantee the etalon class almost alone. We call such a chunk, paired with the perturbed one, a complement text pair (CTP). The number of combinations may be significant when the text paragraph contains even a few sentences, so this step may be stopped immediately after the first CTP is found, or limited to only a few combinations to find the one with minimal length, etc.

The second sampling strategy was the removal of n-grams for n = 1..5 (formed without grams intersection in the scope of the single sentence) in the same testing pipeline as sentences.

6. Results and discussion

The numerical evaluation included calculating the explainability index and searching for a complementary text pair (CTP) during the classification of the particular text paragraph. As the main problem of all AIG content classifiers and detectors is false positives: they often detect human-written text as AIG, especially for non-native speakers, we were interested in the explainability of the text pieces that were classified as AI-generated by the average decision of all 5 nets, with the probability greater than 0.75.

According to [17], all ANNs were trained on the dataset in a relatively narrow IT domain in Ukrainian, so we evaluated the explainability using one of our unpublished drafts of the research from 2020 about the application of shallow convolutional neural networks for decision-making.

We modified the initial document by replacing paragraphs in various sections with AI-generated ones. We used the chatbot GUI of ChatGPT (GPT5), Gemini 2.5 Flash, and Claude Sonnet 4.5, applied the same prompt “напиши подібний до наступного фрагменту текст” (English: “write a text similar to the following fragment”), and replaced some paragraphs.

Thus, we obtained an initial human-written document (18 pages in total) and three versions of it with AI-generated pieces in comparable form.

The entire processing of the document included:

– splitting text into pieces of the appropriate size (about 1000 symbols) as all ANNs were trained for this size;

– classification of each piece with all 5 networks, averaging their outputs;

– if the average decision is greater than 0.5, it means the text is classified as AI-generated, and we are interested in cases when the decision is greater than 0.75; we didn’t control whether the classification result is correct according to ground-truth, but we remember that the accuracies of all these networks are about 0.88;

– each such potentially AI-generated chunk was processed with the explanation module, which perturbed the text to explore how the particular ANN model reacts to changes, searching for the influence of every sentence, calculated explainability index, and CTP (if available);

– all results were combined into an HTML web page for further visual investigation.

An example of an explanation representation is shown in Fig. 2.

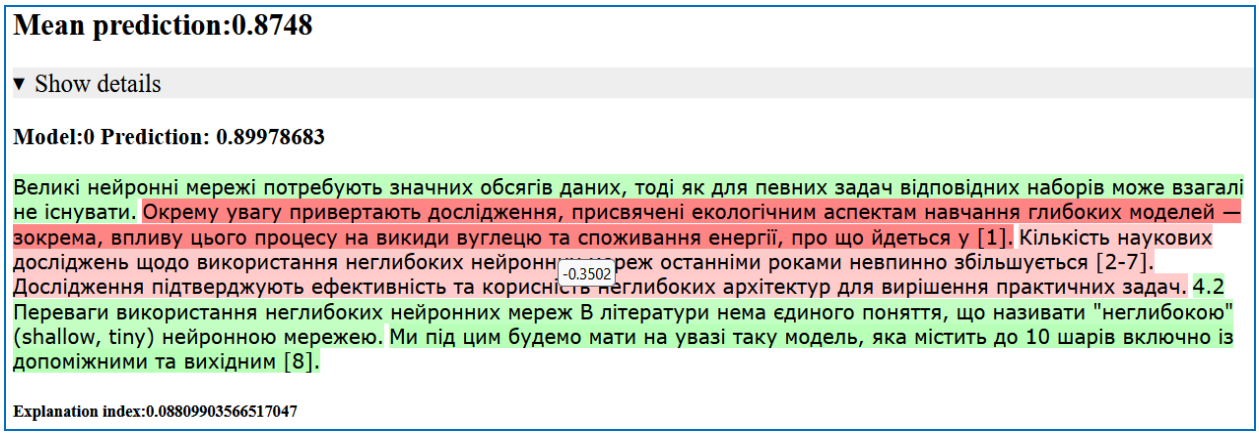


Fig. 2. Example of explanation representation

One can see the mean prediction calculated from the outputs of all 5 nets participating in the processing. The first model (having ID=0) classified the text sample as AI-generated with a probability of about 0.9. The text sample follows next. Different sentences are highlighted in gradient red and green, depending on whether removing the sentence increases (green) or decreases (red) the neural network output. If the prediction decreases after removing it, it means that the decision becomes closer to 0, which corresponds to a human-written class, so the sentence

being removed votes for the AI-generated prediction. Hovering the mouse over each sentence shows the change in the overall classification decision. In this example, removing the second sentence reduced the network's prediction by -0.3502, which is why it has a strong red background. Finally, the explanation index that is about 0.088 for this example is shown under the text fragment. The decomposition of this index according to (1) shows that there are 6 sentences in the fragment, their lengths are 127, 190, 120, 106, 147, and 109 symbols respectively, and the overall length with trailing spaces is 804. The absence of each sentence in the paragraph changes the classification result by such absolute values: 0.058, 0.3502, 0.0056, 0.0039, 0.0425, 0.095, so just the second sentence has a significant influence on the model's output. Summing up all these values with respect to their inverse length coefficients makes the overall normalized explanation index approximately 0.088.

This confirms that the explanation index proposed by (1) is strict enough and may be revised, e.g., taking into account only the sentence with the greatest impact. Its high value shows that there are some short parts of the signal (sentences, words) that affect the ANN prediction.

The example of the CTP is shown in Fig. 3. The sentence on the red background is marked as bold; it means its removal is so significant that it flips the overall classification results, reducing it by 0.4316 from 0.9095. Now the complementary pair for this text is shown below: it still contains this important second sentence and preserves the initial classification prediction (output is greater than 0.99), but without the last sentence of the text chunk. There are multiple such complementary pairs without some other sentences, but all of them contain the sentence in bold and confirm its importance for the decision.

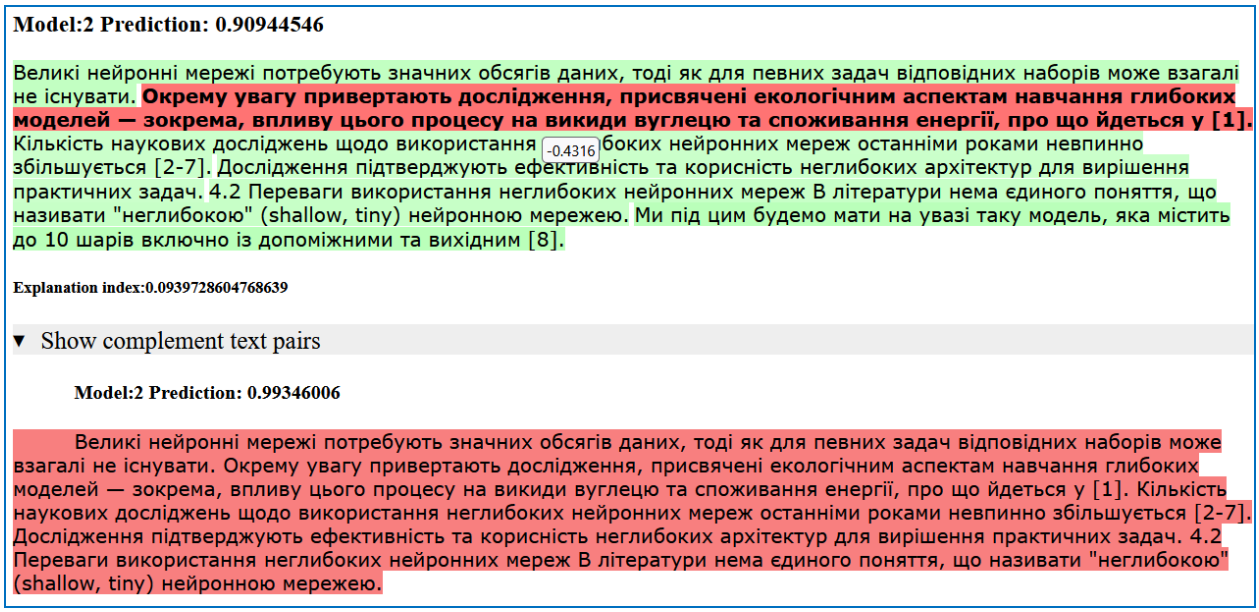


Fig. 3. Example of explanation with CTP

Let's evaluate the qualitative properties for sentence-based and n-grams sampling.

During all experiments, the 9 chunks were classified as AI-generated (the average output by 5 models was greater than 0.75) for the document with GPT 5 insertions, 7 chunks for the document with Gemini 2.5 Flash, and only 6 pieces for the document modified with Claude Sonnet 4.5. For each model, we searched for the explanations, making thus 5x9 = 45 attempts for GPT, 35 for Gemini, and 30 for Claude, respectively. Taking into account that the dataset all models were trained on was created using GPT 4o-mini, the explanations also seem to be more sensitive to GPT-generated fragments.

We were able to find CTP during the hiding of sentences only in 4 cases out of 45 (2 chunks of the text for 2 models) for a document with GPT pieces, 2 different paragraphs for the same model for Gemini (out of 35), and just 1 case for the test with Claude.

The experiments with searching for CTP when hiding n-grams for n = =1..5 were successful only in 1 same case for Model 5 (ID=4) and documents with Gemini-generated pieces. As one can see from Fig. 4, here is the case when just removing of one word (e.g., first one – "Цей") dramatically reduces the classification result by value 0.313 and changes the output class for this chunk from AI-generated class to human-written. These results confirm that finding CTP is more a lucky case for short sequences like words and sentences.

Let's look at the explainability measured according to (1) for different models and text chunks generated by GPT 5, Gemini 2.5 Flash, and Claude Sonnet 4.5 in our test document.

The Table 1 contains the average accuracy (AA) value and average explainability (AE) for the text chunks recognized as AI-generated (output >0.75) per model for each LLM. The correlation coefficient between average output and average explainability is -1 for text chunks generated with GPT, and Gemini, and -0.8 for Claude LLM, meaning the more confident model is on prediction the less it is on explainability (in average) measured according to (1).

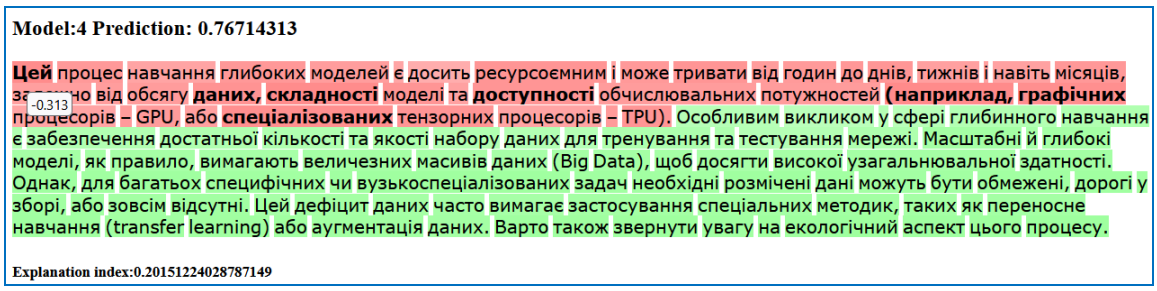


Fig. 4. Example with removing two words

Table 1 – **Average accuracy and explainability values for different models and LLMs**

	Text with GPT 5 AI		Text with Gemini 2.5 Flash		Text with Claude Sonnet 4.5	
Model	AA	AE	AA	AE	AA	AE
1	0.872	0.064	0.863	0.069	0.867	0.064
2	0.886	0.058	0.854	0.075	0.897	0.052
3	0.867	0.066	0.842	0.079	0.891	0.054
4	0.879	0.062	0.893	0.058	0.863	0.076
5	0.803	0.085	0.828	0.082	0.841	0.068

The same negative correlation (<=-0.8) also holds across all LLMs and models when explanations were generated by removing single words, 2-grams, 3-grams, 4-grams, and 5-grams. But taking into account that the explainability index includes the size of the perturbed chunk, and n-grams are shorter compared to sentences, the indices are higher for n-gram sampling. They vary between 0.075 and 0.13, and model 5 was the most explainable for the test document containing GPT insertions, model 3 leads the explainability rating for text with Gemini fragments, and finally, models 1 and 4 have close leading explainability indices for the test documents containing Claude-generated pieces. So, the overall explainability landscape is complicated and may vary from model to model and from sample to sample.

Conclusions

The results presented in the research are ambiguous, highlighting the complexity of AI-generated content classification and the decision-making problems it raises.

We considered searching for the explanations of the classification results for our own shallow ANN models, which are used as black boxes to classify the text chunks in Ukrainian in the IT domain. The explanation module utilizes the perturbation-based idea and modifies the initial text fragment by removing the separate sentences, separate words, and non-intersecting n-grams, $n = \overline{1,5}$.

The experimental modeling showed that finding complementary text pairs is a complex and rare case, and even rarer for n-grams. It is possible to find the contribution of the particular sentence (n-grams) into the overall classification result, but there were just a few cases in our experiments when this contribution was significant enough to change the output class. Probably, the exhaustive search of all combinations of sentences/n-grams could show different results, but this was not the topic of this paper.

We have proposed an explainability index that values the removal of short sentences (n-grams) over long ones and accounts for the significance of this perturbation on the overall classification result.

It would be interesting to investigate not removing sampling but replacing words and/or sentences with synonyms. However, smooth replacement requires a powerful Ukrainian language model, so it could be a possible topic for further work.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Acknowledgements

We dedicate this paper to the Armed Forces of Ukraine and everyone who stands with Ukraine.

The work was supported by the Ministry of Education and Science of Ukraine in the scope of project 0126U002246.

REFERENCES

1. P. Fantozzi and M. Naldi, “The Explainability of Transformers: Current Status and Directions,” *Computers*, vol. 13, no. 4, p. 92, 2024. doi: https://doi.org/10.3390/computers13040092
2. A. Ali, T. Schnake, O. Eberle, G. Montavon, K. R. Müller, and L. Wolf, “XAI for Transformers: Better Explanations through Conservative Propagation,” *Proc. Machine Learning Research (PMLR)*, vol. 162, 2022, pp. 436–451. [Online]. Available: https://proceedings.mlr.press/v162/ali22a/ali22a.pdf
3. A. Dugăeșescu and A. M. Florea, “Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study,” *Neural Computing and Applications*, vol. 37, no. 20, p. 14935-14970, 2025. doi: https://doi.org/10.1007/s00521-025-11282-7
4. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2921-2929, 2015. doi: .1109/ https://doi.org/10CVPR.2016.319

5. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618-626, doi: https://doi.org/10.1109/ICCV.2017.74
6. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *2017 International Conference on Machine Learning*, vol. 70, p. 3319 – 3328. doi: https://doi.org/10.5555/3305890.3306024
7. A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” *2017 International Conference on Machine Learning*, vol. 70, p. 3145–3153. doi: https://doi.org/10.48550/arXiv.1704.02685
8. M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, p. 1135–1144. doi: https://doi.org/10.1145/2939672.2939778
9. S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions”, *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774, doi: https://doi.org/10.48550/arXiv.1705.07874
10. D. Mardaoui and D. Garreau, “An Analysis of LIME for Text Data,” *International Conference on Artificial Intelligence and Statistics AISTATS 2021*, vol. 130, doi: https://doi.org/10.48550/arXiv.2010.12487
11. A. Aghababaei, J. Nikadon, M. Formanowicz, M. Bettinsoli, C. Cervone, C. Suitner and T. Erseghe, “Application of integrated gradients explainability to sociopsychological semantic markers,” Available at: https://arxiv.org/pdf/2503.04989
12. E. Mendez Guzman, V. Schlegel, and R. Batista-Navarro, “From outputs to insights: A survey of rationalization approaches for explainable text classification,” *Frontiers in Artificial Intelligence*, vol. 7, 2024. doi: https://doi.org/10.3389/frai.2024.1363531
13. M. Saarela and V. Podgorelec, “Recent Applications of Explainable AI (XAI): A Systematic Literature Review,” *Applied Sciences*, vol. 14, no. 19, p. 8884, 2024. doi: https://doi.org/10.3390/app14198884
14. B. Wei and Z. Zhu, “ProtoLens: Advancing Prototype Learning for Fine-Grained Interpretability in Text Classification,” *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 4503–4523. doi: https://doi.org/10.18653/v1/2025.acl-long.226
15. H. Yan, L. Gui, and Y. He, “Hierarchical Interpretation of Neural Text Classification,” *Computational Linguistics*, vol. 48, no. 4, p. 987–1020, 2022. doi: https://doi.org/10.1162/coli_a_00459
16. H. Moraliyage, G. Kulawardana, D. De Silva, Z. Issadeen, M. Manic and S. Katsura, “Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers,” *Applied System Innovation*, vol. 8, no. 1, p. 17, 2025. doi: https://doi.org/10.3390/asi8010017
17. O. Peredrii, “Shallow ANN models to classify Ukrainian AI-generated text,” *Control, Navigation and Communication Systems*, no. 4(82), 2025, pp. 108–113. doi: https://doi.org/10.26906/SUNZ.2025.4.108-113
18. O. Gorokhovatskyi, O. Peredrii, and O. Teslenko, “Multiple recursive division explanations for image classification problems,” *Advanced Information Systems*, vol. 9, no. 3, 2025, pp. 5–13. doi: https://doi.org/10.20998/2522-9052.2025.3.01
19. O. Gorokhovatskyi and O. Peredrii, “Recursive Division Explainability as a Factor of CNN Quality,” *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*, vol. 219, 2024, pp. 308–325. doi: https://doi.org/10.1007/978-3-031-70959-3_16

Received (Надійшла) 25.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Передрій Олена Олегівна – кандидат технічних наук, доцент кафедри інформатики та комп’ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Olena Peredrii – PhD, Associate Professor, Department of Informatics and Computer Technology, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: olena.peredrii@hneu.net; ORCID Author ID: https://orcid.org/0000-0003-0390-1931;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202751577.

Гороховатський Олексій Володимирович – кандидат технічних наук, доцент кафедри інформатики та комп’ютерної техніки, Харківський національний економічний університет імені Семена Кузнеця, Харків, Україна;
Oleksii Gorokhovatskyi – PhD, Associate Professor, Department of Informatics and Computer Technology, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine;
e-mail: oleksii.gorokhovatskyi@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-3477-2132;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=23099879900

Пояснюваність мілких моделей класифікації тексту, згенерованого ШІ, через видалення частин

О. О. Передрій, О. В. Гороховатський

Анотація. У цій роботі ми розглядаємо проблему пояснюваності результату класифікації штучними нейронними мережами (ШНМ) текстових фрагментів, згенерованих ШІ та написаних людиною, у текстах українською мовою в IT-домені. Метою є дослідити, чи можуть модифікації текстових фрагментів на основі пертурбацій, які включають видалення речень, слів та словосполучень, бути корисними для пошуку пояснень. Ми використали п'ять мілких моделей ШНМ (із середньою точністю близько 0.88) і протестували їх на вибірці документів, що містять як написані людиною тексти, так і згенеровані ШІ-фрагменти, створені за допомогою GPT-5, Gemini 2.5 Flash і Claude Sonnet 4.5. Експериментальне моделювання показало, що не завжди можна знайти окреме речення або слово, яке змінює результат класифікації. Ми запропонували індекс пояснювальності, який вимірює загальний вплив усіх змінених зразків на результат класифікації, враховуючи, що короткі пертурбації є більш цінними.

Ключові слова: пояснювальність, «чорний ящик», мілка ШНМ, пертурбація, згенерований ШІ контент, написаний людиною контент, текстовий фрагмент, класифікація тексту, індекс пояснюваності.

УДК 004.896:004.932.62:623.458.7 **doi: 10.26906/SUNZ.2026.2.160**

А. О. Подорожняк, О. В. Скорлупін

Національний технічний університет «Харківський політехнічний інститут», Харків, Україна

ВИЯВЛЕННЯ МІН ЗА ДОПОМОГОЮ РОБОТОТЕХНІЧНИХ СИСТЕМ ІЗ ВИКОРИСТАННЯМ МУЛЬТИСПЕКТРАЛЬНИХ ВІДЕОЗОБРАЖЕНЬ

Анотація. Актуальність. Виявлення мін за допомогою робототехнічних систем із використанням мультиспектральних відеозображень зумовлено критичною потребою в безпечних, ефективних та високоточних методах розмінування постконфліктних територій, де традиційні підходи не забезпечують достатньої швидкості й надійності виявлення вибухонебезпечних об'єктів. **Об'єкт дослідження**: процес дистанційного пошуку та виявлення протипіхотних та протитанкових мін, здійсненого за допомогою робототехнічних комплексів, які використовують комп'ютерний зір у видимому та тепловому діапазонах. **Мета статті:** розробка методології багатоспектрального аналізу простору, яка базується на синергії просторових, структурних (текстурних) та теплових характеристик об'єктів. **Результати дослідження**. У статті проаналізовано детально фізичну природу формування відмінностей у видимості (контрасту) між вибуховим пристроєм та його оточенням, принципові схеми конфігурації сенсорного обладнання, алгоритмічні етапи опрацювання відеопотоку в бортовій комп'ютерній системі у межах оперативних часових рамок, а також ступінь впливу зовнішніх умов на результативність ідентифікації. Встановлено, що спільне використання інформації з RGB-камер та тепловізорів (RGB-ІЧ злиття) забезпечує значне покращення частки коректно виявлених цілей у несприятливих умовах спостереження. **Висновки.** Представлені дані свідчать про вагомий потенціал застосування безпілотних наземних та літальних апаратів мультиспектрального моніторингу для проведення операцій з гуманітарного розмінування. Сфера використання отриманих результатів: мобільні робототехнічні системи мультиспектрального виявлення вибухонебезпечних об'єктів.

Ключові слова: виявлення мін; робототехнічна система; мультитспектральне зображення; комп'ютерний зір; комп'ютерна система; RGB-ІЧ злиття; гуманітарне розмінування.

Вступ. Постановка задачі

Замінування територій залишається однією з найбільш затяжних і важких за наслідками загроз, що триває роками та десятиліттями після припинення активних боїв. Ділянки землі, забруднені небезпечними вибуховими пристроями, накладають суттєві обмеження на відновлення транспортних, енергетичних і житлових комунікацій, унеможливлюють повноцінне використання сільськогосподарських площ і є постійним джерелом ризику для мирного населення [1]. За прогнозами міжнародних інституцій, повне очищення значних територій від мінних зарядів та боєприпасів що вибухнули може зайняти десятиліття, навіть за умови застосування передових технологій та значних фінансових вкладень [2].

Окрім прямих людських жертв, мінна загроза тягне за собою довготривалі соціально-економічні проблеми. Забруднення територій вибухонебезпечними предметами фактично виводить ці землі з господарського обігу, що спричиняє погіршення стану ґрунтів, зменшення обсягів виробництва їжі та змушення людей до переселення. Відновлення таких регіонів вимагає не просто механічного звільнення території від небезпеки, але й налагодження стійких систем моніторингу та контролю, здатних запобігти повторному виникненню загроз. Саме тому своєчасне виявлення небезпечних об'єктів набуває першорядного значення на етапі планування заходів з гуманітарного розмінування.

Класичні методи знешкодження мін, що ґрунтуються на ближньому пошуку саперними підрозділами з використанням ручного інструментарію або приладів для виявлення металу, залишаються вкрай ризикованими та вимагають багато часу. Необхідність присутності фахівця безпосередньо у зоні загрози суттєво уповільнює темпи роботи та збільшує ймовірність нещасних випадків. Навіть застосування механізованих установок чи захищеної техніки не спроможне повністю прибрати всі ризики. У світлі цього, дистанційні методи пошуку, які дозволяють проводити первинну розвідку місцевості без прямого залучення людини до небезпечного середовища, стають усе більш актуальними [3].

Одним із найбільш перспективних шляхів розв'язання цієї проблеми є впровадження різноманітних роботизованих комплексів, включно з безпілотними літальними апаратами (БПЛА) та безпілотними наземними апаратами (БПНА). Такі системи здатні обстежувати великі площі, працювати у важкодоступній місцевості та передавати дані оператору у режимі реального часу. Їхня результативність значною мірою залежить від характеристик сенсорів, якими вони оснащені. Оптична апаратура, що оперує у видимому та інфрачервоному спектрах, може бути одним з найбільш доступних та універсальних засобів для дистанційного зондування поверхні, оскільки не потребують контакту з ґрунтом і забезпечує високу деталізацію просторових даних [4].

Зображення у видимому діапазоні дає змогу оцінити габарити об'єктів, їхній колір, текстуру та взаємне розташування елементів сцени, тоді як інфрачервоне випромінювання фіксує теплові характеристики поверхні та дозволяє ідентифікувати приховані або замасковані аномалії. Проте використання лише одного спектрального каналу не гарантує високої надійності виявлення у реальних польових умовах, де суттєво впливають такі фактори, як освітленість, наявність рослинності, вологість ґрунту та інші зовнішні змінні. Це створює потребу у мультиспектральному підході, який інтегрує інформацію з різних джерел, забезпечуючи більш повне розуміння обстежуваної картини рис. 1.

Рис. 1. Структурна схема системи

Фізичні та технічні основи мультиспектрального виявлення мін

Визначення характеристик, які дозволяють розпізнати міну, зумовлене складним сплетінням фізичних явищ. Візуальне представлення видимого та інфрачервоного діапазонів у загальному електромагнітному випромінювання представлено на рис. 2.

У видимому діапазоні світла ключовим елементом даних є відбиття сонячної енергії. Протипіхотна міна може проявлятися через локальні зміни рельєфу ґрунту, або ж зміну його кольорових чи фактурних властивостей на поверхні. Однак, зброя, що використовується нині, нерідко створюється з компонентів, які майже не відрізняються від оточення, або ж піддається цілеспрямованому камуфлюванню [5].

Теплове випромінювання нижнього діапазону безпосередньо залежить від температури поверхні та її здатності випромінювати тепло. Присутність стороннього елемента порушує теплову рівновагу ділянки внаслідок розбіжностей у питомій теплоємності, здатності проводити тепло та тепловій стійкості матеріалів [6, 7].

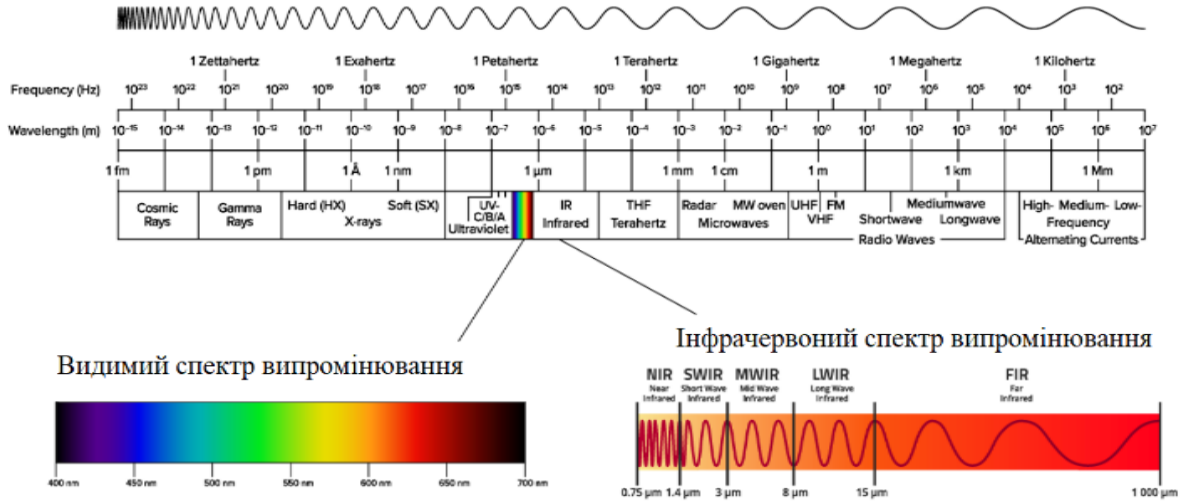


Рис. 2. Візуальне представлення видимого та інфрачервоного діапазонів

Навіть якщо сам об'єкт знаходиться під земною поверхнею, він здатний викликати появу температурного зміщення, яке стає помітним на верхньому шарі ґрунту.

Величина різниці температур обумовлена низкою обставин, серед яких глибина розташування, насиченість ґрунту водою, густота рослинного покриву та метеорологічні умови.

Найбільш помітними відхилення в температурі стають тоді, коли поверхневі шари ґрунту швидко змінюють свій тепловий режим – у час одразу після світанку або безпосередньо перед заходом сонця рис. 3.

Метод одночасного аналізу багатьох спектрів передбачає зведення докупи оцінки просторових і теплових ознак сцени.

Це дає змогу усунути обмеження, властиві кожному окремому інформаційному потоку [8, 9], і тим самим збільшити шанси на ідентифікацію об'єктів з різними властивостями рис. 4.

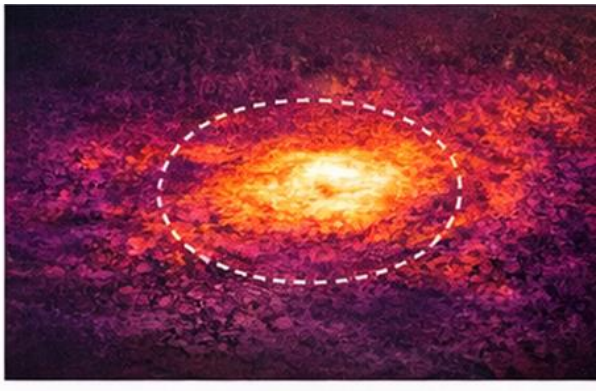


Рис. 3. Термографічне зображення міни

Архітектура робототехнічної системи та обробка відеоданих

Типова апаратура для дистанційного пошуку мін складається з блоку сенсорів, підсистеми

обробки даних, апаратури орієнтування та засобу комунікації із залученим фахівцем [10].

Фотоапарати, що працюють у різних частинах спектра, розміщуються таким чином, аби їхні поля огляду мали спільні зони та дозволяли одночасну фіксацію зображень [11].

Відеодані підлягають попередній обробці, яка може охоплювати стабілізацію візуального ряду, усунення викривлень форми, відсіювання завад та приведення рівня освітлення до стандартного. Після цього відбувається об'єднання інформації з усіх джерел та автом атизоване дослідження отриманих кадрів [12].

Методи виявлення можуть ґрунтуватися як на традиційних підходах комп’ютерного зору, так і на моделях на базі нейронних мереж, придатних для функціонування у реальному часі [13]. Виявлені предмети прив'язуються до просторових координат, і отримані відомості згодом можуть бути надіслані фахівцю або застосовані для формування плану потенційно небезпечних територій рис. 5.

Рис. 4. Робототехнічна платформа

Рис. 5. Схема об’єднання RGB та IR зображень

Невід'ємною частиною апаратури є механізм геопросторового прив'язування, що дає змогу приписати результати розпізнавання до географічних координат та включити їх у геоінформаційні системи [14].

Результати досліджень та вплив умов середовища

Ступінь успішності знаходження мін та інших вибухонебезпечних об’єктів, зокрема нерозірваних летальних боєприпасів, що не спрацювали після застосування і залишаються потенційно небезпечними (НЛБО), істотно зумовлений умовами спостереження [15]. Випробування у реальних умовах доводять, що використання лише інформації з RGB-каналу часто призводить до значної кількості пропусків, особливо у випадках маскування об’єктів або низького контрасту з фоном. Інфрачервоний діапазон дозволяє виявляти частину таких предметів за рахунок температурних відмінностей, однак його ефективність знижується за несприятливих погодних умов або незначного теплового контрасту [16].

Поєднання обох методів аналізу демонструє найбільш надійні результати, оскільки дає змогу враховувати різні типи інформативних ознак. Навіть якщо об’єкт не має виражених візуальних характеристик, його можна зафіксувати за наявністю локальної температурної аномалії, і навпаки рис. 6.

Рис. 6. Приклади детекції

На кінцеві показники також чинять вплив склад ґрунту, ступінь зволоження, наявність рослинності та конфігурація ландшафту. Скажімо, зволожений ґрунт володіє більшою теплоємністю, що здатно знижувати контрастність [17]. Густий рослинний шар створює перешкоди як для оптичного, так і для теплового моніторингу рис. 7.

Рис. 7. Залежність точності від зовнішніх умов

Під точністю виявлення у даному дослідженні розуміється ступінь відповідності результатів автоматичної ідентифікації фактичній наявності вибухонебезпечних об’єктів на обстежуваній території, що кількісно оцінюється за допомогою стандартних метрик бінарної класифікації [18]. Нехай TPTPTP – кількість істинно позитивних спрацювань (коректно виявлених об’єктів), TNTNTN – істинно негативних результатів, FPFPFP – хибно позитивних спрацювань, а FNFNFN – хибно негативних результатів (пропущених об’єктів). Тоді:

загальна точність – $Acc = \frac{TP+TN}{TP+TN+FP+FN}$;

прецизійність (Precision) – $Prec = \frac{TP}{TP+FP}$;

повнота або чутливість (Recall) – $Rec = \frac{TP}{TP+FN}$;

гармонічна оцінка якості – $F1 = 2 \cdot \frac{Prec \cdot Rec}{Prec+Rec}$.

У задачах дистанційного виявлення мін та нерозірваних боєприпасів пріоритетним показником є саме Recall, оскільки пропуск небезпечного об’єкта створює значно більший ризик, ніж помилкове спрацювання системи. Для комплексної оцінки здатності алгоритму відокремлювати ціль від фону додатково застосовується площа під ROC-кривою (ROC-AUC), що характеризує якість класифікації при варіюванні порогу прийняття рішення.

Висновки та перспективи подальших досліджень

Отримана у процесі роботи інформація демонструє, що роботизовані комплекси, здатні працювати у кількох спектрах, слугують дієвим засобом для дистанційного знаходження вибухонебезпечних предметів. Злиття даних візуального ряду, отриманих як у видимому, так і в інфрачервоному діапазонах, дає змогу формувати більш ґрунтовне уявлення про об'єкт розвідки, що підвищує імовірність виявлення у типових польових ситуаціях.

Отримані висновки підтверджують слушність застосування подібних апаратно-програмних рішень для початкового огляду місцевості перед тим, як розпочинати операції з гуманітарного розмінування. Наступні етапи наукового пошуку варто зосередити на відшліфуванні способів об'єднання інформації з різних датчиків, збільшенні незалежності функціонування рухомих платформ, а також адаптуванні програмних засобів до змінних умов їхнього застосування.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. International Mine Action Standards (IMAS). United Nations Mine Action Service, 2023, 23 p. URL: https://www.mineactionstandards.org/fileadmin/uploads/imas/Standards/English/TEP_09.10.01.2023_Ed.1.pdf.
2. Habib, M. K. (2007). “*Humanitarian Demining: Reality and the Challenge of Technology – The State of the Arts,”* International Journal of Advanced Robotic Systems, 2007, 4(2), doi: https://doi.org/10.5772/5699.
3. Gonzalez, R. C., Woods, R. E. (2018). “*Digital Image Processing*,” 4th ed., Pearson, 2018, 1022 p. URL: https://www.pearson.com/en-us/subject-catalog/p/digital-image-processing/P200000003224/9780137848560
4. Szeliski, R. (2022). *“Computer Vision: Algorithms and Applications,”* Springer, 2022, 925 p., doi: https://doi.org/10.1007/978-3-030-34372-9.
5. Murphy, R. R. (2014). “Disaster Robotics,” MIT Press, 2014, 224 p., doi: https://doi.org/10.7551/mitpress/9407.001.0001.
6. Deans, J., Gerhard, J., & Carter, L. J. (2006). *“Analysis of a thermal imaging method for landmine detection, using infrared heating of the sand surface*,” Infrared Physics & Technology, 48 (3), pp. 202–216. doi: https://doi.org/10.1016/j.infrared.2005.06.003.
7. Gallagher, J. E., Oughton, E. J. (2023). *“Assessing thermal imagery integration into object detection methods on air-based collection platforms*,” Scientific Reports, 2023, 13, 8491, doi: https://doi.org/10.1038/s41598-023-34791-8.

8. Wang, P., Wu, J., Fang, A., Zhu, Z., and Wang, C. (2024). *"Multi-spectral image fusion for moving object detection*," Infrared Physics & Technology, 2024, vol. 141, 105489, doi: https://doi.org/10.1016/j.infrared.2024.105489.
9. Gallagher, J. E., Oughton, E. J. and Kosecka, J. (2025). *"Multi-temporal Adaptive Red-Green-Blue and Long-Wave Infrared Fusion for You Only Look Once-Based Landmine Detection from Unmanned Aerial Systems,"* ArXiv, arXiv:2512.20487 [cs.CV], 2025, 21 p., doi: https://doi.org/10.48550/arXiv.2512.20487.
10. Shklyar, S., Andreiev, A., & Golubov, S. (2025). *"Accuracy assessment of landmine detection by infrared aerial imaging,"* Ukrainian Journal of Remote Sensing, 2025, 12(4), pp. 16–20. https://doi.org/10.36023/ujrs.2025.12.4.294.
11. Skorlupin, O., & Podorozhniak, A. (2024). "*Optical methods for detecting explosive objects using autonomous unmanned systems,"* Problems of informatization: Proceedings of 12-th International Scientific and Technical Conference, 2024, p. 132. URL: https://repository.kpi.kharkov.ua/handle/KhPI-Press/87663.
12. Skorlupin, O., & Podorozhniak, A. (2025). *"Mobile explosive detection system for civil defense needs,"* Problems of Informatics and Modeling (PIM-2025): Proceedings of 25-th International Scientific and Technical Conference, 2025, p. 89. URL: https://repository.kpi.kharkov.ua/handle/KhPI-Press/87663.
13. Podorozhniak, A., Liubchenko, N., Skorlupin, O., Korolenko, S., & Stas, A. (2025). *"Mobile explosive object detection system for humanitarian demining needs,"* 2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek), 06-10 October 2025, Kharkiv, Ukraine, 2025, pp. 1-6, doi: https://doi.org/10.1109/KhPIWeek61436.2025.11288620.
14. Popov, M., Stankevich, S., Mosov, S., Dugin, S., Golubov, S., Andreiev A., Lysenko A., & Saprykin, I. (2024). *"Concept of a geoinformation platform for landmines and other explosive objects detection and mapping with UAV,"* Radioelectronic and Computer Systems, 2024, vol. 106, no. 2, pp. 107–118, doi: https://doi.org/10.32620/reks.2024.4.17.
15. Levchenko, D., Podorozhniak, A., & Liubchenko N. (2025). *"Tools and methods for explosive objects detection using artificial intelligence and computer vision,"* Control, Navigation and Communication Systems, 2025, No. 3 (81), pp. 117–121, doi: https://doi.org/10.26906/SUNZ.2025.3.117.
16. Kim, J.-H., & Kwon, G.-R. (2025). *"Image-Level Anti-Personnel Landmine Detection Using Deep Learning in Long-Wave Infrared Images,"* Appl. Sci., 2025, 15 (15), 8613, doi: https://doi.org/10.3390/app15158613.
17. Malizia, M., Hasselmann, K., Miuccio, A., Haelterman, R., Tsiogkas, N. and & Demeester, E. (2025). *"PFM-1 Landmine Detection in Vegetation Using Thermal Imaging with Limited Training Data,"* 2025 25th International Conference on Control, Automation and Systems (ICCAS), Incheon, Republic of Korea, 2025, pp. 1864-1869, doi: https://doi.org/10.23919/ICCAS66577.2025.11301116.
18. Ameyaw, D. A., Deng, Q., & Söffker, D. (2019). *"Probability of Detection (POD)-based Metric for Evaluation of Classifiers Used in Driving Behavior Prediction,"* Annual Conference of the PHM Society, 2019, 11 (1), pp. 1-7, doi: https://doi.org/10.36001/phmconf.2019.v11i1.774.

Received (Надійшла) 28.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Подорожняк Андрій Олексійович – кандидат технічних наук, доцент, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Andrii Podorozhniak – Candidate of Technical Sciences, Associate Professor, Professor of Computer Engineering and Programming Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: andrii.podorozhniak@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-6688-8407;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202229410.

Скорлупін Олександр Володимирович – аспірант кафедри комп'ютерної інженерії та програмування, Національний технічний університет "Харківський політехнічний інститут", Харків, Україна;
Oleksandr Skorlupin– PhD student, Computer Engineering and Programming Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine;
e-mail: oleksandr.skorlupin@cit.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0001-4295-4115;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=60435005700.

Landmine detection using robotic systems based on multispectral video imaging

Andrii Podorozhniak, Oleksandr Skorlupin

Abstract. Relevance. Landmine detection using robotic systems with multispectral video imagery is driven by the critical need for safe, efficient, and highly accurate demining methods on post-conflict territories, where traditional approaches fail to provide sufficient speed and reliability in detecting explosive devices. **Object of research**: the process of remote search and detection of anti-personnel and anti-tank mines performed by robotic complexes utilizing computer vision in visible and thermal ranges. **Purpose of the article**. Development of a methodology for multispectral spatial analysis based on the synergy of spatial, structural (textural), and thermal characteristics of objects. **Research results**. The article details the physical nature of visibility differences (contrast) formation between explosive devices and their surroundings, principal configurations of sensor equipment, algorithmic stages of video stream processing in the onboard computer system within operational time frames, and the degree of external conditions' impact on identification effectiveness. It is established that the combined use of RGB cameras and thermal imagers (RGB-IR fusion) significantly improves the proportion of correctly detected targets under adverse observation conditions. **Conclusions**. The presented data demonstrate substantial potential for applying unmanned ground and aerial vehicles for multispectral monitoring in humanitarian demining operations. Scope of application of the obtained results: mobile robotic systems for multispectral detection of explosive objects.

Keywords: landmine detection; robotic system; multispectral imaging; computer vision; computer system; RGB-IR fusion; humanitarian demining.

УДК 004.934 **doi: 10.26906/SUNZ.2026.2.165**

Daniil Raptanov, Olesia Barkovska, Mykhailo Shylenko, Oleksandr Holovchenko, Diana Ivakhnenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

A STUDY OF THE ACCURACY OF BIMFORMING METHODS IN THE CONTEXT OF AN INCLUSIVE INTERNAL NAVIGATION SYSTEM

Abstract. Relevance. Voice control of elements in inclusive navigation systems is critical for ensuring the independence and safe mobility of people with visual impairments in public spaces, particularly in large retail spaces. However, existing speech-to-text (STT) systems face a significant drop in recognition accuracy due to the highly dynamic and non-stationary acoustic noise in supermarkets. **The object** of this study is audio stream preprocessing and spatial filtering (beamforming) in a voice control system under conditions of dynamic, non-stationary noise. The problem lies in the insufficient selectivity of standard audio signal processing algorithms under conditions of background noise in a store, which leads to a critical increase in the word error rate (WER) and makes the smart cart control system vulnerable. The objective of the article is to evaluate the impact of external factors (number, spatial topology of placement, and power level of acoustic noise sources) on the accuracy of spatial filtering (beamforming) methods for subsequent voice command recognition through computer simulation. **As a result** of the study, the acoustic environment and microphone array were simulated using the Pyroomacoustics library. A comparison was conducted between three methods: Delay-and-Sum (DAS), Max-UDR, and Max-SINR. The study showed that the Max-SINR algorithm provides the highest signal-to-noise ratio gain (Delta SNR from 7.9 to 9.1 dB) and is mathematically robust to changes in the distance to interference sources and their power. The DAS method proved to be the least effective (5.35–5.95 dB) and demonstrated sensitivity to changes in distance. It was established that the key factor in signal degradation is the configuration of noise sources, among which the cross topology is the most difficult to filter.

Keywords: inclusive navigation system, visual impairment, speech recognition, spatial filtering, beamforming, Delay-and-Sum, Max-UDR, Max-SINR, dynamic noise.

Introduction

Problem Statement. Existing speech-to-text (STT) systems suffer from low recognition accuracy due to the presence of noise, reverberation, and multi-channel audio sources [1]. The use of microphone arrays in combination with direction-of-arrival (DoA) and beamforming algorithms significantly improves the signal-to-noise ratio in noisy environments [2]. Modern solutions often combine classical DSP algorithms, such as Delay-and-Sum (DAS), MVDR, or GSC, with deep learning methods. However, neural network models require significant computational resources and are sensitive to the acoustics of the rooms in which they were trained [3].

Despite advances in spatial filtering algorithms, their application in inclusive systems for retail spaces remains understudied [4, 5]. In the highly dynamic and non-stationary acoustic noise conditions of a retail floor, standard cloud-based and local STT methods exhibit significant degradation in accuracy. Accordingly, the adaptation of beamforming methods at the hardware level for smart shopping carts designed to assist visually impaired people in supermarkets is critically necessary to improve the signal-to-noise ratio (SNR) prior to the semantic processing stage.

The problem lies in the insufficient selectivity of standard audio signal processing algorithms under conditions of dynamic non-stationary noise, which leads to a critical increase in word error rates and makes the shopping cart control system vulnerable to background store noise. The subject of the study is the preprocessing of the audio stream and spatial filtering in a voice control system under conditions of dynamic non-stationary noise. The subject of the study is the effectiveness of spatial filtering algorithms (DAS, UDR, SINR) for improving the recognition of user voice commands in an inclusive smart cart system.

The scientific novelty lies in the introduction of an analytical stability metric (robustness index), which allows for a quantitative assessment of the degradation in the performance of spatial filtering algorithms as the topology of nonlinear disturbances becomes more complex (e.g., an increase in the number of shoppers, carts, or other noise sources). The practical significance is determined by the optimal configuration of the spatial filtering module, which maximizes the ASNR parameter and stabilizes the entire speech-to-text conversion pipeline. This ensures the reliable operation of an inclusive voice interface for the shopping cart without requiring significant computational resources, thereby promoting the independence of visually impaired users while shopping.

Analysis of Recent Research and Publications. The field of voice command recognition is explored in this paper in the context of using inclusive navigation systems with voice interfaces for people with visual impairments, underscoring the relevance of the chosen topic. Given the current situation in Ukraine, three key user groups can be identified for whom standard voice assistants are inaccessible:

- military personnel and civilians who have lost their sight due to trauma face an urgent need for spatial and psychological adaptation to new conditions and require tools to restore social autonomy in public spaces;

- elderly people with age-related vision impairments (cataracts, glaucoma, etc.) are unable to read small print on products, price tags, and navigation signs in large retail spaces;

- people with congenital or acquired blindness and low vision are critically dependent on reliable assistive technologies for spatial orientation, avoiding obstacles, and independent self-care.

Voice control aids in the socialization and safe mobility of people with disabilities in public spaces. Therefore, the processing and analysis of voice commands to

control elements of an inclusive system must be implemented for these groups of people with disabilities to enhance their independence, autonomy, and sense of security, which is a pressing task.

Examples of control scenarios for different population groups are provided in the table below (Table 1). The system must adapt to the user's specific needs, ensuring the performance of vital functions even with impaired vision.

Table 1 – **Management Scenarios for People with Limited Mobility**

Command category	Query example	System action
Navigation query	" Show me the way to …"	Calculating the optimal route to the desired department or shelf, with voice guidance and obstacle warnings.
Subject query	" I need …"	Searching for products in the product database, announcing prices and expiration dates, and providing assistance right at the shelf.
Emergency assistance	"Help me get outside", "We need help from healthcare workers "	Calculating the shortest route to the exit, automatically calling an assistant or the sales floor manager via the internal communication system.

Current research has already demonstrated that the use of microphone arrays, combined with sound source localization and mixed signal separation algorithms, can significantly improve the signal-to-noise ratio in multichannel environments. Existing studies show that the use of beamforming and DoA (direction of arrival) methods [6, 7, 13] significantly improves noise suppression while using omnidirectional and relatively inexpensive sensors. The geometry of the microphone array plays an important role and is directly related to the hardware component of the module. Some studies indicate a significant improvement in audio quality when using two- and three-dimensional arrays, with the number of sensors being the primary factor affecting the method's performance.

These methods are particularly critical for embedded real-time speech recognition systems, given limited computational resources and the need for rapid system response. Methods such as beamforming and DoA analysis are actively combined with classical DSP algorithms (e.g., Wiener filters, spectral subtraction) and deep learning methods (GRU, CNN). This allows for effective reduction of background noise, amplification of speech, and ensures stable operation of STT modules even in noisy environments [8, 9, 14]. This study focuses on the audio stream preprocessing module and noise suppression in the presence of user speech disturbances in the system described above. Furthermore, given the variety of speech impairments, microphone array configurations, background noise, and other environmental characteristics, the question arises: which beamforming method and/or signal arrival direction estimation method is optimal under specific conditions. Since the beamforming method provides a significant improvement in the output audio in the presence of background noise, and the hardware component requires a microphone array (two or more microphones), while the software component includes complex mathematical algorithms such as FaS (Filter And Sum), MVDR (Minimum Variance Distortionless Response), GSC (Generalized Sidelobe Canceller), etc.–the best and most accurate method for conducting the experiment will be the use of actual hardware and algorithms optimized for it [10–12]. The Pyroomacoustics utility is proposed as a modeling environment for the room and microphone array, as it provides classic implementations of the most popular algorithms.

Table 2 – **An Overview of Classical Beamforming and DoA Methods**

Method name	Features	Areas of application
Delay–and–Sum (DAS) [15]	- low computational load; - low resolution.	simple hearing aids; educational projects; arrays with a large number of microphones.
Functional Beamforming [16]	higher resolution compared to DAS; better suppression of side lobes and specular sources; requires calibration/correction for accurate dB measurements; sensitive to the choice of power exponent.	anechoic chambers; wind tunnel tests; industrial noise diagnostics.
GCC–PHAT (Generalized Cross–Correlation + Phase Transform) [17]	high resistance to reverberation; low computational complexity; works primarily with a single pair of microphones (TDOA); it is difficult to localize multiple speakers simultaneously.	smart speakers; IoT devices; basic video conferencing systems.
SRP–PHAT (Steered Response Power) [18]	reliable in noisy/reverberant environments; does not require knowledge of the number of sources; high computational cost.	robotics; professional conference systems; acoustic monitoring of premises.
MVDR / Capon (Minimum Variance Distortionless Response) [19]	adaptive method; better noise and interference suppression than DAS; sensitive to microphone calibration errors.	noise cancellation systems; telecommunications; speech processing.
MUSIC (Multiple Signal Classification) [20]	ultra-high resolution; can distinguish sources that are close to one another; high computational complexity; performs poorly in reverberant environments.	laboratory acoustic studies; high-precision measurements; radars.
ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) [21]	less computationally intensive than MUSIC; requires a specific array geometry.	specialized instruments with fixed geometry; radars.

In addition to the classical methods listed in Table 2, it is worth highlighting deep machine learning, which is based on standard DoA techniques. Modern voice assistants, smartphones, and AR/VR headsets use such adaptive methods to improve recognition quality. A neural network can learn to account for complex room acoustics, nonlinear distortions, and noise; however, at its core, it is a "black box" that requires large datasets for training and is dependent on the specific acoustics under which the network was trained. The Pyroomacoustics library contains implementations of popular beamforming methods, some of which feature advanced noise suppression algorithms (Table 3).

Table 3 – **An overview of the methods implemented in the Pyroomacoustics library**

Method name (API)	Description	Applications	Requirements
Delay–and–Sum (DAS) rake_delay_and_sum_weights	A basic algorithm that compensates for signal propagation delays to the microphones and sums them up.	Basic SNR enhancement and spatial filtering.	Coordinates of the target source and the array microphones.
MVDR rake_mvdr_filters	Minimizes noise variance while maintaining a unit gain in the direction of the target.	Suppression of directional interference and uncorrelated noise without distorting the useful signal spectrum.	Coordinates of the target source; estimation of the spatial covariance matrix of noise.
Max–SINR rake_max_sinr_filters rake_max_sinr_weights	Maximizes the ratio of the useful signal including selected early reflections to the sum of noise and interference.	Optimizing reception in reverberant rooms in the presence of strong competing sound sources.	The coordinates of the useful and interfering sources, or the interference covariance matrix.
Max–UDR rake_max_udr_filters rake_max_udr_weights	It separates energy into useful (direct sound + early reflections) and harmful (late reverberation + noise).	Speech reverberation and noise suppression (counteracting the harmful blurring of the spectrum).	Source coordinates and room parameters for separating early and late reflections.
Perceptual rake_perceptual_filters	Calculates filters in the time domain, taking psychoacoustics into account. Relaxes the suppression constraints within a short window (30 ms).	Improved speech perception (the Haas effect, integration of early reflections by the auditory system).	Source coordinates and specified time window.
One–Forcing rake_one_forcing_filters rake_one_forcing_weights	Calculates the time-domain filters of a beamformer with a single response to multiple sources.	Signal extraction with strict response constraints for multiple specified directions/reflections.	The exact coordinates of the target sources and their images.
Distortionless rake_distortionless_filters	Ensures undistorted transmission of the target signal in a multipath environment by imposing strict constraints on the phase and amplitude of the target paths.	Signal extraction in reverberant environments where even the slightest phase or amplitude distortion is unacceptable.	Knowledge of the room's impulse response (RIR) or the exact coordinates of the image sources.

Three methods were selected for the experiment: Delay and Sum, Max–SINR, and Max–UDR. The latter two are suitable for reverberant environments and have straightforward requirements. The Delay and Sum method was chosen as a simple and computationally efficient beamforming method.

The research is conducted in a computer simulation environment to model the room and the microphone array. This brings the experiment as close as possible to real-world conditions: reverberation, diffraction, stationary noise, and room modes.

It is worth noting that computer simulation is not the only valid way to test the research hypothesis, namely, to find a correlation between the beamforming method and the clarity of the output audio (removed noise) in the presence of significant speech disturbances from the user. The next step in working on the chosen topic will be an empirical study using real hardware with real microphone arrays for subsequent statistical processing; however, this requires preliminary empirical validation under controlled conditions.

Since the accuracy of the second stage depends on the clarity of the text obtained in the first stage, it becomes necessary to implement spatial filtering (beamforming) methods even before the STT stage. In environments with high acoustic noise, such as a retail store, standard STT methods suffer from a decline in accuracy. Therefore, a spatial filtering (beamforming) stage is required at the hardware level to improve the signal-to-noise ratio (SNR).

To this end, this paper proposes to investigate the change in the difference between the input and output signal-to-noise ratios (SNR) for various beamforming methods depending on:

- the topology of noise source placement;
- the distance of noise sources from the microphone array;
- noise power.

Thus, **the aim of the study is** to evaluate the influence of external factors (number, placement topology, and power level of acoustic noise sources) on the accuracy of beamforming methods for subsequent voice command recognition, using computer modeling.

To achieve the stated goal, the following tasks must be addressed:

- analysis of beamforming methods for use under conditions of variable external factors;
- development of a model for a voice command processing and analysis subsystem in an inclusive indoor navigation system;
- evaluation of the accuracy of the Delay–and–Sum (DAS), Max–SINR, and Max–UDR methods for

determining the speaker's location based on the number, spatial distribution, and power levels of acoustic noise sources;

- analysis of the obtained results.

Future research will focus on conducting field experiments using real hardware under **variable** external conditions.

Main Content

The proposed voice command processing subsystem, shown in Figure 1, is based on an interconnected sequence of steps: first, it isolates and cleans the speaker's voice of noise, converts it into text, and then, using a pre-trained language model, analyzes the semantics, verifies the request based on context, and generates a final command for execution or a voice response, thereby renewing the interaction cycle.

In the first stage of the system's operation (speech-to-text conversion), an unstructured analog or digital audio stream is converted into a formalized text sequence. This process is based on Voice Activity Detection mechanisms, which allow the useful signal to be separated from the background noise of the sales floor. The use of a specialized hardware interface ensures stable data transmission to the Speech-to-Text module, where phoneme recognition and the formation of a textual representation of the query occur using acoustic and linguistic models. The quality of the output data at this level is critical for the entire system, since recognition errors directly determine the accuracy of the subsequent interpretation of commands.

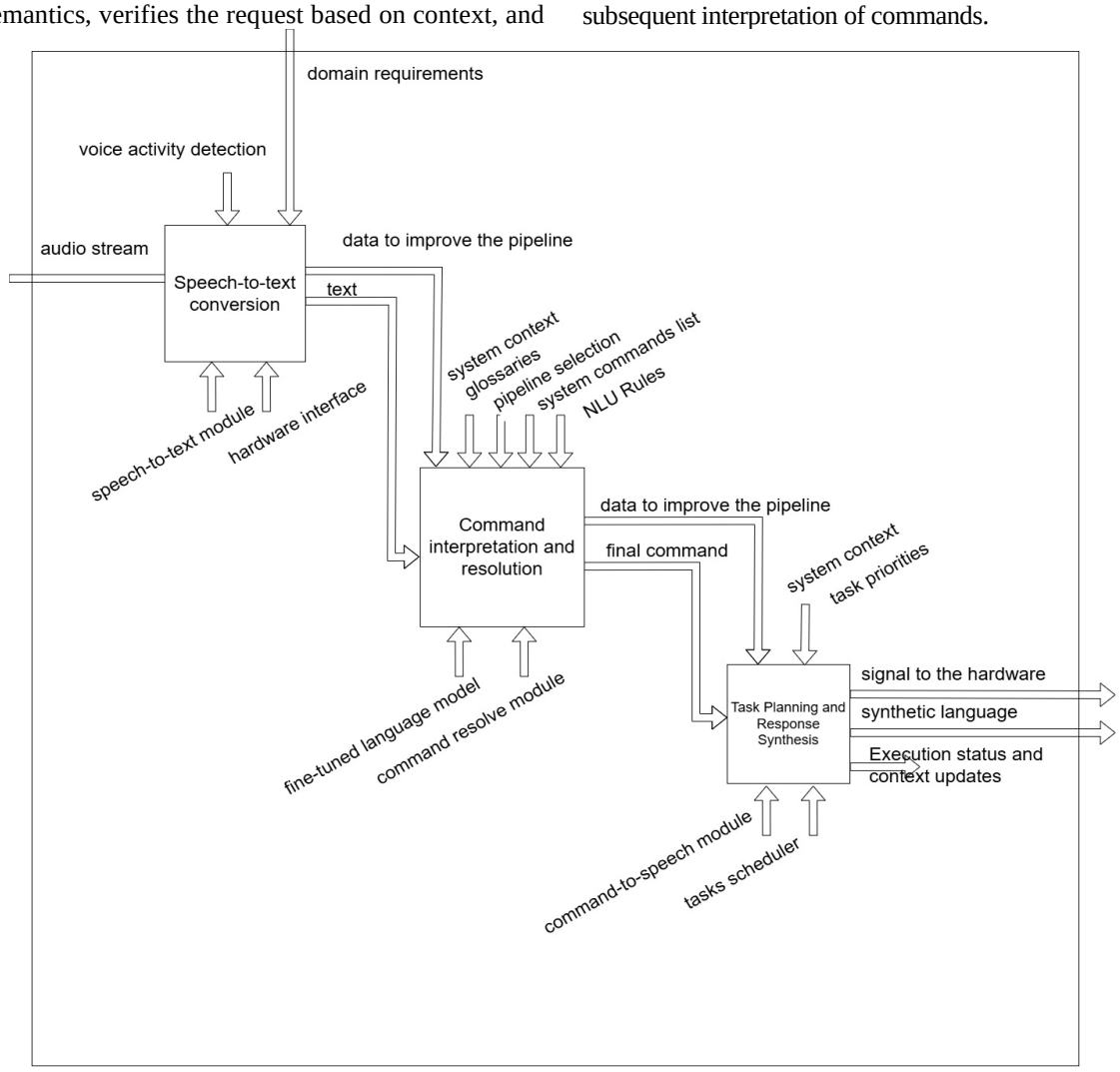


Fig. 1. General model of a voice command processing system

The central component of the system (command interpretation and resolution) is responsible for transforming the received text into structured, machine-readable commands through semantic resolution mechanisms. The interpretation process is based on the use of fine-tuned language models (Fine–tuned LLMs) and natural language understanding rules (NLU Rules), which allow for the identification of user intent within a specific domain. To minimize ambiguity, the system leverages external knowledge in the form of system context and specialized glossaries of terms related to the product assortment and navigation within the marketplace. The result of this stage is the generation of a final command that accurately reflects the user's true intent, taking into account the current situation and the selected processing pipeline.

The final stage of the architecture implements the logic for controlling the intelligent cart's actions and generating feedback. The task scheduling module ranks the received commands according to set priorities and coordinates their execution through context updates. In parallel with sending signals to the hardware, the system generates a voice response using the Command-to-

Speech module, ensuring natural interaction with the user. This approach allows for a closed-loop control cycle, where each action not only fulfills a request but also updates the system state for the correct processing of subsequent dialogue iterations.

The experimental studies were conducted taking into account the following potential influences, limitations, and requirements:

- determining the effect of the distance to noise sources (1 m, 3 m, 5 m) and their intensity relative to the user's voice (quieter than the voice: +10 dB; equal to the voice: 0 dB; louder than the voice: –10 dB) on the input signal;

- calculation of the SNR for beamforming methods under various topologies (a “cross” topology for 4 noise sources, a “ring” topology to simulate diffuse noise from 8 noise sources, and a point-source noise configuration, Fig. 2);

- calculating a coefficient to compare the ability of methods to suppress multiple noise sources without significantly degrading the output signal.

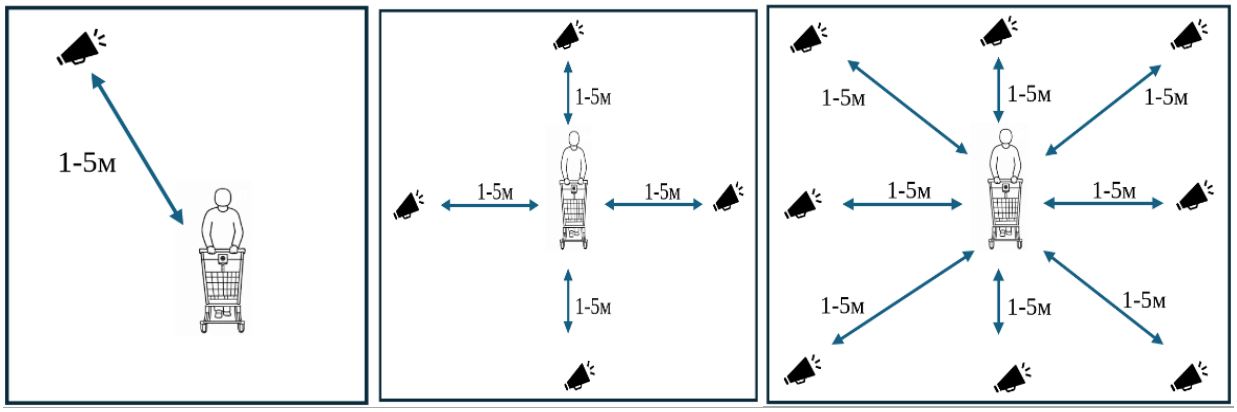


Fig. 2. Modeling of acoustic noise source configurations relative to the useful signal source: a – point source, b – “cross” configuration, c – “ring” configuration

The ΔSNR metric is calculated based on the measurements at the microphone input and output, SNR_{in} та SNR_{out} respectively. The input signal-to-noise ratio is calculated using the formula:

$$SNR_{in} = 10\,log_{10}(\frac{P_{signal}}{P_{noise}}),$$

where P_{signal} - clear voice power, P_{noise} - noise power.

Measurements at the SNR_{out} are determined in the same way, but after applying the beamforming method.

The performance metrics for evaluating the beamforming method under conditions of dynamic non-stationary interference are determined by the formula:

$$SNR_{improvement} = \Delta SNR = SNR_{out} - SNR_{in}.$$

The resulting ΔSNR value (expressed in decibels) serves as a quantitative measure of the effectiveness of spatial signal selection in the presence of additive noise. An increase in the ΔSNR value directly correlates with an increase in the probability of correct operation of the STT module. In particular, high values of this indicator indicate successful compensation for acoustic interference, which allows minimizing the level of phonetic distortions at the hardware interaction level. Conversely, low values of ΔSNR indicate insufficient selectivity of the algorithm, which leads in the future to degradation of the input stream and a critical increase in word error rate, making the system vulnerable to non-stationary noise in the trading floor. In other words, maximizing ΔSNR will lead to the stabilization of the entire pipeline.

The system’s robustness to changes in the noise topology is determined by an analytical metric introduced in this paper to compare beamforming methods. The metric shows how much the algorithm’s performance “drops off” when conditions in the store become more complex (the number of shoppers, carts, or noise increases). It is calculated as the ratio of the average signal gain under complex conditions to the gain under baseline (ideal) conditions.

Results of the research conducted. Discussion

To systematize the research results, taking into account all the variable parameters described in the previous section, this study includes the following experiment: evaluating performance as a function of distance and noise level, as well as assessing robustness to the topology and number of noise sources.

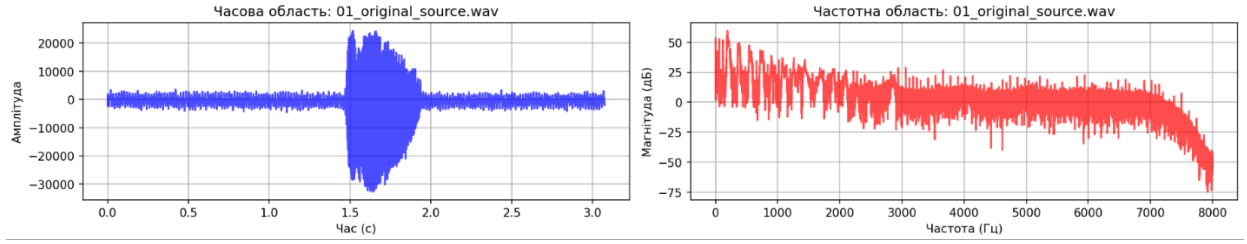
For the experiment, healthy speech recordings were selected from the specialized TORGO dataset, created for the study of dysarthria [21]. Typically, such recordings are clearer and cleaner than those in standard general-purpose datasets. This approach allows us to focus on working with noise and testing hypotheses, and simplifies the overall data preparation process.

For the experiment, 1,000 healthy speech sequences were selected, along with 3 types of noise geometry (point, cross, ring) at 3 different distances (1, 3, and 5 meters) and with varying signal-to-noise ratios (–10, 0, and 10 dB). Each recording was simulated using all three beamforming methods (DaS, UDR, and SINR). The output consisted of 81,000 records in a .csv file with the following columns: method, geometry, distance, SNR, input SNR, output SNR, and delta SNR.

Next, graphs were generated to demonstrate the dependencies of the variables on external factors. Figures 3 - 5 shows the dependence of delta SNR on distance for the DaS, UDR, and SINR methods, respectively.

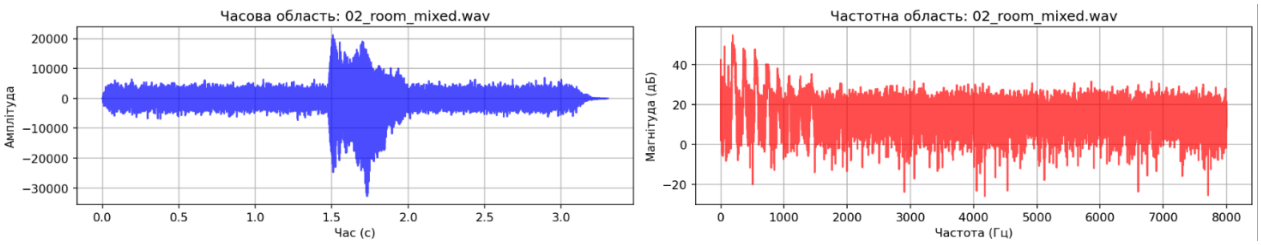
Fig. 3 shows an example of an audio signal from the dataset in the time and frequency domains.

Next, using the Pyroomacoustics library, a simulation of a room with dimensions of 15x15x5 was created, corresponding to the average size of open spaces in supermarkets or shopping malls. After adding noise and the desired signal, the overall audio signal changes (Fig. 4). The speech signal and noise signal are shown separately in Fig. 5 and 6, respectively. After applying the beamforming algorithm, noise is removed from the mixed signal and audio quality is improved (Fig. 7). The microphone array and algorithms ensure a stable spatial gain (array gain). This gain is a characteristic of the system's geometry and the mathematics of the algorithm, and it does not degrade when the total noise power changes. This manifests as the Delta SNR being independent of the initial signal-to-noise ratio. The graphs for conditions of –10 dB, 0 dB, and 10 dB are identical for each individual method.



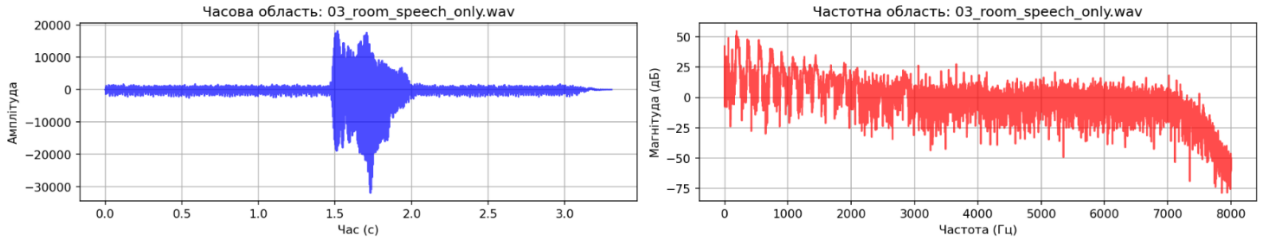
a b

Fig. 3. Visualization of the input audio signal: a – time domain, b – frequency domain



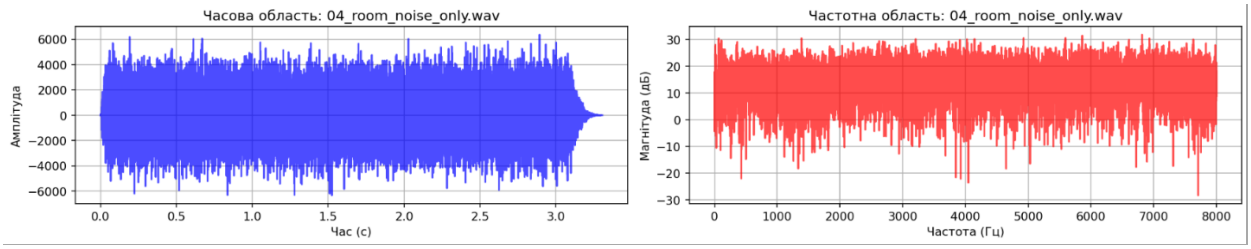
a b

Fig. 4. Visualization of a mixed audio signal in a simulation: a – time domain, b – frequency domain



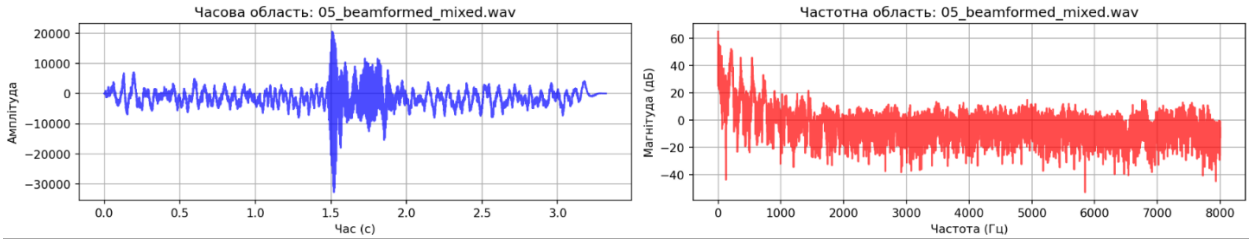
a b

Fig. 5. Visualization of a broadcast audio signal in a simulation: a) time domain, b) frequency domain



a b

Fig. 6. Visualization of a noise audio signal in a simulation: a – time domain, b – frequency domain



a b

Fig. 7. Visualization of the audio signal after applying the beamforming algorithm: a – time domain, b – frequency domain

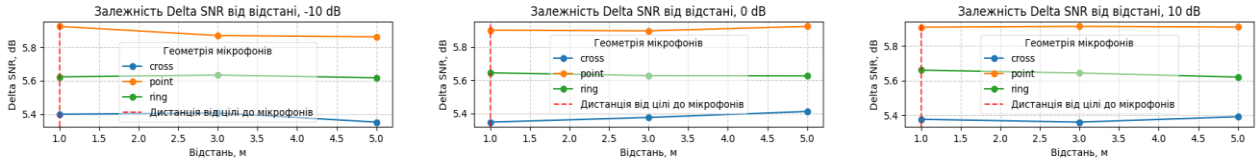
The results of experiments conducted for various beamforming methods under varying conditions in simulation mode are presented below.

All three algorithms demonstrate the same hierarchy of performance depending on how the noise sources are distributed. Point achieves the best result (highest delta SNR) among all methods. It is easiest for beamformers to focus the spatial "zero" in the beam pattern to suppress a single localized source. The ring configuration has average performance: noise surrounds the array, complicating the task compared to a single point, but the algorithms are still capable of filtering it reasonably well. The most challenging scenario for all methods is the cross configuration, which yields the smallest gain in the useful signal. This distributed configuration creates the most complex interference field for processing.

The algorithms differ significantly in terms of overall performance and response to the distance of noise sources. DaS shows the smallest improvement (delta SNR ranging from ~5.35 to ~5.9 dB, Fig. 8). It is the only method that demonstrates a dependence on the distance to the noise source. As point noise is moved from 1 to 5 meters away, performance slowly decreases, whereas for cross noise, it tends to increase slightly.

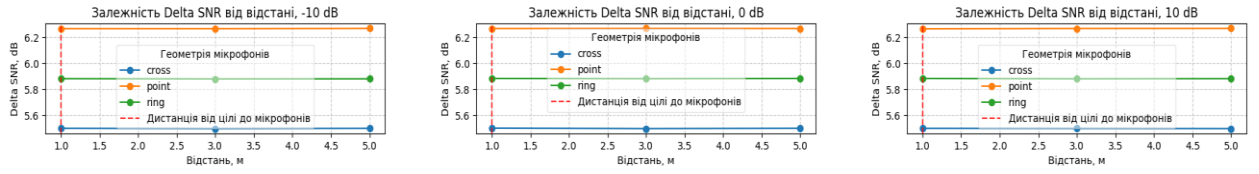
UDR shows average results (ranging from ~5.5 to ~6.25 dB), which are noticeably better than those of the classic DaS (Fig. 9). The graphs consist of horizontal lines. This means that the suppression performance is completely independent of the distance to the noise sources and depends solely on their spatial type (point, ring, cross).

SINR demonstrates the highest performance (delta SNR ranging from ~7.9 to ~9.1 dB, Fig. 10). The algorithm directly maximizes the signal-to-interference-plus-noise ratio and, as expected, performs best at optimizing the weights for interference suppression. As in the previous case, the result is maximally stable and does not depend on the distance to the noise source in the studied range.



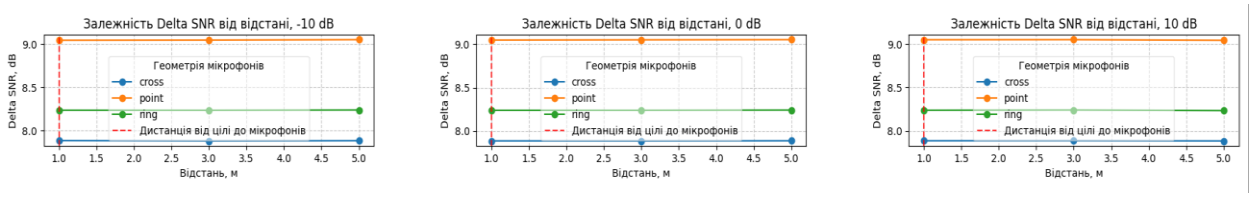
a b c

Fig. 8. Delta SNR as a function of distance, DaS method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB



a b c

Fig. 9. Delta SNR as a function of distance, UDR method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB



a b c

Fig. 10. Dependence of delta SNR on distance, SINR method: a – louder, –10 dB; b – equal, 0 dB; c – quieter, 10 dB

Next, the dependence of delta SNR on the noise level (SNR dB) was analyzed at distances of 1, 3, and 5 meters (Fig. 11–13). The distance from the target to the noise is 0, 2, and 4 meters, respectively.

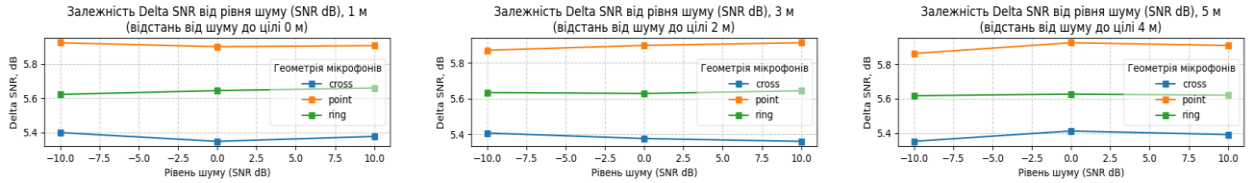
For all methods, the delta SNR value remains stable as the noise level changes. This confirms the linear nature of signal processing in these algorithms: the output signal-to-noise ratio changes in direct proportion to the input, so their difference (delta SNR) remains constant.

At any distance and at any noise level, the pattern of suppression efficiency remains consistent depending on the interference geometry: a point source is filtered best, a ring source is filtered worse, and a cross-shaped source is filtered worst. DaS demonstrates the smallest spatial gain (5.35–5.95 dB, Fig. 11). In the first graph (noise distance of 1 m, coinciding with the coordinates of the useful signal), slight fluctuations in delta SNR are noticeable. This is because DaS is a data-independent method with fixed weighting coefficients; it does not adapt to the environment, so nonlinear interference effects may occur when signal and noise sources are spatially coincident.

UDR demonstrates higher efficiency (5.5–6.25 dB, Fig. 12). The graphs appear as horizontal lines at all distances. Adaptive weight estimation relies on a normalized spatial covariance matrix, which depends on the location of the sources rather than their absolute power. Therefore, the algorithm is robust to changes in noise intensity.

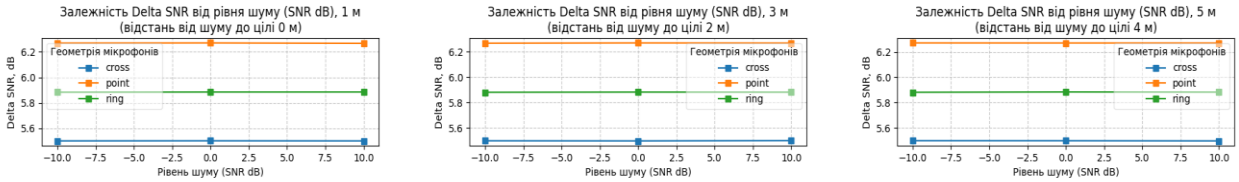
The SINR method demonstrates the highest efficiency (7.9–9.05 dB, Fig. 13). The algorithm also exhibits perfect linearity and stability regardless of the input

noise level and the distance to the noise source. Analytical maximization of the SINR criterion allows achieving the theoretical efficiency limit for a given microphone configuration.



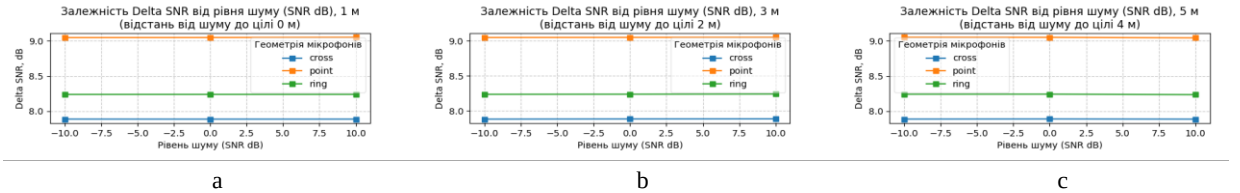
a b c

Fig. 11. Dependence of delta SNR on noise level, DaS method: a – 1 m, b – 3 m, c – 5 m



a b c

Fig. 12. Dependence of delta SNR on noise level, UDR method: a – 1 m, b – 3 m, c – 5 m

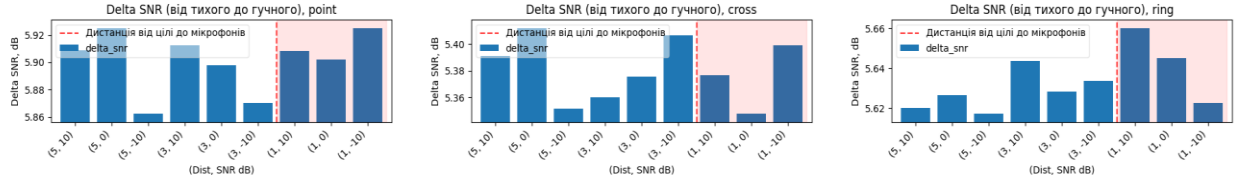


a b c

Fig. 13. Dependence of delta SNR on noise level, SINR method: a – 1 m, b – 3 m, c – 5 m

To confirm the previously established stability of adaptive algorithms for the given speech recognition task, the following histograms were created (Fig. 14–16).

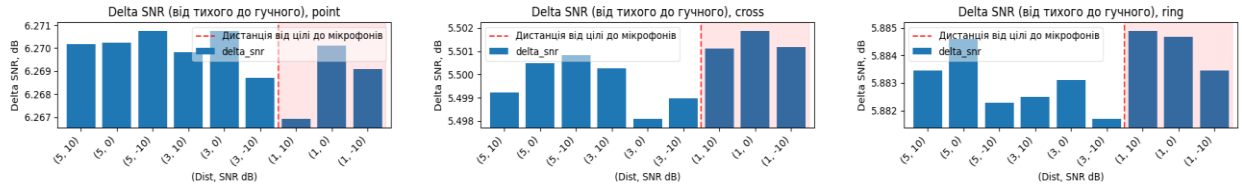
For UDR, Delta SNR fluctuations occur in the range of thousandths of a decibel (for example, from 6.267 to 6.271 dB for the point geometry, Fig. 15). For SINR, fluctuations are also within the range of thousandths or hundredths of a decibel (for example, from 9.0425 to 9.0525 dB for the point geometry, Fig. 16). Such changes are negligibly small. Visual fluctuations in the histograms are a consequence of the scaling of the Y-axis and reflect not a fundamental instability of the methods, but the limits of floating-point calculation accuracy and minor artifacts of the digital simulation of covariance matrices.



a b c

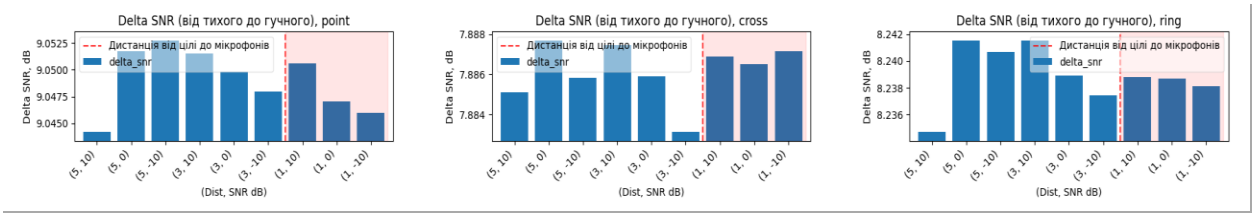
Fig. 14. The dependence of delta SNR on noise level and distance, DaS method: a – point interference topology, b – cross, c – ring

DaS is the only one to exhibit more noticeable fluctuations, reaching hundredths and tenths of a decibel (for example, from 5.86 to 5.92 dB, Fig. 14). The absence of adaptive weight calculation makes the algorithm sensitive to phase shifts that occur when the distance to noise sources and their intensity change. The area highlighted in pink (a distance of 1 m, coinciding with the location of the useful signal) exhibits specific behavior. When the interference is at the same radius as the target, spatial interference complicates the operation of the fixed directional pattern.



a b c

Fig. 15. The dependence of delta SNR on noise level and distance, UDR method: a – point interference topology, b –cross, c – ring



a b c

Fig. 16. The dependence of delta SNR on noise level and distance, SINR method: a – point interference topology, b – cross, c – ring

For the UDR and SINR methods, the baseline efficiency remains constant regardless of the combination of distance (1, 3, 5 meters) and input SNR (–10, 0, 10 dB). Even when noise is located in close proximity to the useful signal (at a distance of 1 m), adaptive algorithms retain their effectiveness. Minor deviations in this region (within thousandths of a dB) confirm the ability of spatial "zeros" to isolate interference regardless of their proximity to the useful signal's focal radius.

The results of the experiments for a single noise source are presented in Table 4. The results of the experiments for the four noise sources are presented in Table 5. The results of the experiments for eight noise sources are presented in Table 6.

Table 4 – **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power in the presence of a point noise source**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	44.67	5.91	6.27	9.05
1 m	0 (Equal)	34.67	5.90	6.27	9.05
1 m	–10 (Loudly)	24.67	5.93	6.27	9.05
3 m	+10 (Quiet)	44.67	5.91	6.27	9.05
3 m	0 (Equal)	34.67	5.90	6.27	9.05
3 m	–10 (Loudly)	24.67	5.87	6.27	9.05
5 m	+10 (Quiet)	44.67	5.91	6.27	9.04
5 m	0 (Equal)	34.67	5.92	6.27	9.05
5 m	–10 (Loudly)	24.67	5.86	6.27	9.05
The average change in SNR in a chaotic environment		34.67	5.90	6.27	9.05

Table 5 – **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power (cross topology)**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	45.14	5.38	5.50	7.89
1 m	0 (Equal)	35.14	5.35	5.50	7.89
1 m	–10 (Loudly)	25.14	5.40	5.50	7.89
3 m	+10 (Quiet)	45.14	5.36	5.50	7.89
3 m	0 (Equal)	35.14	5.38	5.50	7.89
3 m	–10 (Loudly)	25.14	5.41	5.50	7.88
5 m	+10 (Quiet)	45.14	5.39	5.50	7.89
5 m	0 (Equal)	35.14	5.41	5.50	7.89
5 m	–10 (Loudly)	25.14	5.35	5.50	7.89
The average change in SNR in a chaotic environment		35.14	5.38	5.50	7.89

Table 6 - **Evaluation of the effectiveness of beamforming methods as a function of distance and noise power (ring topology)**

Distance to the obstacle, m	Noise level, dB	SNR_{in} (Baseline)	Δ*SNR* (DAS)	Δ*SNR* (UDR)	Δ*SNR* (SINR)
1 m	+10 (Quiet)	45.44	5.66	5.88	8.24
1 m	0 (Equal)	35.44	5.64	5.88	8.24
1 m	–10 (Loudly)	25.44	5.62	5.88	8.24
3 m	+10 (Quiet)	45.44	5.64	5.88	8.24
3 m	0 (Equal)	35.44	5.63	5.88	8.24
3 m	–10 (Loudly)	25.44	5.63	5.88	8.24
5 m	+10 (Quiet)	45.44	5.62	5.88	8.23
5 m	0 (Equal)	35.44	5.63	5.88	8.24
5 m	–10 (Loudly)	25.44	5.62	5.88	8.24
The average change in SNR in a chaotic environment		35.44	5.63	5.88	8.24

Conclusions

This paper presents a system for processing and analyzing voice commands for inclusive smart carts, designed to operate in the dynamic, non-stationary noise environment of retail spaces.

An approach is proposed that uses microphone arrays and spatial filtering (beamforming) methods for continuous preprocessing of the audio stream to improve the signal-to-noise ratio (SNR) prior to the speech-to-text conversion stage.

Using spatial filtering algorithms, specifically Delay-and-Sum (DAS), Max-UDR, and Max-SINR, the system achieves high accuracy in extracting the useful signal in the presence of acoustic disturbances with various spatial topologies.

A comparative analysis of algorithm performance was conducted, the results of which showed that the Max-SINR method demonstrated the highest noise suppression efficiency (an SNR gain of 7.9 to 9.1 dB) compared to Max-UDR (5.5–6.25 dB) and DAS (5.35–5.95 dB).

The results confirm the effectiveness of the proposed approach for spatial signal selection, which is critically important in ensuring the reliability of voice control. The developed subsystem automatically compensates for the effects of background noise at the hardware level, significantly minimizing phonetic distortions and the vulnerability of the recognition system.

The use of adaptive beamforming algorithms, in particular Max-SINR, ensures mathematical stability of operation regardless of the strength of interference and the distance to it, making the solution promising for implementation in assistance and spatial orientation systems for people with visual impairments. Further research will focus on conducting field experiments using real hardware in variable outdoor conditions.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. O. Barkovska, A. Havrashenko and P. Botnar, "The influence of reverberation, equalization and compression methods on speaker recognition," *2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek)*, Kharkiv, Ukraine, 2025, pp. 1-5, doi: https://doi.org/10.1109/KhPIWeek61436.2025.11288718
2. Kulkarni, S., Thakur, A., Soni, S., Hiwale, A., Belsare, M. H., & Raj, A. B. (2025). A comprehensive review of direction of arrival (DoA) estimation techniques and algorithms. *Journal of Electronics and Electrical Engineering*, 138-186. https://doi.org/10.37256/jeee.4120255708
3. H. A. Kassir, Z. D. Zaharis, P. I. Lazaridis, N. V. Kantartzis, T. V. Yioultsis and T. D. Xenos, "A Review of the State of the Art and Future Challenges of Deep Learning-Based Beamforming," in *IEEE Access*, vol. 10, pp. 80869-80882, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3195299
4. Barkovska Olesia, Vitalii Serdechnyi. Intelligent Assistance System for People with Visual Impairments. *Innovative technologies and scientific solutions for industries*, no. 2(28), June 2024, pp. 6–16. https://doi.org/10.30837/2522-9818.2024.28.006
5. Barkovska, O., Holovchenko, O., Storchai, D., Kostin, A., & Lehezin, N. (2025). Investigation of computer vision techniques for indoor navigation systems. Innovative technologies and scientific solutions for industries, (2(32), 5–15. https://doi.org/10.30837/2522-9818.2025.2.005
6. Xi, J., Xu, Z., Zhang, W., Xie, Y., & Zhao, L. (2025). Speech Enhancement Algorithm Based on Microphone Array and Multi-Channel Parallel GRU-CNN Network. Electronics, 14(4), 681. https://doi.org/10.3390/electronics14040681
7. Wang, J.-H., Le, P. T., Bee, W.-S., Putri, W. R., Su, M.-H., Li, K.-C., Chen, S.-L., He, J.-L., Pham, T., Li, Y.-H., & Wang, J.-C. (2024). Implementation of Sound Direction Detection and Mixed Source Separation in Embedded Systems. Sensors, 24(13), 4351. https://doi.org/10.3390/s24134351
8. Wang, J.-H., Le, P. T., Kuo, S.-J., Tai, T.-C., Li, K.-C., Chen, S.-L., Wang, Z.-Y., Pham, T., Li, Y.-H., & Wang, J.-C. (2024). Audio Pre-Processing and Beamforming Implementation on Embedded Systems. Electronics, 13(14), 2784. https://doi.org/10.3390/electronics13142784
9. Huang, P., Ullah, I., Wei, X., Ahamed, A. T., Hassan, N., & Shah, Z. H. (2025). Towards Energy-Efficient and Low-Latency Voice-Controlled Smart Homes: A Proposal for Offline Speech Recognition and IoT Integration. ArXiv.org. https://arxiv.org/abs/2506.07494
10. Ciccarelli, G., Barber, J., Nair, A., Cohen, I., & Zhang, T. (2022). Challenges and Opportunities in Multi-device Speech Processing. ArXiv.org. https://arxiv.org/abs/2206.15432
11. Haeb-Umbach, R., Heymann, J., Drude, L., Watanabe, S., Delcroix, M., & Nakatani, T. (2020). Far-Field Automatic Speech Recognition. ArXiv.org. https://arxiv.org/abs/2009.09395
12. Rascon, C. (2021). A Corpus-Based Evaluation of Beamforming Techniques and Phase-Based Frequency Masking. Sensors, 21(15), 5005. https://doi.org/10.3390/s21155005
13. Rowe, H. P., Gutz, S. E., Maffei, M. F., Tomanek, K., & Green, J. R. (2022). Characterizing Dysarthria Diversity for Automatic Speech Recognition: A Tutorial From the Clinical Perspective. Frontiers in Computer Science, 4. https://doi.org/10.3389/fcomp.2022.770210
14. Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3025453.3025786
15. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," in IEEE ASSP Magazine, vol. 5, no. 2, pp. 4-24, April 1988, doi: https://doi.org/10.1109/53.665

16. C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320-327, August 1976, doi: https://doi.org/10.1109/TASSP.1976.1162830
17. Rakerd, B., Hartmann, W.M. (2005). Localization of noise in a reverberant environment. In: Pressnitzer, D., de Cheveigné, A., McAdams, S., Collet, L. (eds) Auditory Signal Processing. Springer, NY. https://doi.org/10.1007/0-387-27045-0_51
18. J. Capon, "High-resolution frequency-wavenumber spectrum analysis," in Proceedings of the IEEE, vol. 57, no. 8, pp. 1408-1418, Aug. 1969, doi: https://doi.org/10.1109/PROC.1969.7278
19. R. Schmidt, "Multiple emitter location and signal parameter estimation," in IEEE Transactions on Antennas and Propagation, vol. 34, no. 3, pp. 276-280, March 1986, doi: https://doi.org/10.1109/TAP.1986.1143830
20. R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 7, pp. 984-995, July 1989, doi: https://doi.org/10.1109/29.32276
21. Rudzicz, F., Namasivayam, A.K. & Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang Resources & Evaluation 46, 523–541 (2012). https://doi.org/10.1007/s10579-011-9145-0

Received (Надійшла) 15.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Раптанов Даніїл Андрійович - магістрант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Daniil Raptanov - master's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: daniil.raptanov@nure.ua, ORCID Author ID: http://orcid.org/0009-0001-9564-0080 .

Барковська Олеся Юріївна – кандидат технічних наук, доцент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Olesia Barkovska – Candidate of Technical Sciences, Associate Professor at the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: olesia.barkovska@nure.ua; ORCID Author ID: http://orcid.org/0000-0001-7496-4353;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=24482907700

Шиленко Михайло Павлович – магістрант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Mykhailo Shylenko - master's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: mykhailo.shylenko@nure.ua; ORCID Author ID: http://orcid.org/0009-0009-4084-4711.

Головченко Олександр Сергійович - аспірант кафедри Електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleksandr Holovchenko – Phd student of Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: oleksandr.holovchenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0002-7582-1746.

Івахненко Діана Сергіївна - бакалавр кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Diana Ivakhnenko - bachelor's student of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: diana.ivakhnenko@nure.ua; ORCID Author ID: https://orcid.org/0009-0005-0989-0016.

Дослідження точності роботи методів бімформінгу в контексті інклюзивної системи внутрішньої навігації

Д. А. Раптанов, О. Ю. Барковська, М.П. Шиленко, О. С. Головченко, Д. С. Івахненко

Анотація. Актуальність. Голосове керування елементами інклюзивних навігаційних систем є критично важливим для забезпечення автономності та безпечної мобільності людей з порушеннями зору в громадських місцях, зокрема у великих торговельних залах. Однак існуючі системи перетворення мовлення на текст (STT) стикаються із суттєвим зниженням точності розпізнавання через високодинамічний та нестаціонарний акустичний шум супермаркетів. **Об'єктом дослідження** є препроцесинг аудіопотоку та просторова фільтрація (бімформінг) у системі голосового керування за умов динамічного нестаціонарного шуму. Проблема полягає у недостатній вибірковості стандартних алгоритмів обробки аудіосигналу в умовах фонових завад магазину, що призводить до критичного зростання частки помилок у розпізнаних словах (WER) та робить систему управління розумним візком вразливою.. **Метою статті** є оцінка впливу зовнішніх факторів (кількості, просторової топології розміщення та рівня потужності джерел акустичного шуму) на точність методів просторової фільтрації (бімформінгу) для подальшого розпізнавання голосових команд шляхом комп'ютерного моделювання. **В результаті** роботи за допомогою бібліотеки Pyroomacoustics було змодельовано акустичне середовище та мікрофонну решітку. Проведено порівняння трьох методів: Delay-and-Sum (DAS), Max-UDR та Max-SINR. Дослідження показало, що алгоритм Max-SINR забезпечує найвищий приріст співвідношення сигнал/шум (delta SNR від 7,9 до 9,1 дБ) і є математично стійким до змін відстані до завад та їхньої потужності. Метод DAS виявився найменш ефективним (5,35–5,95 дБ) і продемонстрував чутливість до зміни дистанції. Встановлено, що ключовим фактором деградації сигналу є конфігурація джерел шуму, серед яких перехресна топологія (cross) є найскладнішою для фільтрації.

Ключові слова: інклюзивна система навігації, порушення зору, розпізнавання мовлення, просторова фільтрація, бімформінг, Delay-and-Sum, Max-UDR, Max-SINR, динамічний шум.

UDC 004.942 **doi: 10.26906/SUNZ.2026.2.176**

Olha Rybak

Odesa Polytechnic National University, Odesa, Ukraine

DEVELOPMENT OF A DECISION SUPPORT SYSTEM USING ARTIFICIAL NEURAL NETWORK AND GENETIC ALGORITHM

Abstract. Relevance. Nowadays artificial intelligence technologies are developing rapidly making it possible to automate the most routine component of data processing. AI is based on the computing architecture of a neural network which applies modeling of biological processes that occur in human brains. To improve the structure of neural network for a decision support system and determine its key parameters, such as the number of inputs, the quantity of layers and neurons within each of them, and choosing a training method, this study suggests to use evolutionary methods. **The purpose of this research** is to investigate the principle of operation of an artificial neural network, whose parameters and structure are determined using genetic algorithm, and to design a decision support system on the basis of the developed model. **Research results.** Taking into consideration that genetic algorithms software implementation requires a good random number generator and that the basis for the correct functioning of a neural network is the training sample that describes the presented task, it was decided to use the source database of learning materials which can provide parameter values for this purpose. Along with databases, important parts of the developed system are new phenotypes generation block, the block for evaluating them and the neural network training block. The process of a neural network training is preceded by determining a set of training samples and adding noise to them, since the output signals of a well-trained neural network should be insensitive to variations of input values within certain acceptable limits in order to implement monotonic data display. The main criterion when choosing the optimal network architecture is its ability to generalize different types of tasks. **Conclusions.** Defining parameters of artificial neural network using genetic algorithm allows to simplify the design of its structure and to develop a decision support system on its basis. Experimental results prove that after the training phase is complete the processed data is divided into clusters that correspond to either solution.

Keywords: artificial neural network; decision support system; genetic algorithm; artificial intelligence; clustering; evolutionary methods.

Introduction

Rapid advancement of artificial intelligence technologies in recent years has fundamentally changed the approach to many types of activities. AI tools make it possible to automate the most routine component of digital tasks and data processing. AI is based on the computing architecture of a neural network which applies modeling of biological processes that occur in the brains of humans and animals. Such a network, created within a computer program, is capable of self-learning, performing specific tasks, correcting errors, etc. The concept of an artificial neural network was originally proposed by American scientists Warren McCulloch and Walter Pitts in 1943. Their model of the complex neural connections of the human brain represented a network of vacuum tubes. Over the years expanding technological capabilities of computer engineering led to the creation of more complex mathematical algorithms based on machine learning. It was a new approach to solving decision-making problems as well as modeling, identification and signal processing. Today problems of classification and image recognition, forecasting and control of dynamic processes are solved using AI. To improve the structure of neural network and determine its key parameters, such as the number of inputs, the quantity of layers and neurons within each of them, choosing a training method, etc., this study suggests to use evolutionary methods, specifically the genetic algorithm.

Review of Recent Studies and Publications. Over the past few years research focus has shifted from a simple increase of the number of artificial neural network layers towards understanding their scalability, security and logical capabilities. Scaling laws investigation [1] has become a major scientific breakthrough and proved that model performance is a function of both number of parameters and training data size. A designed compute-optimal model named Chinchilla dramatically improved efficiency of the training process and allowed to predict its performance beforehand. It revealed that most of neural network models had been undertrained earlier, setting a new standard for large models training.

The study of large language models, presented by OpenAI company [2], proposed to configure neural network models using approach of reinforcement learning from human feedback. Therefore, models upgraded significantly giving more useful, safer and better response to following complex user instructions, which became the basis for all modern chatbots including ChatGPT and Claude.

An attempt of adapting neural networks to solve difficult tasks was made in paper [3] on the basis of Chain-of-Thought. It enables models to solve multi-step logic, mathematics and coding problems due to intermediate reasoning steps generated before the final answer is found.

Creating photorealistic images using artificial neural network received a boost in the study [4]. It is focused on diffusion models that enable to remove noise gradually from a random set of information, particularly visual data. A step towards Artificial General Intelligence was made in paper [5], where single neural network architecture is used to handle various inputs/outputs variables (text, images, control signals). As a result, it can perform hundreds of different tasks and reduces the need for maintaining specialized models.

Massive increase in capacity of a neural network with only a fractional increase in computational cost was presented in [6], which allowed to create huge network models. Instead of activating the entire network, the developed model routes queries to specific sub-networks.

Authors of [7] carried out training to perform an internal "thought process" of neural network before its responding. This approach improves performance on complex reasoning tasks and makes it possible for the model to correct itself.

Therefore, the improvement of neural network architectures and data processing models based on them became a foundation for today's significant progress of artificial intelligence technologies.

The purpose of this research is to investigate the principle of operation of an artificial neural network, whose parameters and structure are determined using genetic algorithm, and to design a decision support system on the basis of the developed model.

Main part

Artificial neural network consists of the elements called neurons which have the similar structure and summarize received signals. If the total sum exceeds threshold level, an output signal is produced, otherwise neuron does not respond to input signals. Therefore, transfer function of the neuron can be expressed as:

$$f_{tr} = \begin{cases} 0, \ for \ \sum_i x_i \le \gamma; \\ 1, \ for \ \sum_i x_i > \gamma, \end{cases} \quad (1)$$

where x_i – input signals of the neuron; γ – threshold value. Hence, output of a binary neuron is determined according to the equation:

$$F = f_{tr}\left(\sum_i x_i w_i, \gamma\right), \quad (2)$$

where w_i – input weights. Fig. 1 descibes operating principle of binary neurons with *n* inputs and one output.

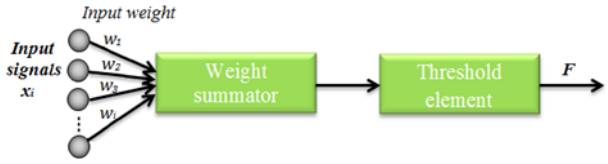


Fig. 1. Model of a binary neuron with n inputs and one output

At the beginning of the learning process neurons have equal or randomly distributed weights of input contacts with a summator. To resolve more complicated problems, several neurons can be combined into neural network. Learning of an artificial neural network means that identified data in the form of combination of signals x_i come to the input of the system. Each of the signals is binary and can take one out of two values – either 0 or 1. Reference signal F_0 is given to the output of a neuron. It defines what should appear at the output of trained neuron. In the case neuron returns the signal different from reference one, the process of machine learning begins.

Nowadays, there are a large number of neural network architectures and methods for training them. Combination of a multilayer perceptron and a backpropagation algorithm is one of the most widespread approaches. This technique is based on gradient estimation, and its disadvantages are the significant time required for calculations and the fact that it is not always possible to obtain accurate results. Another way to train neural networks is using probabilistic methods, in particular the principle of stochastic modeling. Metropolis-Hastings algorithm, simulated annealing, Gibbs sampling can be distinguished among them. Stochastic neural network learning procedure is developed in Bayesian networks, restricted Boltzmann machine, Helmholtz machine, deep belief network. A separate class of neural network training strategies is the search for the weight of synaptic connections and the network structure organization which can be performed using genetic algorithms.

First of all, it is necessary to note that genetic algorithms software implementation requires a good random number generator, since this approach is largely determined by probabilistic selection. Besides, the basis for the correct functioning of a neural network is the training sample that describes the presented task. A standard version of the source database of learning materials can provide parameter values for this purpose. Architecture of decision support system that uses a genetic algorithm to find the best neural network structure for the problem being solved is shown in Fig. 2.

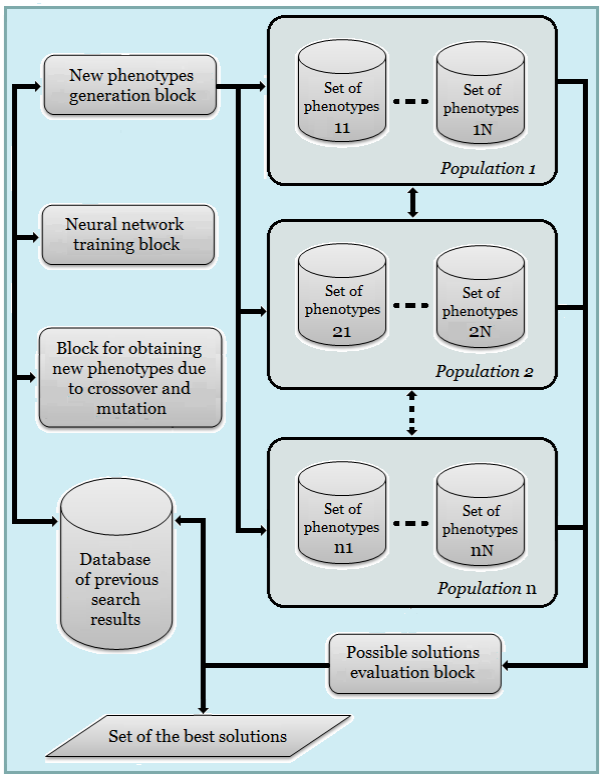


Fig. 2. Decision support system architecture based on genetic algorithm and neural network

Presented system implies the use of databases in its topology, which are mostly dedicated to store necessary data about the results of previous successful and unsuccessful decisions. Therefore, the time required to determine an appropriate neural network structure during further operation of the algorithm can be reduced. The database also stores protocols of all neural network parameters search results and the solutions of previous tasks. Along with databases, important parts of the developed system are new phenotypes generation block, the block for evaluating them and the neural network training block. In addition to direct testing of the new population, the evaluation block also uses a specific

algorithm to determine individuals for the next generation. Selection can be implemented using the roulette wheel or the tournament principle [8]. This subsystem should be independent of other blocks of the considered topology. New phenotypes generation block operates both at the stage of forming the initial set of possible solutions and for further obtaining subsequent populations. Applying databases in the presented scheme allows to determine starting position during phenotypes generating, so that the final solution of the problem can be discovered faster [9]. Meanwhile the random nature of the required parameters search using genetic algorithm remains unchanged.

Neural network training process takes place in a separate subsystem – the neural network training block, which receives a common chromosome with parameters for the network designer and recommendations for the training process from the database according to the previous runs. These recommendations contain information about the method and speed of learning, its sequence, etc., and at the output of the subsystem matrices of synaptic connection weight values are determined. The objective function of the problem of determining a neural network structure can be considered the formula that evaluates the quality of neural network training:

$$Q = \sum_{j=1}^{m}\left(\sum_{i=1}^{k}\left|F_i^j - F_{0i}^j\right| \Big/ (k \cdot m)\right),, \qquad (3)$$

where F – real output value; F_0 – required output value; m – number of examples in the validation selection; k – number of neural network outputs.

The dynamic architecture of a neural network assumes that neuron layers are consistently generated until the given level of accuracy is achieved. The final decision on which network scheme to choose is made only after a full study of different possible types of its structure. It is followed with reducing error in solving the problem to an admitted value. The process of a neural network training is preceded by determining a set of training samples and adding noise to them, since the output signals of a well-trained neural network should be insensitive to variations of input values within certain acceptable limits in order to implement monotonic data display. The main criterion when choosing the optimal network architecture is its ability to generalize different types of tasks.

To develop a decision support system on the basis of neural network, it is necessary to identify a specific set of outcomes regarding certain decision and solve the clustering problem. Unlike classification, which involves distribution of input space vectors between several classes defined by the user, clustering performs research of the input set of vectors in order to identify and divide them among several groups according to characteristics that determine proximity between the elements of the set. Clustering is carried out automatically, clusters are not set by the initial conditions of the problem, they are formed assuming similarity of feature vectors. The features of the studied objects should be encoded in numeric form and normed using an appropriate algorithm. After preprocessing, N-dimensional feature space is obtained including grouped vectors. The dimensionality of the space N depends on the number of parameters that determine each decision. During the neural network training, the number of feature vectors of the training sample must be greater than the specified number of clusters C, and when the number of vectors exceeds the $N \cdot C$ product, the clustering process provides satisfactory result. In case of linear cluster discreteness, received clusters of inputs can be separated with lines (for $N = 2$) or planes (for $N = 3$). If they are separated with a line or a surface of more complex shape, there is a nonlinear cluster discreteness. If the clusters overlap, there is probabilistic discreteness, which means that a feature vector can be attributed to one or another cluster with a certain probability. Since most of the neural network architectures are unusable for solving problems with probabilistic discreteness, the problem is restricted to condition that decision support system is based on clusters which are linearly separable or nonlinearly separable, or can be reduced to them due to data preprocessing.

Neural network simulation was carried out via Deep Learning Toolbox application, which is part of the Matlab software package. The result of the neural network operation is presented as a graph in two-dimensional space of input features (Fig. 3).

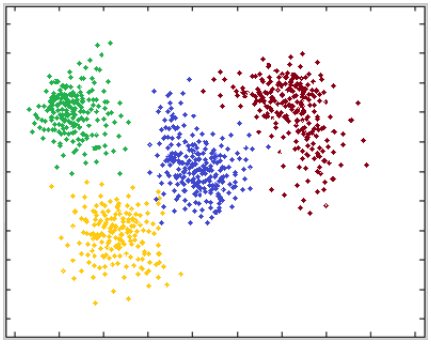


Fig. 3. Visualization for clustering using trained network

For a three-dimensional space of parameters, the graphic image is also three-dimensional, and for higher dimensional vectors, it is practically impossible to visualize the results. In this case, the multidimensional feature space can be reduced to a space of lower dimension using reflection.

Conclusions

Nowadays neural networks are no longer constrained and insufficient methods for solving complex, highly specialized problems. Featuring artificial intelligence advancement, they have become widespread in common applications for data organization, time reduction and decision making. Development of a decision support system on the basis of artificial neural network is carried out using evolutionary methods to create its structure. Parameters of the neural network are determined proceeding from the genetic algorithm results. Such an approach simplifies neural network design and allows to carry out cluster analysis of data sets. Experimental results prove that after the training phase is complete the processed data is divided into clusters that correspond to either solution.

Conflicts of interest. The author declare that he has no conflicts of interest in relation to the current study,

including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence. The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E. et al. (2022). Training Compute-Optimal Large Language Models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 2176. – P. 30016–30030. DOI: https://doi.org/10.48550/arXiv.2203.15556
2. Ouyang L., Wu J., Jiang X., Almeida D. et al. (2022). Training language models to follow instructions with human feedback. OpenAI. Journal «Advances in neural information processing systems». Vol. 35. – P. 27730-27744. DOI: https://doi.org/10.48550/arXiv.2203.02155
3. Wei J., Wang X., Schuurmans D., Bosma M. et al. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800. – P. 24824–24837. DOI: https://doi.org/10.48550/arXiv.2201.11903
4. Sauer A., Boesel F., Dockhorn T., Blattmann A. et al. (2024). Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. In SIGGRAPH Asia 2024 Conference Papers (SA '24). Association for Computing Machinery, New York, NY, USA, Article 106. – P. 1–11. DOI: https://doi.org/10.1145/3680528.3687625
5. Reed S., Żołna K., Parisotto E., Gomez S. et al. (2022). A Generalist Agent. Transactions on Machine Learning Research. – 42 p. DOI: https://doi.org/10.48550/arXiv.2205.06175
6. Fedus W., Zoph B., Shazeer N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research. Vol. 23, No. 1, Article 120. – P. 5232–5270. DOI: https://doi.org/10.48550/arXiv.2101.03961
7. Zelikman E., Harik G., Shao Y., Jayasiri V. et al. (2024). Quiet-star: Language models can teach themselves to think before speaking. First Conference on Language Modeling. arXiv preprint arXiv:2403.09629. – 26 p. DOI: https://doi.org/10.48550/arXiv.2403.09629
8. Рибак О. (2021). Застосування еволюційних методів оптимізації для вибору режимів шліфування. Технічні науки та технології. Чернігів: НУ «Чернігівська політехніка». № 4(26). – С. 97-105. DOI: https://doi.org/10.25140/2411-5363-2021-4(26)-97-105
9. Bayer H.G., Schwefel H.P., Wegener I. (2002). How to analyse evolutionary algorithms. Theoretical Computer Science, 2002, Vol. 287, Is. 1. – P. 101−130. DOI: https://doi.org/10.1016/S0304-3975(02)00137-8

Received (Надійшла) 02.02.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Рибак Ольга Володимирівна – кандидат технічних наук, доцент, доцент кафедри інформаційних технологій проєктування та дизайну, Національний університет «Одеська політехніка», Одеса, Україна;
Olha Rybak – PhD, Associated Professor, Department of Information Technologies and Design, Odesa Polytechnic National University, Odesa, Ukraine;
e-mail: olga.vol.rybak@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-0250-3037;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57208656221.

Розробка системи підтримки прийняття рішень за допомогою штучної нейронної мережі та генетичного алгоритму

О. В. Рибак

Анотація. Актуальність. На сьогоднішній день технології штучного інтелекту зазнали бурхливого розвитку, що дає можливість автоматизувати рутинну частину обробки даних. ШІ заснований на обчислювальній архітектурі нейронної мережі, яка застосовує моделювання біологічних процесів, що відбуваються в людському мозку. З метою вдосконалення структури нейронної мережі для системи підтримки прийняття рішень та вибору її ключових параметрів, зокрема визначення числа входів, кількості шарів та нейронів у кожному з них, методу навчання тощо у даному дослідженні пропонується використовувати еволюційні методи. **Метою статті** є дослідження принципу функціонування штучної нейронної мережі, параметри і структура якої визначаються за допомогою генетичного алгоритму, і розробка системи підтримки прийняття рішень на основі створеної моделі. **Результати дослідження.** Враховуючи те, що генетичні алгоритми багато в чому спираються на ймовірнісний вибір, при їхній програмній реалізації необхідно використовувати якісний генератор випадкових чисел. Також основою правильного функціонування нейронної мережі є навчальна вибірка, що описує представлену задачу, відтак для зберігання значень параметрів застосовується база даних. Окрім баз даних, важливою частиною розробленої системи є блок генерації нових особин, блок їхньої оцінки і блок навчання неронної мережі. Процесу навчання нейронної мережі передує визначення набору навчальних вибірок та додавання до них шуму, оскільки вихідні сигнали добре навченої нейронної мережі повинні бути нечутливими до варіацій вхідних величин, що знаходяться у певних допустимих межах, задля реалізації монотонного відображення даних. Головним критерієм при виборі оптимальної архітектури нейронної мережі виявляється її здатність до узагальнення різних типів задач. **Висновки.** Визначення параметрів штучної нейронної мережі за допомогою генетичного алгоритму дозволяє спростити процес проєктування її структури, а також розробити на її основі систему підтримки прийняття рішень. Експериментальні результати доводять, що після завершення етапу навчання мережі оброблені дані поділяються на кластери, які співвідносяться з тим чи іншим варіантом вирішення задачі.

Ключові слова: штучна нейронна мережа; система підтримки прийняття рішень; генетичний алгоритм; штучний інтелект; кластеризація; еволюційні методи.

UDC 004.8:004.932.2 **doi: 10.26906/SUNZ.2026.2.180**

Oleksandr Sokolov, Anton Poroshenko, Roman Yaroshevych, Vladislav Kholiev

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

APPLICATION AND ANALYSIS OF MACHINE LEARNING METHODS FOR IMAGE CLASSIFICATION

Abstract. Relevance. Image classification is a key task in computer vision, which has wide applications in medicine, transportation, industry, and security. The use of optimized CNN architectures allows high accuracy to be achieved with limited resources, which is relevant for mobile and embedded systems. **Research subject:** machine learning methods and neural network architectures for image classification. **The purpose of the article** is to develop and evaluate a modified convolutional neural network that provides a balance between classification accuracy and computational efficiency, as well as to compare its results with classical and modern models. **Research results.** The proposed CNN achieved 93.8% accuracy on the Fashion MNIST dataset, exceeding the performance of LeNet-5 (91.2%) and classical algorithms (KNN – 85.3%, Decision Tree – 82.7%, XGBoost – 90.4%). On the more complex CIFAR-10 dataset, the model showed an accuracy of 80.2%, exceeding LeNet-5 but falling short of ResNet and EfficientNet. This confirms the effectiveness of the model for tasks of medium complexity and systems with limited resources. **Conclusions.** Modified CNN is a compromise between simple classical methods and complex modern architectures. It provides an optimal balance between accuracy and learning speed, making it suitable for practical application in mobile and embedded systems. Further research may focus on the use of more complex datasets, automatic hyperparameter selection, and the integration of self-attention mechanisms. Scope of application of the results obtained: medium-complexity computer vision systems, mobile and embedded devices with limited resources, applied image classification tasks.

Keywords: image classification, CNN, Vision Transformer, Fashion MNIST, neural networks, machine learning.

Introduction

Problem statement. Image classification is one of the key tasks in computer vision, which finds application in many areas: from medicine and autonomous vehicles to the fashion industry and face recognition. Recent advances in machine learning have made it possible to create models that can analyze and classify images with high accuracy. The development of deep learning methods, such as convolutional neural networks (CNNs), as well as architectures based on the self-attention mechanism (e.g., Vision Transformer), have opened up new horizons for solving this problem. This article examines the effectiveness of various approaches to image classification based on the Fashion MNIST dataset, which contains images of clothing, footwear, and accessories. Both classical machine learning methods (K-Nearest Neighbors, Decision Tree, XGBoost) and modern deep learning models were selected for the study: LeNet-5, VGG16, ResNet, EfficientNet, Vision Transformer [1, 2]. Particular attention is paid to the presentation and analysis of my implementation, which demonstrates significant advantages over other architectures.

The aim of this work is to study existing machine learning methods and modify the classical convolutional neural network. The proposed model is a modification of the classical convolutional neural network, built on the basis of well-known architectural solutions widely used in image classification tasks. The main focus of the work is on the practical analysis of the influence of architectural parameters (number of convolutional layers, use of MaxPooling, choice of activation functions and optimization algorithm) on the classification accuracy and computational efficiency of the model. Thus, the research is applied in nature and aims to evaluate the feasibility of using simplified CNN architectures in tasks with limited computational resources. A comparative analysis of the proposed model with other approaches allows us to assess how effective my implementation is in visual data analysis tasks

Main material

Analysis of methods. To achieve the set goal, a number of classification methods were applied, covering both classical machine learning algorithms and modern neural networks. Classic approaches include KNN, Decision Tree, and XGBoost. These methods are based on geometric and statistical classification principles, but their application to image processing tasks is limited due to the local nature of feature computation and insufficient generalization ability on complex datasets.

Neural networks, in particular ResNet, EfficientNet, and Vision Transformer, demonstrate higher efficiency due to their ability to consider both local and global features of images. Within the scope of this study, both the modern architectures and the implemented convolutional neural network were used to classify images from the Fashion MNIST dataset.

The Fashion MNIST dataset consists of 28×28 pixel grayscale images and contains 10 classes corresponding to different categories of clothing, footwear, and accessories [3]. The data is divided into a training sample of 60,000 images and a test sample of 10,000 images. In addition, the study used the CIFAR-10 dataset, which contains 32×32 pixel color images belonging to 10 classes, including vehicles, animals, and everyday objects. The use of CIFAR-10 made it possible to evaluate the ability of models to work with more complex and diverse visual data.

Classic machine learning methods have demonstrated limited effectiveness in image classification tasks. The KNN algorithm is simple to implement [4], but its performance decreases on large samples due to significant computational complexity. Decision Tree provides convenient interpretation of results but is prone

to overfitting. XGBoost shows better results thanks to its ensemble approach but requires significant computational resources. In contrast, neural networks are capable of automatically forming complex features from input data, making them more suitable for image analysis.

Modified CNN. The implemented convolutional neural network was optimized (Fig. 1) to work with relatively simple datasets. Its architecture includes several convolutional layers, MaxPooling operations to reduce feature dimensions, as well as modern ReLU activation functions and the Adam optimization algorithm. This combination allows for improved classification accuracy without a significant increase in computational costs. The proposed model is built in accordance with generally accepted principles of CNN development and does not contain fundamentally new architectural components, which ensures the correctness of comparing its results with classical and modern models such as LeNet-5, ResNet, EfficientNet, and Vision Transformer [6].

The general structure of the proposed convolutional neural network is shown in Fig. 1.

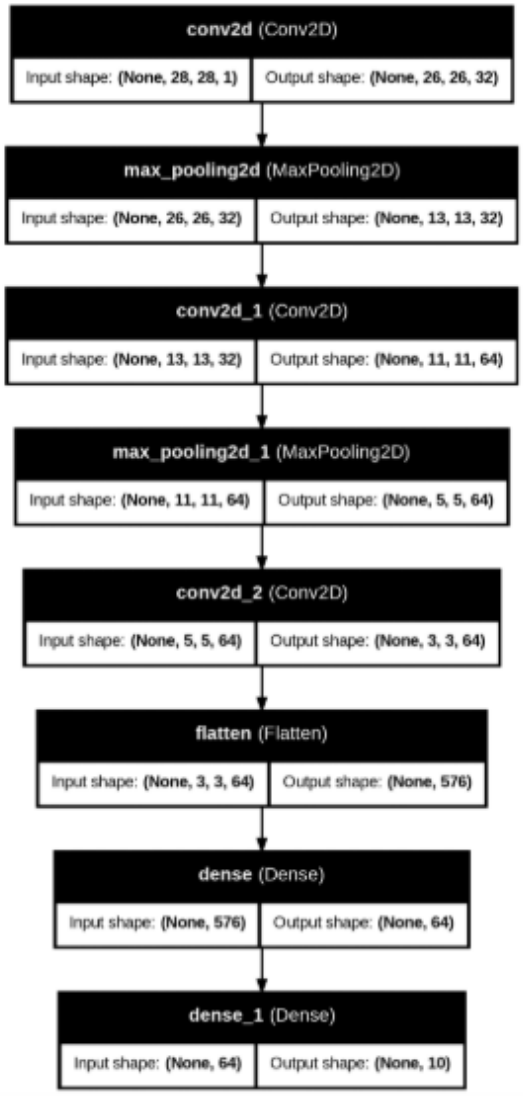


Fig. 1. Model architecture

The architecture was designed with an emphasis on simplicity and computational efficiency while preserving sufficient representational capacity for image classification tasks. Particular attention was paid to the selection of the number of layers and their configuration in order to ensure stable training and reliable performance on benchmark datasets.

The design of the modified convolutional neural network was guided by the need to achieve a balance between model simplicity and classification performance. Unlike deep modern architectures that rely on many layers and complex structural elements, the proposed CNN focuses on a limited number of carefully selected components that are sufficient for effective feature extraction. This approach allows the model to remain computationally efficient while avoiding excessive overfitting, which is especially important when working with relatively small and low-resolution datasets such as Fashion MNIST.

Results. The results of the study showed that this model demonstrates significant improvements in accuracy compared to classical machine learning methods and the classical LeNet-5. On the Fashion MNIST dataset, the developed CNN achieved an accuracy of 93.8%, while the classic version of LeNet-5 achieved only 91.2% (Table 1). This improvement was achieved using additional convolutional layers, modern activation functions (ReLU), and the Adam optimization algorithm.

After the convolutional feature extraction stage, the network includes fully connected layers that perform the final classification. The Adam optimization algorithm was selected due to its ability to adaptively adjust the learning rate, which leads to faster convergence and more stable training. Overall, the proposed architecture follows well-established CNN design principles and serves as a representative baseline for comparison with both classical and modern deep learning models.

Classic machine learning methods such as KNN, Decision Tree, and XGBoost demonstrated significantly lower performance. KNN achieved 85.3% accuracy, which can be explained by its high computational complexity and low ability to work with high-dimensional data. Decision Tree demonstrated 82.7% accuracy but suffered from overfitting. XGBoost achieved 90.4% accuracy thanks to its powerful ensemble approach but required significant computational resources and did not outperform neural networks.

Modern neural networks such as ResNet, EfficientNet, and Vision Transformer showed the highest results. ResNet achieved 95.5% accuracy by using residual connections that ensure stable training of deep models [8]. EfficientNet achieved 96.1% accuracy by optimizing the width, depth, and resolution of the model. Vision Transformer showed the best accuracy 96.3% using a self-attention mechanism that allows it to analyze the global context in images. However, these models require significantly more computational resources, which limits their application on less powerful hardware systems.

Discussion. It should be noted that the results of this study were obtained from a limited number of datasets, which may affect the generalizability of the conclusions.

Table 1 – **Test results**

	KNN	Decision tree	XG-Boost	Dence NN	LeNet-5	**My CNN**	VGG16	Res-Net	Efficient Net-80	Vision Transformer
Accuracy FASHION	85	75	88	90	92	**94**	94	96	96	97
Loss FASHION	-	-	0.6	0.4	0.35	**0.3**	0.3	0.25	0.2	0.18
Accuracy CIFAR-10	40	30	50	65	75	**80**	88	93	94	94
Loss CIFAR-10	-	-	1.2	1.1	0.8	**0.44**	0.4	0.4	0.35	0.3

In particular, the Fashion MNIST dataset contains relatively simple images with low resolution, which makes the classification task easier compared to real-world application scenarios. In addition, the selection of model hyperparameters was done empirically and did not cover the entire space of possible configurations. In future studies, it would be advisable to expand the set of experiments using more complex datasets and methods for automatic hyperparameter selection.

Although CNN is inferior in accuracy to ResNet, EfficientNet, and Vision Transformer, it has several key advantages. It provides an optimal balance between accuracy and computational efficiency, making it ideal for tasks of medium complexity and systems with limited resources. For example, the training time for this network is significantly shorter than for EfficientNet and Vision Transformer, allowing the model to be adapted to new tasks more quickly.

On the CIFAR-10 dataset, the implementation showed an accuracy of 80.2%, which surpasses the classic LeNet-5 (78.1%) but is inferior to modern models such as ResNet (91.3%) and EfficientNet (92.5%). This indicates that it is effective for tasks of low and medium complexity, but for high complexity, it is recommended to use deeper models with global data processing.

It should be noted that the results of this study were obtained from a limited number of datasets, which may affect the generalizability of the conclusions. In particular, the Fashion MNIST dataset contains relatively simple images with low resolution, which makes the classification task easier compared to real-world application scenarios.

In addition, the selection of model hyperparameters was performed empirically and did not cover the entire space of possible configurations. In further research, it is advisable to expand the set of experiments using more complex datasets and methods of automatic hyperparameter selection.

Conclusions

Thus, the proposed model demonstrates practical effectiveness as a compromise solution between simple classical machine learning methods and complex modern deep models. This confirms the feasibility of using similar CNN architectures in applied image classification tasks of medium complexity.

In this article, the following results were obtained:

1. A comparative analysis of classical machine learning methods and modern deep learning models for image classification was conducted using the Fashion MNIST and CIFAR-10 datasets, demonstrating the limitations of traditional approaches when applied to visual data.

2. A modified convolutional neural network based on standard CNN architectural principles was implemented and evaluated. The proposed model showed improved classification accuracy compared to the classical LeNet-5 architecture while maintaining lower computational complexity than deeper modern models.

3. Experimental results confirmed that the proposed CNN provides a balanced trade-off between accuracy and computational efficiency, making it suitable for image classification tasks of low to medium complexity, especially in environments with limited computational resources.

Future research may focus on extending the proposed approach by exploring more complex datasets, applying automated hyperparameter optimization techniques, and integrating advanced architectural components such as attention mechanisms or lightweight residual connections to further improve model performance and generalization ability.

From a practical perspective, the modified CNN can be effectively applied in mobile applications, embedded vision systems, and industrial monitoring tasks, where a compromise between accuracy, inference speed, and memory consumption is required. This confirms the relevance of lightweight convolutional models in modern computer vision pipelines.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Lebedev G. S., et al. (2020). Deep machine learning (artificial intelligence) in ultrasound diagnostics. Journal of Telemedicine and E-Health. Vol. 2020, No. 2, pp. 22–29. URL: https://doi.org/10.29188/2542-2413-2020-6-2-22-29
2. Polischuk M., Kostyuchko S., Khrystynets M. (2019). Porivnyannya metodiv optimizaciyi neyronnyh merezh na prykladi zadachi klasifikaciyi zobrazen'. [Comparison of Neural Network Optimization Methods on the Example of an Image Classification Task]. Komp'yuterno-integrovani texnologyi: osvita, nauka, vyrobnytstvo. No. 37, pp. 43–52. URL:https://doi.org/10.36910/6775-2524-0560-2019-37-7

3. Image classification using multiple convolutional neural networks on the fashion-mnist dataset / O. Nocentini та ін. *Sensors*. 2022. Т. 22, № 23. С. 9544. URL: https://doi.org/10.3390/s22239544
4. Liu Q. The development of image classification algorithms based on cnns. *Highlights in science, engineering and technology*. 2023. Т. 34. С. 275–280. URL: https://doi.org/10.54097/hset.v34i.5484
5. Mukhamediev R. I. State-of-the-Art Results with the Fashion-MNIST Dataset. *Mathematics*. 2024. Т. 12, № 20. С. 3174. URL: https://doi.org/10.3390/math12203174
6. Improved residual networks for image and video recognition / I. C. Duta та ін. *2020 25th international conference on pattern recognition (ICPR)*, м. Milan, Italy, 10–15 січ. 2021 р. 2021. URL: https://doi.org/10.1109/icpr48806.2021.9412193
7. An intelligent fashion object classification using CNN / D. Swain та ін. *EAI endorsed transactions on industrial networks and intelligent systems*. 2023. Т. 10, № 4. С. e2. URL: https://doi.org/10.4108/eetinis.v10i4.4315
8. Understanding robustness of transformers for image classification / S. Bhojanapalli та ін. *2021 IEEE/CVF international conference on computer vision (ICCV)*, м. Montreal, QC, Canada, 10–17 жовт. 2021 р. 2021. URL:https://doi.org/10.1109/iccv48922.2021.01007
9. A comprehensive survey of convolutions in deep learning: applications, challenges, and future trends / A. Younesi та ін. *IEEE access*. 2024. С. 1. URL: https://doi.org/10.1109/access.2024.3376441
10. Shermin T. Enhancing deep transfer learning for image classification : thesis. 2021. URL: http://researchonline.federation.edu.au/vital/access/HandleResolver/1959.17/179551

Received (Надійшла) 11.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Соколов Олександр Олександрович – студент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Oleksandr Sokolov – Student, Department of Electronic Computing Machines, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.
e-mail: oleksandr.sokolov@nure.ua; ORCID Author ID: https://orcid.org/0009-0005-5648-9663.

Порошенко Антон Ігорович – доктор філософії, старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Anton Poroshenko – PhD, Senior Lecturer, Department of Electronic Computing Machines, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: anton.poroshenko@nure.ua; ORCID Author ID: https://orcid.org/0000-0001-7266-4269;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57250025600.

Ярошевич Роман Олександрович – доктор філософії, старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Roman Yaroshevych – PhD, Senior Lecturer, Department of Electronic Computing Machines, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: roman.yaroshevych@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-7949-1513;
Scopus Author ID https://www.scopus.com/authid/detail.uri?authorId=58624172500.

Холєв Владислав Олександрович – доктор філософії, асистент кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vladyslav Kholiev – PhD, Assistant Professor, Department of Electronic Computing Machines, National University of Radio Electronics, Kharkiv, Ukraine;
e-mail: vladyslav.kholiev@nure.ua; ORCID Author ID: https://orcid.org/0000-0002-9148-1561;
Scopus Author ID https://www.scopus.com/authid/detail.uri?authorId=57224189723.

Застосування та аналіз методів машинного навчання для класифікації зображень

О. О. Соколов, А. І. Порошенко, Р. О. Ярошевич, В. О. Холєв

Анотація. Актуальність. Класифікація зображень є ключовим завданням у комп'ютерному зорі, яке має широке застосування в медицині, транспорті, промисловості та безпеці. Використання оптимізованих архітектур CNN дозволяє досягти високої точності з обмеженими ресурсами, що є актуальним для мобільних та вбудованих систем. **Тема дослідження:** методи машинного навчання та архітектури нейронних мереж для класифікації зображень. **Мета статті**. розробити та оцінити модифіковану конволюційну нейронну мережу, яка забезпечує баланс між точністю класифікації та обчислювальною ефективністю, а також порівняти її результати з класичними та сучасними моделями. **Результати дослідження.** Запропонована CNN досягла точності 93,8% на наборі даних Fashion MNIST, перевищивши продуктивність LeNet-5 (91,2%) і класичних алгоритмів (KNN – 85,3%, Decision Tree – 82,7%, XGBoost – 90,4%). На більш складному наборі даних CIFAR-10 модель показала точність 80,2%, перевищивши LeNet-5, але поступившись ResNet і EfficientNet. Це підтверджує ефективність моделі для завдань середньої складності та систем з обмеженими ресурсами. **Висновки.** Модифікована CNN є компромісом між простими класичними методами та складними сучасними архітектурами. Вона забезпечує оптимальний баланс між точністю та швидкістю навчання, що робить її придатною для практичного застосування в мобільних та вбудованих системах. Подальші дослідження можуть зосередитися на використанні більш складних наборів даних, автоматичному виборі гіперпараметрів та інтеграції механізмів самоуваги. Сфера застосування отриманих результатів: системи комп'ютерного зору середньої складності, мобільні та вбудовані пристрої з обмеженими ресурсами, прикладні завдання класифікації зображень.

Ключові слова: класифікація зображень, CNN, Vision Transformer, Fashion MNIST, нейронні мережі, машинне навчання.

UDC 004.832:004.932.2 **doi: 10.26906/SUNZ.2026.2.184**

Svitlana Shapovalova, Olga Mazhara, Yurii Moskalenko, Vladyslav Titov

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

RULE EXTRACTION FROM A KOHONEN SELF-ORGANISING MAP FOR EQUIPMENT CONDITION ASSESSMENT USING NOISY DIAGNOSTIC SIGNALS

Abstract. This paper proposes a method for extracting classification rules for one-dimensional (1D) diagnostic signals from a trained Kohonen self-organising map (SOM). To validate equipment condition assessment from time-series data, a classification problem involving second-order curves with similar fragments was formulated. A balanced dataset was generated from a mathematical model, the SOM was trained, and the corresponding clusters were identified using a multilayer perceptron (MLP). Rules were extracted using a decompositional approach in IF–THEN form, where conditions were defined in terms of the best-matching units (BMUs) of the input signal. The proposed approach enables robust classification of noisy signals. A software suite for populating a knowledge base with extracted rules was developed. Computational experiments on signal classification were conducted using both the SOM and a CLIPS-based rule-based system, with the number of antecedent conditions and the noise factor serving as simulation parameters. The results show that, with the maximum number of antecedent conditions, the classification accuracy of the rule-based system decreases by 1–3 percentage points depending on the noise factor.

Keywords: rule extraction, neural networks, Kohonen self-organising map, CLIPS, rule-based system, best-matching unit.

Introduction

Problem statement. Neural networks have been widely adopted to solve applied artificial intelligence problems, including system identification, process control, decision-making, pattern recognition, data mining, and medical and technical diagnostics. However, a well-known limitation of this approach is the lack of interpretability mechanisms for the inference process, which is particularly critical in safety-sensitive systems.

One of the most common applied tasks in equipment condition monitoring is identifying operating conditions from time-series data. Real-time diagnostic systems must be computationally efficient while also providing explanations of their outputs. Extracting rules from trained neural networks is therefore both a timely and practically important research area.

Analysis of recent research and publications sources. A systematic review of rule extraction methods from feedforward neural networks, along with contemporary classification criteria for implementing explanations, is presented in [1].

Three main approaches to rule extraction from neural networks can be distinguished:

pedagogical – rule extraction for the network as a whole (the black-box principle);

decompositional – rule extraction for each individual component of the neural network (the white-box principle);

eclectic – a combination of the decompositional and pedagogical approaches.

In [2], an analysis and evaluation of the effectiveness of rule extraction algorithms across all three approaches using three datasets was conducted. Examples of rule extraction algorithms are presented in [3, 4] (pedagogical), [5, 6] (decompositional), and [7] (eclectic). A method for generating explanations of neural network inference is presented in [8].

For state identification from diagnostic signals, the decompositional approach was applied to the Kohonen self-organising map (SOM). A key property of the SOM is its ability to preserve topology by mapping high-dimensional input data onto a two-dimensional representation. Rule extraction from the SOM is discussed in [9] and [10]. In those studies, cluster identification within the SOM constituted a distinct extraction stage. In the present work, class labels (corresponding to equipment operating modes) are assigned to SOM clusters beforehand using a multilayer perceptron (MLP) for classification. This SOM–MLP neural network ensemble enables rapid diagnostics even from signals with missing data [9].

In addition, extracting classification rules from diagnostic signals raises the problem of overly large rule antecedents produced by the SOM. Addressing this requires a separate stage devoted to identifying the most informative signal features.

To formalise the rules derived from the mapping between diagnostic signals and SOM clusters, the CLIPS expert system language [12] was adopted.

Task statement. This paper proposes a method for extracting classification rules for 1D diagnostic signals from a trained Kohonen self-organising map.

To this end, the following objectives are pursued:

1) define the applied task of rule extraction;

2) determine the rule representation format;

3) develop a method for extracting classification rules from the SOM;

4) experimentally validate the effectiveness of the proposed method.

The rule extraction task for second-order curve classification

The task was formulated as the extraction of rules from a self-organising map trained to classify time-series data corresponding to second-order curves with similar fragments, namely the upper arcs of a circle, an ellipse, and a parabola. To ensure maximum similarity, the parameters of the analytical equations for the second-order curves and their domains were carefully selected. The input signal to the neural network is a set of discrete function values defining the corresponding curve, with additive Gaussian noise. The mathematical model for generating training examples in the dataset is presented in [11]. Each example

comprises the neural network input signal and the class (curve label) to which the signal corresponds.

Rule construction is based on matching each component of the input vector to the grid nodes of the Kohonen network. Under the unsupervised learning paradigm, the SOM requires no target vector: it learns to cluster the data from unlabelled examples. For each input signal, the best-matching unit (BMU) is computed – the grid node whose weight vector is closest to the input signal:

$$Dist = \sqrt{\sum_{i=1}^{i=n}(x_i - w_i)^2}, \quad (1)$$

where x_i is the i-th value of the input vector, w_i is the value of the i-th weight of the BMU, and n is the length of the input vector.

Here, the Euclidean distance serves as the metric.

Extracting classification rules from the SOM requires:

1) a trained self-organising map;

2) a balanced dataset of input signal examples for rule construction;

3) a known mapping between classes and SOM clusters.

Rule Representation

Rule extraction algorithms for neural networks can represent knowledge as mathematical expressions, symbolic logic expressions, fuzzy logic expressions, or decision trees. In practice, however, the two most common types of logical rules are:

IF–THEN (conjunctive): IF condition1 AND condition2 AND condition3 THEN RESULT;

M-of-N (subset selection): IF (M of the following N antecedents are TRUE) THEN RESULT.

Under the decompositional approach, the rule conditions directly reflect characteristics of the neural network.

An M-of-N rule treats its antecedent as true when M out of N Boolean expressions representing the network inputs are satisfied; the consequent is then interpreted from the neuron outputs. Notably, M-of-N rules can always be converted into IF–THEN form.

A more general approach to rule representation relies on first-order predicate logic (typically with certain restrictions). For example, in [13] the Gyan methodology is proposed, which encodes the knowledge of a trained network as restricted first-order predicate rules.

In this work, the extracted rules are formalised using the syntax of expert system shells to make them executable.

Among existing knowledge representation models, the production model aligns most closely with the IF–THEN rule format. Rule-based systems based on this model are commonly viewed as an adaptation of classical logic for artificial intelligence. Accordingly, the rules extracted from the SOM are expressed here in the CLIPS language [12]. CLIPS was chosen for its ability to construct antecedent logical expressions consistent with first-order logic, including existential (exists) and universal (forall) quantifiers, as well as for its support of basic mathematical and user-defined functions.

In the CLIPS environment, rule conditions are represented as fact templates. Facts can be either predefined (ordered) or generated dynamically during rule construction (unordered).

To simplify the generation procedure, rule antecedents may be built from unordered fact structures; however, preprocessing the data to identify significant structures improves rule readability.

For tasks with a well-defined output format (e.g., classification), output values are represented as ordered facts. A similar concept appears in the predicate-based rule generation approach of [13], where the target predicate maps to an ordered fact template in CLIPS.

Method for extracting classification rules from SOM

Classification rules are extracted from the SOM in the following stages:

Stage 1. Identifying the BMU for every input vector in the dataset. All vectors from the dataset are fed to the trained network, and the BMU is determined for each:

$$\overline{X_j} \rightarrow BMU_k, if;$$
$$Dist_k = min(Dist_1, Dist_2, \ldots, Dist_m), \quad (2)$$

where $\overline{X}_j$ is the j-th vector in the dataset, BMU_k is the k-th SOM grid node that provides the best match for the input vector $\overline{X}_j$, $Dist_k$ is the distance between the vector $\overline{X}_j$ and the weight vector of the k-th SOM grid node (1), and m is the number of nodes in the SOM grid.

This yields a set of mappings $\overline{X}_j \rightarrow BMU_k$.

Stage 2. Determining the range of input vector component values for each BMU. Since a single BMU may correspond to several input vectors, all input signals mapped to each BMU are first identified:

$$\{\overline{X_1^k}, \overline{X_2^k}, \ldots, \overline{X_j^k}, \ldots, \overline{X_v^k}\} -> BMU_k, \quad (3)$$

where BMU_k is the BMU corresponding to the *k*-th grid node, belonging to the set of best-matching units identified in Stage 1; $\overline{X_j}^k$ is the j-th input vector from the set of vectors corresponding to the k-th BMU; and *v* is the number of input vectors corresponding to the *k*-th BMU.

All vectors on the left-hand side of (3) share the same dimensionality. The maximum and minimum values among all i-th elements must be found:

$$x_{ji\ min}^k = min(x_{j1}^k, x_{j2}^k, \ldots, x_{ji}^k, \ldots, x_{jn}^k), \quad (4)$$

where $x_{ji}{}^k$ is the *i*-th element of the j-th input vector corresponding to the *k*-th BMU, and *n* is the length of the input signal.

The maximum value is determined analogously.

This yields the following set of value ranges:

$$\{(x_{j1\ min}^k, x_{j1\ max}^k), (x_{j2\ min}^k, x_{j2\ max}^k), \ldots, (x_{ji\ min}^k, x_{ji\ max}^k), \ldots,$$
$$(x_{jn\ min}^k, x_{jn\ max}^k)\} -> BMU_k, \quad (5)$$

where x_{ji}^k is the i – th element of the *j*-th input vector $\overline{X_j^k}$ corresponding to the *k*-th BMU, BMU_k is the *k*-th SOM grid node that provides the best match for all vectors X_j^k (6), n is the length of the input signal.

Stage 3. Determining the classes for all BMUs identified in Stage 1. At this stage, the right-hand side of (5) is replaced with the class label from the predefined class-to-cluster mapping:

$$BMU_k := A_r, \quad (6)$$

where A_r is the class identifier from the set SA.

The resulting set of mappings takes the form:

$$\{(x^k_{j1\ min}, x^k_{j1\ max}), (x^k_{j2\ min}, x^k_{j2\ max}),\ldots,$$
$$(x^k_{ji\ min}, x^k_{ji\ max}),(x^k_{jn\ min}, x^k_{jn\ max})\}\text{->}A_r \quad (7)$$

Expression (7) is a formal rule suitable for inclusion in an expert system knowledge base.

Stage 4. Reducing the antecedent dimensionality. The left-hand side of rule (8) has the same dimensionality as the input vector; consequently, high-dimensional input signals produce rules with a correspondingly large number of conditions. This increases the size of the knowledge base and slows down inference. It is therefore beneficial to reduce the rule antecedent at this stage by removing input signal fragments that are nearly identical across classes.

For the second-order curve classification task, regions were identified where the difference in curve values did not exceed a specified threshold ε (Fig. 1).

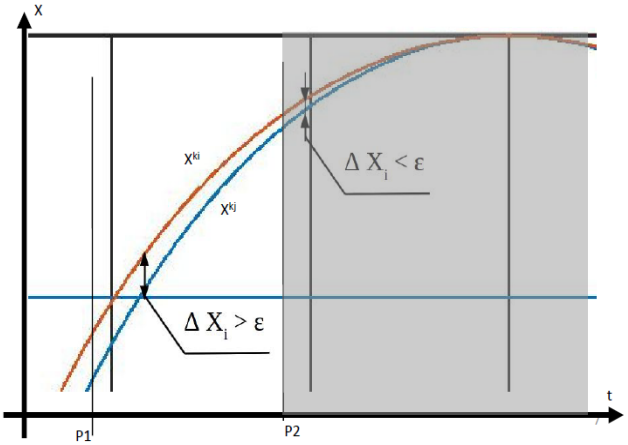


Fig. 1. Identification of curve regions for antecedent dimensionality reduction

The values x_i falling within these regions were removed from all rules (8).

Stage 5. Knowledge base rule representation. In production form, rules (8) are written as:

if
$x_1 \in [x^k_{j1\ min}, x^k_{j1\ max}]$ & $x_2 \in [x^k_{j2\ min}, x^k_{j2\ max}]$ & ... & $x_{n^*} \in [x^k_{jn^*\ min}, x^k_{jn^*\ max}]$ (8)
then
where x_i is the i-th value of the input vector;

x_{ji} min^k and x_{ji} max^k are the minimum and maximum values among all i-th elements of the input vectors;
A_r is the class identifier from the set SA; and n* is the number of input vector values remaining after the removal of signal convergence regions (Stage 4).

Stage 6. Knowledge base entry generation. The knowledge base entries were generated in the CLIPS language following formula (9).

Fact template representation:

```
(deftemplate figure
  (slot x1 (type Float) * (range − 3.0 9.0))
  (slot x2 (type Float) * (range − 3.0 9.0))
  (slot x3 (type Float) * (range − 3.0 9.9))
  ...
  (slot x100 (type Float) * (range − 3.0 9.9)))
```

Example of an ellipse classification rule:

```
(defrule ellipse
  (figure (x1 ? x1) * (x2 ? x2) * (x3 ? x3) ... (x50 ? x50)
  (test (and (≥ ? x1 2.277) * (≤ ? x1 2.357)))
  (test (and (≥ ? x2 2.321) * (≤ ? x2 2.401)))
  (test (and (≥ ? x3 2.342) * (≤ ? x3 2.422)))
  ...
  (test (and (≥ ? x50 2.342) * (≤ ? x50 2.422)))
  =>
  (assert (response ellipse))
  (printout t The figure is ellipse" clrf))
```

Computational experiments on rule extraction

Fig. 2 shows the components of the software system developed to extract classification rules for second-order curves from the Kohonen SOM under noisy input conditions, together with the intermediate data.

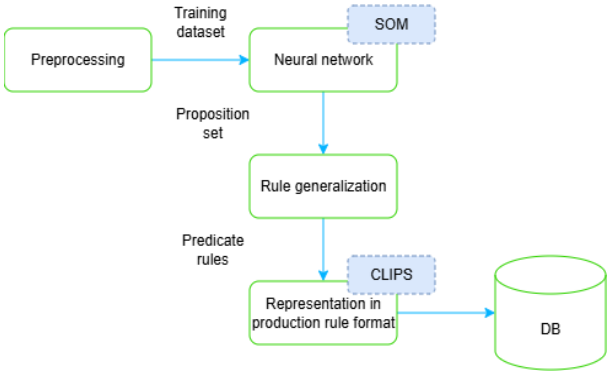


Fig. 2. Schematic diagram of rule extraction for the knowledge base

The neural network ensemble used for classification is described in [11]; it consists of the SOM (for clustering) and the MLP (for classification). Table 1 summarises the characteristics and corresponding values of the neural network, the training sample, and the training process.

Table 2 presents the conditions and results of the computational experiments. The computational experiments reveal that, compared with the neural network, the classification accuracy declines with increasing rule complexity (i.e., the number of antecedent conditions):

when 50 features remain after removing similar signal segments, accuracy drops by 1–2 percentage points;

for 40 remaining features, the drop ranges from 3 to 8 percentage points;

for 30 remaining features, the drop ranges from 3 to 10 percentage points.

Overall, the results demonstrate that rules produced by the proposed method yield rule-based systems whose inference accuracy is comparable to the neural network's.

Table 1 – **Characteristics of the computational experiments**

#	Category	Parameter	Value
1	Training sample parameters	Abscissa range	*[–3; 3]*
		Ordinate range	*[0; 9]*
		Noise	*δ = Rand(–x_max/k, x_max/k)*
2	Training characteristics	Number of SOM training examples	*100*
		Number of SOM test examples	*100*
		Number of MLP training iterations	*800*
3	Neural network characteristics	Input vector size	*100*
		SOM grid dimensionality	*10×10*
		Output vector size	*3*

Table 2 – **Results of the computational experiments**

Noise factor δ	Classification accuracy, %			
	Neural network	50 conditions	40 conditions	30 conditions
10	99	98	96	96
20	98	96	92	91
30	95	93	88	86
40	90	89	82	80

Conclusions

1. To validate equipment condition assessment from time-series data, a classification problem involving second-order curves with similar fragments was formulated. A balanced dataset was generated from a mathematical model, a Kohonen self-organising map was trained, and the corresponding clusters were identified using an MLP.

2. Rules were extracted using a decompositional approach in IF–THEN form, with conditions defined in terms of the best-matching units (BMUs) of the input signal.

3. A method is proposed for extracting 1D signal classification rules from the output of a trained Kohonen SOM, enabling robust classification of noisy signals.

4. A software suite for populating a knowledge base with extracted rules was developed. Computational experiments on signal classification were conducted using both the SOM and a CLIPS-based rule-based system, with the number of antecedent conditions and the noise factor serving as simulation parameters. The results show that, with the maximum number of antecedent conditions, the classification accuracy of the rule-based system decreases by 1–3 percentage points depending on the noise factor.

Since most contemporary expert system tools employ a syntax similar to that of CLIPS, the proposed method is readily applicable to other knowledge bases and can be used to benchmark inference performance.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Mekkaoui, S.E ., Benabbou, L., Berrado A. (2023). Rule-Extraction Methods From Feedforward Neural Networks: A Systematic Literature Review. arXiv preprint. https://doi.org/10.48550/arXiv.2312.12878
2. Augasta, M. G., Kathirvalavakumar, T. (2012). Rule extraction from neural networks — A comparative study. PRIME-2012, 404-408. https://doi.org/10.1109/ICPRIME.2012.6208380
3. Dattachaudhuri, A., Biswas, S.K., Chakraborty, M. et al. (2021). A transparent rule-based expert system using neural network. Soft Comput 25, 7731–7744. https://doi.org/10.1007/s00500-020-05547-7
4. Augasta, M.G., Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. Neural Process Lett 35, 131–150. https://doi.org/10.1007/s11063-011-9207-8
5. Zilke, J.R., Loza Mencía, E., Janssen, F. (2016). DeepRED – Rule Extraction from Deep Neural Networks. In: Calders, T., Ceci, M., Malerba, D. (eds) Discovery Science. DS 2016. Lecture Notes in Computer Science. 9956. Springer, Cham. https://doi.org/10.1007/978-3-319-46307-0_29
6. Hongjun Lu, Setiono R., Liu H. (1996). Effective data mining using neural networks. IEEE Transactions on Knowledge and Data Engineering. vol. 8. no. 6, 957-961. https://doi.org/10.1109/69.553163

7. Iqbal R. (2011). Eclectic Extraction of Propositional Rules from Neural Networks. ICCIT 2011, 234-239. https://doi.org/10.1109/ICCITechn.2011.6164790
8. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. arXiv. https://doi.org/10.48550/arXiv.1805.10820
9. Hammer, B., Rechtien, A., Strickert, M., Villmann, T. (2002). Rule Extraction from Self-Organizing Networks. In: Dorronsoro, J.R. (eds) Artificial Neural Networks — ICANN 2002. Lecture Notes in Computer Science, 2415. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-46084-5_142
10. Malone, J., Mcgarry, K., Wermter, S., Bowerman, C. (2006). Data mining using rule extraction from Kohonen self-organising maps. Neural Computing and Applications. 15. 9-17. https://doi.org/10.1007/s00521-005-0002-1
11. Shapovalova S., Moskalenko Yu. (2019). Increasing the share of correct clustering of characteristic signal with random losses in self-organizing maps. Eastern-European Journal of Enterprise Technologies. 2. 4 (98). 13-21. DOI: https://doi.org/10.15587/1729-4061.2019.160670
12. Riley, G. (2022). Adventures in Rule-Based Programming: A CLIPS Tutorial. Kindle Edition, 200. https://clipsrules.net/airbp.html
13. Nayak R. (2009). Generating rules with predicates, terms and variables from the pruned neural networks. Neural Networks. 22, 4, 405–414. https://doi.org/10.1016/j.neunet.2009.02.001.

Received (Надійшла) 18.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Шаповалова Світлана Ігорівна – кандидат технічних наук, доцент, доцент кафедри автоматизації проектування енергетичних процесів і систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна;
Svitlana Shapovalova – Candidate of Technical Sciences, Associate Professor, Associate Professor of Department Digital Technologies in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine;
e-mail: lanashape@gmail.com; ORCID Author ID: http://orcid.org/0000-0002-3431-5639;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=35367847700.

Мажара Ольга Олександрівна – кандидат технічних наук, доцент кафедри автоматизації проектування енергетичних процесів і систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна;
Olga Mazhara – Candidate of Technical Sciences, Associate Professor of Department Digital Technologies in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine;
e-mail: yuramuv@gmail.com; ORCID Author ID: https://orcid.org/0000-0001-7887-6764;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57190385872.

Москаленко Юрій Володимирович – доктор філософії, старший викладач кафедри автоматизації проектування енергетичних процесів і систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна;
Yurii Moskalenko – PhD, Senior Lecturer of the Department Digital Technologies in Energy, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine;
e-mail: yuramuv@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-0824-9201;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57210336448.

Тітов Владислав Миколайович – аспірант, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна;
Vladyslav Titov – Post-Graduate Student, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine;
e-mail: 19Titov.vlad.Titov19@gmail.com; ORCID Author ID: https://orcid.org/0009-0000-5780-5596;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59729089700.

Екстракція правил із самоорганізувальної карти Кохонена для визначення стану обладнання за зашумленим діагностичним сигналом

С. І. Шаповалова, О. О. Мажара, Ю. В. Москаленко, В. М. Тітов

Анотація. Об'єктом дослідження статті є методи екстракції правил з нейронних мереж. Метою статті є представлення методу екстракції правил класифікації 1D діагностичних сигналів на основі навченої нейронної мережі Кохонена. Conclusions: для тестування задачі визначення стану обладнання за часовим рядом показників поставлено задачу класифікації кривих другого порядку за подібними фрагментами. За математичною моделлю створено збалансований датасет, проведено навчання Kohonen self-organising map, за допомогою MLP визначено відповідні кластери; правила екстракції визначалися за декомпозиційним підходом у форматі IF-THEN, де умови визначалися на основі одиниць найкращої відповідності вхідного сигналу мережі BMU (Best Matching Unit); запропоновано метод екстракції правил класифікації 1D сигналів на основі результатів класифікації навченої нейронної мережі Кохонена, який дозволяє класифікувати зашумлені сигнали; розроблено програмний комплекс для екстракції правил у базу знань. Проведено обчислювальні експерименти з класифікації сигналів SOM та Rule based system на основі CLIPS, параметрами моделювання яких була кількість умов в правилах бази знань та коефіцієнт шуму сигналу. Визначено за максимальної розмірності умовної частини в залежності від рівня шуму точність класифікації на Rule based system знижується від 1 до 3%.

Ключові слова: екстракція правил, нейронні мережі, Kohonen self-organising map, CLIPS, Rule based system, Best Matching Unit.

УДК 004.27:511.1 **doi: 10.26906/SUNZ.2026.2.189**

А. С. Янко, О. І. Крук

Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна

МЕТОД ПРИСКОРЕНОЇ РЕАЛІЗАЦІЇ МОДУЛЬНИХ ОПЕРАЦІЙ У СПЕЦІАЛІЗОВАНИХ КОМП'ЮТЕРНИХ ЗАСОБАХ НА ОСНОВІ РЕВЕРСИВНОГО КІЛЬЦЕВОГО ЗСУВУ

Анотація. У статті розроблено та досліджено метод прискореного виконання базових модульних операцій у системі залишкових класів (СЗК), що базується на використанні непозиційних кодових структур. Наукова новизна роботи полягає у впровадженні принципу реверсивного (зворотного) кільцевого зсуву, який дозволяє адаптивно обирати мінімальну траєкторію перетворення станів кільцевого зсувного регістра.

Доведено, що використання кільцевих регістрів забезпечує високий рівень апаратної надійності за рахунок спрощення логічного базису, а розроблений метод реверсивного зсуву додає до цього необхідну високу продуктивність, створюючи базу для побудови відмовостійких систем реального часу. На відміну від існуючих підходів, запропонований метод дозволяє скоротити тривалість циклу обробки інформації до 90% у випадках, коли значення операнда наближається до величини модуля. Доведено, що отриманий часовий ресурс може бути ефективно використаний для проведення процедур самодіагностики та повторних обчислень, що безпосередньо підвищує рівень відмовостійкості та достовірності функціонування СКЗОІ. Визначено математичні умови вибору напрямку зсуву та представлено функціональну схему операційного пристрою. Результати дослідження є актуальними для проектування високопродуктивних систем управління безпілотними платформами, де критично важливим є поєднання швидкодії з надійністю обчислювального процесу.

Ключові слова: система залишкових класів, непозиційна кодова структура, кільцевий регістр зсуву, реверсивний зсув, висока продуктивність, швидкодія, апаратна надійність, спеціалізовані комп'ютерні засоби обробки інформації.

Вступ

Сучасний стан розвитку спеціалізованих комп'ютерних засобів обробки інформації (СКЗОІ), зокрема для безпілотних платформ [1, 2] та систем критичної інфраструктури, вимагає безперервного пошуку нових архітектурних рішень для підвищення їхньої продуктивності та надійності. Широке впровадження СКЗОІ в усі сфери державного управління та військову справу практично усуває вплив людського фактора на процеси контролю та управління складними технічними об'єктами [3]. Це ставить функціонування систем у повну залежність від надійності, відмовостійкості та достовірності засобів обробки інформації.

Широке впровадження СКЗОІ в усі сфери практично усуває вплив людського фактора на процеси контролю та управління складними технічними об'єктами, що ставить процес управління в повну залежність від надійності та достовірності засобів обробки інформації, насамперед для БПЛА.

Дана обставина обумовлює необхідність розробки принципово нових методів підвищення відмовостійкості.

Сучасна тенденція розвитку СКЗОІ спрямована на збільшення довжини машинного слова (зокрема використання 64-розрядних сіток у системах реального часу), де недоліки позиційних систем виявляються особливо гостро [4, 5].

Одним із найбільш перспективних напрямків у цій галузі є використання непозиційних кодових структур системи залишкових класів (СЗК), що дозволяє реалізувати паралельну обробку даних на рівні окремих модулів без міжрозрядних зв'язків [6]. Застосування непозиційних кодових структур системи СЗК дозволяє суттєво підвищити ефективність засобів обробки цифрової інформації завдяки незалежності залишків за обраною системою основ.

Малорозрядність залишків у СЗК відкриває широкі можливості для вибору варіантів системотехнічних рішень при реалізації базових операцій [7]. Аналіз наукових джерел дозволяє виокремити чотири основні принципи реалізації арифметичних операцій у модульних кодах:

1. Суматорний принцип — на базі малорозрядних двійкових суматорів.

2. Табличний принцип — на основі використання ПЗП або матричних схем.

3. Прямий логічний принцип — заснований на описі операцій системами перемикальних функцій (реалізується на ПЛІС або програмованих логічних матрицях).

4. Принцип кільцевого зсуву — заснований на використанні кільцевих регістрів зсуву (КРЗ).

Суматорний варіант реалізації модульних операцій має низку критичних недоліків:

– складність синтезу та великий час перетворення інформації для значних розрядних сіток;

– неефективне використання двійкових елементів через надмірність щодо величини основ;

– низька достовірність обчислень за рахунок виникнення помилок у процесі міжрозрядних переносів.

Головною перевагою табличного принципу є повна відсутність переносів між розрядами операційного пристрою (ОП), що кардинально відрізняє їх від традиційних позиційних систем числення (ПСЧ) [8]. У ПСЧ наявність міжрозрядних зв'язків обмежує швидкодію, ускладнює апаратуру та знижує загальну надійність через можливість лавиноподібного поширення помилок переносу. Однак для досить великої розрядної сітки СКЗОІ (для великих

за величиною модулів СЗК) при табличному принципі різко збільшується кількість обладнання ОП.

У цьому контексті особливої актуальності набуває проміжний варіант реалізації арифметичних операцій у СЗК, заснований на застосуванні принципу кільцевого зсуву шляхом використання КРЗ. Використання матричних схем [9] у поєднанні з КРЗ дозволяє створювати пристрої з низьким енергоспоживанням та підвищеними показниками надійності.

Проте традиційні методи кільцевого зсуву потребують значної кількості тактів при великих значеннях модуля.

Розробка методу прискореної реалізації модульних операцій на основі реверсивного кільцевого зсуву дозволить мінімізувати кількість кроків обробки, забезпечуючи максимальну швидкість виконання операцій додавання та віднімання при збереженні високого рівня відмовостійкості системи.

Теоретичне обґрунтування та метод реверсивного кільцевого зсуву

В основі реалізації модульних операцій за допомогою кільцевих регістрів лежить відома теорема Келі, яка встановлює ізоморфізм між елементами скінченної абелевої групи та елементами групи підстановок.

Нехай G – циклічна група (порядок якої дорівнює m) з елементами $g_0, g_1, \ldots, g_{m-1}$, де g_0 – нейтральний елемент.

У цьому випадку матриця додавання порядку m задається таблицею Келі (табл. 1).

Таблиця 1 – **Таблиця Келі для довільного модуля *m***

β_i	a_i				
	0	1	2	…	m_i-1
0	0	1	2	…	m_i-1
1	1	2	3	…	0
2	2	3	4	…	1
…	…	…	…	…	…
m_i-1	m_i-1	0	1	…	m_i-2

Перший операнд a_i вказує на номер розряду КРЗ, що визначає результат модульної операції, а другий операнд β_i — на необхідну кількість тактів зсуву вмісту розрядів регістра.

Очевидно, що застосування принципу кільцевого зсуву суттєво підвищує достовірність виконання арифметичної операції модульного додавання завдяки усуненню помилок, які могли б виникнути як у процесі формування результату, так і внаслідок поширення міжрозрядних переносів, через їхню повну відсутність (аналогічно табличному принципу).

Слідством теореми Келі є висновок про те, що відображення групи G на групу всіх цілих чисел є гомоморфним. Це дозволяє розглядати циклічний зсув вмісту КРЗ як відображення деякої множини M у себе:

$$f : M \to M.$$

Для скінченних множин класи відображень (ін'єкція, бієкція, сюр'єкція) збігаються. Якщо оперувати не елементами множини, а їх номерами, то зсув вмісту КРЗ представляється як бієкція множини на себе у вигляді перестановки:

$$\pi = \begin{pmatrix} 0 & 1 & 2 & \ldots & m-1 \\ 1 & 2 & 3 & \ldots & 0 \end{pmatrix}.$$

Будь-яка модульна

$$A \pm B \pmod m$$

може бути інтерпретована як степінь z такого перетворення. При цьому перший операнд A вказує на номер початкового розряду КРЗ, а другий операнд B визначає необхідну кількість тактів зсуву вмісту регістрів.

Операція додавання в R множині СЗК, породжених ідеалом J, утворює кільце СЗК R/J. Його можна представити у вигляді z/m_i, де z – множина цілих чисел $0, \pm 1, \pm 2, \ldots$, а m_i – основа СЗК. Якщо основа m_i є простим числом, то таке кільце є полем $GF(m_i)$. Дана обставина обумовлює можливість реалізації арифметичної операції додавання в модульній арифметиці без міжрозрядних переносів шляхом використання принципу кільцевого зсуву за допомогою КРЗ.

На рис. 1 представлена вихідна інформаційна структура вмісту КРЗ для методу прямого зсуву.

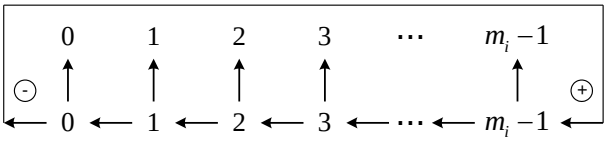


Рис. 1. Вихідна інформаційна структура вмісту КРЗ для методу прямого зсуву

Розглянемо приклад, де $m = 5$. У цьому випадку таблиця значень модульної $A + B \pmod m$ для кільця СЗК подається у вигляді числових даних, наприклад, першого рядка (стовпця) таблиці Келі (табл. 1), а на рис. 1 знаком «+» позначено додатний (проти годинникової стрілки) напрямок зсуву вмісту розрядів КРЗ.

Залежно від форми представлення вмісту розрядів КРЗ, виокремлюємо два методи технічної реалізації пристрою:

1. Метод двійкового позиційно-залишкового кодування: вміст КРЗ представляється у стандартному двійковому коді, що мінімізує кількість ліній зв'язку.

2. Метод унітарного позиційного кодування: кожному значенню залишку відповідає окремий розряд регістра, що забезпечує максимальну швидкодію та простоту дешифрації результату.

Метод реверсивного кільцевого зсуву для реалізації модульних операцій

У СЗК операнд *A* представляється набором залишків $\{a_i\}$ від розподілу його за набір простих (загалом взаємно попарно простих) чисел $\{m_i\}$, $i=\overline{1,n}$, що становить поле Галуа і є прямою сумою своїх підполів, тобто цей набір залишків можна ототожнити безпосередньо із сумою *n*-полів Галуа $\sum_{i=1}^{n} GF(m_i)$ [10].

Під ступенем перетворення *z* – показника оператору кільцевого зсуву (ПОКЗ) в даній статті розуміємо добуток φ^z, тобто $\underbrace{\varphi^0\ \varphi^0\ \dots\ \varphi^0}_{z}$, де $\varphi^0=\varepsilon$ – перетворення, яке всі елементи множини $\{a_i\}$ (вміст розрядів КРЗ) залишає на вихідному місці.

Оскільки зсув вмісту КРЗ можна здійснювати як у додатному, так і у від'ємному напрямках, то для довільних перестановок (або циклічних зсувів) поняття степеня *z* можна узагальнити на випадок цілих від'ємних чисел φ^{-z}, тобто

$$\underbrace{\varphi^{-1}\ \varphi^{-1}\ \dots\ \varphi^{-1}}_{z}=\left(\varphi^z\right)^{-1}.$$

Зазначимо, що для кожної перестановки $\varphi\in S(M)$ (*M* – скінченна множина) знайдеться таке натуральне *S*, що $\varphi^S=\varepsilon$.

Вочевидь, що у разі використання принципу кільцевого зсуву $\varphi^{m_i}=\varepsilon$, де m_i – порядок перестановки φ.

Ступені циклічної перестановки $(P_0\ P_1\ \dots\ P_{n-1})$ вихідного вмісту КРЗ (при двійковому поданні об'єктів)

$$P_{m_i}=\left[\log_2(m_i-1)+1\right]$$

можна визначити за формулами:

$$\begin{cases}\left(P_0\|P_1\|\dots\|P_{n-1}\right)^z=\left(P_z\|P_{z+1}\|\dots\|P_{n-1}\|P_0\|\dots\|P_{z-1}\right);\\ \left(P_0\|P_1\|\dots\|P_{n-1}\right)^{-z}=\left(P_{n-z}\|P_{n-z+1}\|\dots\|P_0\|P_1\|\dots\|P_{n-z-1}\right);\\ \left(P_0\|P_1\|\dots\|P_{m_i-1}\right)^{m_i}=\varepsilon.\end{cases} \quad (1)$$

Зазначені властивості структури поля $GF(m_i)$ вирахувань свідчать про те, що будь-яка модульна операція додавання або віднімання може бути інтерпретована як циклічна перестановка елементів множини *M*.

В апаратному аспекті це дозволяє реалізувати обчислювальний процес за допомогою кільцевих регістрів зсуву, де результат операції визначається кількістю тактів зсуву, що відповідає значенню операнда.

Однак традиційна реалізація передбачає рух вмісту регістра лише в одному (прямому) напрямку. Це призводить до того, що для великих значень операнда *A*, який наближається до значення модуля m_i, кількість тактів обробки стає $(t\to m_i-1)$, що суттєво обмежує швидкодію спеціалізованих комп'ютерних засобів обробки інформації.

Враховуючи симетрію циклічної групи та можливість узагальнення степеня перетворення на випадок від'ємних чисел, доцільно впровадити принцип реверсивного (зворотного) кільцевого зсуву.

Наукова ідея методу полягає в тому, що для будь-якого елемента поля існує два шляхи досягнення цільового стану регістра:

1. Прямий шлях — через виконання *k* кроків зсуву.

2. Реверсивний шлях — через виконання m_i-k кроків у протилежному напрямку.

Один з способів підвищення швидкодії виконання операції модульного додавання (віднімання) є метод реверсивного шляху, заснований на властивості наступної тотожності:

$$(a_i+\beta_i)=\left[a_i-(m_i-b_i)\right]\bmod m_i, \quad (2)$$

тобто зсув вмісту кільцевих регістрів зсуву можна здійснити як у позитивну, так і в негативну сторону (для рис. 2, де операції модульного додавання ПОКЗ подаються у вигляді:

$$z=\begin{cases}+\beta_i, \text{ если } 0\le\beta\le(m_i-1)/2;\\ -(m_i-\beta_i), \text{ если } (m_i+1)/2\le\beta_i\le m_i-1,\end{cases} \quad (3)$$

а для операції модульного віднімання ПОКЗ подається у вигляді:

$$z=\begin{cases}+\beta_i, \text{ если } 0\le\beta\le(m_i-1)/2;\\ +(m_i-\beta_i), \text{ если } (m_i+1)/2\le\beta_i\le m_i-1.\end{cases} \quad (4)$$

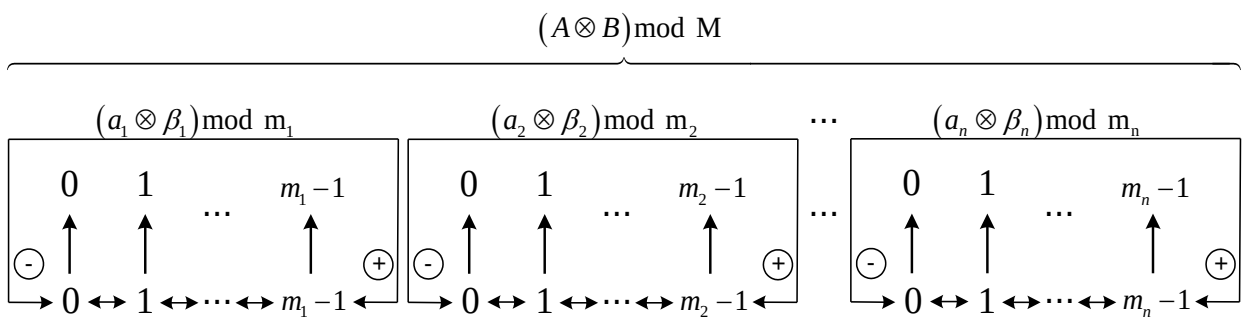


Рис. 2. Функціональна схема операційного пристрою СКЗОІ в СЗК

Для реалізації операцій додавання (віднімання) $(a_i \pm b_i) \bmod m_i$ вибір кількості тактів N та напрямку зсуву D здійснюється за наступним:

$$N = \begin{cases} A, & \text{якщо } A \le \frac{m}{2} \text{ (прямий зсув)} \\ m - A, & \text{якщо } A > \frac{m}{2} \text{ (реверсивний зсув)} \end{cases}. \quad (5)$$

Таким чином, для мінімізації часових витрат необхідно здійснювати адаптивний вибір напрямку зсуву на основі порівняння величини операнда з половиною значення модуля.

Це дозволяє обмежити максимальну кількість тактів обробки значенням $\frac{m}{2}$, що в середньому вдвічі прискорює виконання модульних операцій, а у випадках, коли операнд близький до модуля — забезпечує прискорення до 90%.

Висновки

У роботі запропоновано та науково обґрунтовано метод прискореної реалізації модульних операцій у спеціалізованих комп'ютерних засобах обробки інформації, що базується на використанні непозиційних кодових структур системи залишкових класів.

Встановлено, що використання реверсивного (зворотного) кільцевого зсуву дозволяє подолати головний недолік традиційних пристроїв на основі кільцевих регістрів зсуву – лінійну залежність часу виконання операції від величини операнда.

Завдяки адаптивному вибору найкоротшої траєкторії зсуву (прямої або реверсивної) вдалося обмежити максимальну кількість тактів обробки значенням $\frac{m}{2}$.

Доведено, що впровадження запропонованого методу дозволяє скоротити тривалість циклу виконання модульних операцій додавання та віднімання в середньому у два рази, а для операндів, величини яких наближаються до значення модуля — до 90%. Використання кільцевих регістрів забезпечує високий рівень апаратної надійності за рахунок спрощення логічного базису, а розроблений метод додає до цього необхідну високу продуктивність, створюючи базу для побудови відмовостійких систем реального часу.

Запропонована функціональна схема Запропонована функціональна схема операційного пристрою спеціалізованих комп'ютерних засобів обробки інформації забезпечує високу швидкість обробки даних при збереженні ключових переваг систем залишкових класів: відсутності міжрозрядних зв'язків та високої відмовостійкості обчислювального процесу.

Отриманий часовий ресурс може бути використаний для проведення процедур самодіагностики та контрольних обчислень, що є критично важливим для систем, які працюють у реальному часі.

Використання запропонованого підходу дозволяє суттєво підвищити достовірність визначення результату за рахунок повного усунення міжрозрядних переносів (аналогічно табличному принципу), але при значно менших апаратних витратах для великих модулів.

Практична значущість результатів полягає у можливості створення енергоефективних обчислювальних модулів для бортових систем безпілотних літальних апаратів (БПЛА) та автономних роботизованих платформ.

Оскільки метод базується на використанні простих кільцевих регістрів, він є ідеальним для реалізації на сучасній елементній базі (ПЛІС/FPGA), забезпечуючи баланс між мінімальними витратами апаратних ресурсів та максимальною продуктивністю.

Подальші дослідження будуть спрямовані на:

1. Розширення принципу реверсивного зсуву на операції модульного множення з використанням табличних методів та логарифмічних перетворень у СЗК.

2. Розробку алгоритмів автоматичного контролю та виправлення помилок у структурі кільцевих регістрів зсуву безпосередньо під час виконання зсуву, що дозволить ще більше підвищити надійність функціонування критичних інформаційних систем.

3. Інтеграцію розроблених пристроїв у склад складних ієрархічних систем управління для забезпечення національної цифрової резильєнтності.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Лактіонов, О., Педченко, Н., Янко, А., & Боряк, Б. (2024). Моделювання базової конструкції робототехнічної платформи. *Measuring and computing devices in technological processes*, (3), 95–99. https://doi.org/10.31891/2219-9365-2024-79-13
2. Yanko, A., Pedchenko, N., & Kruk, O. (2024). Enhancing the protection of automated ground robotic platforms in the conditions of radio electronic warfare. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, (6), 136–142. https://doi.org/10.33271/nvngu/2024-6/136
3. Попов, М., Комаровський, І., & Яценко, В. (2023). Інформаційні системи та технології в публічному управлінні. *Теоретичні та прикладні питання державотворення*, (30). https://doi.org/10.35432/tisb302023294963

4. Bo, L., Ruifeng, Z., Jiangang, L., Wenxin, G., & Yang, L. (2021). Control on abnormal data overflow of distribution network management platform. *Journal of Physics: Conference Series*, *1748*(3), 032064. https://doi.org/10.1088/1742-6596/1748/3/032064
5. Meakin, R. L. (2000). Adaptive spatial partitioning and refinement for overset structured grids. *Computer Methods in Applied Mechanics and Engineering*, *189*(4), 1077–1117. https://doi.org/10.1016/S0045-7825(99)00369-2
6. Mohan, P. V. A. (2016). *Residue number systems: Theory and applications*. Birkhäuser Basel; Springer International Publishing. https://doi.org/10.1007/978-3-319-41385-3
7. Salnikov, D., Karaman, D., & Krylova, V. (2023). Highly reconfigurable soft-CPU based peripheral modules design. *Advanced Information Systems*, *7*(2), 92–97. https://doi.org/10.20998/2522-9052.2023.2.13
8. Yanko, A. S., Krasnobayev, V. A., & Kovalchuk, D. M. (2022). Methods for tabular implementation of arithmetic operations of the residues of two numbers represented in the system of residual classes. *Radio Electronics, Computer Science, Control*, (4), 18–27. https://doi.org/10.15588/1607-3274-2022-4-2
9. Han, S., Ge, Y., Shi, Y., & Yi, R. (2026). A Fast Method for Estimating Generator Matrixes of BCH Codes. *Electronics*, *15*(1), 244. https://doi.org/10.3390/electronics15010244
10. Kavun, S. (2015). Conceptual fundamentals of a theory of mathematical interpretation. *International Journal of Computing Science and Mathematics*, *6*(2), 107–121. https://doi.org/10.1504/IJCSM.2015.069459

Received (Надійшла) 21.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Янко Аліна Сергіївна – кандидат технічних наук, доцент, доцент кафедри комп'ютерних та інформаційних технологій і систем Національного університету «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Alina Yanko – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Computer and Information Technologies and Systems of the National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine;
e-mail: al9_yanko@ukr.net; ORCID Author ID: https://orcid.org/0000-0003-2876-9316;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57094953000.

Крук Олег Іванович – аспірант кафедри автоматики, електроніки та телекомунікацій Національного університету «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Kruk Oleg – PhD student, Department of Automation, Electronics and Telecommunications of the National University «Yuri Kondratyuk Poltava Polytechnic», Poltava, Ukraine;
e-mail: olegkruk1975@gmail.com; ORCID Author ID: https://orcid.org/0009-0004-4241-2676;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59172145800&origin=recordpage.

Method of accelerated implementation of modular operations in specialized computer tools based on reverse cyclic shift

Alina Yanko, Kruk Oleg

Abstract. The article develops and scientifically substantiates a method for the accelerated execution of basic modular operations (addition and subtraction) in the Residue Number System (RNS), based on the use of non-positional code structures. The relevance of the study is driven by the need to increase the performance of specialized computer information processing tools (SCIPT) under conditions of intensive data bit-depth growth and increasing reliability requirements for control systems of unmanned aerial vehicles and robotic platforms. The scientific novelty of the work lies in the implementation of the reverse (backward) cyclic shift principle, which allows for the adaptive selection of the minimum transformation trajectory of the cyclic shift register (CSR) states. It has been established that the traditional implementation of modular operations based on the cyclic shift principle is characterized by a linear dependence of the computation time on the operand value, leading to significant delays at large modulus values. The proposed method negates this drawback by analyzing the input operand and comparing it with half the modulus value. If the operand value exceeds half the modulus, the system automatically initiates a shift in the opposite (reverse) direction, which significantly reduces the number of processing cycles required. It is mathematically proven that the maximum number of shift steps in such an approach does not exceed $\lfloor m/2 \rfloor$, where m is the base of the residue number system. This allows for reducing the duration of the computational cycle by an average of two times, and in boundary cases where the operand approaches the modulus value — by up to 90%. It is proven that the use of cyclic registers ensures a high level of hardware reliability due to the simplification of the logical basis and the complete absence of complex inter-bit carry chains inherent in positional number systems. Special attention in the article is paid to the issue of fault tolerance. It is shown that the time resource freed up as a result of the calculation acceleration can be effectively used for background self-diagnosis procedures of the hardware and for performing repeated (control) calculations. This directly increases the reliability level of the SCIPT functioning without involving additional hardware costs. A functional diagram of the operating device based on a CSR with a controlled shift direction has been developed, creating a basis for building fault-tolerant real-time systems. The results obtained are a fundamental basis for the further design of energy-efficient computing modules, where the combination of maximum performance with the reliability of information processing in a non-positional code basis is critically important.

Keywords: residue number system, non-positional code structure, cyclic shift register, reverse shift, high performance, operating speed, hardware reliability, specialized computer information processing tools.

UDC 614.8:355/359 **doi: 10.26906/SUNZ.2026.2.194**

Ramil Akhundov[1] , Elshan Hashimov[1,2], Aziz Talibov[2]

[1] National Defence University, Baku, Azerbaijan
[2] Azerbaijan Technical University, Baku, Azerbaijan

RISK MANAGEMENT AND MATRIX DECISION MAKING IN EMERGENCY SITUATIONS AT CRITICAL AND MILITARY FACILITIES

Abstract. This study develops a methodological framework for risk management and matrix-based decision making in emergency situations at critical and military facilities. The proposed approach is based on the assumption that the risk of an emergency should be assessed not only by the severity of the devastating event, but also by the criticality of the affected protected area. The framework links these two variables through a decision matrix that connects risk levels and zone classes to standardized operational response packages. The study shows that such an approach makes it possible to move from descriptive emergency assessment to structured management action. Its practical value lies in improving the consistency, speed, and traceability of decisions concerning resource allocation, access regulation, surveillance reinforcement, personnel protection, and continuity measures. The results indicate that the matrix model provides a more rigorous and operationally relevant basis for emergency risk management in high consequence facilities than ad hoc or undifferentiated response logic.

Keywords: risk management, matrix decision making, emergency situations, critical facilities, military facilities, physical protection, protected zone class, operational response package, decision support, emergency response planning.

Introduction

Emergency situations at critical and military facilities create a management environment in which risk is no longer a passive analytical category but an immediate determinant of operational decisions [1, 2]. In such conditions, the consequences of failure extend beyond local damage and may affect personnel safety, continuity of command, technological stability, communication, energy support, and the overall ability of the facility to maintain its core functions [3, 4]. This makes emergency risk management a particularly important scientific and practical problem for physical protection systems, because protective measures must be selected and adjusted under time pressure, incomplete information, and rapidly changing conditions.

The problem becomes more acute when emergency effects interact with the existing security architecture of the facility [5]. Fire, explosion, toxic release, infrastructure damage, or cascading technical failures may not only generate direct losses, but also weaken surveillance, restrict movement, complicate communication, reduce response speed, and create new pathways of unauthorized access or further escalation [6]. As a result, the operational picture becomes multidimensional: the facility must simultaneously contain the emergency, preserve critical functions, protect personnel, and prevent secondary security failures. In such circumstances, decision making cannot rely only on general emergency procedures or on isolated risk indicators. It requires a structured mechanism that translates risk assessment into concrete operational choices.

A further difficulty lies in the differentiated structure of critical and military facilities. Internal zones, command nodes, technical systems, restricted sectors, and support elements do not have equal significance [3, 7, 8]. The same emergency may require fundamentally different management actions depending on where it occurs, which functions are exposed, and how quickly disruption may spread. Therefore, effective decision making must account not only for the severity of the risk itself, but also for the class of the affected zone and the operational importance of the threatened element. Without such differentiation, management responses tend to become either excessively generalized or insufficiently prioritized.

Conventional decision making in emergency conditions often remains descriptive, fragmented, or strongly dependent on expert improvisation [9–11]. Although expert judgment remains essential, reliance on ad hoc interpretation alone reduces consistency, complicates coordination, and makes it more difficult to justify why one package of measures was chosen instead of another. For high consequence facilities, especially military ones, this is a serious limitation. Decisions must be not only fast, but also logically structured, reproducible, and compatible with the internal hierarchy of protected zones and operational priorities [12–14].

In this context, matrix based decision making offers an important methodological advantage. It allows risk interpretation to be linked to predefined management packages through a formal relation between risk level and zone class. Such an approach makes it possible to standardize escalation logic, coordinate actions across subsystems, and support more consistent allocation of personnel, technical resources, access restrictions, response measures, and continuity related interventions. The value of the matrix model lies not in replacing managerial judgment, but in organizing it within a framework suitable for emergency conditions, where the cost of delay or inconsistency may be exceptionally high.

The aim of this study is to develop a methodological framework for risk management and matrix based decision making in emergency situations at critical and military facilities by linking risk levels and protected zone classes to standardized operational response packages. The article focuses on the emergency specific features of risk, the logic of risk management under disrupted conditions, the structure of the decision matrix, and the practical role of

standardized response packages in supporting operationally relevant protective decisions.

Review of Recent Research and Publications

Recent studies on emergency management emphasize that risk assessment is one of the key foundations of operational decision making in high consequence environments [1-4, 15]. In the context of critical infrastructure, researchers have shown that emergency situations should be analyzed not only in terms of direct damage, but also through their impact on continuity of functions, vulnerability of internal systems, and degradation of protective capacity [16, 17]. This perspective is particularly relevant for critical and military facilities, where the effects of an emergency may propagate across command, communication, technical support, and access control structures [3, 4]. A substantial part of the literature addresses the protection of critical infrastructure under disruptive conditions [18, 19, 20]. These studies demonstrate that emergency events often produce compound effects, combining physical damage, functional disruption, and security degradation. For military and other special purpose facilities, the problem is even more complex because the protected object usually includes zones and components of unequal significance, while response measures must be implemented under strict temporal and organizational constraints [4, 21, 22]. As a result, the same level of hazard may require different management actions depending on the criticality of the affected area and the operational role of the exposed element [23, 24].

Another important line of research concerns decision support in emergency situations [25]. Existing works show that rapid response quality depends not only on the accuracy of risk identification, but also on the availability of structured decision mechanisms that can translate risk interpretation into practical action [18, 26]. In many cases, however, emergency decision making remains descriptive or highly dependent on expert improvisation, which reduces reproducibility and complicates coordination [1-4]. This creates a clear need for models that formalize the relation between assessed risk and predefined management responses [17]. Matrix based approaches have been applied in different areas of safety and security management because they allow analytical variables to be connected with operational categories [23]. Their main advantage lies in simplicity, interpretability, and direct applicability under time pressure. At the same time, the available literature rarely offers a sufficiently integrated matrix model for emergency situations at critical and military facilities, where risk level must be interpreted together with the class of the affected zone and the required response package [10]. Most existing models either classify risk without linking it to differentiated action sets, or propose general measures without formalizing their dependence on the internal structure of the protected facility [12, 13].

Thus, the literature confirms the importance of emergency risk assessment, differentiated protection of critical facilities, and decision support formalization, but it does not yet provide a sufficiently unified framework that links emergency risk levels, protected zone classes, and standardized operational response packages [14]. It is this gap that the present study addresses.

Task Statement

The management of emergency situations at critical and military facilities requires a decision-making framework capable of converting risk assessment into operationally relevant and differentiated protective actions. In such environments, emergency effects interact with the internal structure of the facility, the criticality of protected zones, the condition of personnel and technical systems, and the limited time available for intervention. As a result, the central scientific task is not only to identify and interpret risk, but also to formalize the logic by which different levels of risk should lead to different management responses depending on the significance of the affected zone.

The aim of this study is to develop a methodological framework for risk management and matrix-based decision making in emergency situations at critical and military facilities by linking risk levels and protected zone classes to standardized operational response packages.

To achieve this aim, the following tasks are addressed:

- to identify the emergency specific features of risk at critical and military facilities;
- to define the principal variables of risk management under disrupted operating conditions;
- to develop a matrix model that connects risk levels and zone classes with differentiated decision packages;
- to substantiate the practical value of the matrix model for emergency response planning and operational decision support.

The methodological basis of the study includes systems analysis, comparative analysis, structured interpretation of emergency risk conditions, and matrix modeling of management decisions. These methods are used to construct a framework in which emergency risk is treated not only as an analytical indicator, but also as a basis for standardized and operationally consistent protective action.

Basic Material and Results

Emergency situations at critical and military facilities fundamentally transform the logic of risk management because they affect not only the protected object itself, but also the operational conditions under which physical protection and emergency response must be implemented [1, 3, 22]. Under normal conditions, risk may be interpreted through relatively stable parameters of protection, access control, subsystem readiness, and functional zoning. Under emergency conditions, however, this stability is disrupted. Fire, explosion, toxic release, infrastructure failure, or combined emergency effects may simultaneously produce direct damage, reduce the effectiveness of surveillance and control, complicate communication, delay response, and create additional vulnerabilities inside the protected structure. As a result, risk is no longer only a measure of possible loss, but also a dynamic indicator of how quickly and how severely the internal protective logic of the facility may deteriorate. This makes risk management inseparable from the immediate organization of protective and operational decisions [10].

In critical and military facilities, this problem is intensified by the differentiated structure of the protected space. Not all zones of the facility have equal significance, and not all emergency effects produce the same operational

consequences [23]. A local emergency in a secondary service sector and an emergency affecting a command node, communication element, restricted technological segment, or protected access point may differ radically in their management implications even when their nominal physical scale appears similar. For this reason, emergency risk management cannot be based only on general hazard severity. It must combine the interpreted risk level with the class of the affected zone or element. This makes it possible to distinguish situations in which an identical risk indicator requires different operational responses because the threatened part of the facility has a different functional role, a different consequence profile, or a different value for continuity of operations.

Within this methodological logic, risk management in emergency conditions should be understood as the transition from risk identification to risk controlled intervention. The purpose of assessment is not limited to describing the emergency situation. Its real value lies in determining what package of actions is justified under the current combination of danger intensity, protected zone significance, response feasibility, and available resources. The key management variables therefore include the interpreted level of emergency risk, the class or criticality of the protected zone, the condition of personnel and technical systems, the temporal feasibility of response, and the continuity requirements of the facility. When these variables are not organized within a structured decision mechanism, management responses tend to become fragmented, delayed, or weakly justified. Under emergency pressure, this increases the probability of inconsistent actions across protection subsystems and weakens the ability of the facility to preserve its essential functions.

For this reason, the study proposes a matrix-based decision model in which the two primary coordinates of management are the level of risk and the class of the affected zone. The first coordinate reflects the severity and urgency of the emergency condition as interpreted through its probable consequences for personnel, operations, and protection capacity. The second reflects the protected significance of the zone in which the emergency occurs or to which its effects may propagate. Their combination creates a structured basis for management choice. Instead of leaving action selection to ad hoc interpretation, the matrix assigns each combination of risk level and zone class to a predefined package of operational and protective measures. In methodological terms, this transforms emergency risk from an abstract analytical variable into a direct management instrument.

The logic of the proposed decision model is summarized in Fig. 1, which presents the matrix relationship between risk level, protected zone class, and the corresponding package of management actions under emergency conditions.

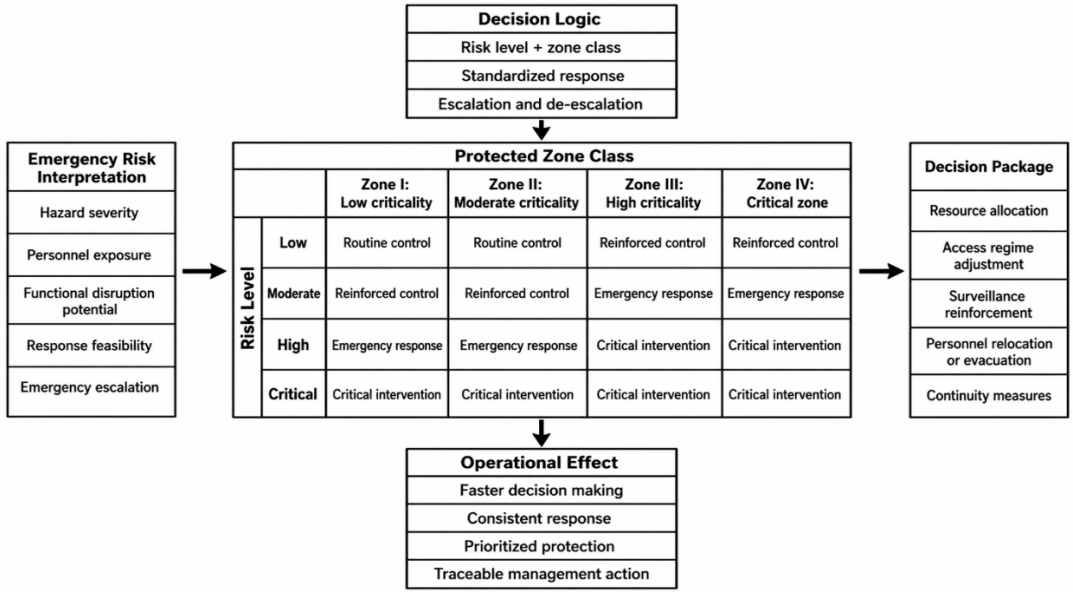


Fig. 1. Matrix framework for risk management and decision making in emergency situations at critical and military facilities

As shown in Fig. 1, the matrix transforms risk interpretation into a structured decision mechanism by linking each combination of risk level and zone class to a predefined set of protective and operational measures. This approach increases consistency of emergency response and creates a common decision language for different subsystems of facility protection and management.

The practical value of the matrix lies in the fact that it does not treat all emergency responses as uniform. Low risk conditions in a peripheral or low criticality zone may justify monitoring reinforcement, local restriction measures, and readiness adjustments without immediate large-scale intervention. The same low risk value in a highly critical zone may require enhanced surveillance, temporary access tightening, or advance deployment of response resources because the functional consequences of escalation are much greater. Medium risk conditions may justify localized intervention, controlled movement restrictions, selective relocation of personnel, or reinforced coordination measures depending on the class of the zone. High and critical risk conditions require progressively stronger decision packages, including broader access lockdown, emergency resource concentration, evacuation or redistribution of personnel, full strengthening of surveillance and response posture, and immediate continuity preserving actions for critical functions. Thus, the same formal risk level does not

automatically correspond to the same action. The operational content of the response is determined jointly by risk and zone criticality.

This logic also supports escalation and deescalation management. In emergency conditions, decisions must often be revised as the situation changes, resources are consumed or restored, and the spread of consequences becomes clearer. A matrix structure facilitates such revision because it allows the operator or decision maker to move from one cell of the matrix to another as the interpreted risk level or affected zone status changes. This ensures continuity of management logic and reduces the chance that response measures will be selected inconsistently from one stage of the emergency to another. For critical and military facilities, where coordinated action across access control, surveillance, technical support, response forces, and command structures is essential, this property is especially important. The matrix does not replace judgment, but it disciplines judgment by placing it within a predefined and operationally interpretable structure.

In order to operationalize this logic, the decision model must specify what measures belong to each matrix cell. These measures may be grouped into four broad classes: intervention intensity, resource allocation, regime regulation, and continuity preserving or alternative operational actions. Intervention intensity reflects how actively the emergency must be contained and how urgently the affected zone must be stabilized. Resource allocation determines the scale and direction of personnel, technical means, and response assets required under the interpreted conditions. Regime regulation covers access restrictions, movement control, internal routing, and special protective modes. Continuity preserving actions include relocation of critical functions, protection of key nodes, backup activation, and temporary operational restructuring. The composition of these packages should vary according to both the risk level and the criticality of the affected zone. To operationalize the proposed matrix, the main decision packages are grouped by risk level and protected zone class, as presented in Table 1.

Table 1 – **Decision packages by risk level and protected zone class**

Risk level	Zone I: Low criticality	Zone II: Moderate criticality	Zone III: High criticality	Zone IV: Critical zone
Low	Routine monitoring, local control, readiness check	Routine monitoring with reinforced control of access points and internal movement	Reinforced control, enhanced surveillance, precautionary standby of response personnel	Reinforced control, temporary access tightening, advance readiness of response and continuity staff
Moderate	Reinforced control, localized restriction measures, technical inspection	Reinforced control, selective movement restriction, partial deployment of response resources	Emergency response at zone level, strengthened surveillance, controlled personnel relocation	Emergency response, strict access regulation, immediate reinforcement of protection and continuity measures
High	Emergency response, localized resource concentration, temporary operational limitation	Emergency response, extended restriction regime, active deployment of response forces	Critical intervention, emergency lockdown of the affected sector, rapid protection of key elements	Critical intervention, full strengthening of regime measures, priority protection of critical functions, immediate continuity actions
Critical	Critical intervention, emergency containment, full local control measures	Critical intervention, emergency containment, evacuation or redistribution of exposed personnel	Critical intervention, full sector isolation, maximum concentration of response resources, continuity activation	Critical intervention, highest priority response, full lockdown, evacuation or protected relocation, immediate activation of continuity and backup mechanisms

Table 1 shows that the same risk level may require different operational responses depending on the criticality of the affected zone, while the same zone class may require different actions as risk escalates. This structure makes it possible to standardize decisions without eliminating the necessary differentiation of emergency management in high consequence facilities.

From a results perspective, the proposed matrix model yields several important outcomes. First, it creates a unified decision framework for emergency conditions at critical and military facilities. Instead of relying on fragmented interpretation, the facility obtains a structured relation between assessed risk and operational response. Second, the model preserves the internal hierarchy of protected zones, ensuring that management responses are not based only on general emergency severity, but also on functional significance. Third, the model improves reproducibility and auditability of emergency decisions because the choice of action package can be traced to a defined matrix position rather than to implicit or purely intuitive reasoning. Fourth, it supports more rational use of limited resources by indicating where reinforcement is most justified under the current combination of risk and zone criticality. Fifth, it provides a framework that is suitable for integration into emergency plans, physical protection procedures, and training scenarios. The model is especially relevant for military facilities because such facilities frequently combine high internal differentiation, strict regime requirements, and limited tolerance for interruption of command and support functions. Under these conditions, emergency management cannot be reduced to generic civil emergency logic. The consequences of poor prioritization are much greater, and the interaction between emergency effects and security degradation is more complex. A matrix-based approach helps address this problem by ensuring that emergency decisions remain anchored in both operational significance and structured risk interpretation. At the same time, the model is also applicable to other critical facilities where protected zones differ sharply in their strategic, technological, or organizational importance.

Thus, the basic material and results of the study show that risk management in emergency situations at critical and military facilities should be formalized as a structured decision process rather than treated as a sequence of isolated reactions. The key methodological result is the development of a matrix model that links interpreted risk levels and protected zone classes to standardized decision packages. The key practical result is that this model strengthens the speed, consistency, traceability, and operational relevance of emergency decisions while preserving the differentiated nature of high consequence protected facilities.

Conclusions

This study has developed a methodological framework for risk management and matrix-based decision making in emergency situations at critical and military facilities. The results show that emergency risk management in such facilities cannot be reduced to general hazard interpretation or to isolated response actions. Under emergency conditions, risk must be treated as a decision variable that reflects not only the severity of the disruptive event, but also the criticality of the affected zone and the operational consequences of delayed or inadequate intervention.

The study demonstrates that the combination of two analytical coordinates, namely risk level and protected zone class, provides a more adequate basis for emergency decision making than generalized response logic. This makes it possible to differentiate management actions according to the functional significance of the affected area and to avoid the methodological weakness of applying identical measures to zones with unequal operational importance. As a result, the proposed matrix model supports more precise prioritization of resources, access restrictions, surveillance reinforcement, personnel protection, and continuity related measures.

A further contribution of the study is the formalization of standardized decision packages for emergency conditions. The matrix framework transforms interpreted risk into a structured operational response mechanism and thereby improves the consistency, traceability, and reproducibility of protective decisions. Its practical value lies in the fact that it supports coordinated action under time pressure, incomplete information, and changing emergency conditions, which is especially important for military and other high consequence facilities. Overall, the proposed approach provides a more rigorous and operationally relevant basis for emergency risk management than descriptive or ad hoc decision making. Future research should focus on empirical validation of the matrix model, adaptation of response packages to specific facility types, and integration of the framework into emergency response plans, training procedures, and decision support systems.

Conflicts of interest. The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence. The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Akhundov, R. Modeling information processes and deriving measurable requirements in physical protection system design / R. Akhundov, E. Hashimov // Management Information System and Devices. - 2026. - No. 1(188). - P. 5-16. - DOI: https://doi.org/10.30837/0135-1710.2026.188.005
2. Akhundov, R. Quantitative categorization of facilities and modeling of potential adversaries / R. Akhundov, E. G. Hashimov // Grail of Science. - 2025. - No. 60. - P. 469-482. - DOI: https://doi.org/10.36074/grail-of-science.26.12.2025.049
3. Akhundov, R. Scenario oriented sufficiency criteria for physical protection systems provide a traceable path from threat classes to design requirements / R. Akhundov, E. G. Hashimov, I. Islamov // Grail of Science. - 2026. - No. 63. - DOI: https://doi.org/10.36074/grail-of-science.06.03.2026.074
4. Akhundov, R. Methodological limitations of normative design of physical protection systems for critical and military facilities in a dynamic threat environment / R. Akhundov, E. G. Hashimov, I. Islamov // Grail of Science. - 2026. - No. 62. - P. 873-889. DOI: https://doi.org/10.36074/grail-of-science.20.02.2026.096
5. Kostin, V., Borovsky A. Definition of basic violators for critically important objects using the information probability method and cluster analysis. CEUR Workshop Proc. - 2020. - Vol. 2667, p.343-347. Available at: https://ceur-ws.org/Vol-2667/paper75.pdf
6. Broder, J. F., Tucker E. Risk Analysis and the Security Survey / - B.-H., 2012. DOI: https://doi.org/10.1016/C2009-0-63855-1
7. Cozens, P. A review and current status of crime prevention through environmental design (CPTED) / P. Cozens, T. Love // Journal of Planning Literature. - 2015. - Vol. 30, No. 4. - P. 393-412. - DOI: https://doi.org/10.1177/0885412215595440
8. Cooper, W. W. Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, and DEA-Solver Software / W. W. Cooper, L. M. Seiford, K. Tone. - Boston : KAP, 2000. DOI: https://doi.org/10.1007/978-0-387-45283-8
9. Yang, J. A 2D-graph model-based heuristic approach to visual backtracking security vulnerabilities in physical protection systems / J. Yang, L. Huang, H. Ma [et al.] // International Journal of Critical Infrastructure Protection. - 2022. - Vol. 38. - P. 100554. - DOI: https://doi.org/10.1016/j.ijcip.2022.100554
10. Kaplan, S. On the quantitative definition of risk / S. Kaplan, B. J. Garrick // Risk Analysis. - 1981. - Vol. 1, No. 1. - P. 11-27. - DOI: https://doi.org/10.1111/j.1539-6924.1981.tb01350.x
11. Garcia, M. L. Design and Evaluation of Physical Protection Systems / M. L. Garcia. - 2nd ed. - Elsevier, 2008. - DOI: https://doi.org/10.1016/C2009-0-25612-1
12. Hashimov, E. Constrained optimization of an integral security indicator for adaptive management of hazardous facilities / E. Hashimov, R. G. Akhundov, A. M. Talibov, I. Islamov // Grail of Science. - 2026. - No. 62. - P. 1003-1014. - DOI: https://doi.org/10.36074/grail-of-science.20.02.2026.109
13. Akhundov, R. Conceptual models of multi-level physical protection systems for special-purpose and critical infrastructure facilities / R. Akhundov, E. G. Hashimov, I. Islamov // Grail of Science. - 2026. - No. 61. - P. 591-608. - DOI: https://doi.org/10.36074/grail-of-science.23.01.2026.066
14. Rehak, D. Strengthening resilience in the energy critical infrastructure: Methodological overview / D. Rehak, S. Slivkova, H. Janeckova [et al.] // Energies. - 2022. - Vol. 15, No. 14. - P. 5276. - DOI: https://doi.org/10.3390/en15145276

15. Reniers, G. L. L. Preparing for major terrorist attacks against chemical clusters: Intelligently planning protection measures with respect to domino effects / G. L. L. Reniers, A. Audenaert // Process Safety and Environmental Protection. - 2014. - Vol. 92, No. 6. - P. 583-589. - DOI: https://doi.org/10.1016/j.psep.2013.04.002
16. Hashimov, E. Research of the efficiency multiservice networks using MIMO technology / E. Hashimov [et al.] // Advanced Information Systems. - 2026. - Vol. 10, No. 1. - P. 66-71. - DOI: https://doi.org/10.20998/2522-9052.2026.1.08
17. Lovecek, T. Critical infrastructure protection systems effectiveness evaluation / T. Lovecek, J. Ristvej, L. Simak // Journal of Homeland Security and Emergency Management. - 2010. - Vol. 7, No. 1. - DOI: https://doi.org/10.2202/1547-7355.1613
18. Akhundov, R. Enhancing the physical protection of critical facilities through the integration of physical process models and machine learning / R. Akhundov, E. Hashimov // Grail of Science. - 2026. - No. 61. - P. 722-731. - DOI: https://doi.org/10.36074/grail-of-science.23.01.2026.083
19. Gündüz, M. Z. Akıllı şebekelerde iletişim altyapısı ve siber güvenlik / M. Z. Gündüz, R. Daş // Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi. - 2020. - Vol. 10, No. 2. - P. 970-984. - DOI: https://doi.org/10.21597/jist.655990
20. Mondal, S., Adak, B. and Mukhopadhyay, S. 11 Functional and smart textiles for military and defence applications. *Smart and Functional Textiles*, Berlin, Boston: De Gruyter, 2023, pp. 397-468. Doi: https://doi.org/10.1515/9783110759747-011
21. Shoop, B. Mobile detection assessment and response systems (MDARS): A force protection physical security operational success / B. Shoop [et al.] // Unmanned Systems Technology VIII. - 2006. - Vol. 6230. - P. 668-678. https://doi.org/10.1117/12.665939
22. Hashimov, E. Decision support for physical protection systems using route-level metrics and simulation-based evaluation / E. Hashimov, R. Akhundov, A. Talibov, I. Islamov // Grail of Science. - 2026. - No. 63. - P. 531-542. - DOI: https://doi.org/10.36074/grail-of-science.06.03.2026.059
23. Kampova, K. Quantitative approach to physical protection systems assessment of critical infrastructure elements: Use case in the Slovak Republic / K. Kampova, T. Lovecek, D. Řehák // International Journal of Critical Infrastructure Protection. - 2020. - Vol. 30. - P. 100376. - DOI: https://doi.org/10.1016/j.ijcip.2020.100376
24. Zou, B. Evaluation of vulnerable path: Using heuristic path-finding algorithm in physical protection system of nuclear power plant / B. Zou, M. Yang, Y. Zhang [et al.] // IJCIP. - 2018. - Vol. 23. - P. 90-99. - DOI: https://doi.org/10.1016/j.ijcip.2018.08.006
25. Řehák, D. Complex approach to assessing resilience of critical infrastructure elements / D. Řehák, P. Senovsky, M. Hromada, T. Lovecek // IJCIP. - 2019. - Vol. 25. - P. 125-138. - DOI: https://doi.org/10.1016/j.ijcip.2019.03.003
26. El Wely, I. C. Analysis of physical protection system effectiveness of nuclear power plants based on performance approach / I. C. El Wely, A. Chetaine // Annals of Nuclear Energy. - 2020. - Vol. 153. - P. 108051. - DOI: https://doi.org/10.1016/j.anucene.2020.107980

Received (Надійшла) 12.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Ахундов Раміль – доктор філософії з національної безпеки та військових наук, професор Національного університету оборони, Баку, Азербайджан;
Ramil Akhundov – PhD in National Security and Military Sciences, Professor at National Defense University, Baku, Azerbaijan;
e-mail: mr.axundov1@gmail.com; ORCID Author ID: http://orcid.org/0009-0001-8798-8044.

Гашимов Ельшан Гіяс – доктор національної безпеки та військових наук, професор, професор Азербайджанського технічного університету; професор Національного університету оборони, Баку, Азербайджан;
Elshan Hashimov – Doctor in National Security and Military Sciences, Professor at Azerbaijan Technical University, Professor of National Defense University, Baku, Azerbaijan;
e-mail: hasimovel@gmail.com; ORCID Author ID: http://orcid.org/0000-0001-8783-1277;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57195631270.

Талібов Азіз Машалла – доктор національної безпеки та військових наук, професор, професор Азербайджанського технічного університету, Баку, Азербайджан;
Aziz Talibov – Doctor in National Security and Military Sciences, Professor, Professor at Azerbaijan Technical University, Baku, Azerbaijan;
e-mail: talibovaziz65@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-0572-7212.

Управління ризиком і матричне прийняття рішень у надзвичайних ситуаціях на критично важливих і військових об’єктах

Раміль Ахундов, Елшан Гашімов, Азіз Талібов

Анотація. У статті розроблено методологічну основу управління ризиком і матричного прийняття рішень у надзвичайних ситуаціях на критично важливих і військових об’єктах. Запропонований підхід ґрунтується на положенні, що ризик у надзвичайних умовах має визначатися не лише за рівнем небезпеки події, а й з урахуванням критичності ураженої захищеної зони. У межах дослідження ці змінні поєднано в матричній моделі, яка пов’язує рівні ризику та класи зон зі стандартизованими пакетами управлінських і захисних заходів. Показано, що такий підхід забезпечує перехід від описового оцінювання надзвичайної ситуації до структурованого механізму прийняття рішень. Практична цінність моделі полягає в підвищенні послідовності, швидкості та обґрунтованості рішень щодо розподілу ресурсів, регулювання режиму доступу, посилення спостереження, захисту персоналу та забезпечення безперервності функціонування. Отримані результати свідчать, що матрична модель створює більш строгий і практично орієнтований підхід до управління ризиком у надзвичайних ситуаціях на об’єктах із високим рівнем наслідків.

Ключові слова: управління ризиком, матричне прийняття рішень, надзвичайні ситуації, критично важливі об’єкти, військові об’єкти, фізичний захист, клас захищеної зони, пакет оперативного реагування, підтримка прийняття рішень, планування реагування на надзвичайні ситуації.

УДК 621.384.2.002 **doi: 10.26906/SUNZ.2026.2.200**

Я. І. Бірук, Я. А. Підлісний

Київський національний університет будівництва і архітектури, Київ, Україна

ЕЛЕКТРОМАГНІТНА СУМІСНІСТЬ ЕЛЕКТРИЧНОГО ТА ЕЛЕКТРОННОГО ОБЛАДНАННЯ ЯК СКЛАДОВА ЕЛЕКТРОМАГНІТНОЇ БЕЗПЕКИ

Анотація. Досліджено зв'язок між станами електромагнітної сумісності електричного і електронного обладнання та електромагнітною безпекою людей. Визначено, що в результаті несиметричних та несинусоїдальних режимів силової електромережі генеруються електромагнітні поля гігієнічно значущих рівнів. Електричні струми гармонік промислової частоти генерують магнітні поля, які додаються до полів промислової частоти і погіршують електромагнітну обстановку. Показано, що навіть на великих відстанях від сторонніх джерел електричні та магнітні поля струмонесучих джерел можуть мати складну амплітудно-частотну характеристику. Амплітудні значення напруг цих частот значно перевищують нормативні. Найбільш вірогідним є те, що цей ефект обумовлений нелінійністю вольт-амперних характеристик кінцевих споживачів. Значний вплив на електромагнітну безпеку людей мають електричні струми витоку кабельними лініями і заземленими металевими конструкціями. Крім генерації магнітних полів через відсутність їх компенсації струмів протилежного напрямку, такі електричні струми є причиною електрохімічної корозії. Ліквідація цих струмів за рахунок розриву кола припиняє корозію, але без виявлення причин витоку може порушувати надійність системи електроживлення. Перевантаження нульових робочих провідників, крім аварійної ситуації, може привести до проблеми спрацювання пристроїв захисту. Електричні струми витоку і генеровані ними магнітні поля можуть давати індукційні наводки у комунікаційному обладнанні та спричиняти нештатну роботу комп'ютерної техніки. Будь-які негаразди з функціонуванням технічних приладів порушують психоемоційний стан користувачів, тому цей факт можна трактувати як непрямий (опосередкований) вплив електромагнітних полів на людей. Це особливо неприпустимо для персоналу з керування об'єктами критичної інфраструктури. Такий вплив може привести до помилок у роботі і прийняття хибних рішень. Зроблено висновок, що електромагнітна безпека і електромагнітна сумісність у широкому сенсі є двоєдиною задачею. Вирішення задач електромагнітної сумісності, крім заощадження електроенергії та збільшення термінів експлуатації обладнання, сприяє підвищенню електромагнітної і загальної безпеки людей у виробничих та побутових умовах.

Ключові слова: електромагнітна безпека, електромагнітна сумісність, гармоніки, інтергармоніки, струми витоку.

Вступ

У сучасних умовах забезпечення електромагнітної сумісності обладнання є важливим аспектом безперебійного функціонування комплексу технічних засобів на промислових підприємствах, навчальних та медичних закладах точно. У загальному випадку електромагнітна сумісність – здатність електронних пристроїв функціонувати належним чином у даному електромагнітному оточенні, не створюючи неприпустимих радіозавад іншим пристроям і водночас не піддаючись впливу небажаних електромагнітних перешкод від них.

Таким чином запобігаються збої у роботі критично важливого обладнання та гарантується стабільна робота у реальних умовах експлуатації, де існує багато джерел електромагнітних впливів. У галузі електротехніки електромагнітна сумісність забезпечується підтриманням симетричністю і синусоїдальністю навантажень у силовій електромережі. Але за певних умов, наприклад, за несиметричністю навантаження на окремі фази трифазної силової мережі у нульовому робочому провіднику протікають електричні струми частот гармоній промислової частоти, кратні трьом. Такі струми за певних умов генерують некомпенсовані магнітні поля гігієнічно значущих напруженостей.

Аналогічний ефект може проявлятися з боку обладнання з нелінійними вольт-амперними характеристиками навантаження. Ці магнітні поля шкідливо впливають та складно екранується через просторову розосередженість їх джерел. Порушення вимог з електромагнітної сумісності електричного та електронного обладнання може спричинити непрямий, несприятливий вплив на людей.

Нестабільна робота комп'ютерної техніки погіршує психоемоційний стан користувачів, що може бути причиною помилок та прийняття хибних рішень. Особливо це стосується персоналу з керування транспортними потоками, медичних працівників тощо.

Тому доцільно дослідити зв'язки показників електромагнітної сумісності обладнання та прямим і опосередкованим впливом електромагнітних полів на людей.

Огляд літературних джерел

Електромагнітна сумісність електричного та електронного обладнання регламентується низкою міжнародних стандартів. Щодо систем електроживлення промислової частоти, це стандарти серії IEC61000, наприклад, [1].

Сумісність обладнання високочастотних джерел електромагнітних полів регулюється нормативами [2, 3] та іншими.

Більшість міжнародних нормативних документів з електромагнітної сумісності набули чинності в Україні методом підтвердження [4].

Якщо нормативи з електромагнітної сумісності високочастотних джерел спрямовані на підтримання стабільності роботи електронного обладнання, то усі розробки і дослідження у напрямі електромагнітної сумісності обладнання напругами наднизьких частот спрямовані на вирішення задач енергозаощадження.

Дослідження [5, 6] присвячено стійкості бездротових мереж та електронних пристроїв. Зокрема, розглянуто надійність компонентів пристроїв в умовах наявності перехресних перешкод, та параметри модуляції і кодування, взаємний вплив сумісних мереж зв'язку.

Ці питання дуже актуальні для забезпечення стабільного бездротового зв'язку, але існують не передбачувані впливи на мережі. Крім того, доцільно розглянути опосередкований вплив нештатної роботи системи зв'язку на людей.

У роботі [7] визначено засади комп'ютерного моделювання електромагнітних процесів, пов'язаних з електромагнітною сумісністю обладнання. Це дозволяє оптимізувати параметри комп'ютерних мереж, але у реальних умовах експлуатації існують зовнішні впливи на системи, які неможливо передбачати і закласти у відповідні моделі.

Дослідження [8] стосується проблематики електромагнітної сумісності систем автоматизації управління будинками. Втім, єдиним засобом підвищення електромагнітної сумісності електричного та електронного обладнання пропонується екранування комунікаційних кабелів. Але висновок щодо захищеності кабелів металевими екранами за рахунок відбиття низькочастотних електромагнітних хвиль не є коректним. Низькочастотні поля у суцільних екранах індукують струми провідності, що може тільки погіршити ситуацію, якщо не внести додаткових заходів, наприклад, одночасне заземлення екрана великої довжини.

Проблематика електромагнітної сумісності мереж електроживлення розглянута у [9]. Зокрема, проаналізовано засоби зниження несиметричності і несинусоїдальності напруги, але ця робота стосується споживачів з великими електроспоживаннями й акцентується на зниженні втрат енергії.

У роботі [10] розглянуто узагальнені показники якості енергії, зокрема з одночасним врахуванням усіх критичних факторів. Дана робота стосується загальних мереж електропередачі, не розглядає можливих відхилень якості енергії у окремих об'єктах.

В останні роки дуже вагомим фактором впливу на стан електромагнітної сумісності низькочастотних мереж є джерела відновлювальної енергії [11]. Коливання реактивної потужності та використання інверторів на сонячних і вітрових електростанціях призводить до погіршення якості електрогенерації та ускладнює застосування фільтрокомпенсуючих пристроїв. Дослідження спрямоване на зниження втрат електроенергії, не розглядаючи можливість впливів магнітних полів електрострумів гармонік та інтергармонік промислової частоти на людей.

Мета роботи – визначення зв'язку між забезпеченням електромагнітної сумісності і електричного та електронного обладнання й забезпеченням електромагнітної безпеки людей з визначенням заходів її підвищення.

Викладення основного матеріалу

Уся електронна апаратура проходить тестування на предмет відповідності вимогам щодо електромагнітної сумісності.

Наприклад, в країнах Євросоюзу це випробування «EU-Type examination», які визначають відповідність обладнання вимогам Директиви 2014/30/EU (Додаток 1). Згідно вимогам Директиви, апаратура повинна виготовлятися з урахуванням останніх науково-технічних досягнень для забезпечення стабільної роботи – не створювати ненормативних радіозавад і мати відповідну стійкість до них.

Емісійні властивості засобів обчислювальної техніки регламентуються поширеними міжнародними стандартами TCO та MPRII. Описана у них процедура тестування і граничні рівні електромагнітних полів практично виключають збої у роботі техніки та негативний вплив на користувачів. Але випробування зразків технічних засобів виконуються у лабораторних умовах із унеможливлюванням стороннього впливу на тестоване обладнання. У реальних умовах експлуатації за підключення обладнання до силової та інформаційної мережі показники можуть змінюватися. Дослідження свідчать, що у деяких випадках електричне поле навколо сучасного відеомонітора сягає 2000–2500 нТл, а гранично допустимий рівень – 250 нТл. При цьому немає сумнівів у відповідності обладнання чинним вимогам з електромагнітної сумісності.

Причиною підвищення рівнів електромагнітних полів, в основному, є незадовільна якість електроенергії та сторонні магнітні поля наднизької частоти, при цьому останнє обумовлене або хибами у монтажі системи електроживлення, або електричними струмами витоку. Якість електроенергії для кінцевого споживача визначається симетричністю та синусоїдальністю напруги. Обидва ці параметри нормуються. Відмінності навантаження на окремі фази трифазної силової мережі не повинні перевищувати у розподільних щитах 30 %. Коефіцієнт симетричності напруги за нульовою послідовністю – не більше 4 %, за зворотною послідовністю – 2 %.

З точки зору електромагнітної безпеки, несприятливий вплив на людей мають магнітні поля, генеровані некомпенсованими електричними струмами у нульових робочих провідниках.

Аналогічний ефект може виникати через навантаження споживачів з нелінійними вольт-амперними характеристиками. Це може бути причиною генерації електричних полів зі складними амплітудно-частотними характеристиками (рис. 1).

Як видно з рис. 1, структура електричного поля має чітко виражені гармоніки промислової частоти. Вимірювання здійснювалися на великій відстані від сторонніх джерел, тобто цей склад поля обумовлений виключно гармонічним складом напруги у провіднику.

Склад й амплітуди гармонік відносно промислової частоти нормується (табл. 1).

Різні значення вищих гармонік обумовлені різними чинниками їх генерації, тому у енергетиці вони розділені на три групи. Але з точки зору електромагнітної безпеки критичними є їх наявність і амплітуди. У загальному випадку коефіцієнти спотворення для силової мережі напругою 380 В складають 8 % у нормальному режимі і 12 % у максимально допустимій.

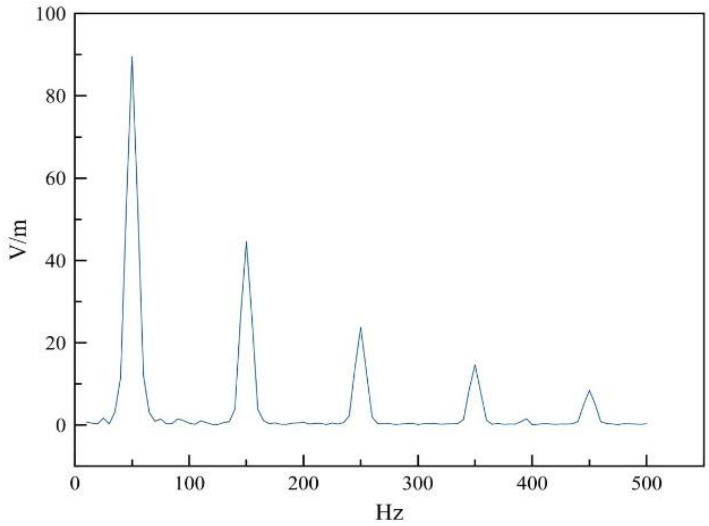


Рис. 1. Електричне поле лінійного струмонесучого провідника
за відсутності зовнішнього впливу на амплітудно-частотні характеристики

Таблиця 1 – **Допустимі значення коефіцієнта гармонік для нормального режиму електромережі напругою 380 В**

Номер гармоніки	2	3	4	5	6	7	8	9	10	11	12
Амплітуда, %	2	2,5	1	6	0,5	5	0,5	0,75	0,5	3,5	0,2

Дані рис. 1 свідчать, що амплітудні значення гармонік промислової частоти можуть значно перевищувати гранично допустимі значення. Крім значних втрат енергії та зниження ресурсу електротехнічного обладнання, електричні та магнітні поля гармонік підвищують несприятливий вплив на людей.

Наявність у силовій електромережі інтергармонік викликає спотворення напруги живлення. Внаслідок цього можливі низькочастотні коливання у рухомих пристроях (двигунах), що підвищує акустичний вплив на людей.

Порушується робота люмінесцентних систем освітлення і електронного обладнання, ймовірні перешкоди для телекомунікаційного обладнання, що є непрямим шкідливим впливом на людей.

Ще одним несприятливим чинником впливу на людей у разі порушення електромагнітної сумісності обладнання є флікери (коливання напруги різної періодичності). Це явище несприятливо впливає на зорову функцію. Провали напруги через спрацювання автоматики захисту негативно впливають на автоматизовані системи моніторингу та управління у сучасних виробничих та житлових комплексах. У таких умовах зростає значення коректної роботи фільтрокомпенсаційних пристроїв.

Слід враховувати, що в Україні системи компенсації реактивної потужності здебільшого мають великі терміни експлуатації та не виконують свої функції у умовах погіршення якості електроенергії. Розрахунки щодо ефективності зменшення не синусоїдальності напруги (параметри компенсуючого пристрою та фільтру) наведено в [9], але вони мають оціночний характер.

Це пояснюється тим, що у реальних умовах необхідно враховувати значення активних опорів навантаження.

Ще одним чинником несприятливого впливу на людей є магнітні поля електрострумів витоку. Порушення електромагнітної сумісності є причиною виникнення струмів витоку у нульових робочих провідниках силової електромережі. Значною мірою це обумовлене застосуванням у багатьох будівлях застарілих систем організації заземлення TN-C. Якщо з якихось причин нульовий робочий дріт має гальванічний зв’язок з нульовим захисним провідником або металевою конструкцією будівлі, частинка електроструму протікає PE-системою. При цьому виникає дисбаланс струмів і генерує у оточуючому просторі магнітне поле. Додаткове магнітне поле генерується електричними струмами, які протікають заземленими металевими конструкціями. Такі поля впливають прямо та опосередковано на людей (рис. 2).

Прямий вплив полягає у безпосередній дії полів на людей, а опосередкований – у порушенні штатної роботи комунікаційних ліній електронного обладнання, що впливає на психоемоційний стан людей. У користувачів засобів обчислювальної техніки це може бути причиною помилок у роботі, що особливо неприпустимо для персоналу з управління об’єктами критичної інфраструктури. На сьогодні ліквідація струмів витоку здійснюється в основному для ліквідації електрохімічної корозії. Найефективнішим методом є застосування діелектричних вставок для розриву електронного кола або заміна металевих комунікацій на полімерні. При цьому ліквідується причина електрохімічної корозії, але не усувається першопричина наявності струмів витоку.

Тому при цьому може суттєво зрости навантаження на нульові робочі провідники і підвищитися опір кола «фаза-нуль», що призводить до зниження струму короткого замикання. Тому можуть не спрацювати пристрої захисту від короткого замикання.

Рис. 2. Прямий та опосередкований вплив електромагнітних полів на людей

Для штатного функціонування систем електроживлення будівель в умовах нестабільності електропостачання та наявності великої кількості електроспоживачів з нелінійними вольт-амперними характеристиками доцільно виконати комплекс робіт з виявлення та усунення електричних струмів витоку.

Необхідно провести повний комплекс перевірок електроустановок будівель. Особливо увагу слід приділити технічному стану обладнання компенсації реактивної потужності. Обов'язковою є ревізія нульових захисних провідників, – перевірка правильності монтажу, перерізів тощо. Також обов'язковим є встановлення сучасних пристроїв захисного відключення електроживлення.

Наведене свідчить, що забезпечення електромагнітної сумісності електричного і електронного обладнання та електромагнітної безпеки людей є двоєдиною задачею. Їх вирішення окремо призводить до певних суперечностей, виходячи з різних цілей, та збільшує час та обсяги робіт. Комплексний підхід до реалізації розглянутих задач зменшує обсяги витрат та дозволить оптимізувати параметри силового електроживлення будівель і споруд різного призначення.

Висновки

1. Показано, що за несиметричного навантаження та нелінійних вольт-амперних характеристик обладнання у будівлях можуть генеруватися магнітні поля частот гармонік промислової частоти з напруженостями гігієнічно значущих рівнів. Такі поля за рахунок індукційних наводок спричиняють збої у роботі електронного обладнання, зокрема комп'ютерної техніки. Рівні гармонік перевищують нормативні значення навіть за межами будівель, що може бути наслідком нелінійності навантаження кінцевих споживачів.

2. Наявність у силовій мережі інтергармонік, провалів напруги та її пульсації (флікера) негативно впливає на людей через малопомітні пульсації штучного освітлення та спричиняє додатковий шум рухомих електричних пристроїв. Наявність електричних струмів витоку на нульові робочі провідники силової мережі та металеві заземлені конструкції підвищує електромагнітний фон у приміщеннях. Фізичне припинення протікання струмів витоку без ліквідації причин витоків може негативно впливати на надійність систем електроживлення.

3. Зовнішні електромагнітні поля і поля технічних засобів мають прямий вплив на людей, але вони можуть впливати на стабільність роботи технічних засобів, наприклад, комп'ютерної техніки, порушуючи психоемоційний стан користувачів. Це може вважати непрямим (опосередкованим) несприятливим впливом на людей. Такий вплив обумовлений порушеннями електромагнітної сумісності технічних засобів. Таким чином, задачі забезпечення електромагнітної сумісності і електромагнітної безпеки можна вважати двоєдиною задачею.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. IEC 61000-3-12:2011. Electromagnetic compatibility (EMC) - Part 3-12: Limits - Limits for harmonic currents produced by equipment connected to public low-voltage systems with input current >16 A and ≤ 75 A per phase. International Electrotechnical Commission. URL: https://cdn.standards.iteh.ai/samples/16966/85cef6616b0f4d06b7a86cfe3e60e808/IEC-61000-3-12-2011.pdf
2. ETSI EN 300 220-2 V2.4.1 (2012-01). Electromagnetic compatibility and Radio spectrum Matters (ERM); Short Range Devices (SRD); Radio equipment to be used in the 25 MHz to 1 000 MHz frequency range with power levels ranging up to 500 mW; Part 2: Harmonized EN covering essential requirements under article 3.2 of the R&TTE. Directive. European Telecommunications Standards Institute. 2012. 20 p. URL: https://www.etsi.org/deliver/etsi_en/300200_300299/30022002/02.04.01_60/en_30022002v020401p.pdf

3. ETSI EN 301 489-1 V2.2.1 (2019-03). ElectroMagnetic Compatibility (EMC) standard for radio equipment and services; Part 1: Common technical requirements; Harmonised Standard for ElectroMagnetic Compatibility. Directive. European Telecommunications Standards Institute, 2019. 36 p. URL: https://www.etsi.org/deliver/etsi_en/301400_301499/30148901/02.02.01_20/en_30148901v020201a.pdf
4. Про затвердження Технічного регламенту з електромагнітної сумісності обладнання. [Чинний від 2018-11-17]: затв. Постановою Кабінету Міністрів України від 16 грудня 2015 р. № 1077. Київ, 2018. URL: https://zakon.rada.gov.ua/go/1077-2015-%D0%BF
5. Лазебний, В. С., & Омельянець, О. О. (2024). Electromagnetic compatibility of wireless networks IEEE 802.11AC. Technologies and Engineering, (1), 67–76. https://doi.org/10.30857/2786-5371.2024.1.7
6. Leonov, S., & Borovik, O. (2020). Дослідження роботи електронних пристроїв з урахуванням електромагнітної сумісності. Вісник Національного технічного університету «ХПІ». Серія: Нові рішення у сучасних технологіях, 4(6), 71–76. https://doi.org/10.20998/2413-4295.2020.04.11
7. Пантелят , М., & Мясоєдов , П. (2025). Електромагнітна сумісність технічних об‘єктів і систем: попередній огляд деяких програмних засобів для комп’ютерного моделювання. Вісник НТУ «ХПІ». Серія: Проблеми удосконалювання електричних машин І апаратів. Теорія І практика, (1 (13), 21–25. https://doi.org/10.20998/2079-3944.2025.1.05
8. Кубай В.С., Зіньковський Ю.Ф. Електромагнітна сумісність системи автоматизації і управління будинками KNX. Міжнародна науково-технічна конференція «Радіотехнічні поля, сигнали, апарати та системи» : матеріали конференції, 16-22 листопада 2020 р., м. Київ, Україна / КПІ ім. Ігоря Сікорського, РТФ. – Київ : КПІ ім. Ігоря Сікорського, 2020. – С. 42-45. URL: https://ela.kpi.ua/items/e5f5a0fc-f3b2-457d-a0aa-f5c39dc643a2
9. Kuznetsov, V. G., Tugay, Y. I., Kuchanskiy, V. V., Lyhovyd, Y. G., & Melnichuk, V. A. (2018). THE RESONANT OVERVOLTAGE IN NON-SINUSOIDAL MODE OF MAIN ELECTRIC NETWORK. Electrical Engineering & Electromechanics, (2), 69–73. https://doi.org/10.20998/2074-272X.2018.2.12
10. В.Г. Кузнєцов. Узагальнений показник якості енергії в електричних мережаї і системах. Електроенергетичні системи та устаткування, 2011, №3, стор. 46-52. URL: https://nasplib.isofts.kiev.ua/server/api/core/bitstreams/51e165b0-9049-4a3f-9a74-59721817a12a/content
11. Papaika, Y., Lysenko, O., Bublikov, A., & Olishevskiy, I. (2021). Problems of electromagnetic compatibility of powerful energy associations during mass connection of renewable energy sources. Electrical Engineering and Power Engineering, (1), 34–45. https://doi.org/10.15588/1607-6761-2021-1-4

Received (Надійшла) 31.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Бірук Яна Ігорівна – доктор філософії, доцент, доцент кафедри фізики, Київський національний університет будівництва та архітектури, Київ, Україна;
Yana Biruk – PhD, Associate Professor, Associate Professor of the Department of Physics, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: vesna0999@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-3669-9744;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57225188391.

Підлісний Ярослав Анатолійович – аспірант кафедри технологій захисту навколишнього середовища та охорони праці, Київський національний університет будівництва і архітектури, Київ, Україна;
Yaroslav Pidlisnyi – PhD student at the Department of Environmental Protection Technologies and Labour Protection, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: Pidlisnyi97@gmail.com; ORCID ID: https://orcid.org/0009-0008-4906-3164

Electromagnetic compatibility of electrical and electronic equipment as a component of electromagnetic safety

Yana Biruk, Yaroslav Pidlisnyi

Abstract. The relationship between the electromagnetic compatibility of electrical and electronic equipment and the electromagnetic safety of people has been investigated. It has been determined that asymmetrical and non-sinusoidal modes of the power grid generate electromagnetic fields of hygienically significant levels. Industrial frequency harmonic electric currents generate magnetic fields that add to industrial frequency fields and worsen the electromagnetic environment. It has been shown that even at large distances from external sources, the electric and magnetic fields of current-carrying sources can have complex amplitude-frequency characteristics. The amplitude values of the voltages of these frequencies significantly exceed the normative ones. The most likely explanation for this effect is the non-linearity of the volt-ampere characteristics of end consumers. Electric leakage currents through cable lines and grounded metal structures have a significant impact on the electromagnetic safety of people. In addition to generating magnetic fields due to the absence of compensation currents in the opposite direction, such electric currents cause electrochemical corrosion. Eliminating these currents by breaking the circuit stops corrosion, but without identifying the causes of leakage, it can compromise the reliability of the power supply system. Overloading of zero working conductors, except in an emergency, can lead to problems with the operation of protective devices. Leakage electric currents and the magnetic fields they generate can cause induction interference in communication equipment and cause computer equipment to malfunction. Any malfunction of technical devices affects the psycho-emotional state of users, so this fact can be interpreted as an indirect (mediated) effect of electromagnetic fields on people. This is especially unacceptable for personnel managing critical infrastructure facilities. Such an impact can lead to errors in work and wrong decisions. It has been concluded that electromagnetic safety and electromagnetic compatibility in a broad sense are two sides of the same coin. Solving electromagnetic compatibility issues, in addition to saving electricity and increasing the service life of equipment, contributes to improving electromagnetic and general safety for people in industrial and domestic environments.

Keywords: electromagnetic safety, electromagnetic compatibility, harmonics, interharmonics, leakage currents.

УДК 621.39.004.021 **doi: 10.26906/SUNZ.2026.2.205**

Н. Б. Бурдейна, Д. Б. Осадчий

Київський національний університет будівництва і архітектури, Київ, Україна

ЗАСОБИ ПІДВИЩЕННЯ НАДІЙНОСТІ І ЕФЕКТИВНОСТІ СИСТЕМ ЕНЕРГОПОСТАЧАННЯ

Анотація. Визначено перелік і вміст організаційно-технічних заходів для підвищення стабільності електропостачання регіону. Дослідження виконано на прикладі найбільш поширених електричних підстанцій 150/35/10 кВ. Визначено необхідність заміни штатного обладнання та обладнання, ушкодженого внаслідок терористичних атак. Розроблено однолінійні принципові схеми підстанцій для забезпечення резервного електроживлення споживачів. Частина електрообладнання, наприклад, трансформатори струму підлягають заміні для підвищення максимального навантаження та керування об'єктом у режимі реального часу. На електричних підстанціях напругами 150–10 кВ відсутній постійний персонал. Тому, для оперативного регулювання навантаження у випадках дисбалансів потоків енергії було впроваджено телекомунікаційне обладнання. Таке обладнання дозволяє оперативно вмикати та вимикати навантаження дистанційно. Це мінімізує ризики виходу з ладу енергетичного обладнання при виникненні нештатних ситуацій. Як частина комплексу заходів з відновлення та регулювання електропостачання запропоновано ремонт трансформаторів, на яких не виникли пожежі. Після фізичного усунення протічок трансформаторна олива піддається регенерації з використанням оригінальної дегазаційної установки. Такий процес дозволяє повторно використовувати трансформаторну оливу. Установка здійснює дегазацію, осушування й фільтрацію оливи. Процес безпечний для довкілля та відновлює діелектричні властивості, охолоджувальну здатність і текучість оливи. При цьому підвищується стабільність до окислення, викликаючи контакт з повітрям у процесі регенерації оливи. Наведені заходи певним чином підвищують стійкість системи електропостачання та підвищують її стабільність у випадках вимушених відключень для балансування об'єднаної енергетичної системи України.

Ключові слова: електропостачання, стабільність, трансформаторна підстанція, резервування.

Вступ

Стабільність енергопостачання промислових підприємств електричного транспорту житлового фонду є вирішальним аспектом життєдіяльності і функціонування всіх складових економіки держави. Енергосистема складається з багатьох ланок – підприємства генерації електроенергії, передачі і перетворення та кінцевих споживачів.

Загалом, уся електрична мережа формує і впливає на якість електроенергії та стан ефективності (відсоток нераціональних втрат). В Україні значна частина енергосистеми зруйнована або ушкоджена внаслідок терористичних атак. При цьому розосереджені ланки енергосистеми – лінії електропередачі, трансформаторні підстанції є найбільш уразливими, які важко захистити фізично.

Проблема загострюється тим, що значна частина обладнання, наприклад, силові трансформатори, мають терміни експлуатації більше 50 років. На них проблематичним є перемикання з метою підвищення напруги у разі потреби.

Вони схильні до перегріву і реагування на зниження якості електроенергії.

У таких умовах потрібні рішення щодо дублювання джерел електроживлення і передачі електроенергії.

Необхідно розробити і впровадити телекомунікаційні системи керування подачею енергії у режимі реального часу, що дозволить максимально уникати дисбалансу мережі і її роботи у нештатному режимі.

Важливим аспектом забезпечення стабільності мережі є максимально швидке відновлення працездатності трансформаторів, які не зазнали критичних пошкоджень. Наведене обумовлює актуальність роботи.

Стан питання

Більшість досліджень і розробок у галузі електропостачання стосуються підвищення якості електроенергії. В основному це обумовлене поширенням відновлювальних джерел енергії [1]. Такі джерела не мають гарантованої генерації і можуть спотворювати синусоїдальність напруги. Уникнення таких явищ для електромереж загалом та на окремих підприємствах досягається за рахунок гнучкого керування електропостачання [2, 3]. Загалом оцінка енергоефективності системи енергопостачання здійснюється з допомогою спеціально визначених індексів надійності [4]. В основному заходи з підвищення надійності електроживлення і якості електрогенерації обмежуються впровадженням системи придушення гармонік та інтергармонік напруги промислової частоти [5, 6]. Традиційні засоби забезпечення стабільності роботи енергетичної системи, описані у роботах [7, 8] на сьогодні недостатні. Загалом стабільність енергопостачання визначається національним стандартом України [9], але у нинішніх умовах забезпечити необхідні показники практично неможливо Актуальною задачею є задовільнення мінімальних потреб у електроенергії з урахуванням можливих навмисних втручань у роботу енергосистем та аварійних ситуацій, викликаних перенавантаженнями, погодними умовами тощо.

Необхідно передбачити дублювання електропостачання, систем аварійного відключення для унеможливлення дисбалансів енергетичних потоків та можливість швидкого відновлення працездатності критично важливого обладнання.

Мета роботи – визначення засобів забезпечення надійності енергопостачання в умовах можливих навмисних втручань та аварійних ситуацій.

Викладення основного матеріалу

Найбільш поширеними електропідстанціями є підстанції класу 150/35/10 кВ. Вони забезпечують електроживлення більшості споживачів, але зазнають найбільших ушкоджень у наслідок аварій та терористичних атак. Тому доцільно розглянути засоби забезпечення стабільності електропостачання та енергетичної безпеки саме на прикладі таких підстанцій в одному з регіонів України.

З міркувань безпеки конкретні назви об'єктів не наводяться.

Схематично діючу систему енергопостачання міста і прилеглих територій та її приєднання до об'єднаної енергетичної системи України наведено на рис. 1.

Така система, у разі потреби, не забезпечує дублювання енергетичних потоків. Крім того, значна частина трансформаторів та відкритих розподільчих пристроїв зазнала ушкоджень (рис. 2).

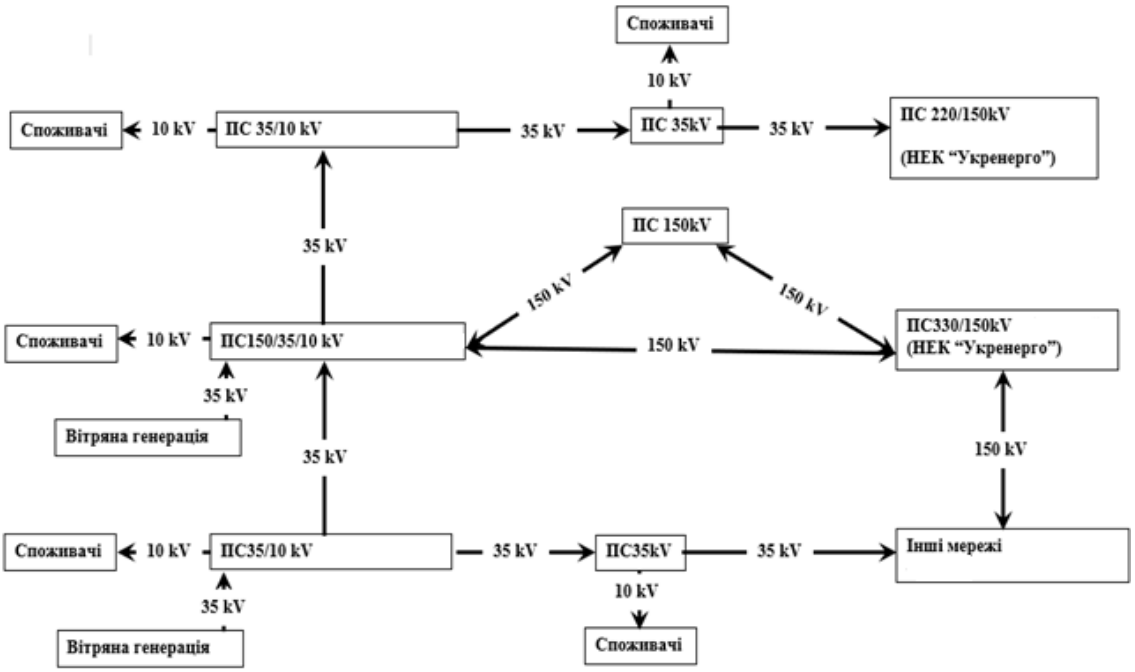


Рис. 1. Схематичне представлення діючої системи енергопостачання міста та прилеглих територій

Рис. 2. Пошкоджений трансформатор та відкритий розподільчий пристрій підстанції 150/35/10 кВ

Враховуючи таку ситуацію, доцільним є вирішення двох задач:

- Відновлення ушкоджених трансформаторів та відкритих розподільчих пристроїв;
- Організація резервної схеми.

Одночасна заміна трансформаторів й частини іншого обладнання обумовлена необхідністю підвищення надійності роботи й робочих потужностей. Фрагмент однолінійної принципової схеми підстанції 150/35/10 кВ наведено на рис. 3.

Особливо важливою є заміна трансформаторів струму.

По-перше вони розраховані на більші навантаження, по-друге – дозволяють у режимі реального часу контролювати навантаження й вчасно реагувати на несприятливі тенденції.

Для швидкого переведення навантажень проводиться реконструкція підстанції 35/10 кВ, а саме – розподільчого пристрою 10 кВ та окремих комунікаційних апаратів 35 кВ на підстанції 150/25/10 кВ (рис. 4).

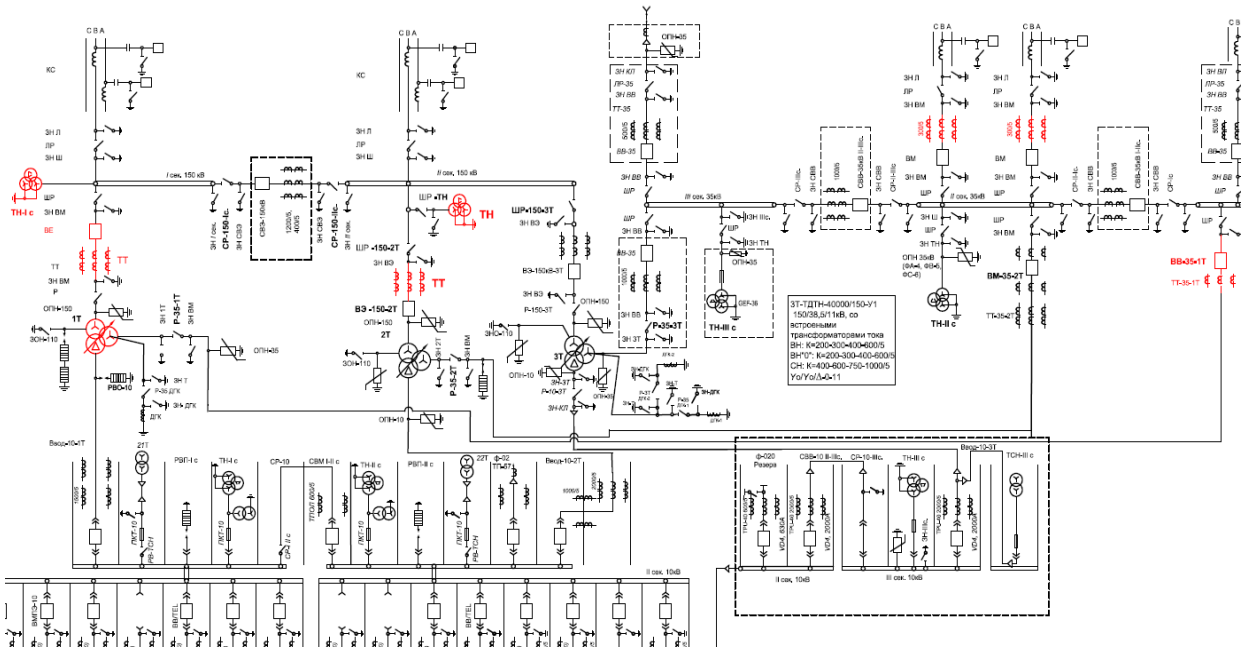


Рис. 3. Фрагмент однорідної схеми підстанції 150/30/10 кВ.
Червоним кольором позначено замінене обладнання

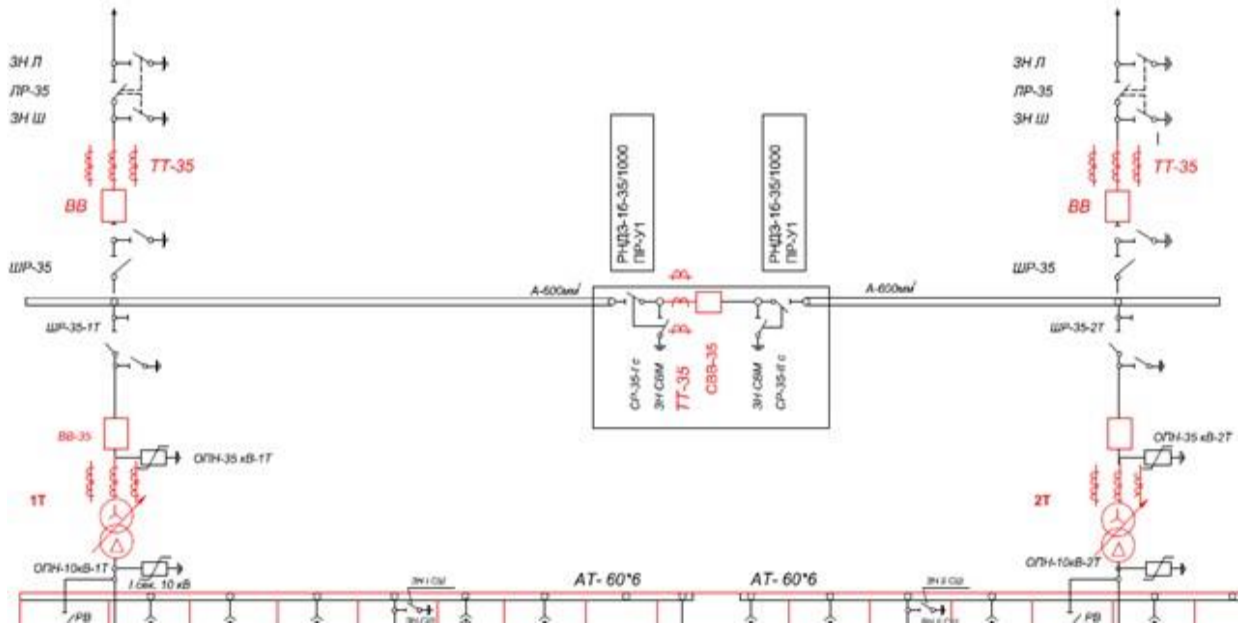


Рис. 4. Фрагмент однолінійної принципової схеми підстанції 35/10 кВ.
Червоним кольором виділене змінене обладнання

Такі підключення надають змогу резервувати електроживлення окремих населених пунктів та територій.

Важливим аспектом забезпечення стабільного електропостачання є підтримання навантаження у штатному режимі. Це вимагає, у разі потреби, вимикання окремих гілок силової мережі. На електропідстанціях 150/35,/10 кВт у штатному режимі роботи відсутній персонал. Тому неможливе оперативне відключення у разі потреби.

З метою оперативного реагування на нештатні ситуації було впроваджено телекомунікаційне обладнання дистанційного керування роботою підстанції (рис. 5).

Рис. 5. Телекомунікаційна шафа дистанційного керування роботою підстанцій

Застосування такого обладнання дозволяє у режимі реального часу здійснювати підтримання електромережі у збалансованому стані. Для надійної роботи дистанційного автоматичного управління режимами роботи мережі шляхом перемикання комутаційних апаратів було замінено застарілі оливні перемикачі, не здатні працювати в системі дистанційного управління на сучасні вакуумні та елегазові. Ще однією актуальною задачею є відновлення трансформаторів. Частина з них зазнала пошкоджень без виникнення пожежі (рис. 6). У цьому випадку крім фізичного ремонту – ліквідації витікання оливи, необхідне її відновлення. Це здійснюється з використанням оригінальної дегазаційної установки (рис. 7).

Рис. 6. Трансформатори з ушкодженими контурами охолодження

Рис. 7. Дегазаційна установка для відновлення трансформаторної оливи

Застосування такої установки дозволяє здійснити дегазацію, висушування та фільтрацію трансформаторної оливи. Процес безпечний для довкілля і повністю відновлює діелектричні властивості трансформаторної оливи, охолоджувальну здатність та текучість, підвищує стабільність до окислення, виключаючи контакт з повітря у процесі регенерації. Такий процес дозволяє повторно використовувати відпрацьовану оливу, економлячи кошти та природні ресурси.

Задачі підвищення надійності електричних мереж, особливо під час військового стану, ще далекі від остаточного вирішення. Але запропоновані заходи дозволяють не тільки відновлювати ушкоджене обладнання і електромережі, а й впроваджувати сучасні технологічні рішення, що підвищують надійність, безпеку та ефективність енергопостачання.

Висновки

1. На прикладні найбільш поширених електричних підстанцій і мереж електроживлення проаналізовано й сформульовано основні задачі підвищення стабільності електрозабезпечення регіону. Показана необхідність часткової модернізації обладнання та впровадження дублювання енергетичних потоків.

2. Розроблено принципову схему модернізації енергопостачання й виявлено перелік обладнання, яке підлягає заміні – трансформатори струму, перемикачі напруги. Заміна оливних перемикачів на елегазові й вакуумні дозволяє застосовувати дистанційне керування об’єктом для оперативного керування енергомережею. Розроблено й впроваджено телекомунікаційне обладнання, яке дозволяє у режимі реального часу контролювати енергетичні потоки та своєчасно вимикати та вмикати обладнання підстанцій, на яких постійна присутність персоналу не передбачено.

3. Частину обладнання електричних підстанцій можливо відновити. У першу чергу це трансформатори, які зазнали пошкоджень без виникнення пожежі. Після ліквідації витікання трансформаторної оливи вона регенерувалася за допомогою оригінальної дегазаційної установки. У процесі регенерації відбувається дегазація, осушування та фільтрація оливи. У результаті відновлюються охолоджувальна властивість, текучість та діелектрична проникність оливи. Підвищується стійкість до окислення та стабільність оливи, виключаючи контакт з повітрям у процесі регенерації.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Papaika, Y., Lysenko, O., Bublikov, A., & Olishevskiy, I. (2021). Problems of electromagnetic compatibility of powerful energy associations during mass connection of renewable energy sources. Electrical Engineering and Power Engineering, (1), 34–45. https://doi.org/10.15588/1607-6761-2021-1-4
2. Інтелектуальні електричні мережі: елементи та режими: за заг. ред. акад. НАН України О.В. Кириленка/ Інститут електродинаміки НАН України. – К.: Ін-т електродинаміки НАН України, 2016. – 400 с, URL: https://www.old.nas.gov.ua/UA/Book/Pages/default.aspx?BookID=0000009008
3. І.В. Жежеленко, Ю.А. Папаїка, О.Г. Лисенко, К.С. Родна. Застосування індивідуальних графіків вищих гармонік в задачах електромагнітної сумісності та енергоефективності гірничих підприємств. Гірнича електромеханіка та автоматика. – 2019. – No 101. – С. 3-7. URL: https://gea.nmu.org.ua/ua/ntz/archive/101.pdf
4. І.В. Жежеленко, Ю.А. Папаїка, О.Г. Лисенко . Оцінка енергетичної ефективності систем електропостачання за допомогою індексу надійності. Гірнича електромеханіка та автоматика. – 2018. – No 100. – С. 25-30, URL: https://gea.nmu.org.ua/ua/ntz/archive/100.pdf
5. Ghorbani, M. J., & Mokhtari, H. (2015). Impact of Harmonics on Power Quality and Losses in Power Distribution Systems. International Journal of Electrical and Computer Engineering (IJECE), 5(1), 166–174. DOI: http://doi.org/10.11591/ijece.v5i1.pp166-174
6. Zhezhelenko, I. V., & Nesterovych, V. V. (2017). Оцінка втрат електричної енергії, що викликані зниженням її якості. Вісник Приазовського ДТУ. Серія: Технічні науки, (34), 119–126. https://doi.org/10.31498/2225-6733.34.2017.105674
7. Г.Г. Півняк, І.В. Жежеленко, Ю.А. Папаїка. Енергетична ефективність систем електропостачання.– Д.: НТУ «ДП», 2018. – 149 с. URL: https://ir.nmu.org.ua/entities/publication/89696801-38ee-481b-8756-f1e3b4f0d9a7
8. Papaika, Y., Lysenko, O., Koshelenko, Y. and Olishevskyi,I.. 2021.Mathematical modeling of power supply reliability at low voltage quality. Naukovyi Visnyk NHUn, (2), pp.97-103. https://doi.org/10.33271/nvngu/2021-2/097
9. ДСТУ EN 50160:2023 Характеристики напруги електропостачання в електричних мережах загальної призначеності (EN 50160:2022, IDT), URL: https://online.budstandart.com/ua/catalog/doc-page.html?id_doc=106226

Received (Надійшла) 04.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Бурдейна Наталія Борисівна – доктор технічних наук, професор, професор кафедри фізики, Київський національний університет будівництва та архітектури, Київ, Україна;
Nataliia Burdeina – Doctor of Technical Sciences, Professor, Professor of the Department of Physics, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: burdeina.nb@knuba.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-2812-1387;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57220047954 .

Осадчий Дмитро Борисович – аспірант кафедри технологій захисту навколишнього середовища та охорони праці, Київський національний університет будівництва і архітектури, Київ, Україна;
Dmytro Osadchyi – PhD student at the Department of Environmental Protection Technologies and Labour Protection, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: osadchyi_db-2023@knuba.edu.ua; ORCID Author ID: https://orcid.org/0009-0002-9977-8738;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59413002900

Means of improving the reliability and efficiency of power supply systems

Nataliia Burdeina, Dmytro Osadchyi

Abstract. A list and content of organisational and technical measures to improve the stability of electricity supply in the region have been determined. The study was carried out using the example of the most common 150/35/10 kV electrical substations. The need to replace standard equipment and equipment damaged as a result of terrorist attacks has been determined. Single-line diagrams of substations have been developed to ensure backup power supply to consumers. Some electrical equipment, such as current transformers, needs to be replaced to increase the maximum load and control the facility in real time. There is no permanent staff at electrical substations with voltages of 150–10 kV. Therefore, telecommunications equipment has been installed for operational load control in cases of energy flow imbalances. Such equipment allows for remote switching of loads. This minimises the risk of power equipment failure in the event of an emergency. As part of a set of measures to restore and regulate the power supply, it is proposed to repair transformers that have not caught fire. After the physical elimination of leaks, the transformer oil is regenerated using the original degassing unit. This process allows the transformer oil to be reused. The unit degasses, dries and filters the oil. The process is environmentally safe and restores the dielectric properties, cooling capacity and fluidity of the oil. At the same time, it increases resistance to oxidation caused by contact with air during the oil regeneration process. These measures increase the stability of the power supply system and improve its stability in cases of forced shutdowns to balance the integrated energy system of Ukraine.

Keywords: power supply, stability, transformer substation, backup.

УДК 628.5/8:614.7 **doi: 10.26906/SUNZ.2026.2.210**

В. А. Глива, Я. О. Галонько

Київський національний університет будівництва і архітектури, Київ, Україна

ТЕОРЕТИЧНІ ТА ЕКСПЕРИМЕНТАЛЬНІ ПІДХОДИ ДО ЗАБЕЗПЕЧЕННЯ МІКРОКЛІМАТИЧНИХ ПАРАМЕТРІВ У ПРИМІЩЕННЯХ СПЕЦІАЛЬНОГО ПРИЗНАЧЕННЯ

Анотація. На основі аналізу чинної нормативної бази та наукових досліджень встановлено що концептуальні підходи до забезпечення мікрокліматичних параметрів у приміщеннях спеціального призначення потребують доопрацювання. До таких приміщень належать приміщення диспетчеризації та управління й сховища цивільного захисту різних типів і конструкцій. Приміщення диспетчеризації містить велику кількість обладнання моніторингу і керування, а сховища можуть мати велике скупчення людей, що впливає на параметри середовища. У таких умовах генерується велика кількість аерозолей. Запропоновано простий у використанні розрахунок основних показників динаміки аерозолей у залежності від швидкості руху повітря, обумовленого вентиляцією. Визначено коефіцієнт релаксації частинок – час, за який вони набувають швидкості повітряного потоку. Розраховано стаціонарний показник осідання частинок. Надано значення цих показників для спектра розмірів аерозольних частинок. Показано, що за відсутності вентиляції або її малої продуктивності співвідношення температури та відносної вологості незадовільні. Запропоновано застосування ультразвукового іонізатора повітря. Експериментально визначено, що за використання ультразвукового організатора повітря нормалізуються концентрації аероіонів обох полярностей та відносна вологість повітря. При цьому за рахунок спрямованого руху повітря з нормативною швидкістю відбувається винос аерозолей, що утворюється внаслідок дихання і можуть містити хвороботворні бактерії. У реальних умовах потоки повітря можуть бути частково нестаціонарні. Це особливо стосується приміщень спеціального призначення зі складними системами природньої та примусової вентиляції. Показано, що в підвальних та напівпідвальних приміщеннях можливе накопичення радону. Визначення його присутності складне і потребує спеціального обладнання. Про його присутність може свідчити підвищена концентрація аероіонів, порівняно з концентраціями у зовнішньому повітрі. Забезпечення нормативних показників мікроклімату у приміщеннях спеціального призначення комплексна задача. Її вирішення залежить від різних додаткових чинників впливу – електростатичних магнітних полів, деіонізації повітря у каналах вентиляції тощо. Тому на стадії проєктування організаційно-технічних заходів щодо забезпечення необхідних параметрів необхідним є моніторинг вихідних умов у конкретних приміщеннях, визначення переліку основного обладнання, яке експлуатується у таких приміщеннях й оціночне визначення прогнозованих умов перебування людей.

Ключові слова: мікроклімат, сховище, аерозолі, аероіони, радон.

Вступ

Мікрокліматичні умови є головним показником якості середовища перебування людей. До таких показників належать температура та відносна вологість повітря, швидкість його спрямованого руху, концентрація іонів. Усі ці параметри прямо або опосередковано пов’язані між собою. Відносна вологість залежить від температури повітря, концентрації аероіонів – від продуктивності вентиляції та відносної вологості тощо. Не дивлячись на те, що сучасні системи кліматичного контролю вважаються такими, що забезпечують комфортність середовища, у реальних умовах вони забезпечують температурний режим і частково – нормативну вологість повітря. Крім цих факторів, потребують підтримання й концентрації аероіонів, що на сьогодні практично неможливо у автоматичному режимі. Крім того, на цей параметр впливає низка випадкових і змінних чинників. Це наявність у повітрі приміщень дрібнодисперсної субстанції – пилу та аерозолей. Особливо це критично для приміщень спеціального призначення, таких як приміщення диспетчеризації житлових комплексів та сховищ цивільного захисту. У приміщеннях диспетчеризації розташовано багато контрольного обладнання та перебуває персонал, який виконує відповідальні функції. У сховищах цивільного захисту через скупчення людей генерується велика кількість аерозолей через їх дихання. При цьому такі приміщення зазвичай розташовуються в напівпідвальних або підвальних приміщеннях з мінімальними функціональними системами примусової вентиляції або природньої вентиляції. Прогнозування змін усіх параметрів середовища таких приміщень потребує теоретичного обґрунтування та експериментальних досліджень. Такі дослідження можуть бути використані для проєктування або облаштування приміщень спеціального призначення.

Стан питання

Мікрокліматичні параметри середовища регламентуються національними і міжнародними нормативами [1]. Але вони стосуються приміщень загального призначення. В Україні чинний норматив щодо приміщень сховищ цивільного захисту [2]. В усіх цих документах висуваються вимоги до конкретних параметрів з посиланням на інші будівельні санітарні норми [3]. Але цей норматив на сьогодні переглядається через недостатню обґрунтованість деяких технічних рішень, зокрема, для нестандартних об’єктів та приміщень. Вдосконаленню засобів нормалізації мікрокліматичних параметрів середовища присвячено багато досліджень. Зокрема, це стосується нестандартних будівель [4]. У цих роботах визначено вплив організації вентиляції на температуру і відносну вологість повітря. Втім у багатьох випадках відсутня обґрунтованість саме пропонованих режимів вентиляції. У дослідженні [5] визначено можливість нормалізувати аероіонний режим приміщень засобами штучної іонізації повітря. Перевагою роботи є визначення ступенів деонізації повітря аерозолями та керування цими параметрами. Дослідження [6] пропонує засоби керування якістю повітря у приміщеннях інно-

ваційним пристроєм іонізації та зволоження повітря. Його перевагою є відсутність побічних ефектів (генерація сполук азоту та озону) та керованість переважних полярностей іонів. Але поза увагою залишаються необхідні продуктивності іонізаторів, виходячи з наявних чинників деіонізації. Це ж стосується дослідження [7]. Такі чинники можуть бути неоднозначними. За наявності деяких технічних засобів та розташування приміщень, додаткова іонізація може виникати через електричний розряд та проникнення радону. Останнє актуально для напівпідвальних та підвальних приміщень [8]. Але на сьогодні недостатньо досліджено зв'язки мікрокліматичних параметрів з наявністю аерозолей. Утворення аерозолей досліджується, в основному, з точки зору абсорбції ними шкідливих речовин у довкіллі [9]. Але актуальною задачею є визначення зв'язку механізмів переносу аерозолей з режимами вентиляції та оптимізація цих процесів. Наведене вимагає проведення певних теоретичних досліджень з метою оптимізації цих процесів та забезпечення нормативних умов перебування людей у приміщеннях спеціального призначення.

Мета роботи – розробити практично значущі теоретичні засади підтримання належних мікрокліматичних параметрів у приміщеннях спеціального призначення та верифікувати отримані результати.

Викладення основного матеріалу

На сьогодні підтримання температурного і вологогісного режиму приміщень будь-якого призначення добре опрацьовано, але за обмежених просторів та наявності великої кількості людей ці співвідношення можуть бути неоднозначними. У багатьох випадках це обумовлене наявністю у повітрі аерозолей, які змінюють співвідношення вмісту вологи, за даної температури і поглинають аероіони. Швидкості осідання (седиментацію) аерозолей можна оцінити з рівняння Стокса. Але сам по собі без прив'язки до конкретних умов такий підхід не дає практичного результату. Тому доцільно оцінити основні параметри, які впливають на шукані показники. Треба враховувати, що у будь-яких системах вентиляції забезпечується максимально можлива ламінарність потоку повітря, тобто $R_e < 1$. Також вважаємо, що аерозольні частинки мають сферичну форму і рухаються по однорідному повітряному потоці. Рівняння руху частинки в цьому випадку відповідає співвідношенню.

$$m\frac{d\upsilon}{dt} = 3\pi\mu d\left(\upsilon_n - \upsilon_r\right) + mg,$$

де m – маса частинки, υ_r – швидкість частинки, υ_n – швидкість повітряного потоку, d – діаметр частинки, μ – динамічна в'язкість повітря. Це рівняння можна перетворити на співвідношення:

$$\tau\frac{d\upsilon}{dt} + \upsilon_r = \upsilon_n + \tau g, \text{ де } \tau = \frac{m}{3\pi\mu d}.$$

Цей показник уявляється важливим і характеризує, з якою швидкістю частинка аерозолю набуває швидкості повітряного потоку. Він вимірюється у секундах і показує, за який час частинка починає рухати зі швидкістю повітря за наявності повітрообміну. Відомо, що аерозольні частинки, які утворюються під час дихання, мають час життя у нерухому повітрі до 9 годин. При цьому вони інтенсивно поглинають аероіони непередбачуваної полярності [5]. Підвищення повітрообміну знижує час перебування частинок у повітрі до кількох хвилин. Час релаксації можна отримати як

$$\tau = \frac{d^2\left(\rho_e - \rho_n\right)}{18\mu},$$

де ρ_r та ρ_n – густини аерозольної частинки та повітря.

Враховуючи, що $\rho_n \ll \rho_r$: $\tau = \frac{d^2}{18\mu}\rho_r$.

Якщо стоківский опір руху дорівнює силі тяжіння, можна приблизно оцінити швидкість осідання. Але для точного визначення стаціонарного показника осідання τ_0 необхідно враховувати прискорення та уповільнення аерозольних частинок. Швидкість досягнення стаціонарного показника осідання визначається чинником $e^{-t/\tau}$. Наприклад, для нормальних умов ($101\cdot10^3$ Па та 20 °С) для сферичної водяної частинки відповідні показники наведені у табл. 1.

Таблиця 1 – **Залежність часу релаксації та швидкості осідання від діаметрів аерозольних частинок**

d, мкм	τ, с	τ_0, с
0,05	$4{,}00\cdot10^{-8}$	$2{,}8010^{-7}$
0,10	$9{,}15\cdot10^{-8}$	$6{,}40\cdot10^{-7}$
0,50	$1{,}03\cdot10^{-6}$	$7{,}20\cdot10^{-6}$
1,00	$1{,}56\cdot10^{-6}$	$2{,}50\cdot10^{-5}$
5,00	$7{,}87\cdot10^{-5}$	$5{,}50\cdot10^{-4}$
10,00	$7{,}10\cdot10^{-4}$	$2{,}20\cdot10^{-3}$
50,00	$7{,}64\cdot10^{-3}$	$2{,}30\cdot10^{-2}$

Як видно, частинки розмірами до 1 мкм дуже швидко прискорюються або уповільнюються до стаціонарного показника осідання. В даному випадку вважалося, що аерозольні частинки складаються з води, що відповідає умові їх виникнення внаслідок дихання людей. Вище зазначалося, що співвідношення температури і відносної вологості повітря не завжди однозначне. Вимірювання упродовж кількох годин у приміщенні зі стандартною системою охолодження повітря наведено у табл. 2.

Таблиця 2 – **Залежність відносної вологості повітря від температури у приміщенні при відсутності повітрообміну**

t, °C	15	16	17	18	19	20	21	22	23	24	25
φ, %	32	35	37	38	40	43	45	46	48	51	54

Наведені результати відповідають умовам відсутності повітрообміну і людей у приміщенні. Отримані дані свідчать, що за оптимальних температур відносна вологість перебуває на мінімальній межі. Цей показник підвищується внаслідок перебування людей, але одночасно відбувається деіонізація повітря через осідання аероіонів на аерозольній частинці. У вентильованому приміщенні можна розрахувати швидкість потоків повітря для видалення аерозолей, що має позитивний ефект. Але при цьому змінюється і відносна вологість повітря та знижуються концентрації аероіонів. Тому доцільно застосувати штучну іонізацію повітря. Найбільш прийнятним є ультразвуковий іонізатор повітря [10]. Такий іонізатор іонізує й частково підвищує вологість

повітря. Іонізація відбувається за рахунок балоелектричного ефекту і не дає небажаних побічних ефектів. Результати експериментальної нормалізації мікрокліматичних показників наведено у табл. 3.

Таблиця 3 – **Зміна концентрацій іонів і відносної вологості повітря при ультразвуковому розпиленні води упродовж 3 годин**

T, h	t, °C	φ, %	n, $см^{-3}$	
			n-	n+
0	23,0	43	320	390
1	22,5	46	720	660
2	22,0	50	1120	960
3	21,0	55	1450	1120

Наведені дані свідчать, що підвищення концентрації аероіонів в обох полярностях разом з відносною вологістю повітря відбувається поступово і рівномірно. Для часткового очищення повітря від аерозолей, які утворилися наслідок дихання людей і можуть нести хвороботворні бактерії, доцільно здійснювати штучну вентиляцію повітря з урахуванням стаціонарного показника осідання. Слід враховувати, що за реальних концентрацій іонів у повітрі (табл. 2) взаємодія аерозольних частинок з іонами не впливає на осідання частинок на поверхні через незначну напруженість електричного поля зарядів.

Експериментальне визначення змін концентрації аерозолей у повітрі складне через їх малі розміри і рідинну основу. У стаціонарних умовах закон зниження концентрації відповідає експоненціальній залежності. Слід враховувати, що у реальних (нормативних) умовах швидкість повітряного потоку не перевищує 0,1 м/с. Тому наведені співвідношення достатньо точно відповідають середовищу, яке рухається у просторі зі сталою швидкістю. Але це спрощення може бути неприйнятним для деяких задач, пов'язаних з динамікою аерозолей, наприклад, у каналах вентиляції. Обирання рівнянь для розрахунків руху частинок ускладнюється через неможливість точно описати розподіли повітряних потоків. Це особливо стосується приміщень спеціального призначення зі складною системою вентиляції, що притаманне, наприклад, пристосованим сховищам цивільного захисту. Не дивлячись на те, що наведені рівняння стосуються одномірного випадку зі зростанням числа Рейнольдса нелінійність сили опору буде змінювати кінцеві результати. Тому у процесі проєктування систем життєзабезпечення приміщень спеціального призначення необхідно намагатися створити умови руху повітряних потоків з малими числами Рейнольдса. Слід враховувати, що на півпідвальних та підвальних приміщеннях можуть спостерігатися концентрації аероіонів вищі за показники зовнішнього повітря, яке подається системою вентиляції. Вимірювання каліброваним вимірювачем концентрації радону AlphaE свідчить, що за активності 80–100 $Бк/м^3$, концентрації аеронів обох полярностей складали 1100–1200 $см^{-3}$. При цьому зовнішнє повітря за відсутності техногенних впливів – 500–600 $см^{-3}$. Через те, що радон має α-розпад, його присутність неможливо визначити поширеними приладами радіаційного контролю. Тому за відсутності спеціального обладнання про присутність радону можна отримати інформацію з порівнянням показника приміщення за відсутністю примусової та природньої вентиляції з показниками зовнішнього повітря. Слід зазначити, що забезпечення нормативних мікрокліматичних показників у приміщеннях спеціального призначення – комплексна і багатопланова задача. Її вирішення у кожному окремому випадку може мати певні особливості. Це наявність електростатичних полів, які виникають внаслідок трибоелектричного ефекту, дрібнодисперсного пилу через пересування людей тощо. Потужне електричне обладнання може генерувати магнітні поля, які викликають дрейф всіх заряджених частинок. Тому на етапі проєктування відповідних організаційно-технічних заходів необхідним є аналіз переліку обладнання, яке планується для використання параметрів внутрішнього і зовнішнього середовища, електромагнітного оточення тощо.

Висновки

1. Визначено, що для підтримання нормативних показників мікроклімату у приміщеннях спеціального призначення необхідно врахувати низку додаткових параметрів. Це обумовлено генерацією аерозолей внаслідок дихання людей у обмежених просторах, складністю вентиляційних систем, можливим надходженням радону.

2. Запропоновано простий у використанні розрахунок часу релаксації частинок аерозолей та показника осідання частинок. Надано значення цих параметрів у залежності від діаметрів аерозольних частинок для нормальних умов.

3. Наведено експериментальні дані щодо зміни відносної вологості повітря у залежності від температури. Встановлено, що у обмежених просторах оптимальні співвідношення температури й вологості повітря можуть відрізнятися. Отримані дослідні дані щодо застосування ультразвукового іонізатора повітря для нормалізації концентрацій аероіонів та відносної вологості повітря для нормального температурного режиму.

4. Розрахунки і визначення змін мікрокліматичних параметрів здійснювалося за умови ламінарності повітряних потоків з малими числами Рейнольдса. У реальних умовах стаціонарність потоків може порушуватися. Крім того, на мікрокліматичні параметри можуть впливати додаткові чинники – електростатичні і магнітні поля, які викликають осідання й дрейф усіх заряджених частинок. У підвальних та напівпідвальних приміщеннях може накопичуватися радон. Його визначення складне, але про наявність радону може свідчити підвищена іонізація повітря, порівняно з зовнішнім повітрям.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Standard of Building Biology Testing Methods. SBM-2015/ Building biology evaluation guidelines for sleeping areas. Baubiologie Maes. Institut für Baubiologie + Nachhaltigkeit IBN. GUIDELINES. 4 p. URL: https://buildingbiology.com/building-biology-standard/
2. ДБН В.2.2-5:2023. Будинки і споруди. Захисні споруди цивільного захисту. – Затверджено Наказом Мінрегіону України № 702 від 10.08.2023. – Чинний з 01.11.2023. – Діє зі Зміною № 2 з 01.04.2025. – Київ: Міністерство розвитку громад і територій України, 2023, URL: https://online.budstandart.com/ua/catalog/doc-page.html?id_doc=104666
3. ДБН В.2.5-67:2013 Опалення, вентиляція та кондиціонування. Наказ 25.01.2013 № 24 Про затвердження ДБН В.2.5-67:2013 Опалення, вентиляція та кондиціонування. Наказом від 28.08.2013 р. № 410 дата введення в дію змінена на 01.01.2014 р., URL: https://online.budstandart.com/ua/catalog/doc-page.html?id_doc=50154
4. Laurini, E., De Vita, M., & De Berardinis, P. (2021). Monitoring the Indoor Air Quality: A Case Study of Passive Cooling from Historical Hypogeal Rooms. Energies, 14(9), 2513. https://doi.org/10.3390/en14092513
5. Tykhenko, O., Glyva, V., Levchenko, L., Burdeina, N., Biruk, Y., Zozulya, S., Krasnianskyi, G., Nikolaiev, K., Aznaurian, I., & Zozulia, L. (2024). Study of air deionization factors. Eastern-European Journal of Enterprise Technologies, 2(10 (128), 26–33. https://doi.org/10.15587/1729-4061.2024.300909
6. Bolibrukh, B., Glyva, V., Kasatkina, N., Levchenko, L., Tykhenko, O., Panova, O., Bogatov, O., Petrunok, T., Aznaurian, I., & Zozulya, S. (2022). Monitoring and management ion concentrations in the air of industrial and public premises. Eastern-European Journal of Enterprise Technologies, 1(10(115), 24–30. https://doi.org/10.15587/1729-4061.2022.253110.
7. Ченчевой В.В., Сукач С.В., Ченчева О.О., Федорова Н.С., Григор'єва Д.С. Дослідження параметрів гідроаероіонного складу повітря робочого приміщення з ультразвуковою іонізацію. Вісті Донецького гірничого інституту. 2020. Вип. № 2(47). С. 168−174. https://doi.org/10.31474/1999-981x-2020-2-168-175.
8. Nunes, L. J. R., Curado, A., & Lopes, S. I. (2023). The Relationship between Radon and Geology: Sources, Transport and Indoor Accumulation. Applied Sciences, 13(13), 7460. https://doi.org/10.3390/app13137460
9. Kirkby, J., Amorim, A., Baltensperger, U., Carslaw, K. S., Christoudias, T. et al. (2023). Atmospheric new particle formation from the CERN CLOUD experiment. Nature Geoscience, 16 (11), 948–957. https://doi.org/10.1038/s41561-023-01305-0
10. Ультразвуковий іонізатор повітря: пат. України 138020, МПК G12B 17/00. Заявл. 23.05.2019, опубл. 11.11.2019. Бюл. № 21. 2 с. URL: https://sis.nipo.gov.ua/uk/search/detail/1391281/

Received (Надійшла) 31.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Глива Валентин Анатолійович – доктор технічних наук, професор, завідувач кафедри фізики, Київський національний університет будівництва та архітектури, Київ, Україна;
Valentyn Glyva – Doctor of Technical Sciences, Professor, Head of the Department of Physics, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: hlyva.va@knuba.edu.ua, ORCID Author ID: https://orcid.org/0000-0003-1257-3351;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57210185162.

Галонько Ярослав Орестович – аспірант кафедри технологій захисту навколишнього середовища та охорони праці, Київський національний університет будівництва і архітектури, Київ, Україна;
Yaroslav Halonko – PhD student at the Department of Environmental Protection Technologies and Labour Protection, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: halonko_yo-2024@knuba.edu.ua, ORCID Author ID: https://orcid.org/0009-0000-4932-4338.

Theoretical and experimental approaches to ensuring microclimatic parameters in special-purpose premises

Valentyn Glyva, Yaroslav Halonko

Abstract. Based on an analysis of the current regulatory framework and scientific research, it has been established that conceptual approaches to ensuring microclimatic parameters in special-purpose premises need to be refined. Such premises include dispatch and control rooms and civil defence shelters of various types and designs. Control rooms contain a large amount of monitoring and control equipment, and shelters can have large crowds of people, which affects environmental parameters. In such conditions, a large amount of aerosols is generated. An easy-to-use calculation of the main indicators of aerosol dynamics depending on the air velocity caused by ventilation is proposed. The particle relaxation coefficient is determined – the time it takes for them to reach the air flow velocity. The steady-state particle settling rate is calculated. The use of an ultrasonic air ioniser is proposed. It has been experimentally determined that the use of an ultrasonic air ioniser normalises the concentrations of air ions of both polarities and the relative humidity of the air. At the same time, due to the directed movement of air at a standard speed, aerosols formed as a result of breathing and which may contain pathogenic bacteria are removed. In real conditions, air flows may be partially unsteady. This is especially true for special-purpose premises with complex natural and forced ventilation systems. It has been shown that radon can accumulate in basements and semi-basements. Determining its presence is difficult and requires special equipment. Its presence may be indicated by an increased concentration of air ions compared to concentrations in the outside air. Ensuring standard microclimate indicators in special-purpose rooms is a complex task. Its solution depends on various additional influencing factors – electrostatic magnetic fields, air deionisation in ventilation ducts, etc. Therefore, at the stage of designing organisational and technical measures to ensure the necessary parameters, it is necessary to monitor the initial conditions in specific rooms, determine the list of basic equipment operated in such rooms, and estimate the predicted conditions for human presence.

Keywords: microclimate, storage facility, aerosols, air ions, radon.

УДК 628.5/8:614.7 **doi: 10.26906/SUNZ.2026.2.214**

Л. О. Левченко[1], Т. В. Шабатура[2]

[1]Національний технічний університет України «КПІ імені Ігоря Сікорського, Київ, Україна
[2]Київський національний університет будівництва і архітектури, Київ, Україна

МОДЕЛЮВАННЯ ПРОЦЕСІВ ПОШИРЕННЯ АЕРОІОНІВ ТА ОЧИЩЕННЯ ПОВІТРЯ У ПРИМІЩЕННЯХ

Анотація. Здійснено моделювання динаміки концентрації аероіонів та завислих частинок у приміщеннях з примусовою вентиляцією. Визначено найбільш прийнятні співвідношення й перелік критичних факторів, які впливають на достовірність результатів моделювання. Такими факторами є продуктивність пристрою штучної іонізації повітря, іонізація зовнішнього повітря, коефіцієнти рекомбінації аероіонів, коефіцієнти осідання аероіонів на завислі частинки та кратність повітрообміну у приміщенні. Враховано наявність електричного поля усієї сукупності аероіонів. Отримана залежність концентрацій аероіонів обох полярностей у повітрі від кратності повітрообміну у приміщенні внаслідок дії системи примусової вентиляції. Показано, що суттєві зміни концентрації аероіонів відбуваються за трикратного повітрообміну. При збільшенні цього показника концентрації аероіонів обох полярностей залишаються стабільними. Концентрації завислих частинок у повітрі приміщення стабільно знижуються з збільшенням кратності повітрообміну. Це пояснюється виносом частинок потоком повітря. Проведено моделювання динаміки концентрацій аероіонів з урахуванням їх осідання на завислі частинки. Складна динаміка цього показника спостерігається за трикратного повітрообміну. Зі збільшенням кратності повітрообміну цей показник стабільний і має тенденцію до незначного підвищення. Усі розрахунки здійснювалися для високих генерацій аероіонів іонізатором повітря. Встановлено, що навіть за таких умов інтенсивність осідання аероіонів на поверхні незначна. Слід очікувати, що процес очищення повітря є ефективним за умови дуже високих концентрацій аероіонів та електризації поверхонь. Тільки за таких умов можна очікувати нейтралізації поверхневих електростатичних зарядів. У іншому випадку такі нейтралізації доцільно здійснювати іншими методами, наприклад, підвищенням відносної вологості повітря до верхньої нормативної межі.

Ключові слова: моделювання, аероіони, завислі частинки, кратність повітрообміну.

Вступ

Концентрація аероіонів обох полярностей та дрібнодисперсного пилу є важливим показником якості повітря. Ці показники є взаємно пов'язаними через осідання іонів на завислі частинці. У результаті відбувається деіонізація повітря. У разі набуття частинками заряду вони дрейфують у повітрі внаслідок наявності власного електричного поля та електричного поля Землі. Більшість облицювальних матеріалів полімерні і відбувається часткове осідання пилу на поверхні, які або мають поверхневі заряди, або набувають їх через осідання заряджених частинок. Очевидно, що ці процеси достатньо складні. Враховуючи, що як концентрації аеронів, так і пилу регламентуються відповідними стандартами та санітарними нормами, потребують дослідження динаміки цих фізичних чинників у просторі і часі та розроблення заходів і засобів їх нормалізації та підтримання їх на нормативному рівні. Це особливо актуально для приміщень, у яких постійно або тимчасово перебуває велика кількість людей. Внаслідок дихання та змін температури і вологості у повітрі утворюються аерозольні частинки, які взаємодіють з аероіонами аналогічно дрібнодисперсному пилу. Тому доцільним є дослідження цих процесів і розроблення проєктних рішень для забезпечення належних умов перебування людей. Ефективним методом проєктування будівель і приміщень з нормативними параметрами середовища є моделювання поширення і змін концентрації аероіонів, що може бути основою для проєктування системи життєзабезпечення будівель і споруд.

Стан питання

Концентрації аероіонів регламентуються міжнародними та національними нормативними документами [1]. У Європі мінімально допустимою концентрацією аероіонів вважається концентрація 500 $см^{-3}$ обох полярностей. В Україні нижньою межею є 400 $см^{-3}$ негативних та 600 $см^{-3}$ позитивних аероіонів. Ці відмінності не є суттєвими через те, що це нижні допустимі межі, а похибка вимірювання найсучасніших лічильників аероіонів складає не менше 20 %. При цьому концентрації дрібнодисперсного пилу регламентуються для конкретних виробничих умов. Концентрації пилу виробничих приміщень не повинні перевищувати 6 $мг/м^3$. Для громадських та житлових приміщень цей показник чітко не регламентується, а існує вимога мінімізації концентрацій завислих частинок.

Дослідженню динаміки концентрації аероіонів присвячено багато робіт [2]. В основному у них констатується факт значної деіонізації повітря під впливом техногенних факторів. Частина досліджень стосується нормалізації аероіонного складу повітря методом штучної іонізації. Розробки [3] описують ефективність іонізації та очищення повітря іонізаторами, які використовують коронні розряди. Але такий спосіб дає побічні ефекти – неконтрольовану генерацію озону та сполук азоту. Ці речовини є шкідливими для людей, а їх концентрації нормуються. Роботи [4] пропонують підвищувати концентрації аероіонів за рахунок ультразвукового розпилення води. Необхідний ефект досягається за рахунок балоелектричного ефекту. Такий спосіб достатньо ефективний, але виникають проблеми з регулюванням переважної полярності генерації. Найбільш прийнятним є спосіб іонізації повітря з використанням світлодіодних систем ультрафіолетового випромінювання [5]. Іонізація починається в усьому об'ємі приміщення, тому визначати розподіл концентрацій аероіонів не потрібно .Але у вентильованих приміщеннях спостерігаються градієнти концентрації у

залежності від кратності повітрообміну. На стадіях проєктування прогнозування концентрації аероіонів та знепилення повітря можливе лише засобами моделювання. Цій проблематиці присвячено як класичній роботи [6], так і сучасні дослідження [7]. При здійсненні моделювання поширення аеронів головною задачею є застосування коректних математичних функцій, які враховують природну іонізацію повітря, штучну іонізацію, рекомбінацію аеронів та осідання на завислі частинки. При цьому у вентильованих приміщеннях важливою є кратність повітрообміну. У деяких роботах враховується електричне поле, яке утворюється з сукупністю іонів та заряджених частинок, але ускладнення відповідних рівнянь має наслідком підвищення обсягів обчислень та збільшення похибки. При цьому важко достатньо точно визначити відповідні вихідні дані. Для створення моделей поширення аероіонів обох полярностей застосовують рівняння на Нав'є-Стокса або розраховують динаміку аероіонів із застосуванням рівняння переносу з визначенням вектору швидкості повітряного потоку (поля потенціалу швидкості). Але на коректність моделювання впливають багато додаткових чинників, наприклад, деіонізація повітря у повітропроводах вентиляційних систем, що врахувати можна тільки для конкретного випадку. Тому доцільно здійснити оціночне моделювання поширення аероіонів у приміщенні та визначити межу концентрації, яка впливає на вміст завислих частинок у повітрі. Для цього достатньо використати спрощені рівняння балансу аероіонів та визначити вплив аероіонів на поведінку завислих частинок. Це дозволить одночасно оцінити швидкість видалення аероіонів й осадження завислих частинок різних розмірів.

Мета дослідження – на основі моделювання процесів поширення аероіонів оцінити динаміку їх концентрацій та очищення повітря від дрібнодисперсного пилу.

Викладення основного матеріалу

Усі співвідношення щодо балансу аероіонів у повітряному середовищі є варіантами рівняння неперервності для слабоіонізованої плазми [8]:

$$\frac{\partial n(r,t)}{\partial t} = g(r,t) + \frac{n(r,t) - n_0}{\tau} - \frac{1}{q} div\, j - \frac{\partial n(r,t)}{\partial x} \cdot \upsilon(r,t),$$

де $n(r, t)$ – концентрація іонів у точці з радіус-вектором r у момент часу t, $g(r, t)$ – продуктивність генерації іонів у одиниці об'єму за 1 секунду, n_0 – вихідна концентрація іонів у повітрі приміщення, τ – середній час життя іонів за даних умов, q – заряд іона, j – густина електричного струму за рахунок руху іонів, x – напрямок, у якому відбувається зміна концентрації іонів, υ – швидкість потоку повітря.

Для зручності використання зазвичай розглядають баланс аероіонів двох полярностей, тобто розглядається система двох диференційних рівнянь.

Розроблено багато варіантів цих рівнянь з урахуванням різної кількості критичних факторів впливу на концентрації аероіонів. Для оціночного визначення концентрацій в умовах дії припливно-витяжної вентиляції за основу розрахунків доцільно обрати співвідношення, наведені у [3]:

$$\frac{dn}{dt} = q_n - \alpha np - \beta nA + n_0 \frac{L}{V} - n\frac{L}{V} - \lambda n,$$

$$\frac{dp}{dt} = q_p - \alpha np - \beta pA + p_0 \frac{L}{V} - p\frac{L}{V} - \lambda p,$$

де n, p – концентрації негативних та позитивних аероіонів (м$^{-3}$), q_n, q_p – швидкість генерації негативних і позитивних іонів у приміщенні (с$^{-1}$), n_0, p_0 – концентрації аероіонів у зовнішньому повітрі (м$^{-3}$); α – коефіцієнт, який характеризує швидкість рекомбінації аероіонів протилежних полярностей (м^3/с), β – швидкість осідання іонів на завислі частинки (м^3/с), A – концентрація завислих частинок у повітрі приміщення (м$^{-3}$), L – продуктивність системи вентиляції (м^3/с), V – об'єм приміщення.

Коефіцієнт λ характеризує осадження аероіонів на поверхні й може бути визначений як [9]:

$$\lambda = \frac{b}{\varepsilon_o}\left(q_e - q_c eA\right),$$

де b – рухливість аероіонів (м^2/В·с), ε_0 – діелектрична стала (8,85·10^{-12} Ф/м), e – елементарний заряд (1,6·10^{-19} Кл), q_e – об'ємна густина електричного заряду аероіонів, q_c – кількість елементарних зарядів, які набувають завислі частинки внаслідок дрейфу у електричному полі аероіонів. При цьому q_e визначається просто: $q_e = en - ep$, а вираз для розрахунку q_c досить складний:

$$q_c = \frac{4\pi\varepsilon_0 dkT}{e^2}\left[\ln\left[1 + \frac{dcpe^2 t}{4\varepsilon_0 kT}\right] - \ln\left[1 + \frac{dcne^2 t}{4\varepsilon_0 kT}\right]\right],$$

де k – стала Больцмана (1,38·10^{-23} Дж/К), T – абсолютна температура (К), d – середній діаметр частинок (м), c – теплова швидкість аероіонів (м/с), t – середній час перебування частинки у приміщенні (с), Більшість коефіцієнтів відомі [6, 9] $\alpha \approx 1{,}5\cdot10^{-12}$ м^3/с, $\beta \approx 1{,}2\cdot10^{-12}$ м^3/с.

Вважається, що розміри завислих частинок менші за 20 мкм, що відповідає експериментальним даним. Температуру вважають близькою до 3000 К. Теплова швидкість легких аероіонів – 300 м/с. Рухомість легких аероіонів близька до 2,5·10^{-6} м^2/В·с.

Виходячи з наведених співвідношень, було проведено моделювання динаміки концентрації аероіонів та завислих частинок у повітрі приміщень. Розрахунки здійснювалися для різних продуктивностей вентиляційної системи (кратність повітрообміну за годину) та продуктивностей пристрою штучної іонізації повітря. Загально визначено, що негативні аероіони більш корисні для людини, тому більшість досліджень стосується забезпечення нормативних концентрацій аероіонів від'ємної полярності. При цьому для однополярної іонізації повітря використовували високовольтні іонізатори, у яких іони генеруються коронним розрядом [7, 8]. Але негативна «корона» має недоліки. За від'ємним зарядом електроду відбувається неконтрольована генерація сполук азоту та озону. Тому у сучасних умовах доцільно покласти у модель генерацію іонів біполярним іонізатором повітря, наприклад, ультразвуковим [9]. Потрібні кон-

центрації аеронів кожної полярності забезпечуються поглинанням частин іонів протилежної полярності.

В усіх нормативах концентрація аеронів визначається в одиницях $см^{-3}$, але у відповідних рівняннях фігурують об'єм приміщення та швидкість повітрообміну, які оперують метрами кубічними, тому для коректності розрахунків концентрації аероіонів та завислих частинок вимірюємо у $м^{-3}$. Початкові умови для моделювання наведені у табл. 1.

Таблиця 1 – **Початкові умови, які використовуються для моделювання динаміки концентрацій аероіонів та завислих частинок у повітрі приміщення**

n_0	$5\cdot10^8$ $м^{-3}$	q_n	$3\cdot10^7$ $м^{-3}$	A	$2\cdot10^6$ $м^{-3}$
p_0	$5\cdot10^8$ $м^{-3}$	q_p	$3\cdot10^7м^{-3}$	n	$1\cdot10^9$ $м^{-3}$
				p	$1\cdot10^9$ $м^{-3}$

Розрахунки здійснювалися для 3, 6, 9, 12 повітрообмінів (ПО) за годину для концентрації аероіонів та завислих частинок. Результати моделювання наведено у табл. 2.

Таблиця 2 – **Зміна концентрацій аероіонів за різних повітрообмінів у приміщенні**

ПО	n^-*10^6, $м^{-3}$	n^+*10^6, $м^{-3}$	ПО	n^-*10^6, $м^{-3}$	n^+*10^6, $м^{-3}$
3	1000–3600	1000–3500	9	3400–3200	3450–3250
6	3600–3400	3500–3450	12	3200–3000	3250–3150

Наведені результати свідчать, що за високої продуктивності іонізатора повітря концентрації аероіонів майже не залежать від продуктивності системи примусової вентиляції. При цьому деякі відмінності змін концентрації аероіонів різної полярності ймовірно обумовлені різними рухливостями μ негативних та позитивних аероіонів. Відомо [9], що $\mu^- = 1{,}66$ $см^2/В\cdot с$, $\mu^+ = 1{,}19$ $см^2/В\cdot с$. Аналогічні розрахунки було виконано для зміни концентрацій завислих частинок.

Таблиця 3 – **Зміна концентрацій завислих частинок за різних повітрообмінів у приміщенні**

ПО	3	6	9	12
$A*10^6$,$м^{-3}$	2,0–2,6	2,6–1,5	1,5–1,0	1,0–0,6

У даному випадку враховувалися частинки мікророзмірів. Навіть оціночне визначення концентрації наночастинок повітря дуже складне, а існуючі експериментальні дані суперечливі. Дані табл. 3 свідчать, що на початковому етапі концентрації завислих частинок зростають, що є результатом притоку частинок із зовнішнім повітрям. У подальшому спостерігається ефективне видалення завислих частинок за рахунок повітрообміну Це надає можливість обрати оптимальний режим роботи систем примусової вентиляції у залежності від початкових умов.

Якщо враховувати вплив завислих частинок на концентрації аероіонів, то середні співвідношення для аероіонів обох полярностей наведені у табл. 4.

Фактично, за умов, наведених у табл. 1, підвищення концентрації аероіонів з наступним її зниженням відбувається за трикратного повітрообміну. У подальшому цей показник стабілізується і не залежить від продуктивності системи примусової вентиляції.

Таблиця 4 – **Прогнозований вплив продуктивності системи примусової вентиляції на концентрації аероіонів із врахуванням впливу завислих частинок**

ПО	3	6	9	12
$n*10^6$, $м^{-3}$	1000–2500–2300	2300–2350	2350–2400	2400–2400

Важливим питанням є визначення взаємного впливу концентрацій аероіонів у повітрі й інтенсивності їх осідання на електризовані поверхні.

Розрахунки свідчать, що за зростання продуктивності іонізатора повітря з $1\cdot10^2$ до $1\cdot10^{10}$ $с^{-1}$ концентрації аероіонів підвищуються від 10^6 до 10^{13} $м^{-3}$ лінійно. За врахуванням, впливу осідання аероіонів на поверхні верхня межа концентрації становить $1{,}5\cdot10^{11}$ $м^{-3}$. При цьому відхилення від лінійної залежності починаються з продуктивності іонізатора порядка 10^6 $с^{-1}$. Таким чином, застосовувати підвищену іонізацію повітря для нейтралізації поверхневих електростатичних зарядів ефективно за великих продуктивностей іонізатора, що не завжди потрібно, виходячи з чинних нормативів. Можна дійти висновку, що процеси осадження суттєві за наявності об'ємного електричного поля в сукупності аероіонів. Тому для нейтралізації поверхневих електростатичних зарядів за прийнятних концентрацій аероіонів доцільно застосовувати інші засоби, наприклад, підвищення відносної вологості повітря до вищої нормативної межі – 60 %.

Слід враховувати, що усі моделі динаміки концентрації аероіонів та завислих частинок у повітрі приміщень мають низьку спрощень і припущень. Тому відповідні моделі доцільно використовувати тільки для оціночного прогнозування аероіонних режимів приміщень.

Висновки

1. Визначено математичні функції, найбільш прийнятні для моделювання динаміки концентрацій аероіонів та завислих частинок у повітрі приміщень. При здійснені моделювання обов'язковим є врахування осідання іонів на частинки. Такий процес, поряд з осіданням на поверхні є джерелом деіонізації повітря.

2. Моделювання процесів змін концентрацій аероіонів та завислих частинок у повітрі здійснювалося за умови функціонування іонізатора повітря та роботи примусової вентиляції різної продуктивності (кратності повітрообміну).

3. В результаті моделювання встановлено, що за сталої генерації аероіонів при повітрообмінах, більших за три за годину, концентрації аероіонів суттєво не змінюються, й мають постійні значення, виходячи з продуктивності іонізатора повітря. У той же час за рахунок зростання продуктивності системи вентиляції спостерігається значний винос завислих частинок з повітря у приміщенні. Сумарний вплив усіх факторів на концентрації аероіонів (осідання на завислі частинки поверхні) не є критичним за кратності повітрообмінів, більших за три на годину. Визначено межу продуктивності іонізатора повітря, до якої електродинамічні явища можна не враховувати.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослід-

женння, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Standard of Building Biology Testing Methods. SBM-2015/ Building biology evaluation guidelines for sleeping areas. Baubiologie Maes. Institut für Baubiologie + Nachhaltigkeit IBN. GUIDELINES. 4 p. URL: https://buildingbiology.com/building-biology-standard/
2. Глива В.А., Ніколаєв К.Д., Тихенко О.М., Тимошенко О.П. Дослідження рівнів фізичних факторів у приміщеннях диспетчерських служб аеропортів цивільної авіації. Системи управління, навігації та зв'язку. Полтава, 2019. Вип. 1(53). С. 32−35. https://doi.org/10.26906/SUNZ.2019.1.032
3. Електростатичний повітряний фільтр-іонізатор: пат. 87189 Україна, МПК: B03C 3/08. № 2013100086; заявл. 14.08.2013; опубл. 27.01.2014, Бюл. № 2. 4 с., URL: https://sis.nipo.gov.ua/uk/search/detail/1108315/
4. Bolibrukh, B., Glyva, V., Kasatkina, N., Levchenko, L., Tykhenko, O., Panova, O., Bogatov, O., Petrunok, T., Aznaurian, I., & Zozulya, S. (2022). Monitoring and management ion concentrations in the air of industrial and public premises. Eastern-European Journal of Enterprise Technologies, 1(10(115), 24–30. https://doi.org/10.15587/1729-4061.2022.253110
5. Glyva, V., Nazarenko, V., Burdeina, N., Leonov, Y., Kasatkina, N., Levchenko, L. et al. (2023). Determining the efficiency of using led sources of ultraviolet radiation for ionization and disinfection of room air. Eastern-European Journal of Enterprise Technologies, 3 (10 (123)), 23–29. https://doi.org/10.15587/1729-4061.2023.282784
6. Noakes C.J., Sleigh P.A., Beggs C.B. Modelling the air cleaning performance of negative air ionisers in ventilated rooms. Proceeding of the 10 th Int.Conference on Air Distribution in Rooms (Roomvent 2007), 13 – 15 June 2007. – Helsinki, 2007. – 11 p., URL: https://eprints.whiterose.ac.uk/id/eprint/7700/1/Noakes_roomvent_07.pdf
7. Levchenko, L., Burdeina, N., Glyva, V., . Kasatkina, N., Biliaiev, M., Biliaieva, V., Tykhenko, O., Petrunok, T., Biruk, Y., Bogatov, O. (2023). Identifying regularities in the propagation of air ions in rooms with artificial air ionization Eastern-European Journal of Enterprise Technologies, 4(10(124)), pp. 6–14. https://doi.org/10.15587/1729-4061.2023.285967
8. Зозуля С. В. Засоби і заходи контролю та нормалізації аероіонного складу повітря виробничих і навчальних приміщень : дис. … техн. наук: 05.26.01. Київ, 2023. 130 с. URL: https://drive.google.com/file/d/1oFMFpQZTrrPOxIlCWdp_14QshfKeaN0i/view
9. Fletcher LA, Noakes CJ, Sleigh PA, Beggs CB, Shepherd SJ. Air Ion Behavior in Ventilated Rooms. Indoor and Built Environment. 2008;17(2):173-182. https://doi.org/10.1177/1420326X08089622

Received (Надійшла) 24.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Левченко Лариса Олексіївна – доктор технічних наук, професор, професор кафедри цифрових технологій в енергетиці, Національний технічний університет України «Київський політехнічний інститут імені І. Сікорського, Київ, Україна;
Larysa Levchenko – Doctor of Technical Sciences, Professor, Professor of Department Digital Technologies in Energy, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine;
e-mail: larlevch@ukr.net; ORCID Author ID: http://orcid.org/0000-0002-7227-9472;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57194577942 .

Шабатура Тарас Вікторович – аспірант кафедри технологій захисту навколишнього середовища та охорони праці, Київський національний університет будівництва і архітектури, м. Київ, Україна;
Taras Shabatura – PhD student at the Department of Environmental Protection Technologies and Labour Protection, Kyiv National University of Construction and Architecture, Kyiv, Ukraine;
e-mail: Tarasshabatura@gmail.com; ORCID Author ID: https://orcid.org/0009-0009-7525-5212

Modelling of the processes of aerion dispersion and air purification in rooms

Larysa Levchenko, Taras Shabatura

Abstract. Modelling of the dynamics of air ion concentration and suspended particles in rooms with forced ventilation has been carried out. The most acceptable ratios and a list of critical factors affecting the reliability of modelling results have been determined. These factors include the performance of the artificial air ionisation device, the ionisation of outside air, the recombination coefficients of air ions, the sedimentation coefficients of air ions on suspended particles, and the air exchange rate in the room. The presence of an electric field of the entire set of air ions is taken into account. The dependence of the concentrations of air ions of both polarities in the air on the air exchange rate in the room due to the action of the forced ventilation system is obtained. It is shown that significant changes in the concentration of air ions occur with a threefold air exchange. When this indicator increases, the concentrations of air ions of both polarities remain stable. The concentrations of suspended particles in the air of the room steadily decrease with an increase in the air exchange rate. This is explained by the removal of particles by the air flow. Modelling of the dynamics of air ion concentrations was carried out, taking into account their precipitation on suspended particles. The complex dynamics of this indicator is observed at a threefold air exchange. With an increase in the air exchange rate, this indicator is stable and tends to increase slightly. All calculations were performed for high generations of air ions by an air ioniser. It has been established that even under such conditions, the intensity of air ion deposition on surfaces is insignificant. It should be expected that the air purification process is effective under conditions of very high concentrations of air ions and electrification of surfaces. Only under such conditions can the neutralisation of surface electrostatic charges be expected. Otherwise, such neutralisation should be carried out by other methods, for example, by increasing the relative humidity of the air to the upper regulatory limit.

Keywords: modelling, air ions, suspended particles, air exchange rate.

УДК 614.84 **doi: 10.26906/SUNZ.2026.2.218**

О. О. Ченчева, Є. Є. Лашко, Д. В. Рєзнік

Кременчуцький національний університет імені Михайла Остроградського, Кременчук, Україна

КЛАСТЕРНИЙ ПІДХІД ДО ОЦІНЮВАННЯ ПОЖЕЖНИХ РИЗИКІВ У КАР'ЄРАХ ГРАНІТНОГО ВИДОБУТКУ В СИСТЕМІ ЦИВІЛЬНОЇ БЕЗПЕКИ

Анотація. Актуальність. Гірничодобувна промисловість посідає важливе місце у структурі української економіки, а пожежні ризики у кар'єрах мають комплексний і складний характер. **Об'єкт дослідження**: кластерний підхід до оцінювання пожежних ризиків у кар'єрах гранітного видобутку в системі цивільної безпеки. **Мета статті**: розробка методики комплексної оцінки ризиків на основі ймовірнісних, наслідкових й оперативних показників із подальшою кластеризацією зон ризику та просторовим відображенням за допомогою мапування. **Результати дослідження**. Наукова цінність роботи полягає у систематизації показників пожежного ризику, застосуванні сучасних статистичних методів і геоінформаційних технологій для інтегрованого аналізу, що дозволяє підвищити точність оцінювання та прогнозування небезпечних ситуацій. Практична цінність статті полягає у можливості використання розробленої методики для формування карт пожежної небезпеки, оптимізації розміщення пожежної техніки, планування заходів реагування та підвищення стійкості об'єктів до надзвичайних ситуацій. **Висновки.** Впровадження кластерного підходу та мапування дозволяє підвищити ефективність управління пожежною безпекою, скоротити час реагування та зменшити соціально-економічні втрати.

Ключові слова: цивільна безпека, пожежний ризик, кластеризація, управління ризиками, кар'єр, надзвичайна ситуація.

Вступ

Постановка проблеми. Гірничодобувна промисловість України посідає важливе місце у структурі національної економіки, зокрема у сфері видобутку будівельного каменю та граніту. Кар'єри відкритого типу, що функціонують переважно у центральних і північних регіонах країни, характеризуються високою концентрацією техніки, паливно-мастильних матеріалів, електрообладнання та вибухових речовин. Незважаючи на те, що граніт як гірська порода є негорючим матеріалом, сам виробничий процес супроводжується значними пожежними ризиками. У зв'язку з цим особливої актуальності набуває впровадження сучасних методів оцінювання й управління ризиками в межах державної системи цивільного захисту, координацію якої здійснює Державна служба України з надзвичайних ситуацій.

Пожежні ризики у кар'єрах мають комплексний характер. Основними джерелами потенційного займання є склади паливно-мастильних матеріалів, автозаправні пункти, трансформаторні підстанції, кабельні лінії, великогабаритна техніка (екскаватори, бурові установки, автосамоскиди), а також ремонтні майстерні. Окрему групу становлять природні фактори – суха рослинність на бортах кар'єру, пилові відкладення, високі температури повітря та сильний вітер, що сприяють швидкому поширенню вогню. Додатковими чинниками є людський фактор, порушення технологічної дисципліни, зношеність обладнання та недостатній контроль за станом електромереж.

Аналіз останніх досліджень і публікацій. Аналіз літературних джерел і нормативних документів, присвячених проблематиці оцінювання пожежних ризиків у кар'єрах гранітного видобутку, свідчить про міждисциплінарний характер дослідження, що поєднує положення цивільного захисту, пожежної безпеки, гірничої справи та теорії управління ризиками. Формування науково обґрунтованого кластерного підходу потребує врахування як національної нормативної бази, так і сучасних методологічних розробок у сфері аналізу ризиків. Правові засади функціонування системи цивільного захисту визначає Кодекс цивільного захисту України, який встановлює обов'язок суб'єктів господарювання здійснювати ідентифікацію й оцінювання ризиків виникнення надзвичайних ситуацій. У документі наголошується на пріоритетності превентивних заходів, прогнозування та мінімізації наслідків техногенних подій [1]. Це створює нормативне підґрунтя для впровадження аналітичних методів, зокрема багатовимірного статистичного аналізу та кластеризації, як інструментів системного управління пожежними ризиками.

Специфічні вимоги до забезпечення пожежної безпеки встановлюють Правила пожежної безпеки в Україні [2]. У них визначено порядок утримання територій, виробничих приміщень, електроустановок, складів паливно-мастильних матеріалів, а також вимоги до оснащення об'єктів первинними засобами пожежогасіння. Для кар'єрів гранітного видобутку особливо важливими є положення щодо експлуатації електрообладнання високої потужності, організації безпечного зберігання пального та дотримання протипожежного режиму в ремонтних зонах. Проте зазначені правила мають переважно регламентуючий характер і не пропонують кількісних моделей оцінювання рівня ризику, що зумовлює потребу в науковому обґрунтуванні додаткових аналітичних підходів.

Галузеві аспекти безпеки гірничих робіт регламентує Гірничий закон України, який визначає правові й організаційні засади ведення гірничої діяльності [3]. Закон встановлює вимоги щодо створення безпечних умов праці, запобігання аваріям і катастрофам, а також відповідальність підприємств за порушення норм безпеки. У контексті пожежної небезпеки цей документ акцентує увагу на необхідності системного контролю за станом виробничого обладнання та технологічних процесів. Однак, як і попередні нормативні акти, він не деталізує методику інтегрованого аналізу багатофакторних ризиків.

Контроль і координацію заходів у сфері техногенної та пожежної безпеки здійснює Державна слу-

жба України з надзвичайних ситуацій, яка розробляє методичні рекомендації щодо оцінювання ризиків і проводить державний нагляд [4]. У сучасних умовах діяльність служби орієнтована на впровадження ризик-орієнтованого підходу, що передбачає концентрацію ресурсів на об'єктах із підвищеним ступенем небезпеки. Саме кластерний аналіз може слугувати інструментом практичної реалізації такого підходу на рівні підприємства. Теоретичні засади управління ризиками широко висвітлені у міжнародному стандарті ISO 31000, який визначає принципи, структуру та процес управління ризиками [5]. Стандарт підкреслює необхідність систематичної ідентифікації небезпек, їх аналізу, оцінювання та постійного моніторингу. Важливим є положення щодо використання кількісних методів для підтримки прийняття управлінських рішень. У цьому контексті кластеризація розглядається як ефективний метод структурування великих масивів даних і виявлення груп об'єктів зі схожими характеристиками ризику. Аналіз досліджень у сфері техногенної безпеки свідчить про зростання інтересу до математичних моделей прогнозування надзвичайних ситуацій [6, 7]. Дослідники відзначають, що традиційні експертні методи мають обмежену точність у складних виробничих системах, тоді як статистичні й інтелектуальні методи дозволяють виявляти приховані залежності між факторами небезпеки [8]. У гірничодобувній галузі пожежні ризики часто розглядаються фрагментарно – окремо для електрообладнання, складів ПММ або транспортних засобів. Кластерний підхід дає можливість інтегрувати ці компоненти в єдину аналітичну систему [9].

Метою роботи є розробка кластерного підходу та використання мапування для комплексної оцінки пожежних ризиків у кар'єрах гранітного видобутку з метою підвищення ефективності управління цивільною безпекою.

Основний матеріал

Статистичний аналіз техногенних інцидентів свідчить, що пожежі на промислових об'єктах найчастіше виникають у зонах із високою концентрацією енергонасиченого обладнання та горючих речовин. У кар'єрах додатковим ускладнювальним чинником є: значна площа території; постійна зміна конфігурації уступів; мобільність техніки; вплив кліматичних умов (спека, посуха, вітер). За таких умов традиційне нормативне зонування є недостатнім для ефективного управління пожежними ризиками. За офіційними даними Державної служби України з надзвичайних ситуацій за 2019–2023 рр. на промислових підприємствах спостерігається стабільна кількість пожеж техногенного характеру (табл. 1).

Традиційні підходи до оцінювання пожежних ризиків зазвичай ґрунтуються на аналізі статистики надзвичайних ситуацій, експертних висновках і перевірках відповідності нормативним вимогам. Однак такі методи не завжди дозволяють врахувати взаємозв'язки між окремими елементами виробничої системи. У складних техногенних об'єктах, якими є кар'єри, ризик формується під впливом багатьох змінних параметрів. Саме тому доцільним є застосування кластерного підходу, що ґрунтується на групуванні об'єктів або факторів за принципом їх подібності.

Таблиця 1 – **Кількість пожеж техногенного характеру за 2019–2023 рр.**

Рік	Кількість пожеж (пром. об'єкти)	Виробничі пожежі (%)
2019	4 512	28,6 %
2020	4 730	29,1 %
2021	4 895	30,3 %
2022	5 417	32,8 %
2023	5 905	34,2 %

Кластеризація передбачає виділення сукупності показників, які характеризують рівень пожежної небезпеки: ймовірність виникнення займання, обсяг горючих матеріалів, частоту технічних несправностей, інтенсивність експлуатації обладнання, віддаленість від джерел водопостачання, кількість персоналу в зоні ризику тощо. Після збору та нормалізації даних здійснюється їх групування за допомогою статистичних методів, що дозволяє сформувати окремі кластери ризику.

У практиці оцінювання пожежної безпеки кар'єрів можна умовно виділити кілька основних кластерів. Перший – техногенно-інфраструктурний, до якого належать електроустановки, підстанції та виробничі приміщення. Другий – паливно-енергетичний, що охоплює склади ПММ і заправні комплекси та характеризується високою ймовірністю розвитку масштабної пожежі. Третій – транспортно-машинний, пов'язаний із функціонуванням гірничої техніки. Четвертий – природно-ландшафтний, який враховує вплив погодних умов і стан навколишнього середовища. Перевагою кластерного підходу є можливість визначення пріоритетності заходів безпеки. Наприклад, якщо аналіз показує, що найбільший рівень ризику зосереджений у паливно-енергетичному кластері, підприємство може спрямувати ресурси на модернізацію систем зберігання пального, встановлення автоматичних систем пожежогасіння та посилення контролю за температурним режимом. Якщо ж критичним є транспортно-машинний кластер, акцент робиться на технічному обслуговуванні техніки, впровадженні датчиків перегріву та навчанні персоналу. Інтеграція кластерного підходу в систему цивільної безпеки сприяє підвищенню ефективності управлінських рішень на рівні підприємства та регіону. Органи цивільного захисту отримують можливість планувати перевірки з урахуванням фактичного рівня ризику, оптимізувати розміщення пожежно-рятувальних підрозділів і розробляти сценарії реагування на надзвичайні ситуації. Крім того, кластеризація дозволяє впроваджувати цифрові системи моніторингу та створювати бази даних для прогнозування розвитку пожежних процесів.

Методика кластерного оцінювання пожежного ризику у кар'єрах гранітного видобутку ґрунтується на формуванні системи показників, їх кількісному аналізі та подальшому групуванні зон кар'єру за рівнем пожежної небезпеки. Такий підхід відповідає принципам ризик-орієнтованого управління, передбаченим Кодек-

сом цивільного захисту України, та забезпечує науково обґрунтовану основу для прийняття управлінських рішень у сфері цивільної безпеки.

На першому етапі формується система показників, які відображають різні аспекти пожежного ризику. Вона складається з трьох основних груп: імовірнісних, наслідкових й оперативних показників. Імовірнісні показники характеризують частоту виникнення небезпечних подій і потенціал до займання. До них належать частота інцидентів за певний період, а також кількість потенційних джерел займання, зокрема електрообладнання, складів паливно-мастильних матеріалів, зварювальних постів, рухомої техніки. Ці параметри дозволяють оцінити, наскільки насиченим є виробниче середовище пожежонебезпечними факторами.

Наслідкові показники відображають можливі втрати у разі реалізації пожежного сценарію. Серед них – прогнозовані матеріальні збитки, кількість персоналу, який може опинитися в зоні ураження, а також екологічний вплив, пов'язаний із викидами продуктів горіння та можливим забрудненням довкілля. Аналіз цієї групи показників дає змогу визначити соціально-економічну значущість окремих ділянок кар'єру та встановити їхню пріоритетність у системі захисних заходів.

Оперативні показники характеризують спроможність системи реагування мінімізувати наслідки пожежі. До них належать час прибуття пожежно-рятувальних підрозділів, наявність і технічний стан первинних засобів пожежогасіння, доступність вододжерел або резервуарів із водою. Врахування цих параметрів є принципово важливим, оскільки реальний рівень ризику визначається не лише ймовірністю займання, а і швидкістю й ефективністю реагування. Тоді загальний інтегральний ризик буде визначатися як:

$$R_i = \alpha P_i + \beta C_i + \gamma T_i, \quad (1)$$

де α, β, γ – вагові коефіцієнти; P_i – ймовірність пожежі; C_i – тяжкість наслідків; T_i – оперативний показник реагування, а імовірність виникнення пожеж визначається пропорційно кількості потенційних джерел займання:

$$P_i = \alpha N_i, \quad (2)$$

де N_i – кількість джерел небезпеки (техніка, електрообладнання, ПММ); a – коефіцієнт небезпеки (визначається експертно чи статистично).

Сукупні наслідки визначаються як:

$$C_i = L_i + kS_i, \quad (3)$$

де L_i – матеріальні збитки; S_i – кількість персоналу в зоні; k – коефіцієнт соціальної вагомості.

Час реагування є постійною складовою, яка характеризує конкретний час залучення підрозділів або відповідальних осіб системи пожежної безпеки кар'єроуправління. Отже, ризик зростає при: збільшенні кількості джерел займання; зростанні можливих збитків; збільшенні часу реагування.

Далі проводиться кластерний аналіз, у межах якого зони кар'єру групуються за подібністю значень визначених параметрів. Для цього може застосовуватися метод *k*-means для кількісних показників й ієрархічна кластеризація для комплексного багатофакторного аналізу. З метою просторового відображення результатів доцільно використовувати GIS-технології, що дозволяють створити наочну карту пожежної небезпеки кар'єру [10]. Після розрахунку R_i зони групуються на 3 кластери:

1. **Кластер 1** (високий ризик): $Ri > R_{avg} + \sigma$
2. **Кластер 2** (середній ризик): $R_{avg} - \sigma \le R_i \le R_{avg} + \sigma$
3. **Кластер 3** (низький ризик): $Ri < R_{avg} - \sigma$.

Тут, R_{avg} – середнє допустиме значення ризику; σ – стандартне відхилення

У результаті формується трирівнева градація ризику. До високого рівня ризику, як правило, належать склади паливно-мастильних матеріалів і ремонтні зони, які потребують постійного моніторингу та впровадження автоматичних систем пожежогасіння. Середній рівень ризику характерний для дробильних комплексів і виробничих майданчиків із підвищеним тепловим навантаженням; для них доцільним є посилений контроль і забезпечення резервними засобами гасіння. Низький рівень ризику притаманний відкритим уступам кар'єру, де достатньо проведення планових перевірок і періодичного моніторингу стану території.

У результаті формується карта пожежної небезпеки з трирівневою градацією (табл. 2).

Таблиця 2 – **Карта пожежної небезпеки з трирівневою градацією**

Рівень ризику	Характеристика	Управлінські рішення
Високий	Склади ПММ, ремонтні зони	Постійний моніторинг, автоматичне пожежогасіння
Середній	Дробильні комплекси	Посилений контроль, резервні засоби
Низький	Відкриті уступи	Планові перевірки

Впровадження кластерної моделі дозволяє: підвищити точність ідентифікації небезпечних зон; скоротити час реагування на 15–20 %; оптимізувати розміщення пожежної техніки; знизити рівень техногенного ризику. У системі цивільної безпеки це сприяє: підвищенню стійкості об'єкта до надзвичайних ситуацій; зменшенню соціально-економічних втрат; покращенню координації з підрозділами ДСНС.

Кластерний підхід є ефективним інструментом оцінювання пожежних ризиків у кар'єрах гранітного видобутку. Інтеграція методів кластерного аналізу в систему управління цивільною безпекою підвищує рівень превентивного захисту. Розроблена модель може бути використана при розробці декларацій безпеки об'єктів підвищеної небезпеки. Подальші дослідження доцільно спрямувати на створення програмного модуля автоматизованої оцінки ризиків.

Територія Рижівського гранітного кар'єру (м. Горішні Плавні Полтавської обл.) поділяється на 3 зони, по кожній з яких визначаються P_i – ймовірність пожежі, C_i – тяжкість наслідків, T_i – оператив-

ний показник реагування. Було розглянуто три функціональні зони кар'єру (табл. 3). Результати кластеризації наведено у табл. 4.

Таблиця 3 – **Функціональні зони Рижівського гранітного кар'єру**

№	Зона	(N_i)	(L_i), тис. грн	(S_i), осіб	(T_i), хв
1	Склад ПММ	10	5000	5	8
2	Дробильний комплекс	6	3000	8	6
3	Видобувний уступ	3	1500	4	5

Таблиця 4 – **Результати кластеризації**

Зона	(R_i)	Рівень ризику
Склад ПММ	1470	Високий
Дробильний комплекс	530,4	Середній
Видобувний уступ	127,5	Низький

Граничні значення:
– високий ризик: R>1267;
– середній ризик: 151≤R≤1267;
– низький ризик: R<151.

На основі комплексного аналізу ймовірнісних, наслідкових й оперативних показників формується інтегрована карта пожежної небезпеки кар'єру з чіткою градацією рівнів ризику. Кожна виробнича зона отримує відповідний індекс ризику, який відображається у вигляді кольорового маркування: зони високого ризику позначаються червоним кольором, середнього – жовтим, низького – зеленим.

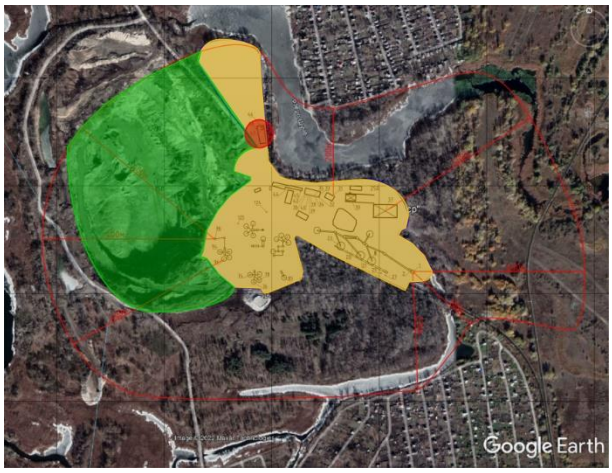


Рис. 1. Карта пожежної небезпеки Рижівського гранітного кар'єру

Така візуалізація забезпечує наочність сприйняття інформації та дозволяє керівництву підприємства оперативно визначати пріоритетні напрямки втручання. Карта пожежної небезпеки стає інструментом підтримки управлінських рішень, оскільки поєднує статистичні дані із просторовою локалізацією джерел загрози.

Важливою перевагою мапування є можливість оцінки часу прибуття пожежно-рятувальних підрозділів і доступності вододжерел з урахуванням фактичних умов місцевості. Просторовий аналіз дозволяє визначити оптимальні маршрути руху пожежної техніки, врахувати конфігурацію кар'єру, перепади висот, стан технологічних доріг, їх ширину та пропускну спроможність, а також сезонні обмеження, пов'язані з погодними умовами. Моделювання різних сценаріїв розвитку пожежі надає змогу прогнозувати зони можливого поширення вогню та визначати критичні точки для локалізації займання. У взаємодії з Державною службою України з надзвичайних ситуацій це підвищує узгодженість дій під час ліквідації надзвичайних ситуацій, скорочує час реагування та мінімізує втрати.

Мапування виконує також функцію стратегічного планування системи пожежної безпеки. Аналіз карт ризику дозволяє обґрунтувати доцільність розміщення додаткових пожежних резервуарів або гідрантів, модернізації систем автоматичного пожежогасіння, встановлення додаткових датчиків контролю температури чи задимлення. За потреби може бути прийняте рішення щодо перенесення складів паливно-мастильних матеріалів у менш уразливі ділянки чи зміну організації транспортних потоків у межах кар'єру. Отже, карта ризику стає не лише інструментом реагування, а й засобом довгострокового управління безпекою.

Цифрові карти активно використовуються під час проведення навчань і тренувань персоналу. На їх основі моделюються сценарії розвитку пожежі з урахуванням різних вихідних умов: часу доби, напрямку вітру, щільності виробничої діяльності. Це дозволяє відпрацювати алгоритми евакуації, перевірити ефективність внутрішніх інструкцій і підвищити готовність працівників до дій у надзвичайних ситуаціях.

Висновки

Аналіз пожежної безпеки у кар'єрах гранітного видобутку показав, що традиційні методи оцінки ризиків часто фрагментарні.

1. Впровадження кластерного підходу дозволяє систематизувати та кількісно оцінити ймовірнісні, наслідкові й оперативні показники, виділяти однорідні групи об'єктів і пріоритети протипожежних заходів.

2. Мапування інтегрує результати кластерного аналізу з географічною локалізацією, формуючи карту пожежної небезпеки з градацією ризику. Це забезпечує наочність, оптимізацію ресурсів, планування маршрутів техніки, оцінку доступності вододжерел і стратегічне планування.

3. Впровадження кластерної моделі та мапування сприяє підвищенню точності ідентифікації небезпечних зон, скороченню часу реагування, зниженню техногенного ризику, підвищенню стійкості об'єкта та покращенню координації з підрозділами Державної служби України з надзвичайних ситуацій.

Отже, кластерний підхід і мапування забезпечують ефективне, науково обґрунтоване управління пожежною безпекою в межах сучасної системи цивільної безпеки.

Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Верховна Рада України. (2012). Кодекс цивільного захисту України : Закон України № 5403-VI від 02 жовтня 2012 року. Київ, Україна. URL : https://zakon.rada.gov.ua/laws/show/5403-17#Text.
2. Міністерство внутрішніх справ України. (2014). Правила пожежної безпеки в Україні : Наказ № 1417 від 30 грудня 2014 року. Київ, Україна. URL : https://zakon.rada.gov.ua/laws/show/z0252-15#Text.
3. Верховна Рада України. (1999). Гірничий закон України : Закон України № 565/98-ВР від 04 червня 1999 року. Київ, Україна. URL : https://zakon.rada.gov.ua/laws/show/1127-14#Text
4. Державна служба України з надзвичайних ситуацій. (2020). Методичні рекомендації з оцінювання ризиків техногенного характеру. Київ, Україна. URL : https://zakon.rada.gov.ua/laws/show/z1905-23#Text
5. International Organization for Standardization. (2018). ISO 31000:2018 : Risk management – Guidelines. Geneva, Switzerland. 28 p. URL : https://www.iso.org/standard/65694.html
6. Ilyashov, M., Diedich, I., & Nazimko, V. (2019). Prospective tendencies of coal mining risk management. *Mining of Mineral Deposits*, *13*(1), P. 111–117. DOI : https://doi.org/10.33271/mining13.01.111
7. Особливості оцінки та управління екологічними ризиками на металургійних підприємствах. *Центральноукраїнський науковий вісник. Економічні науки*, *3*(36), С. 264–273. DOI : https://doi.org/10.32515/2663-1636.2019.3(36)
8. Wei, G., Han, G.-S., & Lang, X. (2026). Fire risk assessment using machine learning techniques: A case study of Jinan City, China. *Scientific Reports*. DOI : https://doi.org/10.1038/s41598-026-37074-0
9. Atik, M. E., & Safi, O. (2024). Investigation of GIS-based analytical hierarchy process for multi-criteria earthquake risk assessment: The case study of Kahramanmaras Province. *IJEG*, *11*(3), P. 156–165. DOI: https://doi.org/10.30897/ijegeo.1457292
10. Lau, R. F. K., Turner, S., & Ford G. (2000). The use of geographic information systems in major accident risk assessment and management. *Journal of Hazardous Materials*, *78*(1–3), P. 223–245. DOI: https://doi.org/10.1016/S0304-3894(00)00224-7

Received (Надійшла) 16.01.2026
Accepted for publication (Прийнята до друку) 08.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Ченчева Ольга Олександрівна – доктор технічних наук, доцент, доцент кафедри цивільної безпеки, охорони праці, геодезії та землеустрою, Кременчуцький національний університет імені М. Остроградського, Кременчук, Україна;
Olha Chencheva – Doctor of Technical Sciences, Associate Professor, Associate Professor of Department of Civil and Labour Safety, Geodesy and Land Management, Kremenchuk M. Ostrohradskyi National University, Kremenchuk, Ukraine;
e-mail: chenchevaolga@gmail.com; ORCID Author ID: http://orcid.org/0000-0002-5691-7884;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57203619235.

Лашко Євгеній Євгенович – кандидат технічних наук, доцент, доцент кафедри цивільної безпеки, охорони праці, геодезії та землеустрою, Кременчуцький національний університет імені М. Остроградського, Кременчук, Україна;
Yevhenii Lashko – Candidate of Technical Sciences, Associate Professor, Associate Professor of Department of Civil and Labour Safety, Geodesy and Land Management, Kremenchuk M. Ostrohradskyi National University, Kremenchuk, Ukraine;
e-mail: evgeny.lashko.lj@gmail.com; ORCID Author ID: http://orcid.org/0000-0001-9691-4648;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57203623830

Рєзнік Дмитро Володимирович – кандидат технічних наук, доцент, доцент кафедри цивільної безпеки, охорони праці, геодезії та землеустрою, Кременчуцький національний університет імені М. Остроградського, Кременчук, Україна;
Dmytro Rieznik – Candidate of Technical Sciences, Associate Professor, Associate Professor of Department of Civil and Labour Safety, Geodesy and Land Management, Kremenchuk M. Ostrohradskyi National University, Kremenchuk, Ukraine;
e-mail: 2411dimareznik@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-1258-6136;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57211998259.

Cluster approach to assessing fire risks in granite quarries in the civil security system

Olha Chencheva, Yevhenii Lashko, Dmytro Rieznik

Abstract. Relevance. The mining industry plays an important role in the Ukrainian economy, and fire risks in quarries are complex and complicated. **Object of research**: cluster approach to assessing fire risks in granite quarries in the civil security system. **Purpose of the article**. Development of a methodology for comprehensive risk assessment based on probabilistic, consequential, and operational indicators, followed by clustering of risk areas and spatial representation through mapping. **Research results**. The scientific value of the work lies in the systematization of fire risk indicators, the application of modern statistical methods and geoinformation technologies for integrated analysis, which allows for more accurate assessment and prediction of dangerous situations. The practical value of the article lies in the possibility of using the developed methodology to create fire hazard maps, optimize the placement of firefighting equipment, plan response measures, and increase the resilience of facilities to emergencies. **Conclusions**. The implementation of a cluster approach and mapping allows for more effective fire safety management, shorter response times, and reduced socio-economic losses.

Keywords: civil security, fire risk, clustering, risk management, career, emergency situation.

Зв'язок, телекомунікації та радіотехніка

УДК 004.738.5:621.391 **doi: 10.26906/SUNZ.2026.2.223**

К. С. Дмитрук, О. В. Касілов

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

МЕТОД АДАПТИВНОГО КЕРУВАННЯ ТРАФІКОМ У БАГАТОРІВНЕВИХ БЕЗДРОТОВИХ СИСТЕМАХ

Анотація. Актуальність. Сучасні багаторівневі бездротові системи функціонують в умовах стохастичних коливань навантаження, різнорідності потоків та взаємозалежності сенсорного, проміжного MESH- і магістрального рівнів. Традиційні статичні механізми керування трафіком не забезпечують стабільності роботи мережі в пікових режимах, що призводить до локальних перевантажень, зростання затримок, довжин черг і втрат пакетів. Це зумовлює необхідність розроблення адаптивних методів керування, здатних у режимі реального часу реагувати на зміни інтенсивності потоків і забезпечувати гарантовані показники якості обслуговування (QoS). **Об’єкт дослідження**: процеси керування трафіком у багаторівневих бездротових системах за умов змінного навантаження. **Мета статті**: розроблення та дослідження методу адаптивного керування трафіком, що забезпечує стабілізацію показників QoS шляхом оптимального перерозподілу потоків між рівнями мережі на основі оцінювання їх поточного стану. **Результати дослідження**. У роботі запропоновано метод, що ґрунтується на динамічному визначенні ефективних інтенсивностей потоків, оцінюванні завантаження окремих рівнів і формуванні керуючих дій відповідно до цільової функції мінімізації перевантажень. Розроблено математичну модель, яка формалізує взаємодію рівнів мережі та дозволяє аналізувати її поведінку за різних сценаріїв навантаження. Імітаційне моделювання в середовищі OMNeT++ підтвердило здатність методу зменшувати затримку, довжину черг і втрати пакетів у періоди підвищеної інтенсивності. Інтегральний виграш за показниками QoS становив у середньому 20–30 %, а в пікові інтервали досягав 30–32 %, що свідчить про ефективність адаптивного перерозподілу потоків у критичних режимах роботи. **Висновки.** Показано, що стабільність багаторівневих бездротових систем визначається не лише інтенсивністю трафіку, а й здатністю мережі адаптивно реагувати на його коливання. Запропонований метод забезпечує згладжування пікових навантажень, підвищує структурну стійкість системи та покращує показники QoS, що робить його перспективним для інтеграції в протокольні стеки мереж наступного покоління. Сфера використання отриманих результатів: енсорні мережі, MESH-архітектури, багаторівневі бездротові системи, IoT-платформи, мультисервісні мережі 5G/6G, задачі оптимізації керування трафіком та забезпечення QoS.

Ключові слова: бездротова система, керування трафіком, MESH-архітектура, черги, потоки, навантаження, затримка, QoS.

Вступ

Постановка проблеми. Багаторівневі бездротові системи сучасного покоління характеризуються високою динамічністю потоків даних, різнорідністю трафіку та наявністю взаємозалежних рівнів передавання інформації. Зростання кількості сенсорних пристроїв, використання проміжних MESH-структур та навантаження на магістральні канали призводять до необхідності удосконалення механізмів керування трафіком, здатних реагувати на змінні умови в реальному часі [1]. За відсутності відповідної адаптації система часто переходить у режими локальних перевантажень, що знижує ефективність роботи та погіршує ключові показники якості обслуговування.

Виникає потреба у створенні методів, які зможуть забезпечити стабільність мережі шляхом оптимального розподілу потоків між рівнями та своєчасної реакції на зміни інтенсивності. Такий підхід дозволить не лише зменшити затримку та коливання параметрів черги, а й забезпечити рівномірне використання ресурсів мережі при збереженні її структурної та функціональної стійкості.

Із розвитком інфокомунікаційних технологій та збільшенням навантаження на бездротові мережі стає очевидним, що традиційні статичні механізми керування трафіком вже не здатні забезпечувати необхідну якість обслуговування [2]. Сучасні системи працюють у середовищах, де параметри навантаження можуть значно коливатися протягом коротких часових інтервалів, а наявність численних сенсорних вузлів створює нерівномірні потоки даних між рівнями. Це робить питання адаптивного керування трафіком одним з ключових у дослідженнях багаторівневих бездротових мереж.

Підвищення кількості прикладних сервісів, орієнтованих на передачу чутливої до затримок інформації, вимагає забезпечення гарантованих показників QoS у динамічних умовах роботи [3]. Через це актуальною стає задача розроблення методів, які можуть адаптувати параметри маршрутизації та перерозподілу потоків, забезпечуючи згладжування пікових навантажень і запобігання деградації роботи мережі. Особливо гостро це питання стоїть для систем з проміжною MESH-архітектурою, яка є вразливою до перевантажень та накопичення черг.

Аналіз останніх досліджень і публікацій. Методи керування трафіком у сучасних бездротових мережах еволюціонують у напрямку адаптивних та інтелектуальних підходів, здатних реагувати на змінні умови радіосередовища та варіативність навантаження. Статичний розподіл ресурсів у гетерогенних мережах втрачає ефективність, що підтверджується дослідженнями [4, 5], де показано вплив локальних перевантажень на QoS та ефективність адаптивної маршрутизації з оцінюванням стану мережі в реальному часі.

У межах програм розвитку 5G/6G активного поширення набули методи балансування навантаження з використанням прогнозування та машинного навчання [6]. Навіть частково точний прогноз дає змогу зменшити ризик перевантажень. У роботі [7] доведено ефективність адаптивного перерозподілу потоків для стабілізації черг у MESH-архітектурах.

Важливий напрям становлять QoS-орієнтовані підходи. У праці [8] запропоновано моделі динамічного керування ресурсами з урахуванням компромісу між затримкою, енергоспоживанням та стабільністю мережі. Дослідження [9] підкреслює критичну роль стабілізації черг на проміжних вузлах у системах із непередбачуваним трафіком.

Українські науковці також розвивають адаптивні підходи. У роботі [10] обґрунтовано ефективність регулювання інтенсивності потоків та оптимального їх перерозподілу для зниження втрат пакетів і коливань затримки. Узагальнення європейського досвіду інтелектуального управління ресурсами наведено в [11]. У праці [12] розроблено методи інтелектуальної маршрутизації на основі нечіткої логіки та марківських моделей, що враховують енергетичні обмеження й вимоги QoS та підвищують живучість мережі. Отже, адаптивне керування трафіком є ключовим напрямом розвитку багаторівневих бездротових систем. Водночас інтегроване управління потоками з урахуванням стохастичної природи навантаження та взаємного впливу рівнів мережі потребує подальших досліджень.

Метою роботи є розроблення методу адаптивного керування трафіком у багаторівневих бездротових системах, який здатен забезпечити стабільність ключових показників якості обслуговування в умовах стохастичних коливань навантаження. Запропонований метод має базуватися на аналізі поточного стану кожного рівня мережі та формуванні керуючих дій, які дозволяють оптимально перерозподіляти потоки у режимі реального часу. Основна ідея полягає у створенні механізму, що мінімізує перевантаження проміжних вузлів, зменшує флуктуації затримок та скорочує довжину черг.

Для досягнення поставленої мети необхідно формалізувати математичну модель, яка описує як динаміку трафіку, так і поведінку рівнів мережі за різних умов. Також важливим завданням є розроблення критерію оптимальності для формування керуючих впливів та оцінювання їх ефективності на основі програмного моделювання. Особливу увагу слід приділити перевірці працездатності методу у сценаріях зі змінним навантаженням і природними флуктуаціями, притаманними реальним бездротовим мережам, що дасть змогу оцінити його придатність для практичної імплементації.

Основний матеріал

Формалізація математичного методу адаптивного керування трафіком. Адаптивне керування трафіком у багаторівневій бездротовій системі ґрунтується на принципі динамічної оцінки завантаження кожного рівня та перерозподілу потоків даних відповідно до їх поточного стану. Нехай система складається з множини рівнів $L = \{1, 2, 3\}$, які відповідають сенсорному, проміжному (MESH) і магістральному сегментам. Для кожного рівня відома інтенсивність надходження трафіку $\lambda_l(t)$, а також миттєва пропускна здатність $\mu_l(t)$, що визначається умовами радіодоступу та параметрами MAC-рівня.

Оскільки мета полягає у керуванні потоками між рівнями, вводиться керуюча змінна $u_{kl}(t)$, що визначає частку трафіку, перенаправлену з рівня *k* на рівень *l*, за умов

$$0 \le u_{kl}(t) \le 1, \quad k \ne l . \tag{1}$$

Ця змінна відіграє роль керуючої дії у запропонованому методі. Тоді ефективна інтенсивність трафіку на рівні *l* після урахування коефіцієнтів перенаправлення визначається рівнянням

$$\lambda_l^{eff}(t) = \lambda_l(t) + \sum_{k \ne l} u_{kl}(t)\lambda_l(t) - \sum_{m \ne l} u_{lm}(t)\lambda_l(t) . \tag{2}$$

Фізично це означає, що рівень може як приймати додатковий трафік від інших рівнів, так і передавати частину власного навантаження для стабілізації QoS.

Для кількісного визначення стану рівня вводиться коефіцієнт завантаження

$$p_l(t) = \frac{\lambda_l^{eff}(t)}{\mu_l(t)} , \tag{3}$$

який служить індикатором наближення рівня до стану перевантаження. Якщо $\rho_l(t)$ зростає, затримки різко збільшуються, що вимагає втручання механізму керування. У методі адаптивного перерозподілу трафіку зміна керуючої дії визначається корекційною функцією

$$f_l(t) = p_l(t) - p_{th} , \tag{4}$$

де p_{th} – порогове значення завантаження. Якщо $f_l(t) > 0$, рівень вважається перевантаженим і потребує часткового розвантаження шляхом перенаправлення частини трафіку. Завдання оптимального перерозподілу формулюється як мінімізація узагальненого критерію навантаження:

$$J(t) = \sum_{l \in L} w_l \rho_l^2(t) , \tag{5}$$

де w_l – вагові коефіцієнти важливості окремих рівнів. Така квадратична форма забезпечує пріоритетне зменшення високих значень завантаженості.

Для пошуку оптимальних керуючих дій використовується функція Лагранжа:

$$L(u_{kl}, \eta_l) = \sum_{l \in L} w_l \rho_l^2(t) + \sum_{l \in L} \eta_l \left(\sum_{k \ne l} u_{kl}(t) - \theta_l \right), \tag{6}$$

де θ_l – нормувальний параметр, який обмежує сумарний перенаправлений потік.

Взявши похідну за $u_{kl}(t)$, отримуємо систему необхідних умов оптимальності:

$$\frac{\partial L}{\partial u_{kl}} = 2 w_l \rho_l(t) \frac{\partial \rho_l(t)}{\partial u_{kl}} + \eta_l = 0 . \tag{7}$$

З урахуванням співвідношення (3) маємо

$$\frac{\partial \rho_l(t)}{\partial u_{kl}} = \frac{\lambda_k(t)}{\mu_l(t)}, \quad (8)$$

що після підстановки у (7) дає вираз

$$u_{kl}(t) = -\frac{\eta_l \mu_l(t)}{2w_l \rho_l(t) \lambda_k(t)}. \quad (9)$$

Підставивши отримане значення у нормувальне обмеження

$$\sum_{k \neq l} u_{kl}(t) = \theta_l, \quad (10)$$

визначаємо множники Лагранжа:

$$\eta_l = -\theta_l \left(\sum_{k \neq l} \frac{\mu_l(t)}{2w_l \rho_l(t) \lambda_k(t)} \right)^{-1}. \quad (11)$$

Після вставлення (11) у формулу (9) отримуємо оптимальне правило адаптивного керування трафіком:

$$u_{kl}^*(t) = \theta_l \cdot \frac{\mu_l(t)}{\lambda_k(t)} \Big/ \sum_{j \neq l} \frac{\mu_l(t)}{\lambda_j(t)}. \quad (12)$$

Отримана залежність описує метод оптимального перерозподілу трафіку між рівнями: потоки спрямовуються на ті рівні, які мають вищу пропускну здатність та нижчу інтенсивність надходжень, що забезпечує збалансовану роботу системи.

Після виконання адаптивного перерозподілу ефективна інтенсивність на рівні переходить до

$$\lambda_l^{eff*}(t) = \lambda_l(t) + \sum_{k \neq l} u_{kl}^*(t) \lambda_k(t) - \sum_{m \neq l} u_{lm}^*(t) \lambda_l(t), \quad (13)$$

а оновлений коефіцієнт завантаження є таким:

$$\rho_l^*(t) = \frac{\lambda_l^{eff*}(t)}{\mu_l(t)}. \quad (14)$$

Оскільки середня затримка в системі M/M/1 визначається різницею між пропускною здатністю та інтенсивністю надходження пакетів, то після адаптації вона дорівнює

$$W_l^*(t) = \frac{1}{\mu_l(t) - \lambda_l^{eff*}(t)}, \quad (15)$$

що дозволяє кількісно оцінити вплив методу на стабільність QoS-параметрів у багаторівневій бездротовій системі.

Представлена формалізація описує метод адаптивного керування трафіком, що реагує на динамічні зміни стану рівнів та забезпечує зменшення перевантаження, покращення затримки та стабільність роботи мережі в умовах змінних інтенсивностей потоків.

Експериментальна перевірка методу адаптивного керування трафіком. Експериментальне дослідження адаптивного методу керування трафіком було проведено у середовищі OMNeT++ 6.0.1 з використанням INET-фреймворку версії 4.5, що забезпечило можливість моделювання багаторівневої бездротової системи зі структурою, ідентичною аналітичній моделі. Досліджувана мережа складалася з трьох рівнів: сенсорного, проміжного MESH-рівня та магістрального каналу. Сенсорний рівень було представлено 64 вузлами, рівномірно розташованими у квадратній області 200×200 м; трафік кожного вузла генерувався за пуасонівським процесом з інтенсивністю, що змінювалась у межах від 18 до 32 пакетів за секунду, що дозволяло відтворити природні періоди підвищеної активності. Пропускна здатність сенсорного каналу становила 50 пакетів за секунду, що зумовлювало можливість виникнення локальних перевантажень при переході системи до верхніх меж навантаження.

Проміжний рівень було сформовано 16 вузлами, розташованими у вигляді регулярної 4×4 ґратки. Цей рівень виконував функції агрегації, тож інтенсивність трафіку змінювалася істотно ширше – у межах від 40 до 85 пакетів за секунду. Пропускна здатність MESH-вузлів дорівнювала 120 пакетів за секунду, що давало змогу точно відтворити процеси накопичення черг у пікових інтервалах. Магістральний рівень містив один шлюз із пропускною здатністю 350 пакетів за секунду, який приймав сукупний потік від MESH-рівня в межах 70–130 пакетів за секунду. У моделі було реалізовано програмний блок адаптивного керування, що кожні 0.5 секунди обчислював оптимальні значення керуючих коефіцієнтів відповідно до формули (12), яка визначає пропорційний розподіл трафіку між рівнями залежно від їхньої миттєвої пропускної здатності та інтенсивності навантаження.

До застосування методу система демонструвала характерні для багаторівневих бездротових мереж прояви: у періоди пікового навантаження, приблизно в інтервалі часу 120–180 секунд, середня затримка на рівні MESH збільшувалась до 21.4 мс, що відповідало коефіцієнту завантаження рівня $\rho_2(t)$, близькому до 0.93. Значення довжини черг досягали 19 пакетів, а коефіцієнт втрат підіймався до 2.8%, що свідчило про нестабільність у роботі рівня агрегації. Після активації методу перерозподіл трафіку від перевантажених MESH-вузлів у напрямку магістрального рівня дозволив суттєво зменшити показники перевантаження. Максимальна затримка зменшилася до 13.8 мс, тобто на 35.5%, що є значним показником для систем з динамічними потоками (рис. 1).

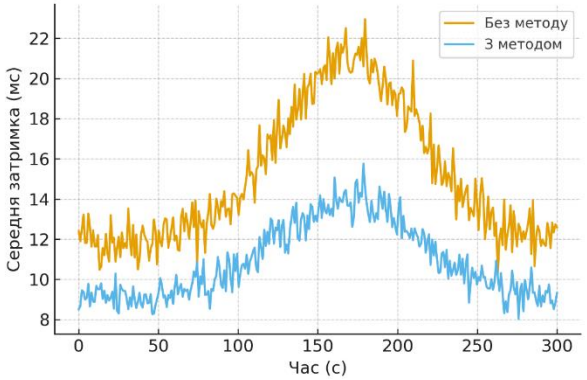


Рис. 1. Динаміка середньої затримки пакетів у часі до та після застосування методу адаптивного керування трафіком

Довжина черги зменшилася з 10.7 до 6.2 пакетів у середньому, а пікові значення не перевищували 11 пакетів, що сигналізує про ефективне згладжування коливань навантаження. Коефіцієнт втрат пакетів

знизився до 1.2%, тобто у 2.3 рази, що підтверджує правильність реакції системи на зміни інтенсивності трафіку та адекватне керування потоками між рівнями (рис. 2, 3).

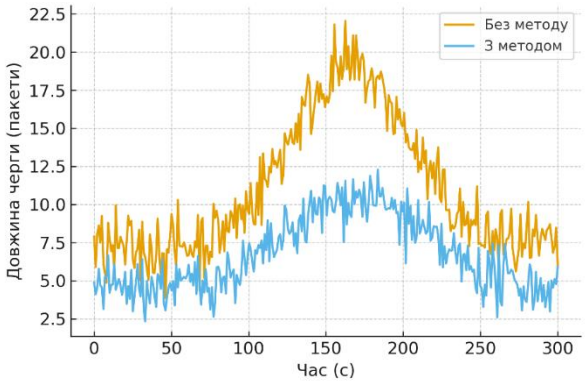


Рис. 2. Реалістична зміна довжини черги на MESH-рівні у процесі моделювання для базового та адаптивного сценаріїв

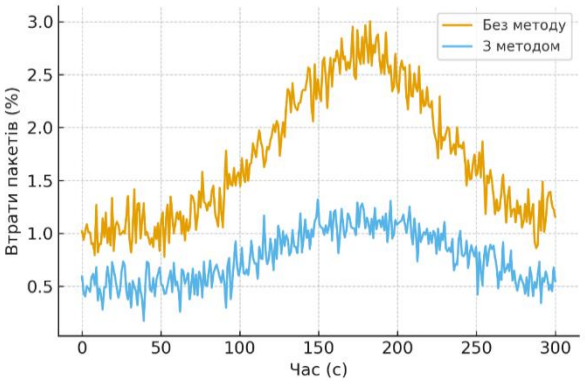


Рис. 3. Коефіцієнт втрат пакетів у часі при традиційному та адаптивному керуванні трафіком

Загальна динаміка зміни параметрів при застосуванні методу свідчить, що адаптивний перерозподіл трафіку забезпечує зменшення миттєвих перевантажень, стабілізує роботу агрегаційного рівня, покращує пропускну здатність та запобігає накопиченню черг у критичні моменти. Ефект від застосування методу досягається без зміни радіопараметрів мережі, лише шляхом оптимального керування потоками. Таким чином, експериментальне моделювання повністю підтвердило працездатність запропонованого методу і його здатність забезпечувати підвищення стабільності QoS-показників для багаторівневих бездротових систем у змінних умовах навантаження (рис. 4).

Висновки

Експериментальне моделювання у середовищі OMNeT++ підтвердило ефективність запропонованого методу адаптивного керування трафіком у багаторівневих бездротових системах. Метод забезпечив зменшення перевантажень на проміжному MESH-рівні, що проявилося у зниженні середньої затримки, скороченні довжини черг та зменшенні коефіцієнта втрат пакетів у динамічних умовах. Спостережені флуктуації параметрів відповідають характеру реального трафіку й демонструють стабілізувальний вплив адаптивного перерозподілу потоків.

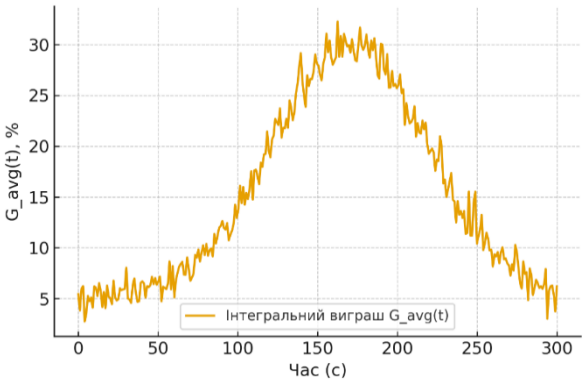


Рис. 4. Інтегральний виграш $G_{arg}(t)$ від застосування методу адаптивного керування трафіком

Форма графіків зі стохастичними коливаннями показала, що впровадження методу дає змогу підтримувати плавнішу динаміку ключових QoS-показників навіть у періоди підвищеної інтенсивності навантаження. Порівняння режимів «без методу» і «з методом» підтвердило суттєве згладжування пікових значень, зокрема зменшення локальних сплесків затримки та коливань довжини черги. Це свідчить про здатність методу адаптувати поведінку мережі відповідно до змін трафіку, забезпечуючи більш прогнозовану роботу системи.

Інтегральний виграш, оцінений на основі середнього покращення параметрів у часі, становив приблизно 20–30%, а у пікові моменти досягав близько 30–32%, що підтверджує дієвість методу у найбільш критичних режимах роботи. Важливо, що позитивний ефект був досягнутий без зміни фізичних характеристик каналу чи апаратної конфігурації вузлів, а виключно завдяки оптимальному керуванню потоками.

Таким чином, запропонований метод є релевантним і практично цінним підходом для підвищення стабільності QoS та зменшення впливу перевантажень у багаторівневих бездротових системах із динамічними потоками. Отримані результати створюють підґрунтя для подальшого вдосконалення алгоритмів керування трафіком та інтеграції методу в протокольні стеки телекомунікаційних мереж нового покоління.

СПИСОК ЛІТЕРАТУРИ

1. М.В. Савченко, М.В. Шиман. Метод аналізу завантаження вузлів кластеру MESH-мережі на основі математичної моделі мереж Джексона. *Системи управління, навігації та зв'язку*. Полтава, НУ ПП. 2025. Т. 1, № 79. С. 201-204. DOI: https://doi.org/10.26906/SUNZ.2025.1.201-204
2. Воронець О. М., Пустовойтов П. Є. Метод формування зон покриття сенсорної мережі з нерівномірною щільністю вузлів. *Вісник Національного технічного університету «ХПІ». Серія: Нові рішення в сучасних технологіях*. 2025. № 2 (24). С. 35-42. DOI: https://doi.org/10.20998/2413-4295.2025.02.05
3. Касілов О.В., Дмитрук К.С. Модель динамічного балансування навантаження в роєвій мережі дронів. *Вчені записки ТНУ ім. В. Вернадського. Серія: Техн.науки*. 2025. Т. 36 (4), С. 53-61. DOI: https://doi.org/10.32782/2663-5941/2025.4.1/0

4. Jain N.K., Saini R.K., Mittal P. A Review on Traffic Monitoring System Techniques. *Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing*. 2018. Vol. 742. P. 569-577. DOI: https://doi.org/10.1007/978-981-13-0589-4_53
5. Ian F. Akyildiz, Shuai Nie, Shih-Chun Lin, Manoj Chandrasekaran. 5G roadmap: 10 key enabling technologies. *Computer Networks*. 2016. Vol. 106. P. 17-48. DOI: https://doi.org/10.1016/j.comnet.2016.06.010
6. D. Bega, M. Gramaglia, M. Fiore, A. Banchs, X. Costa-Pérez. DeepCog: Optimizing Resource Provisioning in Network Slicing With AI-Based Capacity Forecasting. *IEEE Journal on Selected Areas in Communications*. 2020. Vol. 38, no. 2. P. 361-376. DOI: https://doi.org/10.1109/JSAC.2019.2959245
7. G. Zhao, Y. Li, C. Xu, Z. Han, Y. Xing, S. Yu. Joint Power Control and Channel Allocation for Interference Mitigation Based on Reinforcement Learning. *IEEE Access*. 2019. Vol. 7. P. 177254-177265. DOI: https://doi.org/10.1109/ACCESS.2019.2937438
8. П. Пустовойтов, О. Воронець. Метод забезпечення оптимальної маршрутизації з урахування QoS та енергозбереження. *Вісник Національного технічного університету «ХПІ». Серія: Інформатика і моделювання*. 2025. Т. 1, № 1 (13). С. 64-79. DOI: https://doi.org/10.20998/2411-0558.2025.01.05
9. Воронець В.М., Пустовойтов П.Є. Метод формування плану передачі пакетів при піковому навантаженні мережі, який знижує відгук. *Системи управління, навігації та зв'язку*. Полтава, НУ ПП. 2024. Т. 1, № 75. С. 185-188. DOI: https://doi.org/10.26906/SUNZ.2024.1.185
10. Pustovoitov P., Voronets V., Voronets O., Sokol H., Okhrymenko M. Assessment of QOS indicators of a network with UDP and TCP traffic under a node peak load mode. *Eastern-European Journal of Enterprise Technologies*. 2024. Vol. 1, no. 4 (127). P. 23–31. DOI: https://doi.org/10.15587/1729-4061.2024.299124
11. Bithas P. S., Michailidis E. T., Nomikos N., Vouyioukas D., Kanatas A. G. A Survey on Machine-Learning Techniques for UAV-Based Communications. *Sensors.* 2019. Vol. 19, no. 23:5170. DOI: https://doi.org/10.3390/s19235170
12. Jaafari S., Nassiri M., Mohammadi R. Traffic-aware Routing with Software-defined Networks Using Reinforcement Learning and Fuzzy Logic. *International Journal of Computing*. 2022. Vol. 21, no. 3. P. 318-324. DOI: https://doi.org/10.47839/ijc.21.3.2687

Received (Надійшла) 18.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Дмитрук Костянтин Сергійович – аспірант кафедри системи інформації ім. В.О. Кравця, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Kostiantyn Dmytruk – PhD Student, Department of Information Systems named after V. O. Kravets, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: Kostiantyn.Dmytruk@infiz.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0008-1936-3676;

Касілов Олег Вікторович – кандидат технічних наук, доцент, професор кафедри системи інформації ім. В.О. Кравця, Національний технічний університет «Харківський політехнічний інститут»; Харків, Україна;
Oleg Kasilov – Candidate of Technical Sciences, Associate Professor, Professor of Department of Information Systems named after V. O. Kravets, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: oleg.kasilov@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-8524-2345;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55976743600.

Method of adaptive traffic control in multi-level wireless systems

Kostiantyn Dmytruk, Oleg Kasilov

Abstract. Relevance. Modern multi-level wireless systems operate in conditions of stochastic fluctuations in load, heterogeneity of flows and interdependence of sensor, intermediate MESH and backbone levels. Traditional static traffic control mechanisms do not ensure stability of network operation in peak modes, which leads to local overloads, increased delays, queue lengths and packet losses. This necessitates the development of adaptive control methods capable of responding in real time to changes in flow intensity and providing guaranteed quality of service (QoS) indicators. **Object of research**: traffic control processes in multi-level wireless systems under variable load conditions. **Purpose of the article**: development and study of an adaptive traffic control method that ensures stabilization of QoS indicators by optimal redistribution of flows between network levels based on an assessment of their current state. **Research results**. he paper proposes a method based on the dynamic determination of effective intensities of flows, assessment of the loading of individual levels and formation of control actions in accordance with the objective function of minimizing congestion. A mathematical model has been developed that formalizes the interaction of network levels and allows analyzing its behavior under different load scenarios. Simulation modeling in the OMNeT++ environment confirmed the ability of the method to reduce delay, queue length and packet loss during periods of increased intensity. The integral gain in terms of QoS indicators was on average 20–30%, and in peak intervals it reached 30–32%, which indicates the effectiveness of adaptive redistribution of flows in critical operating modes. **Conclusions**. It is shown that the stability of multi-level wireless systems is determined not only by the intensity of traffic, but also by the ability of the network to respond adaptively to its fluctuations. The proposed method provides smoothing of peak loads, increases the structural stability of the system and improves QoS indicators, which makes it promising for integration into protocol stacks of next-generation networks. The scope of application of the obtained results: sensor networks, MESH architectures, multi-level wireless systems, IoT platforms, 5G/6G multi-service networks, traffic management optimization and QoS provisioning problems.

Keywords: wireless system, traffic management, MESH architecture, queues, flows, load, delay, QoS.

УДК 621.396.96 **doi: 10.26906/SUNZ.2026.2.228**

О. Л. Кузнєцов[1], О. В. Коломійцев[2], А. О. Ковальчук[1], А. М. Коржов[1], О. В. Очкуренко[1]

[1] Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна
[2] Національний технічний університет «Харківський політехнічний інститут», Харків, Україна

МОЖЛИВОСТІ ПІДВИЩЕННЯ ТОЧНОСТІ ОТОТОЖНЕННЯ ПЕЛЕНГІВ ПРИ ВИКОРИСТАННІ ТРИАНГУЛЯЦІЙНОГО МЕТОДУ ПАСИВНОЇ РАДІОЛОКАЦІЇ В РЕАЛЬНИХ УМОВАХ РОЗПОВСЮДЖЕННЯ РАДІОХВИЛЬ

Анотація. Об'єднання радіолокаційної інформації є важливим завданням при багатопозиційному прийомі радіолокаційного сигналу. Зокрема, при використанні тріангуляційного методу пасивної радіолокації незалежність оцінювання кутових координат в приймальних пунктах призводить до виникнення хибних об'єктів при ототожненні пеленгів. При цьому функціонування багатопозиційних радіолокаційних комплексів (БПРЛК) часто здійснюється у складних метеорологічних умовах. Оскільки область спостереження БПРЛК є значною, то виконання завдань за призначенням може здійснюватися ними в умовах тропосферної рефракції, що призводить до флуктуацій фазового фронту хвилі прийнятого сигналу. **Предметом** вивчення в статті є вплив флуктуацій фазового фронту хвилі прийнятого сигналу на точність ототожнення пеленгів при використанні тріангуляційного методу пасивної радіолокації. **Метою** є дослідження можливостей використання алгоритму оптимального вимірювання кутових координат з врахуванням флуктуацій фазового фронту хвилі прийнятого сигналу для підвищення точності ототожнення пеленгів при реалізації тріангуляційного методу пасивної радіолокації. **Завданням** є аналіз можливого підвищення вказаної точності в залежності від ступеня викривлення фазового фронту хвилі прийнятого сигналу в реальних умовах його розповсюдження. В ході досліджень використовувалися **методи** математичної статистики та теорії ймовірностей. **Результатами** статті є надання пропозицій щодо підвищення точності ототожнення пеленгів при використанні тріангуляційного методу пасивної радіолокації в умовах впливу флуктуацій фазового фронту хвилі прийнятого сигналу. Отримані результати можуть бути в подальшому використані в ході досліджень спрямованих на підвищення ефективності методів визначення координат в активних та активно-пасивних БПРЛК.

Ключові слова: аеродинамічний об'єкт, пасивна радіолокація, радіолокаційний сигнал, тріангуляційний метод, тропосфера, фазовий метод, флуктуації фазового фронту.

Вступ

Необхідність забезпечення безперервності отримання локаційні інформації про аеродинамічні об'єкти потребує перекриття зон виявлення локаційних засобів по всьому діапазоні можливих висот. Але це призводить до дублювання інформації про об'єкти спостереження та виникнення відповідної надлишкової інформації. Окрім цього, неузгодженість за часом і незалежність вимірювань може призвести до переплутування даних. Вказане обумовлює необхідність застосування алгоритмів об`єднання даних вже на етапі первинної обробки радіолокаційної інформації. В умовах складної цільової та завадової обстановки, використання БПРЛК дозволяє підвищити безперервність отримання інформації про повітряні об'кти. Зокрема, пасивна радіолокація забезпечує максимальну скритність та завадозахищеність радіолокаційних засобів з можливістю пеленгації постановників завад. При цьому одним з найпоширеніших методів пасивної радіолокації є тріангуляційний метод. Вказаний метод реалізується з використанням рознесених у просторі декількох пасивних джерел радіолокаційної інформації.

Недоліком цього методу є наявність хибного виявлення об'єктів в процесі ототожнення пеленгів. Причина даного ефекту полягає в незалежності кутових вимірювань кожним радіолокаційним засобом.

Реальні умови розповсюдження радіохвиль, зокрема наявність тропосферних неоднорідностей, є причиною виникнення флуктуацій фазового фронту хвилі радіолокаційного сигналу, що викликає порушення його просторової когерентності та зниження точності вимірювання кутових координат та висоти аеродинамічного об'єкта.

Оскільки БПРЛК, що реалізують тріангуляційний метод пасивної радіолокації здійснюють спостереження за повітряними об'єктами на значних відстнях від наземних пунктів прийму, вплив середовища розповсюдження радіохвиль може бути значним. У свою чергу, даний вплив може призвести до суттєвого зниження точності кутових вимірювань та, як слід, до відповідного погіршення точності ототожнення пеленгів. Тому, доцільним є визначення умов за якими вплив флуктуацій фазового фронту хвилі прийнятого сигналу потребує врахування в алгоритмі його просторової обробки.

Аналіз останніх досліджень і публікацій. Метеорологічним питанням, що стосуються стану атмосфери та особливостей радіолокації у відповідних умовах присвячено роботи [1]. Питання щодо впливу флуктуацій фазового фронту хвилі радіолокаційного сигналу при застосуванні радіотехнічних систем в нестабільних гідрометеорологічних умовах та стихійних метеорологічних явищах розглянуто у [2]. Зокрема, особливості організації просторових вимірювань в умовах впливу фазових флуктуацій радіохвиль розглянуто у [3]. При цьому, статистичні характеристики даних флуктуацій є експериментально підтвердженими та наведені у [4]. Оцінювання можливих помилок вимірювання кутових координат об'єктів радіолокаційного спостереження, обумовлених впливом флуктуацій фазового фронту хвилі сигналу проведено у [5]. Можливе зниження точності вимірювання кутових координат внаслідок впливу атмосферних неоднорідностей та земної поверхні оцінено у [6].

В роботі [7] доведено, що сучасний розвиток цифрових технологій дозволяє вважати перспективним шлях врахування кореляції фазових флуктуацій безпосередньо у алгоритмах просторової обробки радіолокаційного сигналу. У випадку цифрової обробки радіолокаційного сигналу особливості даного врахування наведені у роботі [8]. Питання обробки радіолокаційної інформації у БПРЛК, зокрема при використанні триангуляційного методу пасивної радіолокації надані у [9]. Представляє практичну користь оцінювання можливостей підвищення точності ототожнення пеленгів за рахунок використання засобами пасивної радіолокації алгоритму оптимального вимірювання кутових координат з врахуванням флуктуацій фазового фронту хвилі прийнятого сигналу при використанні триангуляційного методу пасивної радіолокації в реальних умовах розповсюдження радіохвиль.

Метою статті є дослідження можливостей використання алгоритму оптимального вимірювання кутових координат з врахуванням флуктуацій фазового фронту хвилі прийнятого сигналу для підвищення точності ототожнення пеленгів при реалізації тріангуляційного методу пасивної радіолокації.

Основний матеріал

Розглядається випадок пеленгації джерел радіовипромінювання, що здійснюється двома пунктами прийому рознесеними на відстань *Б*. У кожному з пунктів прийому вимірюються кутові координати – азимут і кут місця. Об'єднання даних здійснюється шляхом визначення точки перетину площин, що відповідають азимутальним і кутомісцевим положенням об'єкта спостереження. Умова формування координатних даних по *i*-му об'єкту спостереження визначається шляхом порівняння різниці висот з величиною припустимого розкиду згідно з виразом [9]:

$$\left|\hat{H}_{i1}-\hat{H}_{i2}\right| \le L\sqrt{\left(\hat{R}_{i1}\mathrm{tg}\sigma_{\varepsilon i1}\right)^2+\left(\hat{R}_{i2}\mathrm{tg}\sigma_{\varepsilon i2}\right)^2}. \qquad (1)$$

де $\hat{H}_{i1}$ і $\hat{H}_{i2}$ – висоти *i*-го об'єкта спостереження, які виміряні у першому і другому пунктах прийому відповідно; L – величина, яка обирається з припустимого значення ймовірності правильного ототожнення; $\hat{R}_{i1}$ і $\hat{R}_{i2}$ – дальності до *i*-го об'єкта спостереження у горизонтальній площині, які виміряні у першому і другому пунктах прийому відповідно; $\sigma_{\varepsilon i1}$ і $\sigma_{\varepsilon i2}$ – середньоквадратичні відхилення помилок вимірювання кута місця *i*-го об'єкта спостереження у першому і другому пунктах прийому відповідно.

Вказані у (1) висоти, які відповідають точці перетину відносно пунктів прийому визначаються як:

$$\hat{H}_{i1}=\hat{R}_{i1}\mathrm{tg}\hat{\varepsilon}_{i1},\quad \hat{H}_{i2}=\hat{R}_{i2}\mathrm{tg}\hat{\varepsilon}_{i2}. \qquad (2)$$

де $\hat{\varepsilon}_{i1}$ і $\hat{\varepsilon}_{i2}$ – значення кутів місця *i*-го об'єкта спостереження, які виміряні у першому і другому пунктах прийому відповідно. Дальності $\hat{R}_{i1}$ і $\hat{R}_{i2}$ визначаються таким чином [9]:

$$\hat{R}_{i1}=\frac{Б\cos\hat{\beta}_{i2}}{\sin(\hat{\beta}_{i1}-\hat{\beta}_{i2})},\ \hat{R}_{i2}=\frac{Б\cos\hat{\beta}_{i1}}{\sin(\hat{\beta}_{i1}-\hat{\beta}_{i2})}. \qquad (3)$$

де $\hat{\beta}_{i1}$ і $\hat{\beta}_{i2}$ – значення азимутів *i*-го об'єкта спостереження, які виміряні у першому і другому пунктах прийому відповідно.

Виконання умови (1) свідчить про належність оцінених у першому та другому пунктах прийому параметрів *i*-му об'єкту спостереження. В іншому випадку вважається, що знайдена точка перетину є хибною і порівняння одержаних оцінок продовжується. При цьому, необхідною умовою є приведення результатів кутових вимірювань до єдиного моменту часу. За відсутності такої можливості необхідна ймовірність правильного ототожнення забезпечується шляхом збільшення величини L, але при цьому збільшується й імовірність хибного ототожнення. Отже алгоритм ототожнення включає такі процедури:

– зведення даних до єдиного моменту часу (синхронізація за часом);

– розрахунок згідно (3) дальностей $\hat{R}_{i1}$ і $\hat{R}_{i2}$ відносно двох пунктів прийому в азимутальній площині по двом одноразово виміряним азимутам $\hat{\beta}_{i1}$ і $\hat{\beta}_{i2}$;

– за оцінними значеннями $\hat{\varepsilon}_{i1}$ і $\hat{\varepsilon}_{i2}$ та розрахованими $\hat{R}_{i1}$ і $\hat{R}_{i2}$ визначення висот $\hat{H}_{i1}$ і $\hat{H}_{i2}$ (2);

– різниця знайдених висот порівнюється з величиною припустимого відхилення по висоті $\left|\hat{H}_{i1}-\hat{H}_{i2}\right|$ (1).

Для оцінювання показників якості ототожнення пеленгів слід враховувати нелінійний зв'язок з помилками первинних кутових вимірювань:

$$\gamma=\left|\hat{H}_{i1}-\hat{H}_{i2}\right|^2-$$

$$-L^2\left\{\left[\hat{R}_{i1}\mathrm{tg}\left(\sigma_{i1}\right)\right]^2+\left[\hat{R}_{i2}\mathrm{tg}\left(\sigma_{i2}\right)\right]^2\right\}\le 0. \qquad (4)$$

При цьому доцільно мати інформацію про закон розподілу випадкової величини $p(\gamma)$.

При нульовому порозі для гіпотези H_1 (точка перетину істинна) та гіпотези H_2 (точка перетину хибна), ймовірності правильного ототожнення P_{no} і хибного ототожнення P_{xo} знаходиться як:

$$P_{no}=\int_{-\infty}^{0} p_{H1}(\gamma)d\gamma, \qquad (5)$$

$$P_{xo}=\int_{-\infty}^{0} p_{H2}(\gamma)d\gamma. \qquad (6)$$

Так, $P_{no}\approx 0{,}997$ відповідає L=3, $P_{no}\approx 0{,}92$ відповідає L=2, а $P_{no}\approx 0{,}67$ відповідає L=1.

Один з методів оцінки значень P_{no} і P_{xo} полягає в лінеаризації залежності γ від випадкових величин $\hat{\beta}_{i1}$, $\hat{\beta}_{i2}$, $\hat{\varepsilon}_{i1}$, $\hat{\varepsilon}_{i2}$. Це дозволить нормалізувати закони розподілу $p_{H1}(\gamma)$ і $p_{H2}(\gamma)$ та оцінити значення P_{no} та P_{xo}.

В [9] наведено два способи оцінювання показників якості ототожнення пеленгів. За першим способом, функція γ підлягає розкладанню в ряд Тейлора

в точках, які відповідають виміряними значеннями. Ця процедура дозволяє лінеаризувати зв'язок γ з випадковими величинами $\hat{\beta}_{i1}$, $\hat{\beta}_{i2}$, $\hat{\varepsilon}_{i1}$, $\hat{\varepsilon}_{i2}$. При цьому дисперсія γ визначається сумою наступного виду:

$$\sigma_\gamma^2 = A^2\sigma_{\Delta\beta_1}^2 + B^2\sigma_{\Delta\beta_2}^2 + C^2\sigma_{\Delta\varepsilon_1}^2 + D^2\sigma_{\Delta\varepsilon_2}^2 . \qquad (7)$$

де $A = \partial\gamma/\partial\beta_{i1}$ при $\beta_{i1} = \hat{\beta}_{i1}$; $B = \partial\gamma/\partial\beta_{i2}$ при $\beta_{i2} = \hat{\beta}_{i2}$; $C = \partial\gamma/\partial\varepsilon_{i1}$ при $\varepsilon_{i1} = \hat{\varepsilon}_{i1}$; $D = \partial\gamma/\partial\varepsilon_{i2}$ при $\varepsilon_{i2} = \hat{\varepsilon}_{i2}$; $\Delta\beta_1 = \beta_{i1} - \hat{\beta}_{i1}$; $\Delta\beta_2 = \beta_{i2} - \hat{\beta}_{i2}$; $\Delta\varepsilon_1 = \hat{\varepsilon}_{i1} - \hat{\varepsilon}_{i1}$; $\Delta\varepsilon_2 = \hat{\varepsilon}_{i2} - \hat{\varepsilon}_{i2}$.

За другим способом будується модель випадкових вимірів кутових координат об'єктів спостереження для фіксованих значень дальності R та висоти H, за наступними співвідношеннями:

$$\hat{\varepsilon}_{i1} = \varepsilon_{i1} + \eta_{\varepsilon 1}, \ \hat{\varepsilon}_{i2} = \varepsilon_{i2} + \eta_{\varepsilon 2}, \qquad (8)$$

$$\hat{\beta}_{i1} = \beta_{i1} + \eta_{\beta 1} \quad \hat{\beta}_{i2} = \beta_{i2} + \eta_{\beta 2}, \qquad (9)$$

де $\eta_{\varepsilon 1}$, $\eta_{\varepsilon 2}$, $\eta_{\beta 1}$, $\eta_{\beta 2}$ – складові випадкових величин з заданою дисперсією, які розподілені за нормальним законом та генеруються датчиком випадкових чисел.

В сучасних цифрових радіолокаторах у якості антенних систем широко використовується фазовані антенні решітки (ФАР), які забезпечують одночасне вимірювання дальності, азимуту та кута місця. В пунктах прийому БПРЛК, кут місця може бути визначений згідно фазового методу, тобто за фазовим зсувом сигналів в елементах антенної решітки. Значення $\hat{H}_{i1}$ і $\hat{H}_{i2}$ можуть бути визначені згідно (2). Отримані оцінки використовуються в алгоритмі ототожнення (1). Корегування коефіцієнту L дозволяє забезпечити потрібну якість ототожнення.

В [10] показано, що при прийомі когерентного сигналу з випадковою рівномірно розподіленою початковою фазою і випадковою, розподіленою за законом Релея амплітудою на фоні внутрішнього шуму, дисперсія помилки оцінювання кута місця для ФАР з рівномірним амплітудним розподілом описується як

$$\sigma_\varepsilon^2 = \frac{12}{q^2(4n^2-1)}\left(\frac{\lambda}{2\pi d}\right)^2, \qquad (10)$$

де q^2 – відношення сигнал/шум по потужності; n – число елементів ФАР.

Точність вимірювання кута місця і як слід висоти аеродинамічного об'єкта залежить від стану атмосфери особливо при локації в умовах тропосферної рефракції, що є причиною виникнення флуктуацій фазового фронту хвилі радіолокаційного сигналу.

Згідно експериментальних даних щодо дослідження фазових флуктуацій, закон їх розподілу є близьким до нормального, а кореляційна функція апроксимуються експонентною або осцилюючою залежностями [4]. Впливом фазових флуктуацій у верхніх шарах тропосфери можна знехтувати внаслідок їх малої інтенсивності та високої кореляції. Однак, вплив фазових флуктуацій, обумовлений нижніми шарами тропосфери, навпаки є дуже суттєвим.

Тобто, лінійні флуктуаційні відхилення точки спостереження, обумовлені впливом тропосферних неоднорідностей можуть досягати сотень метрів.

Висота аеродинамічного об'єкта з урахуванням кривизни Землі визначається за формулою [12]:

$$h = r \cdot \sin\varepsilon + \frac{r^2}{2R_з} + h_a , \qquad (11)$$

де r – похила дальність аеродинамічного об'єкта; $R_з$ – радіус Землі; h_a – висота розташування антени радіолокатора над поверхнею, відносно якої здійснюється відлік висоти аеродинамічного об'єкта.

В табл. 1 [12] наведені розраховані згідно виразів (10) - (11) величини СКП вимірювання висоти за відсутністю σ_h та за наявністю $\sigma_{hфл}$ впливу флуктуацій фазового фронту хвилі прийнятого сигналу з експонентною кореляційною функцією фазових флуктуацій.

Таблиця 2 –**Розрахунки вимірювань**

r, км	50	100	150	200
σ_h, м	8,9	9,8	10,6	11,5
$\sigma_{hфл}$, м	169,7	568,4	1167,3	2617,1

Результати розрахунку, наведені в таблиці 1, отримано для різних значень $r = 50, 100, 150$ та 200 км за наступних умов: $\lambda = 5$ см, n=100, $q^2 = 100$, $d = \lambda/2$, h_a = 8 м та $L_0 = 1$ км.

Як видно, СКП вимірювання висоти $\sigma_{hфл}$ за наявності впливу фазових флуктуацій може в десяткисотні разів перевершувати СКП вимірювання висоти σ_h обумовлену впливом лише внутрішнього шуму приймача. Тобто, на дальностях 50…200 км СКП вимірювання висоти $\sigma_{hфл}$ здатна досягати величин від сотень метрів до одиниць кілометрів.

Отже, необхідним є визначення можливостей врахування впливу флуктуацій фазового фронту хвилі радіолокаційного сигналу при оцінюванні висоти аеродинамічного об'єкта з метою підвищення точності ототожнення пеленгів в пасивному БПРЛК. Це може бути реалізовано шляхом застосування оптимального алгоритму вимірювання кутових координат, у якому враховуються фазові флуктуації сигналу, який надходить на елементи ФАР у пунктах прийому пасивного БПРЛК.

Слід зазначити, що підвищення точності вимірювання кута місця та висоти за рахунок статистичної оптимізації просторової обробки радіолокаційного сигналу за наявності врахування корельованих фазових флуктуацій, обумовлених впливом неоднорідностей тропосфери, може бути ще більшим, якщо фазові флуктуації описуються осцилюючою кореляційною функцією з різною частотою осциляцій.

Окрім впливу неоднорідностей тропосфери, на зниження точності вимірювання висоти повітряного об'єкта суттєвий вплив здійснює земна (морська) підстильна поверхня, оскільки реальна поверхня відбиття не є ідеально рівною. Ступінь "нерівності" шорсткої

поверхні землі або схвильованої поверхні моря визначається співвідношенням між довжиною хвилі і геометричними параметрами нерівностей. При цьому поле в приймальних пунктах пасивного БПРЛК представляє собою результат інтерференції розсіяних хвиль, що також є причиною флуктуацій фазового фронту результуючої хвилі. Дані флуктуації фази є додатковою причиною виникнення складових помилок вимірювання кутових координат і висоти повітряного об'єкта, які є найбільш суттєвими при його спостереженні на малих висотах. Таким чином, вплив реальних умов розповсюдження радіохвиль, тобто неоднорідностей тропосфери та шорсткості підстильної поверхні призводить до виникнення суттєвих флуктуаційних складових СКП вимірювання висоти об'єкта спостереження. Вказане обумовлює розширення областей перетину площин, що відповідають азимутальним і кутомісцевим положенням об'єкта спостереження. При цьому можливою є ситуація, за якою умова формування координатних даних по *і*-му об'єкту спостереження (1) перестає виконуватися.

Особливої актуальності вказана ситуація набуває для БПРЛК приморського та морського базування, для яких флуктуації параметрів прийнятих сигналів є найбільш суттєвими внаслідок сезонної та добової мінливості умов розповсюдження радіохвиль. Наведені результати сприяють вдосконаленню існуючих методів оброблення сигналів в сучасних системах радіолокації спрямованого на адаптацію до зовнішніх умов виконання ними завдань за призначенням.

Висновки

1. Функціонування БПРЛК може здійснюватися у складних метеорологічних умовах, вплив яких призводить до виникнення флуктуацій фазового фронту хвилі прийнятого сигналу, що обумовлює зниження точності кутових вимірювань та відповідне поширення точності ототожнення пеленгів при використанні тріангуляційного методу пасивної радіолокації.

2. За результатами оцінювання, СКП вимірювання висоти за наявності впливу фазових флуктуацій може в десятки-сотні разів перевершувати СКП вимірювання висоти обумовлену впливом лише внутрішнього шуму приймача і досягати величин від сотень метрів до одиниць кілометрів.

3. Використання алгоритму оптимального вимірювання кутових координат з врахуванням флуктуацій фазового фронту хвилі прийнятого сигналу з експонентною кореляційною функцією дозволяє зменшити СКП вимірювання висоти на величину до 100 метрів і більше, що відповідно забезпечує підвищення точності ототожнення пеленгів при реалізації тріангуляційного методу пасивної радіолокації.

4. Підвищення точності вимірювання кута місця та висоти повітряного об'єкта за рахунок врахування корельованих фазових флуктуацій може бути більшим, якщо фазові флуктуації описуються осцилюючою кореляційною функцією з різною частотою осциляцій.

5. Отримані результати можуть бути в подальшому використані в ході досліджень спрямованих на підвищення ефективності методів визначення координат в активних та активно-пасивних БПРЛК.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Климченко В. Й., Камалтинов Г. Г., Місайлов В. Л. Аналіз потенційних можливостей оглядових РЛС сантиметрового діапазону хвиль щодо забезпечення дій авіації Повітряних Сил України метеорологічною інформацією. *Системи озброєння і військова техніка*. – 2011. – №1 (25). – С. 21–27. http://nbuv.gov.ua/UJRN/soivt_2011_1_7.
2. Петрушенко М. М. Особливості застосування радіотехнічних систем Повітряних Сил в нестабільних гідрометеорологічних умовах та стихійних метеорологічних явищах. *Системи управління навігації та зв'язку*. – 2009. – № 2 (10). – С. 54-57. https://journal-hnups.com.ua/index.php/nitps/article/view/317.
3. Кузнєцов О. Л. Оцінювання впливу фазових флуктуацій сигналу на зниження точності вимірювання кутових координат цілі в РЛС з фазованими антенними решітками. Системи обробки інформації. – 2008. – № 1 (68). – С. 38-40. https://scholar.google.com/citations?user=4I8-ivYAAAAJ&hl=en.
4. Карлов В. Д., Родюков А. О, Пічугін І. М. Статистичні характеристики радіолокаційних сигналів відбитих від місцевих предметів в умовах аномальної рефракції. *Наука і техніка Повітряних Сил Збройних Сил України*. – 2015. – Вип. 4 (21). – С. 71-74. http://nbuv.gov.ua/UJRN/Nitps_2015_4_19.
5. Kuznietsov O., Kovalchuk V., Kovalchuk A., Karlov D., Yarovyy S., Vasylyshyn V. Providing the Required Accuracy of Measurements of Spatial Coordinates of Aerial Objects. // 2020 IEEE Ukrainian Microwave Week. On 2020 IEEE 6th International Symposium on Microwaves, Radar and Remote Sensing (MRRS). Kharkiv, Ukraine, September 21-25, 2020. V. 2. P. 226-229. https://doi.org/10.1109/UkrMW49653.2020.9252605.
6. Кузнєцов О. Л., Танцюра О. Б., Мельник О. Л. Обмеження якості просторових вимірювань в РЛС з фазованою антенною решіткою внаслідок впливу атмосферних неоднорідностей та земної поверхні. *Системи управління навігації та зв'язку. – К.: ЦНДІ навігації і управління*. – 2012. – № 1 (21). Том 2. – С. 49-52. https://scholar.google.com/citations?user=4I8-ivYAAAAJ&hl=en.
7. Карлов В. Д., Кузнєцов О. Л., Коломійцев О. В., Красношапка І. В., Петрушенко І. М., Струцінський О. В. Можливості врахування впливу тропосфери при вимірюванні кутових координат та висоти аеродинамічного об'єкта. *Системи управління навігації та зв'язку*. – 2022. – Вип. 3(69). – С. 121-127. doi:https://doi.org/10.26906/SUNZ.2022.3.121
8. Кузнєцов О. Л., Чепурний В. А. Підвищення якості просторових вимірювань в цифрових РЛС з фазованою антенною решіткою. *Системи озброєння і військова техніка*. – 2015. – № 2(42). – С. 113-115. https://scholar.google.com/citations?user=4I8-ivYAAAAJ&hl=en

9. Войтович С. А., Турсунходжаєв Х. А., Траєкторна обробка локаційної інформації. – Х. : ХУПС, 2008. – 112 с. https://knizhkova-skarbnitsya.prom.ua/ua/p1565715779-vojtovich-tursunhodayev-trayektorna.html
10. Сєдишев Ю. М., Карпенко В. І., Атаманський Д. В. та ін. Радіоелектронні системи. – Х. : ХУПС, 2010. – 418 с. https://journal-hnups.com.ua/index.php/nitps/article/view/500
11. Карлов В. Д., Кузнєцов О. Л., Бєлоусов В. В., Тузіков С. А., Олещук М. М., Петрушенко В. М. Точність вимірювання кутових координат аеродинамічних об'єктів в умовах тропосферної рефракції. *Системи управління навігації та зв'язку*. – 2021 – Вип. 1 (63) – С. 146-152. https://journals.nupp.edu.ua/sunz/issue/view/74/41
12. Kuznietsov O., Kolomiitsev O., Kiyko A., Kovalchuk A., Sadovyi K. Analysis of possibilities of providing of necessary exactness of measuring of spatial coordinates of air objects in the radio-location station of accompaniment with phase aerial by a gate. *Сучасні інформаційні системи*. – 2020. – Том 4, № 1. С. 91-96. http://ais.khpi.edu.ua/article/view/2522-9052.2020.1.13

Received (Надійшла) 25.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Кузнєцов Олександр Леонідович – кандидат технічних наук, доцент, доцент кафедри Харківського національного університету Повітряних Сил ім. І. Кожедуба, Харків, Україна;
Oleksandr Kuznietsov – PhD, Associate Professor, Associate Professor of Department of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: SAG2121@ukr.net; ORCID Author ID: https://orcid.org/0000-0002-5915-8107;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202953428.

Коломійцев Олексій Володимирович – доктор технічних наук, професор, професор кафедри комп'ютерної інженерії та програмування, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Oleksii Kolomiitsev – Doctor of Technical Sciences, Professor, Professor of Department of computer engineering and programming, National Technical University is the «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: alexus_k@ukr.net; ORCID Author ID: https://orcid.org/0000-0001-8228-8404;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57211278112.

Ковальчук Андрій Олексійович – кандидат технічних наук, доцент, професор кафедри Харківського національного університету Повітряних Сил ім. І. Кожедуба, Харків, Україна;
Andrii Kovalchuk – PhD, Associate Professor, Professor of Department of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: Inna700nf@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-1269-9368;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57220834484.

Коржов Андрій Миколайович – кандидат технічних наук, доцент, доцент кафедри Харківського національного університету Повітряних Сил ім. І. Кожедуба, Харків, Україна;
Andrii Korzhov – PhD, Associate Professor, Associate Professor of Department of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: kan1972.kan@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-4595-2366;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57216643189 .

Очкуренко Олександр Вікторович – кандидат технічних наук, доцент, доцент кафедри Харківського національного університету Повітряних Сил ім. І. Кожедуба, Харків, Україна;
Oleksandr Ochkurenko – PhD, Associate Professor, Associate Professor of Department of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: kolidor@ukr.net; ORCID Author ID: https://orcid.org/0000-0003-3809-5175;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57209506925.

Possibilities of increasing the accuracy of bearing identification using the triangulation method of passive radar location in real conditions of radio wave propagation

Oleksandr Kuznietsov, Oleksii Kolomiitsev, Andrii Kovalchuk, Andrii Korzhov, Oleksandr Ochkurenko

Abstract. Combining radar information is an important task in multi-position radar signal reception. In particular, when using the triangulation method of passive radar, the independence of estimating angular coordinates at receiving points leads to the appearance of false objects when identifying bearings. At the same time, the operation of multi-position radar complexes is often carried out in difficult meteorological conditions. Since the observation area of the multi-position radar complexes is significant, the performance of tasks for their intended purpose can be carried out by them in conditions of tropospheric refraction, which leads to fluctuations in the phase front of the received signal wave. The subject of study in the article is the influence of fluctuations in the phase front of the received signal wave on the accuracy of bearing identification when using the triangulation method of passive radar. The aim is to investigate the possibilities of using an algorithm for optimal measurement of angular coordinates taking into account fluctuations in the phase front of the received signal wave to increase the accuracy of bearing identification when implementing the triangulation method of passive radar. The task is to analyze the possible increase in the specified accuracy depending on the degree of distortion of the phase front of the received signal wave in real conditions of its propagation. The methods of mathematical statistics and probability theory were used in the research. The results of the article are to provide suggestions for increasing the accuracy of bearing identification when using the triangulation method of passive radar under the influence of fluctuations in the phase front of the received signal wave. The results obtained can be further used in research aimed at improving the efficiency of methods for determining coordinates in active and active-passive multi-position radar complexes.

Keywords: aerodynamic object, passive radar, triangulation method, troposphere, phase method, phase front fluctuations.

УДК 621.391.8:519.21 **doi: 10.26906/SUNZ.2026.2.233**

В. П. Лисечко[1], К. А. Трубчанінова[2], О. С. Жученко[2], Г. В. Шубіна[1]

[1] Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна
[2] Український державний університет залізничного транспорту, Харків, Україна

БАГАТОРІВНЕВА ФУНКЦІОНАЛЬНА МОДЕЛЬ ПОКАЗНИКІВ ЕФЕКТИВНОСТІ СИСТЕМИ МНОЖИННОГО ДОСТУПУ З КОДОВИМ РОЗДІЛЕННЯМ

Анотація. Актуальність. Для оцінювання ефективності системи множинного доступу з кодовим розділенням, зокрема на основі ансамблів різноенергетичних складних сигналів, необхідна узагальнена модель, яка впорядковує рівневі показники ефективності та задає функціональні зв'язки між ними. **Об'єкт дослідження**: показники ефективності та функціональні зв'язки між ними в системі множинного доступу з кодовим розділенням. **Мета статті**: розробити багаторівневу функціональну модель показників ефективності системи множинного доступу з кодовим розділенням на основі апарату системного та функціонального аналізу та показати можливість одержання узагальненої оцінки ефективності системи множинного доступу з кодовим розділенням. **Результати дослідження**. Розроблено багаторівневу функціональну модель показників ефективності системи множинного доступу з кодовим розділенням, яка визначає правила функціонально-рівневої декомпозиції. У межах моделі показники впорядковуються за ступенем узагальнення оцінок ефективності, кожному ступеню узагальнення відповідає рівень моделі, а зв'язки між показниками визначаються функціоналами відповідних рівнів. Кожен рівень моделі формує одну числову оцінку ефективності, а за наявності кількох функціоналів у межах одного рівня показник ефективності цього рівня визначається їх композицією. **Висновки.** Запропоновано варіант рівневої будови моделі з виділенням енергетичного, структурного, потужнісно-канального та системного рівнів. Енергетичний рівень пов'язано з показником енергетичної справедливості, структурний рівень - з показником структурної ефективності, потужнісно-канальний рівень - з відношенням енергії корисного сигналу до сумарної енергії завад множинного доступу та шуму, а системний рівень - з узагальненим показником ефективності системи.

Ключові слова: комп'ютерна система; складний сигнал; показник ефективності; багаторівнева функціональна модель; функціонал; енергетичний рівень; структурний рівень; потужнісно-канальний рівень; системний рівень.

Вступ

Постановка проблеми. Оцінювання ефективності системи множинного доступу з кодовим розділенням належить до задач аналізу складної системи, ефективність функціонування якої характеризується сукупністю взаємопов'язаних показників. Для системи множинного доступу з кодовим розділенням, зокрема на основі ансамблів різноенергетичних складних сигналів, такі показники можуть відображати структурно-енергетичні властивості ансамблю сигналів, властивості каналу зв'язку та інші параметри, істотні для конкретної задачі оцінювання. Тому потрібна багаторівнева функціональна модель, яка впорядковує такі показники, визначає функціональні зв'язки між ними та дає змогу одержати узагальнений показник ефективності для систем множинного доступу з кодовим розділенням на основі ансамблів як рівноенергетичних, так і різноенергетичних складних сигналів.

Аналіз останніх досліджень і публікацій. Перспективним напрямом розвитку надширокосмугових систем доступу є застосування імпульсних складних сигналів. Прямим наслідком розвитку надширокосмугових систем доступу з імпульсними сигналами є перехід до систем множинного доступу з кодовим розділенням, що дасть змогу забезпечити гарантований одночасний доступ певної кількості користувачів. Для такого переходу потрібні відповідні ансамблі імпульсних складних сигналів. У [1] запропоновано ансамблі періодичних імпульсних послідовностей з мінімальною енергетичною взаємодією, для яких при довільних часових зсувах відбувається не більше одного суміщення імпульсів. Характерною особливістю таких ансамблів є конструктивна різноенергетичність, що ускладнює пряме оцінювання ефективності ансамблів різноенергетичних сигналів через різні внески заважаючих сигналів у завади множинного доступу для окремого користувача. У [2] для оцінювання ефективності ансамблів різноенергетичних складних сигналів запропоновано застосовувати еквівалентні перетворення, які дозволяють наближено представити ансамбль різноенергетичних складних сигналів умовним ансамблем рівноенергетичних сигналів, еквівалентним вихідному ансамблю за сумарною енергією сигналів, що забезпечує можливість застосування існуючих співвідношень для оцінки ефективності ансамблів рівноенергетичних сигналів до ансамблів різноенергетичних сигналів.

Оцінювання ефективності системи множинного доступу з кодовим розділенням не вичерпується оцінюванням структурно-енергетичних властивостей ансамблів різноенергетичних складних сигналів. Для одержання узагальненої оцінки ефективності системи необхідна побудова узагальненої моделі, яка відображає функціональні зв'язки між окремими показниками ефективності. Для цього скористаємось апаратом системного та функціонального аналізу [3–5], що дозволяє представити такі зв'язки у виді багаторівневої функціональної моделі показників ефективності системи множинного доступу з кодовим розділенням [6, 7].

Метою роботи є розробка багаторівневої функціональної моделі показників ефективності системи множинного доступу з кодовим розділенням, що дає можливість одержання узагальненої оцінки ефективності системи множинного доступу з кодовим розділенням.

Основний матеріал

Розглянемо показники ефективності системи множинного доступу з кодовим розділенням як числові величини, що визначаються функціоналами, тобто відображеннями множин, елементами яких є числові величини, функції, функціонали, вектори або матриці, у числову величину.

Якщо задати функціональні зв'язки між показниками, визначивши одні показники через інші шляхом композиції функціоналів, якими задаються ці показники, то множину таких показників можна впорядкувати за ступенем узагальнення одержуваних оцінок ефективності й поставити цим ступеням у відповідність рівні багаторівневої моделі показників ефективності, причому таке впорядкування будемо називати функціонально-рівневою декомпозицією.

Багаторівневу функціональну модель показників ефективності системи множинного доступу з кодовим розділенням визначимо як упорядковану за ступенем узагальнення оцінок ефективності множину рівнів, кожному з яких поставлено у відповідність функціонали показників ефективності.

Установимо, що кожен рівень багаторівневої функціональної моделі визначає одну оцінку ефективності, яка є числовою величиною. Якщо в межах рівня задано кілька функціоналів показників ефективності, то показник ефективності рівня визначається композицією таких функціоналів без обмежень щодо виду композиції.

Для моделі з кількістю рівнів H номер рівня позначимо як $h = 1,2, \dots, H$, причому номер рівня H відповідає найвищому рівню моделі. Показник ефективності рівня з номером h визначимо через функціонал F_h. Для першого, тобто найнижчого, рівня показник ефективності визначимо через вектор параметрів, які надходять на цей рівень безпосередньо:

$$P_1 = F_1(C_1),$$

де P_1 - показник ефективності рівня 1; F_1 - функціонал рівня 1, яким задається показник ефективності рівня P_1; C_1 - вектор вхідних параметрів, які надходять на рівень 1 безпосередньо. Для рівнів з номерами $h = 2,3, \dots, H$ показник ефективності P_h визначимо через функціонал F_h, вхідний параметр Q_h, який визначається показником ефективності попереднього рівня $h-1$, та вектор вхідних параметрів C_h, які надходять на рівень h безпосередньо:

$$P_h = F_h(Q_h, C_h), h = 2,3, \dots, H,$$

де h - номер рівня моделі, $h = 2,3, \dots, H$; P_h - показник ефективності рівня h; F_h - функціонал рівня h, яким задається показник ефективності P_h; Q_h - вхідний параметр рівня h, який визначається показником ефективності попереднього рівня $h-1$; C_h - вектор вхідних параметрів, які надходять на рівень h безпосередньо; H - кількість рівнів моделі.

Вхідний параметр Q_h визначається показником ефективності попереднього рівня $h-1$:

$$Q_h = P_{h-1}, h = 2,3, \dots, H,$$

де Q_h - вхідний параметр рівня h; P_{h-1} - показник ефективності рівня $h-1$.

Оскільки показник ефективності P_h рівня h визначається показником ефективності P_{h-1} попереднього рівня $h-1$ та вектором C_h, то є справедливим спрощений вираз:

$$P_h = F_h(P_{h-1}, C_h), h = 2,3, \dots, H.$$

Отже, для рівня h визначено показник ефективності P_h через функціонал F_h, вхідний параметр Q_h та вектор вхідних параметрів C_h, які можуть надходити на рівень безпосередньо.

Введена вище багаторівнева функціональна модель показників ефективності задає правила функціонально-рівневої декомпозиції: показники впорядковуються за ступенем узагальнення оцінок ефективності, кожному ступеню узагальнення ставиться у відповідність рівень моделі, а зв’язки між показниками задаються функціоналами відповідних рівнів. Кожен рівень моделі визначає одну оцінку ефективності як числову величину, а за наявності кількох функціоналів у межах одного рівня показник ефективності цього рівня задається композицією таких функціоналів без обмежень щодо виду композиції.

Кількість рівнів моделі може бути визначена потребами розв’язання задач оцінки ефективності, синтезу або оптимізації з урахуванням мінімально необхідного рівня деталізації, достатнього для розв’язання відповідної задачі. Для відображення структурно-енергетичних властивостей ансамблів складних сигналів, впливу каналу зв’язку та одержання узагальненої оцінки ефективності системи множинного доступу з кодовим розділенням, з урахуванням уведених у роботі показника енергетичної справедливості та показника структурної ефективності, виділимо енергетичний, структурний, потужнісно-канальний та системний рівні, що утворюють чотирирівневу модель показників ефективності. Відповідність між рівнями, показниками ефективності та функціоналами чотирирівневої моделі визначимо співвідношеннями:

$$P_1 = P_{\text{EN}}, F_1 = F_{\text{EN}}, \quad P_2 = P_{\text{STR}}, F_2 = F_{\text{STR}},$$

$$P_3 = P_{\text{CHP}}, F_3 = F_{\text{CHP}}, \quad P_4 = P_{\text{SYS}}, F_4 = F_{\text{SYS}},$$

де P_{EN} - показник ефективності енергетичного рівня; F_{EN} - функціонал енергетичного рівня; P_{STR} - показник ефективності структурного рівня; F_{STR} - функціонал структурного рівня; P_{CHP} - показник ефективності потужнісно-канального рівня; F_{CHP} - функціонал потужнісно-канального рівня; P_{SYS} - показник ефективності системного рівня; F_{SYS} - функціонал системного рівня.

Функціональні зв’язки між рівнями чотирирівневої моделі запишемо у виді:

$$P_{\text{EN}} = F_{\text{EN}}(C_{\text{EN}}), \quad P_{\text{STR}} = F_{\text{STR}}(P_{\text{EN}}, C_{\text{STR}}),$$

$$P_{\text{CHP}} = F_{\text{CHP}}(P_{\text{STR}}, C_{\text{CHP}}), \quad P_{\text{SYS}} = F_{\text{SYS}}(P_{\text{CHP}}, C_{\text{SYS}}),$$

де C_{EN}, C_{STR}, C_{CHP}, C_{SYS} - вектори параметрів, які надходять безпосередньо відповідно на енергетичний, структурний, потужнісно-канальний та системний рівні. Вектори параметрів C_{EN}, C_{STR}, C_{CHP} визначимо як

$$C_{\text{EN}} = \big((E_1, E_2, \dots, E_L), L\big),$$

$$C_{\text{STR}} = (L, R^2_{max}), \quad C_{\text{CHP}} = (N),$$

де N - параметр потужнісно-канального рівня, що визначається моделлю каналу зв'язку; L – об'єм ансамблю; R^2_{max} - максимальне по ансамблю значення квадрата взаємної кореляції; E_i - енергія i-го сигналу ансамблю, $i = 1,2, \dots, L$.

З урахуванням визначення векторів C_{EN}, C_{STR}, C_{CHP} функціональні зв'язки чотирирівневої моделі запишемо так:

$$P_{\text{EN}} = F_{\text{EN}}\big((E_1, E_2, \dots, E_L), L\big),$$
$$P_{\text{STR}} = F_{\text{STR}}(P_{\text{EN}}, L, R^2_{max}),$$
$$P_{\text{CHP}} = F_{\text{CHP}}(P_{\text{STR}}, N),$$
$$P_{\text{SYS}} = F_{\text{SYS}}(P_{\text{CHP}}, C_{\text{SYS}})$$

або безпосередньо через показники енергетичної справедливості $\gamma^{(-)}_{\text{en}}$ та структурної ефективності $\gamma^{(+)}_{\text{str}}$ та відношення енергії корисного сигналу до сумарної енергії завад множинного доступу та шуму SINR:

$$P_{\text{EN}} = \gamma^{(-)}_{\text{en}}, P_{\text{STR}} = \gamma^{(+)}_{\text{str}},$$
$$P_{\text{CHP}} = F_{\text{CHP}}\big(\gamma^{(+)}_{\text{str}}, N\big) = \text{SINR},$$
$$P_{\text{SYS}} = F_{\text{SYS}}(P_{\text{CHP}}, C_{\text{SYS}}),$$

Функціональний зв'язок між показниками структурної ефективності $\gamma^{(+)}_{\text{str}}$ та енергетичної справедливості $\gamma^{(-)}_{\text{en}}$, а також SINR визначається таким чином:

$$\gamma^{(+)}_{\text{str}} = F_{\text{STR}}\big(\gamma^{(-)}_{\text{en}}, L, R^2_{max}\big) = \frac{LR^2_{max}}{\gamma^{(-)}_{\text{en}}},$$
$$\text{SINR} = \frac{1}{\gamma^{(+)}_{\text{str}} + N}, N = \frac{E_N}{E^{(+)}_{\text{eq}}},$$

де E_N - енергія шуму; $E^{(+)}_{\text{eq}}$ - енергія сигналу рівноенергетичного еквівалентного ансамблю. При цьому зв'язок показників ефективності з енергетичним еквівалентним об'ємом $L^{(-)}_{\text{eq}}$, структурним еквівалентним об'ємом $L^{(+)}_{\text{eq}}$, енергіями сигналів ансамблю E_i, енергією сигналу рівноенергетичного еквівалентного ансамблю $E^{(+)}_{\text{eq}}$ та енергією шуму E_N має вид:

$$\gamma^{(-)}_{\text{en}} = \frac{L^{(-)}_{\text{eq}}}{L}, \quad \gamma^{(+)}_{\text{str}} = L^{(+)}_{\text{eq}} R^2_{max}, \quad L^{(-)}_{\text{eq}} = \frac{\left(\sum_{i=1}^{L} E_i\right)^2}{\sum_{i=1}^{L} E_i^2},$$
$$L^{(+)}_{\text{eq}} = \frac{L^2 \sum_{i=1}^{L} E_i^2}{\left(\sum_{i=1}^{L} E_i\right)^2}, \quad L^{(+)}_{\text{eq}} L^{(-)}_{\text{eq}} = L^2, \quad E^{(+)}_{\text{eq}} = \frac{\sum_{i=1}^{L} E_i}{L^{(+)}_{\text{eq}}}.$$

Таким чином, наведені співвідношення визначають чотирирівневу функціональну модель показників ефективності системи множинного доступу з кодовим розділенням, у якій поставлено у відповідність: енергетичному рівню - показник енергетичної справедливості, який відображає зосередження сумарної енергії ансамблю в обмеженій кількості сигналів ансамблю; структурному рівню – показник структурної ефективності, який відображає наближену оцінку величини нормованих завад множинного доступу; потужнісно-канальному рівню - відношення енергії корисного сигналу до сумарної енергії завад множинного доступу та шуму, яке є композицією функціоналів енергетичного та структурного рівнів; системному рівню - функціонал, що визначає показник ефективності системного рівня відповідно до поставленої задачі і є композицією функціоналів енергетичного, структурного та потужнісно-канального рівнів. Отже, модель упорядковує введені в роботі показники ефективності за рівнями та задає функціональний перехід між рівневими показниками, що створює основу для подальшого визначення мінімально необхідного рівня деталізації в межах уведеної чотирирівневої моделі при розв'язанні задач оцінки ефективності, синтезу або оптимізації.

Висновки

Запропоновано багаторівневу функціональну модель показників ефективності системи множинного доступу з кодовим розділенням з виділенням енергетичного, структурного, потужнісно-канального та системного рівнів. У межах моделі задано відповідність між рівнями, показниками ефективності та функціоналами, а також встановлено функціональні зв'язки між показниками ефективності рівнів моделі.

Встановлено основні правила побудови багаторівневої функціональної моделі: кожен рівень моделі визначає одну оцінку ефективності, яка є числовою величиною; показник ефективності рівня задається функціоналом відповідного рівня; для найнижчого рівня показник ефективності визначається через вектор параметрів, які надходять на зазначений рівень безпосередньо; для рівнів, наступних після найнижчого, показник ефективності визначається через показник ефективності попереднього рівня та вектор параметрів, які надходять на відповідний рівень безпосередньо; якщо в межах одного рівня задано кілька функціоналів, показник ефективності зазначеного рівня визначається композицією таких функціоналів без обмежень щодо виду композиції.

У багаторівневій функціональній моделі показників ефективності системи множинного доступу з кодовим розділенням поставлено у відповідність: енергетичному рівню - показник енергетичної справедливості, який відображає зосередження сумарної енергії ансамблю в обмеженій кількості сигналів ансамблю; структурному рівню - показник структурної ефективності, який відображає наближену оцінку величини нормованих завад множинного доступу; потужнісно-канальному рівню - відношення енергії корисного сигналу до сумарної енергії завад множинного доступу та шуму, яке є композицією функціоналів енергетичного та структурного рівнів; системному рівню - функціонал, що визначає показник ефективності системного рівня відповідно до поставленої задачі і є композицією функціоналів енергетичного, структурного та потужнісно-канального рівнів.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Indyk S. V., Lysechko V. P., Zhuchenko O. S., Kitov V. S. The Formation Method of Complex Signals Ensembles by Frequency Filtration of Pseudo-Random Sequences with Low Interaction in the Time Domain. *Radio Electronics, Computer Science, Control*. 2020. No. 4. P. 7–14. DOI: https://doi.org/10.15588/1607-3274-2020-4-1
2. Zhuchenko O., Panchenko S., Lysechko V., Indyk S. Methodology of Equivalent Transformations of Unequal-Energy Complex Signal Ensembles in Code-Division Multiple Access Systems. *Комп'ютерно-інтегровані технології: освіта, наука, виробництво.* 2026. Вип. 62. С. 337–345. DOI: https://doi.org/10.36910/6775-2524-0560-2026-62-38. URL: https://cit.lntu.edu.ua/index.php/cit/article/view/884.
3. Ладанюк А. П., Смітюх Я. В., Власенко Л. О., Заєць Н. А., Ельперін І. В. Системний аналіз складних систем управління : навчальний посібник. Київ : НУХТ, 2013. 274 с. URL: https://dspace.nuft.edu.ua/handle/123456789/13508
4. Ніколаєв О. Г., Рвачова Т. В., Соловйов О. І. Функціональний аналіз : навчальний посібник. Харків : Національний аерокосмічний університет "ХАІ", 2020. URL: https://dspace.library.khai.edu/xmlui/handle/123456789/8128.
5. Мокін Б. І., Мокін В. Б., Мокін О. Б. Функціональний аналіз, адаптований до прикладних задач в галузі інформаційних технологій : навчальний посібник. Вінниця : ВНТУ, 2020. URL: https://ir.lib.vntu.edu.ua/handle/123456789/34634
6. Panchenko S. V., Lysechko V. P., Zhuchenko O. S., Indyk S. V. Multi-Level Functional Model of Performance Indicators for Ultra-Wideband Code Division Multiple Access Systems. *Сучасні наукові парадигми: інтеграція знань і технологій* : матеріали науково-практичної конференції, м. Вінниця, 26–27 грудня 2025 р. Одеса : Видавництво «Молодий вчений», 2025. С. 121–123. URL: https://molodyivchenyi.ua/omp/index.php/conference/catalog/view/164/2843/5903-1.
7. Лисечко В. П., Жученко О. С., Індик С. В. Система показників ефективності ансамблів різноенергетичних складних сигналів для надширокосмугових систем множинного доступу з кодовим розділенням. Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління : тези доповідей шістнадцятої міжнародної науково-технічної конференції, 29–30 квітня 2026 р. : у 5 т. Т. 5 : секція 6. Баку ; Харків ; Жиліна, 2026. С. 65–68. DOI: https://doi.org/10.32620/ICT.26.t5.

Received (Надійшла) 14.02.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Лисечко Володимир Петрович – доктор технічних наук, професор, Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Volodymyr Lysechko – Doctor of Technical Sciences, Professor, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: lysechko@kart.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-1520-9515.

Трубчанінова Карина Артурівна – доктор технічних наук, професор, професор кафедри транспортного зв'язку, Український державний університет залізничного транспорту, Харків, Україна;
Karyna Trubchaninova – Doctor of Technical Sciences, Professor, Professor of Department of Transport Communication, Ukrainian State University of Railway Transport, Kharkiv, Ukraine;
e-mail: tka2@ukr.net; ORCID Author ID: https://orcid.org/0000-0003-2078-2647.

Жученко Олександр Сергійович – кандидат технічних наук, доцент, доцент кафедри транспортного зв'язку, Український державний університет залізничного транспорту, Харків, Україна;
Oleksandr Zhuchenko – Candidate of Technical Sciences, Associate Professor, Associate Professor of Department of Transport Communication, Ukrainian State University of Railway Transport, Kharkiv, Ukraine;
e-mail: n030201@gmail.com; ORCID Author ID: https://orcid.org/0000-0003-3275-810X.

Шубіна Галина Валеріївна – навчальний відділ, Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна;
Halyna Shubina – Education Department, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine;
e-mail: gshubina105@gmail.com; ORCID Author ID: https://orcid.org/0009-0006-9895-7987.

Multi-level functional model of performance indicators of a code division multiple access system

Volodymyr Lysechko, Karyna Trubchaninova, Oleksandr Zhuchenko, Halyna Shubina

Abstract. Relevance. To evaluate the efficiency of a code division multiple access system, in particular one based on ensembles of different-energy complex signals, a generalized model is needed. Such a model orders level performance indicators and defines functional relations between them. **Object of research:** performance indicators and functional relations between them in a code division multiple access system. **Purpose of the article:** to develop a multi-level functional model of performance indicators of a code division multiple access system using the tools of system analysis and functional analysis, and to show the possibility of obtaining a generalized efficiency estimate for a code division multiple access system. **Research results.** A multi-level functional model of performance indicators of a code division multiple access system has been developed. The model defines the rules of functional-level decomposition. Within the model, indicators are ordered by the degree of generalization of efficiency estimates, each degree of generalization corresponds to a model level, and relations between indicators are defined by the functionals of the corresponding levels. Each model level forms one numerical efficiency estimate. If several functionals are defined within one level, the performance indicator of this level is defined by their composition. **Conclusions.** A variant of the level structure of the model is proposed with energy, structural, power-channel and system levels. The energy level is related to the energy fairness indicator, the structural level is related to the structural efficiency indicator, the power-channel level is related to the ratio of useful signal energy to the total energy of multiple access interference and noise, and the system level is related to the generalized system performance indicator.

Keywords: computer system; complex signal; performance indicator; multi-level functional model; functional; energy level; structural level; power-channel level; system level.

УДК 621.391:519.217 **doi: 10.26906/SUNZ.2026.2.237**

Ю. С. Меркуленко, М. В. Савченко

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

МАРКІВСЬКА МОДЕЛЬ ОПТИМІЗАЦІЇ РОЗПОДІЛУ СПЕКТРА У КОГНІТИВНИХ РАДІОМЕРЕЖАХ

Анотація. Актуальність. В сучасних бездротових телекомунікаційних системах стрімко зростають обсяги трафіку та виникає дефіцит радіочастотного спектра, особливо в діапазонах нижче 6 ГГц. Статичні моделі розподілу частотних ресурсів призводять до неефективного використання спектра. Вони не відповідають вимогам високодинамічних мереж 5G/6G та систем IoT. Когнітивні радіомережі забезпечують динамічний доступ вторинних користувачів до вільних частотних каналів. Проте ефективність такого доступу визначається коректністю, враховується стохастична поведінка первинних користувачів. Виникає необхідність розробки математично обґрунтованих моделей для оптимізації спектрального доступу, що базуються на основі ймовірнісних процесів. **Об’єкт дослідження**: процеси управління розподілом спектра у когнітивних радіомережах з випадковою появою та зникненням первинних користувачів. **Мета статті**: розробка та дослідження марковської моделі оптимізації спектрального доступу, що основана на процесі прийняття рішень Маркова (MDP). Вона забезпечує збільшення ефективності використання спектра при зменшенні інтерференції з первинними системами. **Результати дослідження**. У статті представлено формалізовану модель управління спектром у вигляді MDP. Модель враховує переходи між станами зайнятості каналів та багатоканальну структуру радіосередовища. Враховані також інтенсивності зміни станів, параметри SINR і характеристики радіоканалу. Оптимальна політика доступу визначається функцією винагороди, яка балансує між покращенням спектральної ефективності та зменшенням ризику конфлікту з первинним користувачем. Результати моделювання показали приріст ефективності використання спектра на 13–15% у порівнянні з базовими підходами. Зменшення частоти конфліктів не перевищує 1,8% та збільшення середньої пропускної здатності вторинного користувача на 15–25%. Отримані характеристики залишаються постійними у широкому діапазоні інтенсивностей появи та зникнення первинних користувачів, що свідчить про адаптивність моделі. **Висновки.** Продемонстровано, що використання марковських процесів прийняття рішень дозволяє досягти оптимального балансу між ефективністю використання спектра та інтерференцією в когнітивних радіомережах. Запропонована система перевершує статичну та жадібну стратегії доступу. При цьому покращена спектральна продуктивність та стабільність роботи у динамічному радіосередовищі. Сфера використання отриманих результатів: когнітивні радіомережі, системи динамічного доступу до спектра, бездротові мережі 5G/6G, інфраструктури IoT, завдання оптимізації радіоресурсів, інтелектуальні алгоритми управління спектром.

Ключові слова: когнітивна радіомережа, марківський процес прийняття рішень, стохастична модель, управління радіоресурсами, бездротові системи, канал зв’язку.

Вступ

Постановка проблеми. З розвитком сучасних бездротових систем, що працюють у все більш складних умовах, зростає попит на ефективне використання радіочастотного спектра для задоволення швидкого зростання обсягів трафіку, кількості взаємопов'язаних пристроїв (таких як радіоперемикачі, голосові маршрутизатори та пристрої зберігання даних) та множини технологій радіодоступу. Традиційна технологія управління частотними ресурсами полягає в тому, що спектр повинен розподілятися статично за послугами або операторами (ці застосування розкидані та нерегулярні). Це особливо важливо на частотах нижче 6 ГГц, де працює більшість мобільних платформ та мереж IoT. В результаті, значна частина спектра залишалася невикористаною протягом деякого часу, і зростає попит на частотні ресурси.

Когнітивне радіо є одним з потенційних способів вирішення цієї проблеми — дозволяючи вторинним користувачам отримувати доступ до частотних каналів за відсутності первинних користувачів — що є перспективним застосуванням [2]. Ця модель забезпечує гнучкість у розподілі ресурсів, а також досягнення більшої ефективності у використанні спектра. Однак використання когнітивних радіомереж на практиці є відносно складним, оскільки первинні користувачі з'являються і зникають випадково та мають різні статистичні характеристики. За відсутності адаптивних механізмів прийняття рішень існує ризик, що вторинні користувачі порушать існуючі системи, що призведе до непослідовної поведінки та деградації послуг.

Отже, є очевидною потреба у математичних моделях, які б характеризували фактичні стани спектра та точно передбачали ці стани в реальному часі. Марковські процеси прийняття рішень (MDP) [3] є надійним підходом до цих проблем, оскільки вони можуть представляти ймовірність кожного переходу стану каналу та давати найкраще рішення щодо того, як вторинні користувачі повинні поводитися в динамічному радіоекосистемі. Використання MDP дозволяє максимізувати пропускну здатність вторинних систем та зменшити інтерференцію з первинними послугами, а також оцінити доцільність та своєчасність подальших досліджень з оптимізації розподілу через стохастичні застосування на спектрі..

Аналіз останніх досліджень і публікацій. Моделювання спектральної зайнятості є однією з основних тем досліджень в галузі когнітивного радіо. Найпоширенішими систематичними підходами стали моделі ON/OFF, в яких активність первинного користувача описана за допомогою експоненційних або гіперекспоненційних розподілів [4-5]. Вони досить легко параметризуються і забезпечують аналітичну зручність, але реальні процеси зазвичай мають квазістаціонарну та фрактальну поведінку, яку класична модель не відтворює за умов нерівномірного трафіку [6, 7].

Для покращення опису запропоновані напівмарковські моделі, які враховують змінну інтенсивність та кореляції між станами [4]. Однак їх висока обчислювальна складність обмежує їх використання в реальному часі. Як результат, зростає інтерес до підходів, основаних на MDP, які пропонують як формальну строгість, так і потенціал для оптимізації динаміки вторинного користувача в динамічному середовищі [8]. Методи машинного навчання та підсиленого навчання розробляються з високою точністю прогнозування, особливо у випадку 6G [9–11]. Однак вони покладаються на величезну кількість даних і не забезпечують оптимальності за умов часткової спостережуваності. Підходи MDP залишаються загальноприйнятою, універсальною та математично обґрунтованою основою для оптимізації динамічного доступу до спектра в даному контексті.

Метою роботи є розробка та аналіз моделі Маркова для управління спектральними ресурсами в когнітивній радіомережі, щоб максимізувати ефективність використання спектральних ресурсів за умов стохастичної появи первинних користувачів. Необхідно створити математичну структуру, котра дозволить формувати оптимальну політику вибору каналу, що базується на основі ймовірнісних характеристик станів спектра та поведінці первинних систем.

Досягнення мети залежить від кількох задач. Спочатку модель марковського процесу повинна бути формалізована, що допоможе описати переходи в станах зайнятості каналу і визначити форму можливих дій вторинного користувача. Також потрібно розробити функцію винагороди, яка адекватно відображає баланс між ефективним використанням спектру та потенційними перешкодами. Фінальним етапом є проведення експериментального моделювання для оцінки продуктивності оптимальної політики в порівнянні з базовими методами доступу. Це дозволить підтвердити коректність роботи моделі та визначити її переваги в реальному радіосередовищі.

Основний матеріал

Формування оптимальної стратегії розподілу спектра у когнітивних радіомережах ґрунтується на моделюванні процесу зайнятості частотних ресурсів первинними користувачами та визначенні дій вторинного користувача у змінних умовах радіосередовища. Для цього використано апарат марковських процесів прийняття рішень (Markov Decision Process, MDP), який дозволяє формально описати залежність між станами спектра, діями вторинного вузла та стохастичною природою появи/зникнення первинних користувачів.

Стан когнітивної системи позначено множиною $S = \{s_1, s_2, ..., s_n\}$, де кожен стан визначає конфігурацію доступності спектра. Дії вторинного користувача описуються множиною $A = \{a_1, a_2, ..., a_m\}$, яка включає вибір каналу, зміну потужності чи утримання поточного ресурсу. Ймовірність переходу між станами визначається марковською матрицею $P = \{p_{ij}(a)\}$, де елемент $p_{ij}(a)$ позначає ймовірність переходу зі стану s_i до стану s_j за умови виконання дії a.

Процес описано як чотириелементний кортеж

$$M = \langle S, A, R, P \rangle, \quad (1)$$

де $R(s,a)$ – миттєва винагорода за виконання дії a у стані s, що відображає доцільність вибору певного спектрального ресурсу. Винагорода визначається як різниця між корисним використанням спектра та можливими штрафами за перешкоди первинному користувачу. Для кожного стану вводиться функція оптимальної цінності

$$V(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad (2)$$

де $\gamma \in (0,1)$ – коефіцієнт дисконту, який знижує вагу майбутніх винагород.

Для отримання оптимальної політики $\pi^*(s)$ використано рівняння Беллмана. Оптимальна дія у кожному стані обчислюється за правилом

$$\pi^*(s) = \arg\max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right]. \quad (3)$$

Структура переходів між станами базується на моделюванні появи первинного користувача з інтенсивністю λ та його зникнення з інтенсивністю μ. Для найпростішої одноканальної моделі ймовірності переходів задано

$$p_{01} = \lambda \Delta t, \; p_{10} = \mu \Delta t, \quad (4)$$

де стани 0 і 1 відповідають вільному та зайнятому каналу відповідно, а Δt – крок дискретизації часу. Для багатоканального випадку формується прямий добуток окремих підпроцесів, що зберігає марковську структуру та дозволяє масштабувати модель до великих когнітивних мереж.

Очікувана ефективність використання спектра обчислюється як математичне сподівання сумарної винагороди за нескінченний горизонт:

$$\eta = \sum_{s \in S} \pi^*(s) V(s). \quad (5)$$

Отримана величина використовується як ключовий критерій продуктивності, що дозволяє кількісно порівнювати різні політики керування та оцінювати вплив інтенсивностей появи/зникнення первинних користувачів на доступність спектра.

На рис. 1 представлено граф станів марковської моделі процесу прийняття рішень (MDP), який використовується для опису динамічного розподілу спектра у когнітивній радіомережі. Кожне коло позначає стан системи s_1, s_2, s_3, s_4, що відповідає різним конфігураціям доступності спектральних каналів. Спрямовані стрілки відображають можливі переходи між станами при виконанні вторинним користувачем певної дії a, а позначення $p_{ij}(a)$ вказують на ймовірності переходу зі стану s_i до стану s_j. Наявність петель та перехресних переходів демонструє стохастичну природу радіосередовища, у якому поява або зникнення первинного користувача може спричинити зміну стану у будь-який момент часу. Така структура

графа дозволяє формально описати варіанти поведінки когнітивної системи та побудувати оптимальну політику доступу до спектра.

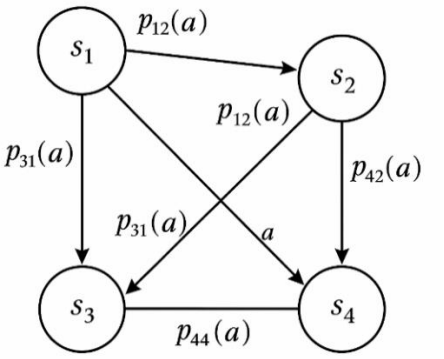


Рис. 1. Граф станів MDP для спектра

Запропонована модель забезпечує формальну основу для побудови оптимальної стратегії динамічного доступу до спектра, дозволяє врахувати стохастичний характер зайнятості каналів та визначити політику, яка мінімізує втрати спектрального ресурсу при збереженні вимог до недопущення інтерференції первинних користувачів.

Моделювання. Моделювання спрямоване на оцінювання ефективності оптимальної політики доступу до спектра, сформованої на основі марковської моделі MDP, у динамічному середовищі когнітивної радіомережі. Для цього створено багатоканальне сценарне оточення, що складається з п'яти незалежних частотних каналів у діапазоні 2,3–2,6 ГГц, кожен із яких може перебувати у станах «вільний» або «зайнятий» первинним користувачем. Сукупний простір станів включає 32 можливі комбінації зайнятості каналів. Вторинний користувач працює з максимальною передавальною потужністю 100 мВт, що відповідає типовим параметрам малопотужних когнітивних пристроїв, і переключає канали на основі оптимальної політики.

Для моделювання поведінки первинних користувачів застосовано пуассонівський процес появи з інтенсивністю у діапазоні $\lambda = 0{,}15...0{,}75c^{-1}$, що відповідає широкому спектру реальних навантажень, від малозайнятих до активно використовуваних каналів. Інтенсивність звільнення каналу визначено інтервалом $\mu = 0{,}5...1{,}3c^{-1}$, що моделює середню тривалість зайнятості від 0,7 до 2 секунд. Такі значення характерні для мобільних первинних систем із варіативним навантаженням. Перехідні ймовірності формуються за стандартною апроксимацією Маркова: $p_{01} = \lambda\Delta t$, $p_{10} = \mu\Delta t$, де крок дискретизації $\Delta t = 0{,}05c$, що забезпечує тонку часову градацію процесу.

Середовище каналів відтворює завади та втрати поширення з урахуванням логнормального федінгу зі стандартним відхиленням 4 дБ та середньою потужністю шуму – 104 дБм при смузі 200 кГц. Для обчислення успішності передачі використовується миттєвий рівень *SINR*, розрахований за співвідношення

$$SINR(t) = \frac{P_t G(d)}{N_0 B + I(t)}, \qquad (6)$$

де $P_t = 100$ мВт, $G(d) = d^{-3,5}$ – модель втрат з експонентою 3,5 при середній відстані 120 метрів, $N_0 B = -104$ дБм, $I(t)$ – інтерференція від інших вторинних пристроїв, що задається стохастично у межах [–110,–95] дБм. Передача вважається успішною, якщо $SINR > \gamma_{th}$, де поріг обрано $\gamma_{th} = 6$дБ. Очікувана пропускна здатність каналу для вторинного користувача визначається за формулою Шеннона

$$C(t) = B\log_2\left(1 + SINR(t)\right), \qquad (7)$$

де B = 200 кГц. Середня пропускна здатність за епізод моделювання є одним із ключових показників продуктивності вторинної системи.

Винагородна функція MDP базується на співвідношенні між корисним використанням спектра та штрафом за інтерференцію. Передача на вільному каналі дає винагороду +1, тоді як створення перешкод первинному користувачу карається штрафом –7, що відповідає стандартам IEEE 802.22 у частині суворості покарання за порушення правил доступу. Дія «утримання каналу» оцінюється як 0. Політика оптимізації формується за допомогою рівнянь Беллмана для нескінченного горизонту з параметром дисконтування $\gamma = 0{,}92$, що забезпечує баланс між миттєвою винагородою та довгостроковою доступністю спектра. Тривалість кожного епізоду становила 300 секунд, що відповідає 6000 дискретним крокам. Для кожної комбінації параметрів (λ,μ) виконувалося 600 незалежних прогонів, що дозволяє усереднити результати з довірчим інтервалом до 1,5%. Вторинний користувач спершу починає з випадкового стану, потім протягом перших 3 секунд відбувається етап «прогріву», після чого результати фіксуються.

Для оцінювання ефективності введено аналітичні метрики. Середня ефективність використання спектра визначається як відношення часу, коли вторинний користувач виконував успішну передачу, до загальної тривалості епізоду:

$$\eta = \frac{1}{T}\sum_{t=1}^{T} 1\left\{SINR(t) > \gamma_{th},\ s \in S_{free}\right\}. \qquad (8)$$

Ймовірність конфлікту з первинним користувачем формально оцінюється як

$$P_{conf} = \frac{1}{T}\sum_{t=1}^{T} 1\left\{s(t) \in S_{busy}\right\}, \qquad (9)$$

а середня пропускна здатність когнітивного каналу:

$$C_{avg} = \frac{1}{T}\sum_{t=1}^{T} C(t). \qquad (10)$$

Порівняння оптимальної політики з базовими стратегіями (random access та greedy) виконується за трьома метриками: η, P_{conf} та C_{avg}. Такий підхід дозволяє об'єктивно оцінити приріст спектральної ефективності та зменшення конфліктів.

Результати моделювання. Результати моделювання демонструють підвищення ефективності використання спектра вторинним користувачем при застосуванні оптимальної політики, отриманої з марковської моделі MDP. У ході експериментів було досліджено поведінку системи в широкому діапазоні інтен-

сивностей появи та зникнення первинних користувачів, а також проаналізовано залежність спектральної ефективності від рівня зайнятості каналів та стохастичних характеристик радіосередовища. Нижче наведені три ключові графічні залежності, що відображають основні закономірності функціонування моделі.

Графік на рис. 2 відображає середню ефективність використання спектра залежно від інтенсивності появи первинного користувача. На рисунку чітко видно, що при незначних інтенсивностях $\lambda < 0,3c^{-1}$ оптимальна політика та жадібна стратегія демонструють близькі результати, оскільки канали тривалий час залишаються вільними. Проте при збільшенні інтенсивності появи первинного користувача ефективність базових методів стрімко знижується. Натомість MDP-політика демонструє суттєво кращу адаптивність, зберігаючи стабільно високі значення ефективності навіть при значеннях λ близьких до 0,7–0,8. Це свідчить про здатність моделі прогнозувати зміни станів каналу та обирати дії, які мінімізують ризик конфлікту.

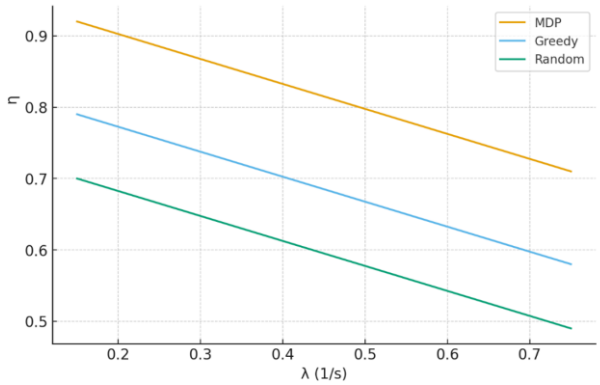


Рис. 2. Залежність ефективності використання спектра від інтенсивності появи первинного користувача λ

Графік на рис. 3 характеризує ймовірність конфлікту вторинного користувача з первинним залежно від інтенсивності звільнення каналу.

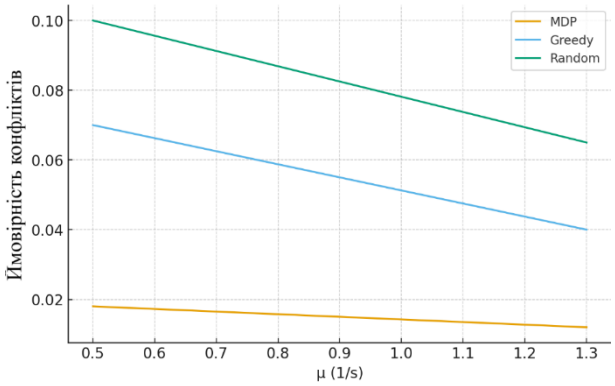


Рис. 3. Ймовірність конфлікту з первинним користувачем залежно від інтенсивності звільнення каналу μ.

Отримані результати показали, що оптимальна політика суттєво знижує частоту конфліктів у всьому діапазоні $\mu = 0,5...1,3c^{-1}$. У випадку жадібної стратегії ймовірність конфлікту становить 4–7%, тоді як оптимальна політика зменшує її до рівня 1,2–1,8%.

У системах когнітивного радіо це критично, оскільки зменшення конфліктів не тільки знижує завади первинним користувачам, але й мінімізує ризик блокування вторинного доступу. Відповідна залежність також підтверджує, що збільшення швидкості звільнення каналу позитивно впливає на стабільність MDP-політики та сприяє підвищенню її точності.

Графік на рис. 4 ілюструє середню пропускну здатність вторинного користувача для трьох підходів: випадкового доступу, жадібної стратегії та оптимальної MDP-політики. Отримані результати засвідчують, що оптимальна політика забезпечує максимальну середню пропускну здатність у всіх розглянутих сценаріях навантаження. За умов помірної інтерференції (середній рівень SINR у межах 6–10 дБ) MDP-політика демонструє приріст пропускної здатності приблизно на 13–15% порівняно з greedy-стратегією та до 20–25% порівняно з випадковим доступом. Навіть у випадках високої інтенсивності появи первинних користувачів середня пропускна здатність за MDP залишається вищою, що свідчить про адекватність моделі як при низькому, так і при високому навантаженні спектра.

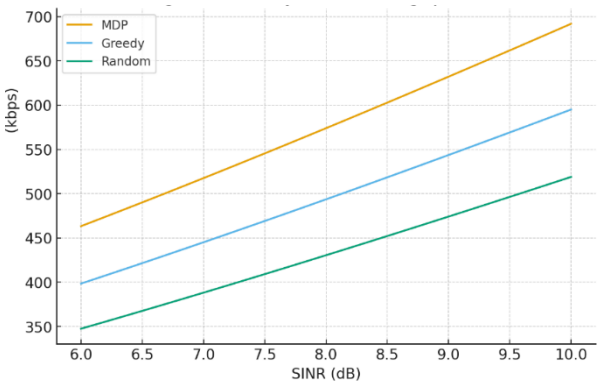


Рис. 4. Середня пропускна здатність вторинного користувача для трьох стратегій доступу: Random, Greedy та MDP

Загалом, результати моделювання вказують на те, що запропонована модель оптимізації доступу до спектру на основі Маркова демонструє високу ефективність. Вона також показала значний приріст ефективності за трьома важливими метриками: зменшення частоти конфліктів, доступність вільних каналів та збільшення середньої пропускної здатності вторинного користувача. Помітний контраст між політикою MDP та базовими підходами ще більш виразний у складних умовах радіо середовища, підтверджуючи життєздатність використання MDP у когнітивних радіомережах з динамічним розподілом спектру.

Аналіз результатів. Отримані результати моделювання дають змогу оцінити ефективність запропонованої моделі Маркова для оптимізації розподілу спектра у когнітивних радіомережах. Також визначити її переваги при порівнянні з традиційними методами доступу. Таким чином, продуктивність оптимальної політики, розробленої на основі MDP, здатна реагувати на зайнятість спектру та передбачати найбільш сприятливі стани для зв'язку, забезпечуючи кращу узагальнену продуктивність у різних радіосередовищах. Поведінка моделі за умов зростання інтенсивності появи первин-

них користувачів продемонструвала важливу закономірність: жадібна стратегія та випадковий доступ відчутно деградує за своїми характеристиками при $\lambda < 0,4c^{-1}$. Це пов'язано з тим, що обидва базові методи реагують на зміни стану спектра лише ретроспективно, не враховуючи ймовірності майбутніх переходів. З іншого боку, політика MDP активно використовує прогнозну природу моделі Маркова, активно вибираючи дії враховуючи очікувані зміни станів. Завдяки цьому забезпечується повільне зменшення спектральної ефективності оптимальної політики, навіть при високому навантаженні. Аналіз графіка ефективності демонструє, що при $\lambda \approx 0,7c^{-1}$ MDP зберігає ефективність на 30–40% вищу за випадковий доступ, це свідчить про її перевагу за умов сильної варіативності радіосередовища.

Вірогідність конфлікту з первинним користувачем є ключовим показником для будь-якої когнітивної системи. Оскільки надмірна інтерференція призводить до обмеження або навіть заблокування вторинного доступу. Порівняльний аналіз показав, що MDP-політика мінімізує частку конфліктів до рівня, що не перевищує 1,8%, що у 3–5 разів нижче порівняно з жадібним методом. Пояснення цьому полягає в тому, що найкраща стратегія налаштовується не лише на поточний стан каналу, але й на передбачену ймовірність знаходження первинного користувача поблизу цього стану. Це підтверджується тим, що політика MDP показує здатність значно зменшувати частку конфліктних ситуацій у сценаріях з високою зміною стану, де вони складають 6–7% небезпечних передач, таким чином виконуючи регуляторні вимоги доступу в когнітивних мережах.

Аналіз середнього виходу вторинного користувача підтверджує вищезазначені закономірності в рамках спектральної ефективності. Висока пропускна здатність досягається не лише шляхом вибору вільних каналів, але й каналів з хорошими характеристиками SINR, коли модель обирає канал з меншою ймовірністю присутності для первинних застосувань. Найкраща політика приймає метод забезпечення тривалішних періодів стабільного зв'язку і таким чином позитивно впливає на поширення даних. Слід зазначити, що MDP при нормальних рівнях втручання забезпечує середню швидкість передачі даних на 15–25% вищу, ніж інші методи, і для рівня шуму також підтримує свою продуктивність завдяки регулярному поверненню до каналів з хорошими характеристиками затухання.

Інтегральний аналіз за трьома ключовими метриками показав, що MDP-політика не лише здатна досягти кращої продуктивності, але й залишається стабільність при будь-яких значеннях параметрів λ та μ. Усі результати показують, що покращення ефективності завдяки моделі Маркова найбільш очевидне в складних умовах, які вимагають використання моделі Маркова, коли загальні статичні та напівстатичні підходи більше не можуть адаптуватися до ситуації. Таким чином, оптимізація Маркова підтверджує свою здатність підвищувати стійкість когнітивної системи, мінімізуючи конфлікти, а також покращуючи пропускну здатність, необхідну для реалізації когнітивного доступу на практиці, особливо в сучасних телекомунікаційних мережах.

Висновки

У роботі розроблено та досліджено марковську модель оптимізації, яка враховує фізичну варіабельність у появі та втраті первинних користувачів. Формалізація моделі, основаної на процесі прийняття рішень Маркова, дала змогу створити оптимальну політику доступу, котра максимізує корисне використання спектра та мінімізує ризик виникнення конфліктів із первинними вузлами. Модель побудована з використанням реалістичних параметрів радіосередовища, таких як багатоканальна структура, інтенсивність зміни стану, рівні шуму, SINR та характеристики радіо каналів, що забезпечує хорошу симуляцію умов роботи когнітивної мережі.

Результати моделювання показали, що оптимальна політика MDP перевершила базові підходи доступу, особливо випадковий розподіл каналів та жадібну стратегію. Використання марковської оптимізації забезпечило підвищення спектральної ефективності на 13–15%, зниження частоти конфліктів з первинними користувачами до рівня не більше 1,8% та збільшення середньої пропускної здатності вторинного користувача на 15–25%. Однорідність результатів аналізу характеристик у широкому діапазоні інтенсивностей появи та зникнення первинних користувачів виявила стійкість запропонованої моделі для досягнення ефективного доступу до спектра в умовах постійно змінюваного радіо ландшафту.

Таким чином, запропонована марковська модель оптимізації розподілу спектра є перспективним механізмом для побудови інтелектуальних алгоритмів когнітивного радіо. Вона досягає компромісу між зменшенням втручання, покращенням продуктивності вторинного доступу та стійкістю до стохастичних змін спектру. Майбутні дослідження полягають у розширенні моделі на багатокористувацький когнітивний доступ, впровадження методів підкріплювального навчання та вивчення співпраці між політикою MDP та стратегіями кооперативного виявлення спектру.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, Shantidev Mohanty. NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer Networks*. 2006. Vol. 50, no. 13. P. 2127-2159. DOI: https://doi.org/10.1016/j.comnet.2006.05.001

2. Arjoune Y., Kaabouch N. A Comprehensive Survey on Spectrum Sensing in Cognitive Radio Networks: Recent Advances, New Challenges, and Future Research Directions. *Sensors*. 2019. Vol. 19, no. 1:126. DOI: https://doi.org/10.3390/s19010126
3. Li H., Wang F., Qian J., Zhu P., Zhou A. Partitioned RIS-Assisted Vehicular Secure Communication Based on Meta-Learning and Reinforcement Learning. *Sensors*. 2025. Vol. 25, no. 18:5874. DOI: https://doi.org/10.3390/s25185874
4. Das D., Das S. A Survey on Spectrum Occupancy Measurement for Cognitive Radio. *Wireless Personal Communication*. 2015. Vol. 85. P. 2581–2598. DOI: https://doi.org/10.1007/s11277-015-2921-1
5. Sumit Kumar Agrawal, Abhay Samant, Sandeep Kumar Yadav. Spectrum sensing in cognitive radio networks and metacognition for dynamic spectrum sharing between radar and communication system: A review. *Physical Communication*. 2022. Vol. 52:101673. DOI: https://doi.org/10.1016/j.phycom.2022.101673
6. Abdelbaset S. E., Kasem H. M., Khalaf A. A., Hussein A. H., Kabeel A. A. Deep Learning-Based Spectrum Sensing for Cognitive Radio Applications. *Sensors*. 2024. Vol.24, no. 24:7907. DOI: https://doi.org/10.3390/s24247907
7. Xu W., Zhang J., Su Z., Jia L. Explainable Multi-Frequency Long-Term Spectrum Prediction Based on GC-CNN-LSTM. *Electronics*. 2025. Vol. 14, no. 17:3530. DOI: https://doi.org/10.3390/electronics14173530
8. Safavinejad Ramin, Chang Hao-Hsuan, Liu Lingjia. Deep Reinforcement Learning for Dynamic Spectrum Access: Convergence Analysis and System Design. *IEEE Transaction on Wireless Communications*. 2024. Vol. 23, no. 12(2). P.18888-18902. DOI: https://doi.org/10.1109/TWC.2024.3414428
9. K. B. Letaief, W. Chen, Y. Shi, J. Zhang, Y. -J. A. Zhang. The Roadmap to 6G: AI Empowered Wireless Networks. *IEEE Communications Magazine*. 2019. Vol. 57, no. 8. P. 84-90. DOI: https://doi.org/10.1109/MCOM.2019.1900271
10. I. F. Akyildiz, A. Kak, S. Nie. 6G and Beyond: The Future of Wireless Communications Systems. *IEEE Access*. 2020. Vol. 8. P. 133995-134030. DOI: https://doi.org/10.1109/ACCESS.2020.3010896
11. Soares M. D., Passos D., Castellanos P. V. G. Cognitive Radio with Machine Learning to Increase Spectral Efficiency in Indoor Applications on the 2.5 GHz Band. *Sensors*. 2023. Vol. 23, no. 10:4914. DOI: https://doi.org/10.3390/s23104914

Received (Надійшла) 11.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Меркуленко Юрій Сергійович – аспірант кафедри системи інформації ім. В.О. Кравця, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Yurii Merkulenko – PhD Student, Department of Information Systems named after V. O. Kravets, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: Yurii.Merkulenko@infiz.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0003-5490-1293;

Савченко Микола Володимирович – кандидат фізико-математичних наук, доцент, доцент кафедри системи інформації ім. В.О. Кравця, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Mykola Savchenko – Candidate of Physical and Mathematical Sciences, Associate Professor, Associate Professor of Department of Information Systems named after V. O. Kravets, National Technical University «KhPI», Kharkiv, Ukraine;
e-mail: Mykola.Savchenko@khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0005-7366-3213;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=7101640966.

Markov model of spectrum allocation optimization in cognitive radio networks

Yurii Merkulenko, Mykola Savchenko

Abstract. Relevance. In modern wireless telecommunications systems, traffic volumes are rapidly increasing and there is a shortage of radio frequency spectrum, especially in the bands below 6 GHz. Static models of frequency resource allocation lead to inefficient use of spectrum. They do not meet the requirements of highly dynamic 5G/6G networks and IoT systems. Cognitive radio networks provide dynamic access of secondary users to free frequency channels. However, the effectiveness of such access is determined by the correctness, and the stochastic behavior of primary users is taken into account. There is a need to develop mathematically based models for optimizing spectral access, based on probabilistic processes. **Object of research**: spectrum allocation management processes in cognitive radio networks with random appearance and disappearance of primary users. **Purpose of the article**: development and study of a Markov model of spectral access optimization, based on the Markov decision process (MDP). It provides increased spectrum efficiency while reducing interference with primary systems. **Research results**. The article presents a formalized spectrum management model in the form of MDP. The model takes into account transitions between channel occupancy states and the multichannel structure of the radio environment. The intensities of state changes, SINR parameters, and radio channel characteristics are also taken into account. The optimal access policy is determined by a reward function that balances between improving spectral efficiency and reducing the risk of conflict with the primary user. The modeling results showed an increase in spectrum efficiency by 13–15% compared to the basic approaches. The reduction in the conflict frequency does not exceed 1.8% and an increase in the average throughput of the secondary user by 15–25%. The obtained characteristics remain constant in a wide range of intensities of the appearance and disappearance of primary users, which indicates the adaptability of the model. **Conclusions**. It is demonstrated that the use of Markov decision processes allows achieving an optimal balance between spectrum efficiency and interference in cognitive radio networks. The proposed system outperforms static and greedy access strategies. At the same time, spectral performance and stability in a dynamic radio environment are improved. The scope of application of the obtained results: cognitive radio networks, dynamic spectrum access systems, 5G/6G wireless networks, IoT infrastructures, radio resource optimization tasks, intelligent spectrum management algorithms.

Keywords: cognitive radio network, Markov decision process, stochastic model, radio resource management, wireless systems, communication channel.

UDC 004.382.2; 004.312 **doi: 10.26906/SUNZ.2026.2.243**

Dmytro Salnikov, Oleg Vasylchenkov

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

AREA-EFFICIENT HARDWARE MODULES FOR FP16/FP8/FP32 FORMAT CONVERSION IN EMBEDDED SYSTEMS

Abstract. The rapid proliferation of neural networks in embedded and edge computing systems has led to an increasing demand for efficient hardware implementations that can support precision-scalable arithmetic. Applications such as autonomous vehicles, intelligent sensors, and industrial automation require high computational performance, low latency, and strict energy constraints. Floating-point arithmetic, defined by the IEEE 754 standard, remains the dominant numerical representation in such systems due to its versatility and broad dynamic range. However, deploying modern deep learning models on resource-limited platforms poses significant challenges in balancing accuracy, throughput, and hardware footprint. To address these challenges, emerging reduced-precision formats such as FP16, BF16, and FP8 (E4M3, E5M2) have gained popularity for both inference and training, enabling decreased memory bandwidth and improved energy efficiency with minimal accuracy degradation. Despite their growing prevalence, many microcontrollers and FPGAs lack native hardware support for these low-precision formats, motivating the need for compact and reconfigurable conversion modules capable of bridging compatibility with conventional FP32 processing units. This work presents the design, implementation, and hardware evaluation of fully synthesizable VHDL modules for converting between FP8, FP16, BF16, and standard IEEE-754 single-precision (FP32) formats. The proposed architecture leverages FPGA Look-Up Tables (LUTs) to perform exponent and mantissa field manipulation, bias adjustment, and classification of special numerical cases such as Infinity and NaN, ensuring full standard compliance. The converters were synthesized using a commercial design flow targeting an Intel Cyclone V device. Experimental results demonstrate exceptionally low resource utilization and high operating frequency, with the FP8E4M3 and FP8E5M2 converters each requiring only 14 ALMs while achieving frequencies exceeding 500 MHz. These outcomes confirm the suitability of the proposed modules for deployment in mixed-precision computing systems and embedded neural network accelerators, providing an efficient hardware foundation for energy-aware and high-performance AI workloads on constrained platforms.

Keywords: floating-point formats, reduced-precision number representation, embedded systems, edge computing, FPGA, VHDL, embedded neural network acceleration, area-efficient architecture.

Introduction

Nowadays, neural networks are increasingly integrated into modern embedded systems, enabling intelligent features such as real-time object detection, speech recognition, and sensor data analysis directly on edge devices. These capabilities underpin a wide range of applications, including autonomous driving, smart cameras, wearables, and industrial IoT, where low latency and energy efficiency are critical. Deploying such models on resource-constrained hardware requires optimized computation and compact data formats making precision-scalable representations like FP16, FP8, and FP32 essential for balancing accuracy, performance, and footprint.

Floating-point arithmetic, standardized by IEEE 754, remains the most versatile and widely used method for numerical computation in modern systems. Among the available formats, “single precision” offers an effective balance between precision, dynamic range, and implementation cost. It serves as the standard choice FP type for most embedded system ICs manufacturers. At the same time, many applications still rely on fixed-point math usage [1].

Background and Related Work

Although FP8 formats are relatively recent (introduced in 2022 in [2]), they have rapidly gained adoption in contemporary neural network architectures. The use of FP8 and FP16 numerical formats has become increasingly prevalent in contemporary neural network workloads.

These low-precision formats allow models to be quantized to a smaller bit-width without a dramatic loss in accuracy, reducing both memory bandwidth and storage requirements.

FP8, in particular, offers a favorable trade-off between dynamic range and computational efficiency, enabling faster inference and lower energy consumption on hardware that supports SIMD or specialized multiply-accumulate units. FP16 remains a common choice for training and fine-tuning tasks because of its wider range and compatibility with existing accelerators. Tradeoffs of INT8 vs FP8 usage are described in [3]. Research in [4] presents a comparison of U-Net performance and memory efficiency across various data representation formats, including FP32, FP16, and INT8 quantization.

Despite their advantages, many microcontroller devices lack native FP8 or even FP16 support, relying instead on integer data types. This limitation motivates the need for dedicated, hardware-optimized conversion modules that can bridge the gap between the neural-network-friendly FP8/FP16 formats and the native FP32 or integer representations found on embedded processors.

A mixed-precision ALU unit and related architectures were introduced in [5], [6] and [7]. Despite demonstrating strong performance characteristics, the use of low-precision computations remains rather limited. Moreover, it leads to increased resource consumption. Most low-precision floating-point operations used in neural network frameworks use FP32 calculations internally.

The main goal of this work is to design and evaluate compact, fully synthesizable "cast-only" hardware modules for efficient conversion between compact floating-point data formats and IEEE-754 single-precision format, targeting resource-constrained embedded and FPGA-based systems.

Architecture of LUT-Based Floating-Point Conversion Modules

Currently, the number of parameters in modern neural network architectures may vary significantly across different model families and configurations (Table 1).

This variability has a direct impact on the overall memory footprint of the system.

Table 1 – **Number of parameters for common neural network topologies**

Model	Number of parameters
wav2vec 2.0 (base)	95 million [8]
wav2vec 2.0 (large)	317 million [8]
Whisper (tiny)	39 million
Whisper (base)	74 million
Whisper (small)	244 million
Whisper (medium)	769 million
Whisper (large-v2 / large-v3)	1.55 billion
YOLOv5 (small)	7.2 million
YOLOv5 (large)	46.5 million
LLaMA 3	from 8M up to 405 billions
Mixtral 8x7B	from 8M up to 140 billions [9]

Moreover, storage requirements are determined not only by the parameters themselves but also by the intermediate activation values generated during computation, both of which contribute substantially to the total memory demand. As a result, understanding parameter count and activation behavior is essential when designing compute- and memory-efficient hardware modules for embedded deployments.

Quantization is often employed to reduce the memory footprint of neural network models by representing parameters and activations with lower-precision numerical formats. While this approach can substantially decrease storage requirements and improve computational efficiency (as was shown in [10]), it is not universally achievable.

Certain models or layers exhibit sensitivity to reduced precision, leading to significant degradation in accuracy or instability during inference. Moreover, some operations require higher numerical fidelity to preserve convergence or maintain representational capacity. As a result, although quantization is a powerful technique for shrinking memory usage, its applicability depends on the model architecture, task constraints, and tolerance for accuracy loss.

At the moment, several floating-point formats are in use across modern machine-learning and hardware platforms (Table 2). Beyond the conventional FP32 format, reduced-precision types such as FP16, BF16, and emerging FP8 variants (E4M3 and E5M2) have gained widespread adoption in both training and inference workloads. This diversity of numerical representations enables significant improvements in performance and energy efficiency, but also introduces new considerations for model stability and hardware support.

Table 2 – **Parameters of floating point types encoding**

Format	Sign	Exponent Bits	Mantissa Bits	Exponent Bias
float32 (FP32)	1	8	23	127
float16 (FP16)	1	5	10	15
bfloat16 (BF16)	1	8	7	127
fp8_e4m3 (143)	1	4	3	7
fp8-e5m2 (152)	1	5	2	15

A representative example is provided in [11], which introduces an adaptive quantization methodology for FP8-based deep neural network training.

Modern FPGAs contain fundamental building blocks for implementing combinational and sequential logic — Look-Up Tables (LUTs). Depending on the vendor, they are organized within slices or adaptive logic modules. Xilinx FPGAs typically use slices containing four 6-input LUTs, along with associated flip-flops and carry-chain logic, while Intel FPGAs employ Adaptive Logic Modules with fracturable LUTs supporting up to 8 inputs.

LUTs are highly flexible, allowing complex logic functions to be implemented efficiently, or partitioned into smaller functions to maximize resource utilization.

For our floating-point type conversion blocks, LUTs provide an ideal substrate for implementing the combinational logic required for exponent and mantissa manipulation, as well as offset adjustments. By leveraging the fracturable or multi-LUT capabilities, these blocks can achieve high throughput and area efficiency, minimizing the consumption of registers and logic resources while supporting multiple precision formats such as FP32, FP16, BF16, and FP8.

All modules share a similar structural organization. To convert a low-precision floating-point number into the IEEE-754 single-precision FP32 format, the sign, exponent, and mantissa fields are first extracted from the input representation. The bit widths of these fields are then adjusted to match the FP32 format, including the appropriate exponent bias transformation and mantissa expansion.

Special numerical cases, such as positive and negative infinity and Not-a-Number (NaN), are explicitly detected and handled to ensure full compliance with the IEEE-754 standard.

The overall architecture of the proposed modules is illustrated in Fig. 1.

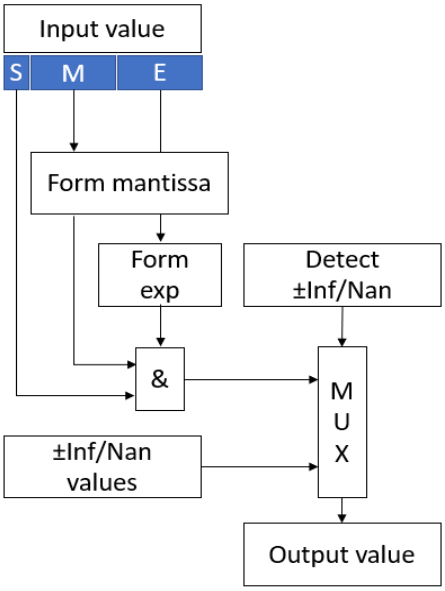


Fig. 1. Conversion block diagram

The diagram presents the structural organization and data flow between the main functional blocks, highlighting the bits extraction, classification, conversion, and value form stages implemented in the design.

Implementation and Results Analysis

In this work, fully synthesizable VHDL modules supporting the FP16, FP8E5M2, FP8E4M3, and BF16 floating-point formats were designed and implemented.

All modules comply with standard hardware design flows, enabling straightforward deployment on both FPGA- and ASIC-based platforms. Functional correctness was verified at the register-transfer level, after which the designs were synthesized using a commercial synthesis tool targeting an Intel Cyclone V FPGA.

The resulting hardware characteristics, including logic utilization and maximum achievable operating frequency, are summarized in Table 3.

These results provide a quantitative comparison of the hardware cost associated with each supported floating-point format and highlight the efficiency of the proposed cast-only conversion architecture.

Table 3 – **Synthesis results for Intel Cyclone V FPGA**

Module	Logic ALMs	Maximum frequency
fp16_to_fp32_bits	42	254
fp8e5m2_to_fp32	14	494
fp8e4m3fn_to_fp32	14	507
bf16_to_fp32_bits	5	497

Overall, the proposed designs demonstrate exceptionally low hardware resource utilization while maintaining high operating frequencies. These characteristics make the modules well suited for integration into modern System-on-Chip (SoC) platforms, including resource-constrained embedded processors and neural network accelerators, where area efficiency and performance are critical.

Conclusions

The proposed conversion modules can be integrated into various parts of an embedded processing pipeline, including the data access stage, ALU pipeline, or as standalone memory-mapped units.

Such flexibility enables efficient adaptation to different system architectures and applications. By offloading precision conversion tasks to dedicated hardware, these modules reduce the overall computational burden, minimize resource duplication, and lower hardware costs.

At the same time, they provide a convenient and scalable mechanism for utilizing low-precision arithmetic, thereby enhancing perfor-mance and energy efficiency in embedded systems without compromising design simplicity.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

REFERENCES

1. Zoni, D., & Galimberti, A. (2022). *Cost-effective fixed-point hardware support for RISC-V embedded systems*. J. Syst. Archit., 126, 102476. https://doi.org/10.1016/j.sysarc.2022.102476.
2. Micikevicius, P., Stosic, D., Burgess, N., Cornea, M., Dubey, P., Grisenthwaite, R., Ha, S., Heinecke, A., Judd, P., Kamalu, J., Mellempudi, N., Oberman, S. F., Shoeybi, M., Siu, M., & Wu, H. (2022). *FP8 formats for deep learning*. arXiv:2209.05433. Machine Learning (cs.LG). https://doi.org/10.48550/arXiv.2209.05433.
3. van Baalen, M., Kuzmin, A., Nair, S. S., Ren, Y., Mahurin, E., Patel, C., Subramanian, S., Lee, S., Nagel, M., Soriaga, J., & Blankevoort, T. (2023). *FP8 versus INT8 for efficient deep learning inference*. arXiv:2303.17951. Machine Learning (cs.LG) https://doi.org/10.48550/arXiv.2303.17951.
4. Tedja, H. A., & Onno W. Purbo. (2024). *Performance and Efficiency Comparison of U-Net and Ghost U-Net in Road Crack Segmentation with Floating Point and Quantization Optimization*. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 8(6), 779-787. https://doi.org/10.29207/resti.v8i6.6089.
5. Chen, J., Hao, H., Wang, S., Li, L., Zhao, X., Yu, F., Wang, J., Xu, G., Sun, Z., & Jiang, K. (2024). *A multiple precision floating-point arithmetic unit based on the RISC-V instruction set*. In Proceedings of the 2024 4th International Conference on

Electronic Information Engineering and Computer (EIECT) (pp. 573–578). IEEE. https://doi.org/10.1109/EIECT64462.2024.10867213.
6. Mach, S., Schuiki, F., Zaruba, F., & Benini, L. (2020). *FPnew: An open-source multi-format floating-point unit architecture for energy-proportional transprecision computing*. arXiv:2007.01530. Hardware Architecture (cs.AR). https://doi.org/10.48550/arXiv.2007.01530.
7. Brand, M., Hannig, F., Keszocze, O., & Teich, J. (2022). *Precision- and Accuracy-Reconfigurable Processor Architectures — An Overview*. IEEE Transactions on Circuits and Systems II: Express Briefs, 69, 2661-2666. https://doi.org/10.1109/TCSII.2022.3173753.
8. Kunešová, M., Zajíc, Z., Šmídl, L. & Karafiát M. (2024) *Comparison of wav2vec 2.0 models on three speech processing tasks*. International Journal of Speech Technology. 27, 847–859. https://doi.org/10.1007/s10772-024-10140-6.
9. Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2024). *Mixtral of Experts*. arXiv:2401.04088. Machine Learning (cs.LG). https://doi.org/10.48550/arXiv.2401.04088.
10. Peng, Z., Budhkar, A., Tuil, I., Levy, J., Sobhani, P., Cohen, R., & Nassour, J. (2021). *Shrinking Bigfoot: Reducing wav2vec 2.0 footprint*. arXiv:2103.15760. Computation and Language (cs.CL). https://doi.org/10.48550/arXiv.2103.15760.
11. Hassani Sadi, M., Sudarshan, C. & Wehn, N. (2024) *Novel adaptive quantization methodology for 8-bit floating-point DNN training*. Design Automation for Embedded Systems, 28, 91–110. https://doi.org/10.1007/s10617-024-09282-2.

Received (Надійшла) 27.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

ABOUT THE AUTHORS / ВІДОМОСТІ ПРО АВТОРІВ

Сальніков Дмитро Валентинович – кандидат технічних наук, старший викладач кафедри автоматики та управління в технічних системах, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Dmytro Salnikov – Candidate of Technical Sciences, Senior Lecturer at the Department of automation and control in technical systems, head of the department, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine
e-mail: dmytro.salnikov@khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0007-6201-5370;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57225126629.

Васильченков Олег Георгійович – кандидат технічних наук, доцент кафедри автоматики та управління в технічних системах, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Oleg Vasylchenkov – Candidate of Technical Sciences, Associate Professor at the Department of Automation and Control in Technical Systems, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine
e-mail: oleh.vasylchenkov@khpi.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-0969-2248);
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57225129501.

Компактні апаратні модулі для перетворення форматів FP16/FP8/FP32 у вбудованих системах

Д. В. Сальніков, О. Г. Васильченков

Анотація. Стрімке поширення нейронних мереж у вбудованих та периферійних обчислювальних системах зумовило зростання попиту на ефективні апаратні рішення, що здатні підтримувати арифметику зі змінною точністю. В таких сферах, як автономні транспортні засоби, інтелектуальні сенсори та промислова автоматизація, вимагається висока обчислювальна продуктивність, мала затримка та суворі обмеження електроживлення. Арифметика з плаваючою комою, визначена стандартом IEEE 754, залишається домінуючим методом числового представлення у подібних системах завдяки своїй універсальності та широкому динамічному діапазону. Водночас розгортання сучасних моделей глибокого навчання на платформах з обмеженими ресурсами створює значні труднощі, пов'язані з досягненням балансу між точністю, пропускною здатністю та апаратними витратами. Для подолання цих обмежень дедалі більшої популярності набувають формати зменшеної точності, такі як FP16, BF16 та FP8 (E4M3, E5M2), які використовуються як під час інференсу, так і під час навчання, забезпечуючи зниження пропускної здатності пам'яті та підвищення енергоефективності одночасно з мінімальною втратою точності. Попри їх зростаюче поширення, багато мікроконтролерів і FPGA не мають нативної апаратної підтримки таких форматів, що зумовлює необхідність розробки компактних і придатних до реконфігурації модулів перетворення для забезпечення сумісності з традиційними обчислювальними блоками FP32. У цій роботі показані результати проєктування, реалізації та оцінки апаратних модулів для перетворення між форматами FP8, FP16, BF16 та стандартним форматом IEEE-754 (FP32). Запропонована архітектура використовує логічні можливості FPGA для виконання операцій перетворення в полях експоненти та мантиси, корекції зсуву та класифікації спеціальних числових випадків, таких як нескінченність і NaN, що забезпечує повну відповідність до стандарту IEEE-754. Синтез перетворювачів виконано з використанням засобів розробки та реалізації на FPGA Intel Cyclone V. Експериментальні результати демонструють надзвичайно низьке використання апаратних ресурсів і високу робочу частоту: перетворювачі FP8E4M3 та FP8E5M2 потребують лише 14 адаптивних логічних модулів (ALM) кожен, досягаючи частот понад 500 МГц. Отримані результати підтверджують придатність запропонованих модулів для застосування в системах змішаної точності та вбудованих прискорювачах нейронних мереж, забезпечуючи ефективну апаратну основу для енергоефективних і високопродуктивних моделей ШІ на платформах з обмеженими ресурсами.

Ключові слова: формати з плаваючою комою, числове представлення зі зменшеною точністю, вбудовані системи, периферійні обчислення, FPGA, VHDL, прискорювачі вбудованих нейронних мереж, компактна архітектура.

UDC 004.7:621.396 **doi: 10.26906/SUNZ.2026.2.247**

Svitlana Sulima

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

LOCAL RECONFIGURATION OF 5G NETWORK SLICES UNDER NODE FAILURES AND OVERLOADS

Abstract. Relevance. The rapid deployment of 5G networks and the widespread use of network slicing create new challenges for ensuring service reliability and resilience. Virtualized infrastructures based on Network Functions Virtualization increase flexibility but also introduce higher failure rates and performance variability. In such environments, centralized recovery mechanisms often fail to meet strict latency requirements, especially for ultra-reliable low-latency communication services. **Research object**: The research object is the process of failure recovery and resource reconfiguration in virtualized 5G network slicing environments under node failures and progressive overload conditions. **Purpose**: The purpose of the study is to develop an efficient distributed method for local reconfiguration of network slices that enables rapid recovery of virtual network functions while considering slice priorities, migration costs, and latency constraints. **Results**. A distributed local reconfiguration (DLR) framework is proposed, based on a hierarchical architecture consisting of a global orchestrator, regional slice managers, and local monitoring agents. The approach introduces a multi-objective optimization model for slice manager placement and a unified migration cost function that considers computational, network, disruption, and SLA penalty components. Localized recovery algorithms are developed to handle both catastrophic node failures and progressive overload scenarios while prioritizing slices according to their service requirements. **Conclusions**. The proposed distributed slice recovery framework enables fast and scalable reconfiguration of 5G network slices under failure conditions. By combining slice-aware prioritization, cost-aware migration decisions, and localized management, the approach improves recovery speed and operational efficiency while preserving service quality for latency-critical applications.

Keywords: 5G networks; network slicing; network function virtualization; failure recovery; distributed network management; virtual network function migration; slice-aware resource management; resilience; service level agreement; distributed local reconfiguration.

Introduction

Formulation of the problem. Visualize a single server residing in a solitary rack located somewhere in the city, currently balancing three unique existences. One existence is that of providing haptic feedback for a surgeon while he utilizes a robotics apparatus to execute a procedure in a faraway healthcare facility. Another existence is that of assisting a fleet of self-driving trucks to communicate to one another to execute a lane change on the freeway. And yet another existence is that of providing streaming HD movies to an endless number of people who are killing time with their phones while waiting for their next task.

Now picture that one of the three servers dies suddenly, and the network is suddenly placed into triage mode regarding life and death. If the network is unable to re-establish the connection to that server within a fraction of a millisecond, the surgeon cannot see in the real world. If the network does not reestablish the connection in that timeframe, the rest of the network will become unstable, and lives will be endangered because of it. On the other hand, individuals streaming movie content from their phones may be able to tolerate a 3- or 4-second delay before they begin to notice something is awry.

Thus, the core problem of managing a 5G network is how to address the same issue with three different outcome requirements, using three different methodologies, all occurring at the same time. Centralized, thinking systems will not be able to share the electronic capacity needed to meet the timing requirement of the robotics issue while maintaining an efficient operating cost to deliver cellular phone data.

The use of operational network slicing in 5G currently is real, and presents an operationally challenging issue yet to be resolved in today's market. The ITU specifies three different service categories within 5G: enhanced mobile broadband (eMBB), ultra reliable low latency communications (URLLC), and massive machine type communication (mMTC) [1]. Each of the three service categories can share physical infrastructure to operate independently while still supporting the requirements of the service categories within each slice. Network slicing accomplishes this through the creation of a logically isolated (virtualized) network for each family of services operating over the same physical infrastructure [2, 3]. The 3GPP Technical Specification 23.501 defines how network slices are logical networks comprised of defined capabilities or features [4] – an example of where there are significant differences in the end-to-end latency associated with services (e.g. autonomous driving requiring ≤1ms maximum end-to-end latency with reliability at 99.9999%, and IoT sensory networks possibly exceeding between 10s to hundreds of billions of users within the same service family) [5, 6].

Recently, mobile data usage has consistently exceeded expectations as new types of services are rolled out that change how data is being used. Network Functions Virtualization (NFV) is now the way in which network slicing will be executed by using general-purpose server hardware instead of specialized devices to implement telecom networks. Although NFV provides substantial savings, it also introduces weaknesses to the reliability of networks that do not exist with traditional carriers. For example, generic server hardware has a significantly greater failure rate than dedicated telecommunication hardware, and the virtual machine layers (hypervisors) create new vulnerabilities for security breaches while using a commodity hardware platform compared with a dedicated hardware platform. Furthermore, Virtual

Network Functions (VNFs) running on the same physical hardware may create unexpected interactions between the VNFs, complicating the service performance or reliability [7][8][9].

In this case, failure is considered a good frequency as it is guaranteed to happen as well as certain measures for mitigation are in place. A single physical node going down can affect dozens of virtual network functions belonging to slices with significantly varying recovery priorities. Relevant approaches to the issue exist in two groups, and neither may be considered satisfactory. Centralized orchestration can produce globally optimal recovery plans, but in doing so, it uses time resources that URLLC slices cannot spare, and it does not scale well with large networks [10]. Proactive redundancy — preplacing backup resources everywhere — keeps recovery fast but exactly wastes the kind of physical capacity, which network slicing exists to consolidate[11]. That is missing is an approach that is distributed enough to be manageable as networks expand, resource-efficient enough to be economically viable, and fast enough to satisfy the most demanding slices.

The paper presents a method to reconstruct data from a location that is close to where a failure occurs by using regional slice managers located throughout a network as part of a recovery process. Each slice manager is responsible for a specific geographical area, and can retrieve and migrate from a host resource to another within that geographical region without waiting for a view of the overall network. If a node fails, the regional slice manager immediately knows what slices are affected, can sort those slices based on the severity of the failure, is able to identify which hosts within the geographical region could be used for the new virtual functions, can calculate the costs associated with migrating, and then will begin to migrate the virtual functions all before the orchestrator is even aware of what has happened. The paper describes three concepts in detail; a hierarchical distributed architecture based on literature on the placement of SDN controllers [12, 13, 14]; a cost model that incorporates factors beyond latency into the determination of the cost of migrating [15, 16, 17]; and a recovery mechanism that treats sudden failures and progressive overload as distinct problems requiring distinct strategies [18, 19].

There are three different interlinked aspects present in the scientific novelty of the Distributed Local Reconfiguration (DLR) approach and not combined so far in any of the provided solutions:

1. New approach for placing managers (RSM placement). All previous works on SDN controller placement either minimized latency to nodes or load balancing. However, the author is first in NFV context to introduce a third criterion – latency between the managers themselves as a separate optimization goal. For failure recovery purposes, this is very important, because if the managers are located far away, the cross-domain recovery will be too long for URLLC.

2. Unified migration cost model. Existing works consider up to at most one or two componentsonly state transfer time, only resource constraints, only reliability constraints. The paper is the first to build a single 4-components cost function (compute + network + migration disruption + SLA penalty), which allows migration decisions to be made considering all these aspects together.

3. Dealing with two failure types in one slice-aware framework. The vast majority of NFV studies consider only catastrophic node failures. As a result, progressive overload is either neglected or handled by unassociated solutions. Also, there are no known works that distinguish the recovery priorities based on the slice type (URLLC vs eMBB vs mMTC). The paper, for the first time ever, combines both scenarios into a single algorithm where the recovery depends explicitly on slice priority as per 3GPP specification.

Current State of Web Accessibility.

1. Network Slicing in 5G

The concept of network slicing has moved from being a theoretical idea in 5G into a concrete part of present-day mobility networks. This has been accomplished primarily through many years of effort from the 3rd Generation Partnership Project (3GPP), which provided the basis for both the system architecture (as outlined in TS 23.501) [4] and the management and orchestration framework (as detailed in TR 28.801) [20]. As a result, operators can now develop end-to-end “slices” that span the last two, three or four layers of the 5G ecosystem: from the Radio Access Network (RAN) to the transport layers (T-Layers) to the core network.

Researchers like Ordonez-Lucena et al. [21] have offered a broad view of how network slicing works in practice. They note that while technologies such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV) serve as essential foundations, real-world deployment depends on intricate orchestration systems. These systems must handle every stage of a slice’s life—from creating and configuring it to continuously monitoring and eventually shutting it down. The key challenges, they argue, often revolve around maintaining strong isolation between slices, managing resources effectively, and ensuring that different slices can work in harmony without competing for resources.

Rost et al. [22] approached slicing from a scalability and flexibility standpoint. Their research focused on how 5G networks can dynamically allocate resources among multiple slices without causing interference. Using simulations, they demonstrated that with proper architecture, it’s possible to meet the needs of very different services—say, a high-speed video stream and a low-latency industrial application—at the same time.

Meanwhile, Foukas et al. [23] placed special emphasis on RAN slicing. They observed that while slicing in the core network has matured thanks to NFV, RAN slicing remains more difficult. That’s largely because the radio environment requires fine-grained control and real-time coordination, making efficient resource distribution a tougher technical challenge.

One noticeable gap in much of the literature is how slicing behaves when things go wrong. Most studies focus on how to set up slices and make them run efficiently, but they rarely address what happens during faults or network failures. For instance, Ksentini and Nikaein [24] discuss RAN slicing with resource abstraction but stop short of failure management, while Zhang et al. [25] explore reinforcement learning methods for dynamic

resource allocation under the assumption that everything operates smoothly. That leaves a clear opportunity for research into recovery and resilience mechanisms within 5G network slicing.

2. NFV Resilience and Failure Recovery

NFV resilience has been examined primarily through the prism of cloud-based deployments, with a strong emphasis on how to maintain service continuity under frequent infrastructure failures. The ETSI NFV architectural framework [26] outlines high-level principles for achieving resilience but intentionally leaves concrete realization choices to individual operators and vendors, reflecting the diversity of deployment environments. In parallel, ETSI NFV-REL 001 [27] refines this view by formalizing reliability and availability requirements and by explicitly recommending redundancy-based protection mechanisms as the baseline approach for sustaining target service levels.

Han et al. [8] provide a detailed discussion of NFV challenges and opportunities, identifying reliability as a central concern rather than a secondary design objective. They argue that while virtualization greatly simplifies rapid deployment, scaling, and flexible placement of network functions, it also introduces novel failure modes compared to traditional, purpose-built telecom appliances. Empirical observations reported in their work indicate that commodity cloud servers experience failures roughly one to two orders of magnitude more frequently than dedicated carrier-grade hardware, which fundamentally shifts the reliability engineering problem toward software- and platform-level mitigation.

This tension is articulated clearly in the foundational NFV white paper by Chiosi et al. [11], where the authors highlight the gap between carrier-grade availability targets (on the order of 99.999%) and the typical enterprise-grade availability (around 99.9%) delivered by commodity hardware platforms. To bridge this gap, they advocate multi-layer redundancy and fast failover mechanisms spanning the NFV infrastructure, VNFs, and management layers, while noting that such designs inevitably increase resource footprint and operational complexity.

Subsequent research proposes more specialized failure recovery mechanisms. Cohen et al. [15] formulate VNF placement under reliability constraints as an optimization problem that aims to minimize expected service disruption in the presence of node failures. Their approach relies on computing backup VNF placements offline so that failover can be executed quickly at runtime, but this strategy requires reserving substantial standby capacity, which may be costly in resource-constrained environments. Gember-Jacobson et al. [28] introduce OpenNF, a control framework that enables fine-grained manipulation of network function state, including live migration of stateful VNFs with disruption on the order of hundreds of milliseconds for typical state sizes, though the reliance on centralized control inherently limits scalability in very large deployments.

Rajagopalan et al. [29] address elasticity through Split/Merge, a system that allows VNFs to be dynamically partitioned across multiple servers or consolidated onto fewer servers to adapt to load variations, focusing primarily on performance and resource efficiency rather than explicit fault tolerance. Their results show that elastic execution of virtual middleboxes can effectively track fluctuating demand, suggesting that similar mechanisms could be extended to support resilience-aware scaling policies.

A common limitation across much of the NFV resilience literature is the implicit assumption that services share homogeneous reliability and recovery requirements. With the advent of network slicing in 5G, this assumption becomes problematic, as slices support heterogeneous service classes with distinct resilience profiles: for example, URLLC slices require sub-second recovery, whereas eMBB slices can tolerate several seconds of degraded performance. This heterogeneity motivates the need for slice-aware recovery mechanisms that explicitly account for per-slice resilience objectives—a gap that the present work is designed to address.

3. Distributed Management in Virtualized Networks

Although theoretically optimal, centralized orchestration suffers in scalability and suffers from latency in large-scale networks. The controller placement problem in SDN has been studied thoroughly on this basis.

Heller et al. [12] initiated research on controller placement in SDN, with the identification of a k-center optimization problem minimizing the maximum node-to-controller latency. On real network topologies, their results show that the communication latency among controllers remains high in large networks even with minimized maximum node-to-controller latency.

Hock et al. [13] extend this work with multi-objective optimization, considering both latency and resilience. They formulate controller placement as a Pareto optimization with three conflicting objectives: minimizing average latency, maximizing resilience from failures, and balancing the load on controllers. Their results show that single objective optimization leads to brittle solutions with poor performance when multiple criteria are important.

Lange et al. [14] propose heuristic algorithms for controller placement in large-scale networks, demonstrating that genetic algorithms and simulated annealing approaches can yield near-optimal solutions in orders of magnitude less time compared to exact optimization methods. The heuristic algorithms provide controller placement solutions within 5% of optimal solutions for networks with a couple of hundreds of nodes in a matter of seconds rather than hours required by exact algorithms.

While these works address control plane placement, they do not explicitly handle data plane VNF recovery or migration cost optimization. More relevant to our work, Baumgartner et al. [30] investigate mobile core network virtualization with combined VNF placement and topology optimization. They formulate a joint optimization problem with initial placement and reconfiguration cases, showing by simulations that considering reconfiguration in advance is more resource-efficient than treating it as a reactive case.

Moens and De Turck [16] propose VNF-P, a model for efficient VNF placement with respect to the resource

constraints and the topology of the service chain. The model includes latency constraints and validates that intelligent placement reduces resource consumption by 20-30% compared to naive placement. However, they focus on the initial placement only and not on the dynamic reconfiguration.

Mehraghdam et al. [17] propose a method to specify and place service chains, which are sequences of VNFs. For this purpose, they propose formal models of service chain specification and algorithms for optimal chain placement. However, this work does not consider failures and migrations at runtime.

4. Virtual Network Reconfiguration

Virtual network reconfiguration is studied in virtual network embedding (VNE). Based on a comprehensive survey of network virtualization, Chowdhury and Boutaba [31] identify reconfiguration as a core challenge which has been minimally addressed relative to embedding.

Fajjari et al. [18] propose a greedy algorithm for virtual network reconfiguration, where the initial embedding can be changed if some nodes and links run out of resources or require rearrangement. Their algorithm greedily migrates virtual nodes and links repeatedly and improves the results in terms of resource utilization, but no failure recovery scenarios are considered.

Beloglazov and Buyya [32] dynamically investigate consolidation of virtual machines in cloud data centers with energy efficiency objectives. Their adaptive heuristics are designed to continuously monitor the physical resource utilization levels, triggering virtual machine migration actions aimed at consolidating the virtual machine workload on a minimum number of physical servers in the data center with a view to reducing energy consumption. Their VM migration techniques are applicable to our work, even though their focus is not on failure recovery.

Qu et al. [33] formulate reliability-aware network service chain provisioning in NFV-enabled enterprise datacenters. They propose algorithms that place service chains proactively considering possible failures, and ensure that backup resources can be used for a quick recovery. While providing a high level of reliability, the proposed approach incurs high resource over-provisioning, ranging between 30-50% overhead.

Some of the major gaps in knowledge still exist despite much research that has been conducted to date:

• Lack of Slice Aware Recovery: All VNFs are treated the same and therefore, do not take into consideration different slice resiliency requirements and recovery priorities as specified by 3GPP standards [4,20].

• Limited Migration Cost Models: Most of the current research has an overly simplified view of migration costs. Gember-Jacobson et al. [28] looked only at state moving times, but did not develop a complete view of the computational overhead and cascading effects associated with those movements.

• Centralized Bottlenecking: Although there are several studies which examine distributed controller placement [12,13,14], all of the NFV orchestration approaches proposed to date are central in nature as noted by the ETSI MANO specification [34]. Therefore, the latencies associated with these centralized solutions would not meet URLLC requirements.

• Inability to Handle Overloads: Most of the research that has been published about failure recovery has focused on catastrophic failures. Failure recovery to date has ignored progressive overload scenarios and as more data centers and shared infrastructures become common, progressive overload recovery will become even more relevant to the state of the art.

This paper fills these gaps with a framework of distributed, cost-aware, slice-specific recovery in 5G network slicing environments.

Key difference from closest competitors are summarized in Table 1.

Table 1 – **Key difference**

Feature	CGO	NAM	DLR (proposed)
No central orchestrator required	✗	partially	✓
Slice-aware prioritization	✗	✗	✓
Full migration cost model	✓	✗	✓
Sub-second URLLC recovery	✗	partially	✓
Progressive overload handling	✗	✗	✓
Inter-manager coordination	—	—	✓ (1 round)

Research Objectives.

Our contributions include:

• Formulation of the slice manager placement problem as multi-objective optimization considering latency, load distribution, and coordination between the managers, considering SDN controller placement work [12,13,14] in the NFV contexts.

• Development of localized reconfiguration algorithms that can work for both forms of reconfiguration: catastrophic failures and progressive overload, based on the lessons from virtual network reconfiguration [17,18] and live migration [28,29].

• Migration cost models, which take into account the computational resources used during migration, the bandwidth costs caused by migration, and any service disruption caused by migration, are introduced on top of VNF placement optimization [15,16,17].

• Experimental validation over different failure scenarios with different slices shows better performance compared to the centralized and benchmarks.

Main material

The optimization problem we consider is inherently complex: it is large-scale, mixed-integer, and multi-objective, and it has to be solved under very tight recovery-time constraints, especially for URLLC slices where delays are unacceptable. In practice, trying to solve the full problem centrally every time a failure occurs is not realistic, because the computation would take too long and would directly conflict with the low-latency guarantees expected in 5G systems.

To cope with this, we introduce a distributed, slice-aware recovery framework that restructures how the network reacts to failures. Instead of relying on a single, global decision point, we assign responsibilities to slice

managers that each have only partial knowledge of the overall topology, allowing them to localize and speed up the failure response. Within this framework, slices are not treated equally: they are prioritized according to the criticality of their SLAs, so more demanding services are handled first. We further separate placement, routing, and migration decisions and organize them in a hierarchical optimization process, which reduces complexity while still enabling coordinated recovery.

This design is intentionally consistent with existing ETSI NFV MANO concepts and with architectures that use multiple distributed SDN controllers. At the same time, it makes explicit room for practical aspects that are often overlooked, such as the cost of VNF migrations, the specific semantics of different failure types, and the heterogeneous resilience requirements of different slices.

Fig. 1 illustrates the proposed architecture, which consists of three logical layers:

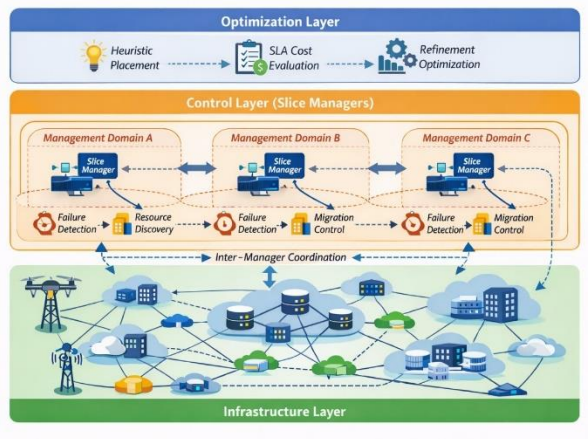


Fig. 1. Architecture for the solution

1. Infrastructure Layer

Physical nodes and links as defined by the substrate network $SN = (N, NE)$.

2. Control Layer (Slice Managers)

A set of distributed slice managers $M \subseteq N$, each responsible for a management domain D_m. Managers perform:

- Local failure detection,
- Candidate resource discovery,
- Migration orchestration,
- Inter-manager coordination for cross-domain recovery.

3. Optimization Layer

A hierarchical decision process combining:

- Local heuristic decisions for fast reaction,
- Lightweight optimization for placement refinement,
- SLA-aware cost evaluation.

Each slice manager operates autonomously for failures within its domain while coordinating with neighboring managers when migration targets lie outside D_m.

Upon detecting an anomaly, the responsible slice manager classifies the event as one of the following:

1. Catastrophic Node Failure

Immediate service disruption for all VNFs placed on n^{fail}.

Recovery must be reactive and fast.

2. Progressive Node Overload

Detected via monitoring of resource utilization trends:

$$\sum_{s,v} x_{n,v}^{s} \cdot D_v^{comp} \to C_n^{comp}.$$

This scenario enables proactive migration, reducing SLA penalties.

The failure type determines the urgency, optimization depth, and migration scope.

The recovery process follows four stages:

Stage 1: Affected Slice Identification.

For a failed or overloaded node n, the manager identifies all impacted slices:

$$S^{aff} = \{s \in S \mid \exists v \in V_s : x_{n,v}^{s} = 1\}.$$

Slices are sorted by priority level π_s, ensuring URLLC slices are handled first.

Stage 2: Candidate Node Filtering.

For each disrupted VNF v, a candidate node set $N_v^{cand} \subseteq N$ is constructed based on:

- Resource feasibility,
- Administrative suitability $\text{suit}_{n,v}^{s}$,
- Latency feasibility with respect to L_s^{max},
- Manager proximity (prefer nodes within D_m).

This pruning step dramatically reduces the solution space.

Stage 3: Cost-Aware Migration Decision.

For each candidate migration $n \to n'$, the manager computes a local reconfiguration cost:

$$\Delta C = \Delta\text{Cost}_{compute} + \Delta\text{Cost}_{network} + \Delta\text{Cost}_{migration} + \Delta\text{Penalty}_{SLA}.$$

Migration decisions are selected using:

- Greedy minimization for URLLC slices,
- Multi-criteria ranking for eMBB and mMTC slices.

For progressive overload, migration is triggered before capacity violation, minimizing disruption time.

Stage 4: Inter-Manager Coordination.

If no feasible candidate exists within D_m, the manager:

1. Requests candidate resources from neighboring managers,

2. Exchanges summarized state information (capacity, latency bounds),

3. Negotiates placement using a lightweight consensus protocol.

This avoids global state synchronization while ensuring feasibility.

To balance optimality and responsiveness, we adopt a two-level optimization approach:

Local Optimization (Fast Reaction)

- Scope: Single failure event
- Variables: Subset of $x_{n,v}^{s}$, $f_{(n_1,n_2),e}^{s}$
- Method: Heuristic + constrained local search
- Time scale: milliseconds to seconds

Global Refinement (Optional)

- Triggered during low-load periods
- Re-optimizes placements to reduce fragmentation
- Improves long-term cost efficiency

This separation ensures SLA compliance without sacrificing overall efficiency.

Algorithm 1: Distributed Slice Recovery

1. Detect failure at node n
2. Classify failure type
3. Identify affected slices S^{aff}
4. Sort S^{aff} by priority π_s
5. For each slice $s \in S^{aff}$:
   - Identify disrupted VNFs
   - Generate candidate nodes
   - Evaluate migration costs
   - Select minimal-cost feasible migration
6. Coordinate with neighboring managers if required
7. Enforce updated placement and routing
8. Monitor post-recovery SLA compliance

The proposed approach is designed to work in realistic, large-scale 5G environments, where many slices and network functions coexist. It does so while explicitly taking into account both the overhead introduced by VNF migrations and the fact that different slices have different performance and resilience requirements. As a result, it can trigger fast, localized recovery actions when failures occur, without violating the SLAs associated with each slice. In addition, the same framework can naturally handle both proactive strategies (anticipating problems before they occur) and reactive strategies (responding after a failure has been detected).

Motivated by distributed SDN control architectures [12,35] and ETSI NFV MANO principles [34], we propose a three-tier hierarchical management architecture (Fig. 2) designed for low-latency and scalable failure recovery in network slicing environments.

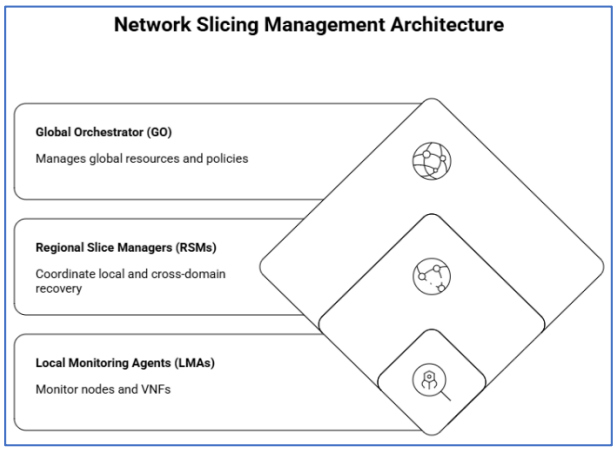


Fig. 2. Control architecture

Tier 1: Global Orchestrator (GO)

The GO maintains a coarse-grained, global view of network resources and performs: Initial network slice instantiation in accordance with ETSI NFV-IFA specifications [36]; Long-term capacity planning and policy enforcement; Therefore, to prevent centralization bottlenecks, the GO has no involvement in the real-time failure recovery process.

Tier 2: Regional Slice Managers (RSMs)

RSMs are deployed at selected nodes with the help of an optimization procedure. Every RSM; It has developed a very precise state of VNFs and physical resources within its domain; Coordinating local failure recovery; It also synchronizes the state with remote peers, i.e., the adjacent RSMs, for cross-domain failure recovery.

Tier 3: Local Monitoring Agents (LMAs)

LMAs are deployed on every physical node and; Continuously track nodes and VNFs for any faults; Perform failure and overload detection; Notify the responsible RSMs and actuation on migration commands.

This hierarchical decomposition makes it possible to perform low-latency recovery required by the URLLC slices [1,21] while alleviating the scalability and reliability challenges of the centralized management of NFV instances [10,8].

Building on SDN controller placement studies [12–14], we formulate RSM placement as a multi-objective optimization problem.

Let the physical network be $SN = (N, N_E)$, and let $|M|$ denote the desired number of RSMs. Binary variable $p_n \in \{0,1\}$ indicates whether node $n \in N$ hosts an RSM, and $\pi_{n,m} \in \{0,1\}$ indicates whether node n is assigned to manager m.

Objective 1: Minimize Maximum Node-to-Manager Latency

$$U^{\text{latency}} = \max_{n \in N} \min_{m \in M} \{L_{n,m} \cdot \pi_{n,m}\}.$$

Objective 2: Balance Manager Load

Following [14], load imbalance is defined as:

$$U^{\text{imbalance}} = \max_{m \in M} \text{load}_m - \min_{m \in M} \{\text{load}_m : \text{load}_m > 0\},$$

where:

$$\text{load}_m = \sum_{n \in N} \pi_{n,m}.$$

Objective 3: Minimize Inter-Manager Communication Latency

To support coordinated failure recovery, we introduce:

$$U^{\text{inter-latency}} = \max_{\substack{m_1, m_2 \in M \\ m_1 \neq m_2}} L_{m_1,m_2}$$

Combined Objective

$$\min_{p_n, \pi_{n,m}} U_{\text{total}} = w^{\text{lat}} U^{\text{latency}} + w^{\text{imbal}} U^{\text{imbalance}} + w^{\text{inter}} U^{\text{inter-latency}}$$

Subject to:

$$\sum_{n \in N} p_n = |M|$$

$$\sum_{m \in M} \pi_{n,m} \geq 1 \forall n \in N$$

$$\pi_{n,m} \leq p_m \forall n \in N, m \in M.$$

The NP-hard problem is solved using a genetic algorithm, following [14]:

- Encoding: Binary vector p of length $|N|$;
- Fitness: U_{total};
- Selection: Tournament selection;
- Crossover: Uniform crossover with constraint-preserving repair;
- Mutation: Random swaps between manager and non-manager nodes.

Unlike prior controller placement formulations [12–14], our approach explicitly optimizes inter-manager latency, which is critical for coordinated, low-latency slice recovery.

Each LMA performs continuous monitoring using threshold-based detection inspired by dynamic VM consolidation systems [32].

Catastrophic Failure:

A node is declared failed if its LMA becomes unresponsive for T^{timeout}(e.g., 100 ms), consistent with high-availability practices [7].

Progressive Overload:

An alert is triggered if:

$$\rho_n^{\text{CPU}} > \theta_{\text{high}}^{\text{CPU}} \text{or } \rho_n^{\text{mem}} > \theta_{\text{high}}^{\text{mem}}.$$

Thresholds are slice-specific, with stricter limits for URLLC slices [4].

The RSM classifies overloads as:

- Transient spike: utilization falls below θ_{low} within $T^{\text{transient}}$;
- Sustained overload: utilization exceeds θ_{high} for $T^{\text{sustained}}$;
- Imminent failure: utilization exceeds θ_{critical}.

Upon detecting failure of node n^{fail}, the responsible RSM executes the following steps.

Step 1: Identify Affected VNFs

$$V^{\text{fail}} = \{(s, v): x^s_{n^{\text{fail}},v} = 1\}.$$

Step 2: Slice-Aware Prioritization

$$V^{\text{fail}}_{\text{sorted}} = \text{sort}(V^{\text{fail}}, (\pi_s, \text{criticality}_v)).$$

Step 3: Localized Candidate Discovery

$$N^{\text{cand}}_{s,v} = \{n: d(n, n^{\text{fail}}) \le r^{\max}, \text{suit}^s_{n,v} = 1, \text{has\_capacity}(n, D_v^{\text{comp}})\}.$$

with $r^{\max} = 2$ hops for URLLC and $r^{\max} = 4$ for eMBB slices.

Step 4: Recovery Cost Computation

$$\text{Cost}_{s,v}(n') = c^{\text{comp}} D_v^{\text{comp}} + c^{\text{net}} \Delta_{\text{BW}}(n', s, v) + \text{Penalty}_{\text{latency}}(n', s, v).$$

Step 5: Greedy Assignment

VNFs are greedily placed on the minimum-cost feasible node; failure triggers regional reconfiguration.

Localized, hop-constrained recovery dramatically reduces decision latency compared to global optimization [15,16], enabling compliance with URLLC recovery bounds [4,20].

For sustained overload on node n^{over}, the RSM performs incremental migration.

VNF Scoring

$$\text{score}_{s,v} = \alpha_1(1 - \pi_s) + \alpha_2 U^{\text{actual}}_{s,v} + \alpha_3(1 - \text{statefulness}_v).$$

Incremental Migration

VNFs are migrated in descending score order until:

$$\rho_{n^{\text{over}}}^{\text{CPU}} \le \theta_{\text{target}}^{\text{CPU}} (\theta_{\text{target}}^{\text{CPU}} = 0.7).$$

Adaptive halting prevents excessive migrations, a known issue in reactive resource management [32].

If local resources are exhausted, RSMs engage in lightweight coordination:

1. Broadcast recovery request (resource, latency, and priority constraints);

2. Parallel candidate search by neighboring RSMs;

3. Minimum-cost selection and coordinated migration.

This single-round protocol enables fast cross-domain recovery without centralized computation.

For complex scenarios, recovery is formulated as an optimization problem.

Decision Variables

- $x'^s_{n,v}$: post-recovery placement;
- $\delta^s_{v,n\to n'}$: migration indicator.

Objective

$$\min \sum_{s,v,n,n'} \delta^s_{v,n\to n'} C^{\text{mig}}_{v,n\to n'} + \lambda \mid \text{Cost}^{\text{new}}_{\text{oper}} - \text{Cost}^{\text{old}}_{\text{oper}} \mid$$

Subject to migration continuity, concurrency limits, and slice-specific recovery deadlines [4].

Solution: MILP for small instances (≤ 20 VNFs); greedy heuristics [14,18] for large-scale deployments.

To evaluate the effectiveness of the proposed distributed local reconfiguration (DLR) method, a synthetic network topology representing a 5G deployment in a metropolitan area was used. The network consists of 10 physical nodes (servers/data centers), including mixed edge and regional nodes. Channel delays vary from 1 ms (intra-regional) to 20 ms (inter-regional). Node capacities are heterogeneous: edge nodes have 8-16 processor cores and 32-64 GB of RAM, while regional nodes have 32-64 cores and 128-256 GB of RAM.

In accordance with 3GPP standards, three types of slices with different resource requirements and SLAs were created:

1. URLLC (Ultra-Reliable Low-Latency Communication): 3 VNFs (MME, SGW, PGW), requirement — 15 cores, maximum latency (Lmax) — 5 ms, priority (π) — 1 (highest).

2. eMBB (Enhanced Mobile Broadband): 5 VNF (including PCRF, HSS), requirement — 30 cores, Lmax — 50 ms, priority — 2.

3. mMTC (Massive Machine-Type Communications): 3 VNFs, requirement — 10 cores, Lmax — 1000 ms, priority — 3 (lowest).

In total, 33 VNFs were deployed in the system (11 VNFs per 3 slice instances). The experiments simulated Single Node Failure (SNF) (sudden catastrophic failure affecting 2-4 VNFs) and Progressive Overload (PO) (gradual depletion of resources) scenarios.

Comparison methods and metrics the proposed approach (DLR) was compared with three baseline methods:

• CGO (Centralized Global Optimization): Complete re-solving of the placement problem using MILP (theoretical optimum).

• NAM (Nearest Available Migration): Greedy migration to the nearest node with available capacity.

• LOO (Latency-Only Optimization): Minimization of end-to-end latency without considering migration costs.

The main performance metrics were: recovery time, resource cost, SLA violations, and migration overhead.

Manager placement results First, the effectiveness of the regional slice manager (RSM) placement algorithm was evaluated. For a topology with 10 nodes, the number of managers is |M|=3. The proposed genetic algorithm showed an improvement in the composite placement quality score of 11-33% compared to the baseline approaches (Random, Latency-Only, K-Center). In particular, it was possible to achieve a 15% reduction in latency between managers compared to the K-Center

method, confirming the importance of explicit modeling of manager-manager coordination.

The results of modeling a single node failure scenario are shown below:

• CGO: Median recovery time – 3.8 s, SLA violation – 0%, cost – 1.00 (baseline).

• NAM: Median recovery time – 0.6 s, SLA violation – 14%, cost – 1.38.

• DLR (Proposed method): Median recovery time – 0.7 s, SLA violation – 4%, cost – 1.12.

Our approach demonstrated the ability to recover 90% of failures within 1.2 s, confirming the possibility of sub-second recovery for URLLC slices. In progressive overload scenarios, DLR performed 44% fewer migrations compared to naive approaches thanks to dynamic threshold management (Fig. 3).

The results show that the proposed DLR achieves a well-balanced tradeoff between the recovery speed, cost, and SLA violations, all essential in latency-critical 5G applications.

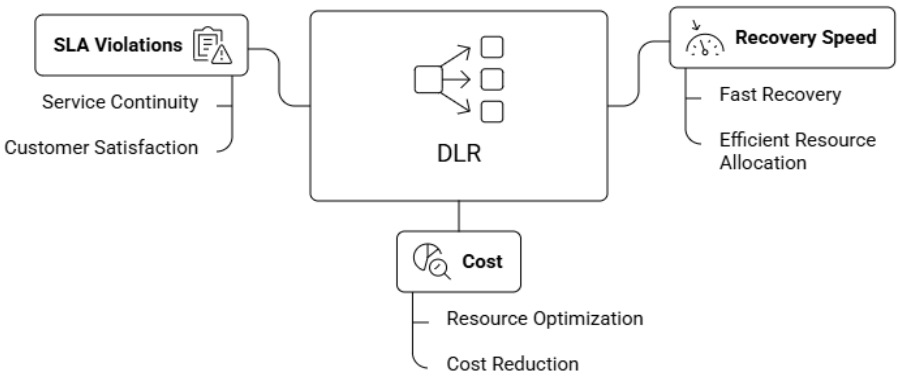


Fig. 3. DLR: Balancing Recovery Speed, Cost, and SLA Violations

Regarding recovery speed, DLR provides a median recovery time of 0.7 seconds, which is about 5.4× faster than the Centralized Global Optimization (CGO). This gain stems directly from confining the search space to a 2–3 hop neighborhood, avoiding the overhead of full network re-optimization and permitting near real-time reconfiguration under dynamic conditions (Fig. 4).

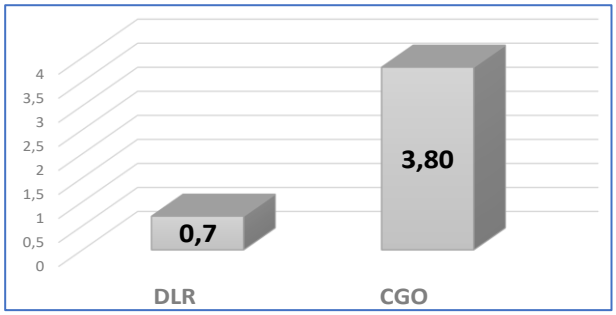


Fig. 4. DLR: median recovery time

In terms of cost, DLR is only 12% more than CGO's optimal solution, which is reasonable considering the speedup, and is 19% less than NAM and 10% less than LOO heuristics.

This is mainly attributed to the candidate selection process that explicitly takes migration and resources costs into consideration to avoid unnecessary or extremely expensive reconfigurations (Fig. 5).

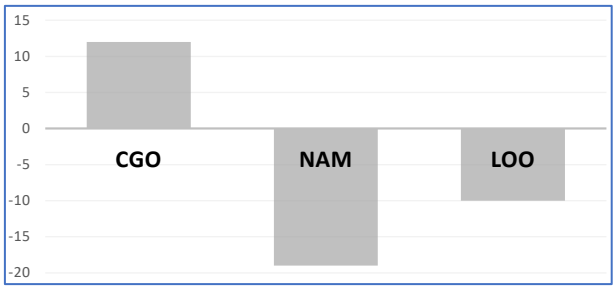


Fig. 5. DLR cost comparison

On SLA compliance, DLR keeps the URLLCs latency violation rate at approximately 4%, with violations arising only during short transients during service migration. (Fig. 6).

Fig. 6. SLA compliance and recovery delay for URLLC solutions

This is compared to the 14% violation rate shown by NAM. In contrast, although CGO produces zero SLA violations, it experiences a recovery delay of 3.8 s, making the solution untenable for URLLC downtime

Finally, towards progressive overload, DLR's use of dynamic thresholds and proactive migration yields effective prevention of cascading failure to contain the extent of cascading failure propagation to at most two rounds of reconfiguration. This body of results combine to show that distributed, cost-aware local reconfiguration does not only stand on theoretical grounds, but is indeed a practically feasible approach towards resilient network slicing in 5G.

The drawbacks of the approach can be summarized as follows.

The evaluation relies solely on synthetic results from a 10-node, handcrafted network. The network characteristics of an actual metropolitan deployment will exhibit far more variance: link latencies will vary widely, traffic patterns will display bursty behavior, and failure correlations will occur among physically co-located nodes.

Because there have not been any validations of the results in either a real testbed or benchmark topologies of record (like GÉANT or Internet2), it is impossible to determine how much the results would differ in practice.

The cost estimation model for VNF (Virtual Network Function) migration assumes that the migration (transfer of state) is a linear function of the state size. This assumption is an oversimplification of the cost of live migration of a stateful function such as an SGW (Serving Gateway) or PGW (PDN Gateway), which includes hypervisor-level checkpointing, dirty-page tracking for memory, and possible sessions involving tunnelling must be re-established; there may be many others as well. All of these introduce non-linear, workload-dependent latency that is not captured by the cost estimates in the model. Therefore, the cost estimates provided in the model and subsequently used to make migration decisions are likely to be systematically optimistic.

The framework will respond to failures and excess loads only after they have exceeded pre-defined thresholds. This means that by the time a recovery process commences, the damage sustained to the SLA has already taken place. A predictive layer could have assisted in pre-migrating workloads before the threshold for the resource is breached. An example of such a predictive layer would be an anomaly detection algorithm on a time-series based on resource usage trends; this would be especially beneficial for URLLC Slices where even a small disruption would constitute a breach of the SLA.

Conclusions

This paper provides an extensive framework for local reconfiguration of 5G network slices due to node failures and overload conditions. The proposed approach solves scalability and performance issues that come with dynamic and large-scale networks.

The proposed architecture has a three-tier hierarchical distribution based upon global orchestrators, regional slice managers, and local management agents to quickly and efficiently recover through the distribution of control to facilitate coordinated decisions throughout the network.

The work also created a multi-objective manager placement problem that considers latency, load balancing, and inter-manager cooperation in one approach. Furthermore, this harbored the concepts of controller placement in a network function virtualization environment to consider inter-manager latency as part of the optimization problem.

Localized recovery algorithms have been established for recovering from two categories of disruptions (catastrophic failures and progressive overload), and also account for slice-specific priorities and for migration costs to provide the basis for adaptive and service-aware decisions.

A comprehensive migration cost framework was proposed that includes the costs of computation, bandwidth consumption, and service disruption. This framework was formalized with a mixed-integer linear programming (MILP) model; an efficient heuristic was developed to make the framework usable in practice.

A light-weight distributed coordination protocol was designed for inter-manager communication and cross-domain recovery. This protocol enables effective collaboration between all management entities, while avoiding the scalability constraints imposed by central control.

Experimental validation showed that: recovery times of up to 6× faster than global optimization; cost savings of 27% in comparison to cost minimization based upon latency; linear scalability (to 200+ nodes); and recovery within less than 1 second for 89% of single-node failures, with less than 5% SLA violations.

The Proposed Framework addresses multiple important operational challenges:

• Lower OPEX - 27% decrease in migration costs provides a great deal of savings on an operational basis to those networks that have numerous daily failures to perform.

• Better User Experience - Sub-second URLLC recovery enables ITU-R M.2083-0 compliant mission critical applications.

• Infrastructure Flexibility - Support for heterogeneous nodes will allow incremental deployment of 5G networks following 3GPP architecture.

• Easier Operations - Automation of failure remediation reduces manual efforts and potential for human error.

Despite its effectiveness, the proposed framework has several limitations that can be investigated in the future. They include more realistic VNF state migration overheads, strong security and isolation mechanisms during migration, and explicitly considering inter-slice dependencies due to shared physical resources. The current recovery model may also be preempted using predictive failure models based on machine learning models and preemptive migration. The extension of the framework to multi-domain and multi-operator settings also remains non-trivial, particularly with respect to cross-domain coordination and enforcement of SLAs.

This work highlights the importance of distributed, and localized intelligence for a large scale network

management; At a higher level, this work is significant for valuing the distributed, localized intelligence for managing large network infrastructures operational overhead and recovery latency.

The more the network improves to 6G, and the more reliable it becomes with stringent demand for latency and scalability, the more critical localized intelligence, slice-aware differentiation discussed, and cost-aware optimization become in realizing fully autonomous mobile networks.

Conflict of interest

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

Using artificial intelligence tools

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

REFERENCES

1. ITU-R (2023) IMT-2030 Framework – Framework and overall objectives of the future development of IMT for 2030 and beyond. Recommendation ITU-R M.2160-0. Geneva: International Telecommunication Union. https://www.itu.int/rec/R-REC-M.2160-0-202311-I/en
2. NGMN Alliance (2020) 5G White Paper 2. Frankfurt am Main: NGMN Alliance. https://www.ngmn.org/publications/5g-white-paper-2.html
3. NGMN Alliance (2021) NGMN 6G Drivers and Vision. Frankfurt am Main: NGMN Alliance. https://www.ngmn.org/work-programme/ngmn-6g-drivers-and-vision.html
4. 3GPP (2022) System architecture for the 5G System (5GS). TS 23.501, Release 17. https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/
5. Singh, D. and Singh, J.P. (2024) 'A review on evolution, expectations and key enabling techniques of 5G', i-Manager's Journal on Communication Engineering and Systems, 13(1), pp. 38–48. https://doi.org/10.26634/jcs.13.1.20859
6. Shafi, M. et al. (2017) '5G: A tutorial overview of standards, trials, challenges, deployment, and practice', IEEE Journal on Selected Areas in Communications, 35(6), pp. 1201–1221. https://doi.org/10.1109/JSAC.2017.2692307
7. 3GPP (2018) Service requirements for the 5G system. TS 22.261, Rel. 16. https://www.3gpp.org/ftp/Specs/archive/22_series/22.261/
8. Di Mauro, M. et al. (2025) 'Reliability and availability in virtualized networks: A survey on standards, modeling approaches, and research challenges', arXiv preprint, arXiv:2503.22034. https://doi.org/10.48550/arXiv.2503.22034
9. Ammar, S., Lau, C.P. and Shihada, B. (2023) 'An in-depth survey on virtualization technologies in 6G integrated terrestrial and non-terrestrial networks', arXiv preprint, arXiv:2312.01895. https://doi.org/10.48550/arXiv.2312.01895
10. Herrera, J.G. and Botero, J.F. (2016) 'Resource allocation in NFV: A comprehensive survey', IEEE Transactions on Network and Service Management, 13(3), pp. 518–532. https://doi.org/10.1109/TNSM.2016.2598460
11. Chiosi, M. et al. (2012) Network Functions Virtualisation: An Introduction, Benefits, Enablers, Challenges and Call for Action. ETSI White Paper. https://portal.etsi.org/NFV/NFV_White_Paper.pdf
12. Heller, B., Sherwood, R. and McKeown, N. (2012) 'The controller placement problem', in Proceedings of the ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN). Helsinki, Finland, pp. 7–12. https://doi.org/10.1145/2342441.2342444
13. Hock, D. et al. (2013) 'Pareto-optimal resilient controller placement in SDN-based core networks', in Proceedings of the 25th International Teletraffic Congress (ITC). Shanghai, China, pp. 1–9. https://doi.org/10.1109/ITC.2013.6662939
14. Lange, S. et al. (2015) 'Heuristic approaches to the controller placement problem in large-scale SDN networks', IEEE Transactions on Network and Service Management, 12(1), pp. 4–17. https://doi.org/10.1109/TNSM.2015.2400758
15. Cohen, R., Lewin-Eytan, L., Naor, J.S. and Raz, D. (2015) 'Near optimal placement of virtual network functions', in Proceedings of IEEE INFOCOM. Hong Kong, China, pp. 1346–1354. https://doi.org/10.1109/INFOCOM.2015.7218483
16. Moens, H. and De Turck, F. (2014) 'VNF-P: A model for efficient placement of virtualized network functions', in Proceedings of the 10th International Conference on Network and Service Management (CNSM). Rio de Janeiro, Brazil, pp. 418–423. https://doi.org/10.1109/CNSM.2014.7014205
17. Mehraghdam, S., Keller, M. and Karl, H. (2014) 'Specifying and placing chains of virtual network functions', in Proceedings of IEEE CloudNet. Luxembourg, pp. 7–13. https://doi.org/10.48550/arXiv.1406.1058
18. Islam, M.S. and Chowdhury, S.A.H. (2024) 'Mobility management in next generation wireless networks', American Journal of Networks and Communications, 13(1), pp. 75–83. https://doi.org/10.11648/j.ajnc.20241301.16
19. Beck, M.T. and Botero, J.F. (2017) 'Scalable and coordinated allocation of service function chains', Computer Communications, 102, pp. 78–88. https://doi.org/10.1016/j.comcom.2016.12.003
20. 3GPP (2018) Telecommunication management; Study on management and orchestration of network slicing for next generation network. TR 28.801, Release 15. https://www.3gpp.org/ftp/Specs/archive/28_series/28.801/
21. Ordonez-Lucena, J. et al. (2017) 'Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges', IEEE Communications Magazine, 55(5), pp. 80–87. https://doi.org/10.1109/MCOM.2017.1600935
22. Rost, P. et al. (2017) 'Network slicing to enable scalability and flexibility in 5G mobile networks', IEEE Communications Magazine, 55(5), pp. 72–79. https://doi.org/10.1109/MCOM.2017.1600920
23. Foukas, X. et al. (2017) 'Network slicing in 5G: Survey and challenges', IEEE Communications Magazine, 55(5), pp. 94–100. https://doi.org/10.1109/MCOM.2017.1600951
24. Ksentini, A. and Nikaein, N. (2017) 'Toward enforcing network slicing on RAN: Flexibility and resource abstraction', IEEE Communications Magazine, 55(6), pp. 102–108. https://doi.org/10.1109/MCOM.2017.1600934
25. Zhang, H. et al. (2017) 'Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges', IEEE Communications Magazine, 55(8), pp. 138–145. https://doi.org/10.1109/MCOM.2017.1600940

26. ETSI (2014a) Network Functions Virtualisation (NFV); Architectural Framework. GS NFV 002 v1.2.1. Sophia Antipolis: ETSI. https://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_nfv002v010201p.pdf
27. ETSI (2015) Network Functions Virtualisation (NFV); Resiliency Requirements. GS NFV-REL 001 v1.1.1. Sophia Antipolis: ETSI. https://www.etsi.org/deliver/etsi_gs/NFV-REL/001_099/001/01.01.01_60/gs_nfv-rel001v010101p.pdf
28. Gember-Jacobson, A. et al. (2014) 'OpenNF: Enabling innovation in network function control', in Proceedings of ACM SIGCOMM. Chicago, IL, USA, pp. 163–174. https://doi.org/10.1145/2619239.2626313
29. Rajagopalan, S. et al. (2013) 'Split/Merge: System support for elastic execution in virtual middleboxes', in Proceedings of USENIX NSDI. Lombard, IL, USA, pp. 227–240. https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/rajagopalan
30. Baumgartner, A., Reddy, V.S. and Bauschert, T. (2015) 'Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization', in Proceedings of IEEE NetSoft. London, UK, pp. 1–9. https://doi.org/10.1109/NETSOFT.2015.7116162
31. Bera, A. et al. (2024) 'Network function virtualization and service function chaining frameworks: A comprehensive review', Electronics, 13(4), Article 748. https://doi.org/10.3390/electronics13040748
32. Sudhamani, C. et al. (2023) 'A survey on 5G coverage improvement techniques', Sensors, 23(4), Article 2356. https://doi.org/10.3390/s23042356
33. Qu, L. et al. (2017) 'A reliability-aware network service chain provisioning with delay guarantees in NFV-enabled enterprise datacenter networks', IEEE Transactions on Network and Service Management, 14(3), pp. 554–568. https://doi.org/10.1109/TNSM.2017.2732343
34. ETSI (2014b) Network Functions Virtualisation (NFV); Management and Orchestration. GS NFV-MAN 001 v1.1.1. Sophia Antipolis: ETSI. https://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_nfv-man001v010101p.pdf
35. Sahu,V., Sahu, N. and Sahu, R. (2024) 'Challenges and opportunities of 5G network: A review of research and development', American Journal of Electrical and Computer Engineering, 8(1), pp. 11–20. https://doi.org/10.11648/j.ajece.20240801.12
36. ETSI (2016) Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Os-Ma-nfvo reference point – Interface and Information Model Specification. GS NFV-IFA 013 v2.1.1. Sophia Antipolis: ETSI. https://www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/013/02.01.01_60/gs_nfv-ifa013v020101p.pdf

Received (Надійшла) 12.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Суліма Світлана Валеріївна – доктор філософії, доцент, доцент кафедри інформаційних технологій в телекомунікаціях, Національний технічний університет України «Київський політехнічний інститут імені І. Сікорського», Київ, Україна;
Svitlana Sulima – PhD, Associate Professor, Associate Professor of Department of Information technologies in telecommunications, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine;
e-mail: itssulima@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-6333-7693;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=55226282100.

Локальна реконфігурація слайсів мережі 5G у разі виходу з ладу або перевантаження вузлів

Світлана Суліма

Анотація. Актуальність. Швидке розгортання мереж 5G та широке застосування технології мережевого сегментування створюють нові виклики для забезпечення надійності та відмовостійкості послуг. Віртуалізовані інфраструктури, засновані на віртуалізації мережевих функцій, підвищують гнучкість, але водночас призводять до збільшення частоти відмов та коливань продуктивності. У таких середовищах централізовані механізми відновлення часто не відповідають суворим вимогам до затримки, особливо для наднадійних послуг зв'язку з низькою затримкою. **Об'єкт дослідження**: Об'єктом дослідження є процес відновлення після збою та реконфігурації ресурсів у віртуалізованих середовищах сегментації мереж 5G в умовах виходу з ладу вузлів та поступового перевантаження. **Мета**: Метою дослідження є розробка ефективного розподіленого методу локальної реконфігурації мережевих сегментів, який забезпечує швидке відновлення віртуальних мережевих функцій з урахуванням пріоритетів сегментів, витрат на міграцію та обмежень щодо затримки. **Результати**. Запропоновано архітектуру розподіленої локальної реконфігурації (DLR), що базується на ієрархічній структурі, яка складається з глобального оркестратора, регіональних менеджерів сегментів та локальних агентів моніторингу. Цей підхід передбачає використання багатокритеріальної моделі оптимізації для розміщення менеджерів сегментів та уніфікованої функції вартості міграції, яка враховує компоненти обчислювальних ресурсів, мережі, перебоїв у роботі та штрафних санкцій за порушення SLA. Розроблено алгоритми локалізованого відновлення для обробки як катастрофічних відмов вузлів, так і сценаріїв прогресивного перевантаження, при цьому пріоритетність сегментів визначається відповідно до їхніх вимог до послуг. **Висновки**. Запропонована архітектура розподіленого відновлення сегментів забезпечує швидку та масштабовану реконфігурацію сегментів мережі 5G в умовах збою. Завдяки поєднанню пріоритетності з урахуванням сегментів, рішень щодо міграції з урахуванням витрат та локалізованого управління цей підхід підвищує швидкість відновлення та операційну ефективність, зберігаючи при цьому якість обслуговування для додатків, для яких критично важлива низька затримка.

Ключові слова: мережі 5G; сегментація мережі; віртуалізація мережевих функцій; відновлення після збою; розподілене управління мережею; міграція віртуальних мережевих функцій; управління ресурсами з урахуванням сегментів; відмовостійкість; угода про рівень обслуговування; розподілена локальна реконфігурація.

УДК 621.396.96:519.872 **doi: 10.26906/SUNZ.2026.2.258**

Є. В. Тарасенко

Національний технічний університет “Харківський політехнічний інститут”, Харків, Україна

ЧАСОВІ ХАРАКТЕРИСТИКИ КАНАЛУ «РАДАР-ОБ’ЄКТ» НА ОСНОВІ GERT-МОДЕЛІ

Анотація. Актуальність. У сучасних радіолокаційних системах, що функціонують в умовах навмисних радіоелектронних перешкод, особливо критичним є початковий етап супроводу повітряних цілей – від моменту первинного виявлення до формування стійкого треку. Саме на цьому інтервалі приймаються рішення щодо підтвердження або втрати цілі, а помилки у виборі часових параметрів супроводу призводять до зростання хибних тривог, нестійкості оцінювання та втрати траєкторії. Традиційні методи налаштування кроку дискретизації, таймаутів і параметрів фільтрації часто мають евристичний характер і не спираються на формалізований аналіз часової структури процесу «радар–об’єкт», що зумовлює необхідність розроблення аналітично обґрунтованого підходу. **Об’єкт дослідження**: процес початкового супроводу повітряних цілей у радіолокаційній системі в умовах навмисних перешкод. **Мета статті**: розробка та статистична валідація інтегрованого методу узгодження часово-ймовірнісної моделі каналу «радар–об’єкт», побудованої на основі GERT-мережі, з параметрами рекурсивного оцінювання стану (фільтра Калмана) для забезпечення відтворюваного та узгодженого супроводу. **Результати дослідження**. У статті сформовано еквівалентну передавальну функцію GERT-мережі, що описує стохастичну структуру переходів між станами виявлення та підтвердження, та відновлено розподіл часу досягнення стану «стійкого супроводу» через аналіз полюсної структури. Показано зв’язок між характеристиками розподілу (математичним сподіванням, дисперсією, квантілями) і параметрами тракту супроводу: кроком дискретизації Δt, ініціалізацією матриці коваріації P_0, адаптивним профілем шуму процесу $Q(t)$, таймаутами підтвердження і втрати треку. Проведено статистичну перевірку узгодженості за критеріями NEES і NIS та порівняльний аналіз точності за метрикою RMSE на серії незалежних прогонів. Отримано статистично значуще зменшення похибки оцінювання та стабілізацію узгодженості після короткої перехідної ділянки. **Висновки.** Запропонований підхід забезпечує формалізований перехід від аналізу часових характеристик у s-площині до практичних інженерних налаштувань рекурсивного супроводу. На відміну від евристичних методів, він базується на полюсній структурі еквівалентної GERT-функції та статистичній валідації, що підвищує відтворюваність і надійність роботи РЛС у завадовому середовищі. Сфера використання отриманих результатів: радіолокаційні системи протиповітряної оборони, системи управління повітряним рухом, багатосенсорні комплекси спостереження та інші системи автоматизованого супроводу об’єктів в умовах активних перешкод.

Ключові слова: радіолокаційна система, GERT-мережа, фільтр Калмана, навмисні перешкоди, супровід цілей, фільтрація сигналів, стохастичне моделювання, завадостійкість.

Вступ

Постановка проблеми. У роботі [1] сформовано ймовірнісно-часову модель взаємодії «радар–об’єкт стеження» на базі експоненціальної GERT-мережі. Для гілок мережі вибрано твірні функції моментів експоненціального розподілу. Це дало змогу аналітично поєднувати послідовні та паралельні з’єднання, будувати еквівалентну *W*-функцію та виводити криві функції розподілу і густини ймовірності часу «взаємодії» (interaction time) на початковому етапі роботи РЛС (радіолокаційної станції) за спостережуваним об’єктом. Така конструкція добре відбиває структуру процесу, у якому присутні повернення до попередніх станів, підтвердження детекції та уточнення траєкторії. І саме тому GERT є зручним інструментом для макрорівневого аналізу, коли важливими є маршрути переходів, їхні ймовірності та часові затримки.

GERT-мережа застосовується для опису стохастичної структури переходів між станами виявлення та підтвердження. Вона дозволяє отримати еквівалентну передавальну функцію та відновити розподіл часу досягнення стану «стійкого супроводу». Однак, залишається відкритою проблема перетворення цієї аналітичної інформації, що подається в s-площині, через полюсну структуру функції W(s). Для конкретних інженерних налаштувань тракту супроводу та параметрів рекурсивного оцінювання. Особливу уваги потрібно приділити впливу кратності полюсів і повільноспадних компонентам розподілу на дисперсію часу підтвердження, а також відповідні порогові рішення. Отже, виникає наукова проблема розробки формалізованого підходу щодо узгодження часово-ймовірнісної моделі каналу «радар–об’єкт», котра будується на основі GERT, з параметрами фільтра Калмана, а також процедурами статистичної валідації (NEES, NIS). Це забезпечить відтворюване й статистично обґрунтоване налаштування РЛС на початковому етапі функціонування при навмисних перешкодах. Розв’язання цієї проблеми забезпечить перехід від аналітичного опису часових характеристик до практичних рекомендацій.

Аналіз останніх досліджень і публікацій. Процеси супроводу в РЛС зазвичай спираються на два напрями, це стохастичне моделювання часової структури процесу, а також рекурсивна оцінювання стану. Праці присвячені GERT-мережам [2–4], показують можливість формалізації складних процесів що розгалужуються та петлями, завдяки передавальній функції *W(s)*, це дозволяє отримувати моменти та розподіл часу завершення події. Перевагою підходу є аналітичне врахування повторних переходів і затримок, проте на практиці застосування результатів до налаштування параметрів супроводу звичай не розглядається.

Праці з теорії фільтра Калмана та радарного трекінгу [5–7] детально описують процедури оціню-

вання стану, ініціалізацію треку, вибір шумів Q і R, а також методи асоціації вимірювань. Перевірка на узгодженість моделей відбувається за критеріями NEES/NIS та χ^2-границями [8, 9]. При цьому, часові параметри підтвердження, а також втрати треку здебільшого визначаються евристично [10].

Аналіз джерел показує, що інтеграція полюсної структури та часові характеристики з інженерними налаштуваннями, залишаються мало дослідженими.

Метою роботи є розробка та обґрунтування моделі каналу «радар-об'єкт», котра побудована на основі GERT-мережі та фільтрі Калмана. Що забезпечує узгоджений та відтворюваний супровід цілей на початковому етапі функціонування РЛС за умов навмисних перешкод. Щоб досягти дану мету, потрібно виконати наступні завдання: зробити розподіл часу досягнення стану для стійкого супроводу, через передавальну функцію та визначити його моменти та квантилі; забезпечити зв'язок між отриманими часовими характеристиками та параметрами супроводу (дискретизація, ініціалізація, шум процесу); провести валідацію узгодженості супроводу за метриками та критеріями; провести порівняльний аналіз запропонованого підходу з базовим макроописом процесу.

Основний матеріал

Одним із завдань систематизації процесу супроводу цілей у РЛС на початковому етапі функціонування в умовах навмисних перешкод є перетворення якісних ймовірнісних висновків мережі GERT у кількісні часові показники. Це безпосередньо потрібно для налаштування тракту «радар–об'єкт» і вибору параметрів дискретизації у рекурсивному супроводженні. Під «часовими характеристиками» розуміємо розподіл випадкової величини часу від моменту первинного опромінення до набуття стану «стійкого супроводу» T_{cc}, її моменти $\mathbb{E}[T]$, $Var(T)$, квантілі t_p (час досягнення довірчої події з імовірністю p), а також похідні інженерні показники (середній лічильник сканів до підтвердження, рекомендований таймаут, розмір вікна гейтінгу тощо).

У формалізмі GERT кожна гілка мережі описується трансмісійною функцією вигляду

$$W_i(s) = p_i M_i(s), \qquad (1)$$

де p_i – імовірність проходження гілкою, $M_i(s) = L\{f_i(t)\}(s)$ – перетворення Лапласа густини часу затримки цієї гілки.

Для експоненційних гілок з інтенсивністю λ_i маємо $M_i(s) = \frac{\lambda_i}{\lambda_i + s}$.

Для послідовних і паралельних з'єднань використовуються добуток і зважена сума відповідно, а для наявних петель використовується узагальнена формула «еквівалентного передавання» (аналог правила Мейсона). В результаті для всієї мережі отримуємо раціональну функцію

$$W(s) = \frac{N(s)}{D(s)}, \qquad (2)$$

де N, D – багаточлени, $\deg D > \deg N$.

Ця функція є перетворенням Лапласа деякої густини $f_T(t)$ (з точністю до нормувальних множників, що фіксуються вимогою $\int_0^\infty f_T(t)dt = 1$. Відновлення часової густини здійснюється через зворотне перетворення Лапласа по контуру Бромвіча:

$$f_T(t) = L^{-1}\{W(s)\}(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} W(s)e^{st}ds, \qquad (3)$$

де γ вибрано правіше за всі сингулярності $W(s)$. На практиці інтеграл обчислюємо як суму залишків у полюсах знаменника $D(s)$.

Для наочності структуру сингулярностей (полюсів) еквівалентної функції $W(s)$ подано на рис. 1.

Рис. 1. Спектр полюсів $W(s)$

Прості полюси відповідають експоненційним складовим у часовій області, повторний полюс породжує поліноміально-експоненційні доданки.

Властивості та інтерпретація цієї реконструкції зручні з точки зору інженерного налаштування. Якщо всі полюси прості, $D(s) = \prod_{j=1}^{J}(s+\alpha_j)$ з $\alpha_j > 0$,

то $$f_T(t) = \sum_{j=1}^{J} c_j e^{-\alpha_j t} 1_{\{t\geq 0\}}, \qquad (3)$$

$$c_j = \operatorname*{Res}_{s=-\alpha_j}\left(W(s)e^{st}\right)\Big|_{t\geq 0}, \qquad (4)$$

і отже $f_T(t)$ – скінченна суміш експонент; функція розподілу $F_T(t) = \int_0^t f_T(u)du$ монотонно наближається до 1.

Якщо трапляються кратні полюси (порядку $m \geq 2$), у часовій області з'являються поліноміально-експоненційні доданки

$$f_T(t) = \sum_{j}\sum_{r=1}^{m_j} c_{j,r} t^{r-1} e^{-\alpha_j t} 1_{\{t\geq 0\}}, \qquad (5)$$

які створюють довші повільноспадні частини розподілу та підвищують дисперсію.

Це безпосередньо відбивається на виборі часових налаштувань. Довгі повільноспадні частини розподілу вимагають або збільшувати таймаут підтвердження треку, або посилювати фільтрацію хибних тривог через суворіше виконування вимірювань.

Наслідок кратності полюса для часової густини проілюстровано на рис. 2.

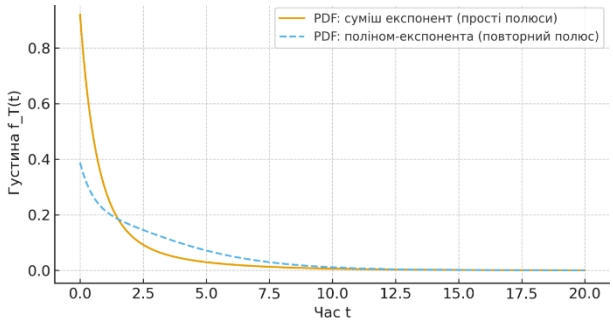


Рис. 2. Порівняння густин $f_T(t)$ для двох сценаріїв: суміш експонент (усі полюси прості) проти поліном-експоненти (повторний полюс)

Як видно з рис. 2, наявність повторного полюса формує повільноспадну кінцеву частину розподілу $f_T(t)$, що збільшує дисперсію часу та зсуває робочі квантілі вправо.

Моменти розподілу зручно добувати з похідних у нулі (моментної/твірної функції): якщо $W(s)$ нормовано як $\Phi(s) = \mathbb{E}\left[e^{-sT}\right]$, то

$$\mathbb{E}[T] = -\Phi'(0),\ Var(T) = \Phi''(0) - \left(\Phi'(0)\right)^2.$$

Ці величини ми використовуємо для первинної ініціалізації та погодження часової шкали з рекурсивним оцінюванням. Середній час $\mathbb{E}[T]$ задає орієнтовне вікно підтвердження (кількість сканів до стійкого треку), а дисперсія задає запас для таймаутів та ширину початкових коваріацій положення у P_0.

Практичний алгоритм чисельного уточнення має чотири кроки.

По-перше, формуємо $W(s) = \dfrac{N(s)}{D(s)}$ для еквівалентної GERT-мережі заданого сценарію (структура переходів, імовірності гілок, розподіли затримок).

По-друге, знаходимо всі корені $D(s) = 0$ у правій півплощині $\Re(s) < \gamma$ – це набір $\{-\alpha_j\}$. На цьому етапі бажано виконати кластеризацію близьких коренів, щоб виявити кратності.

По-третє, обчислюємо коефіцієнти залишків $c_{j,r}$ (для простих полюсів за стандартною формулою, для кратних через похідні знаменника).

По-четверте, збираємо $f_T(t)$ у вигляді суми експонент і поліноміально-експоненційних доданків. Інтегруванням отримуємо $F_T(t)$, а диференціюванням при потребі миттєву інтенсивність $h_T(t) = \dfrac{f_T(t)}{\left(1 - F_T(t)\right)}$.

Серед інженерних наслідків опису через сингулярності (полюси) передавальної функції зручні для параметризації супроводу можно відмітити наступні.

1) Вибір кроку дискретизації Δt. Найменший з характерних часових масштабів $\tau_{\min} \approx 1/\alpha_{\max}$ визначає верхню межу Δt. Якщо $\Delta t \gg \tau_{\min}$, рекурсивний фільтр не визначає швидкі події.

Якщо $\Delta t \ll \tau_{\min}$, зростають обчислювальні витрати без помітного виграшу.

2) Таймаути підтвердження/втрати треку. Квантілі $t_{0,9}$, $t_{0,95}$ доцільно брати як основу для верхніх меж очікування підтвердження. Для втрати використовувати характерну кінцеву частину розподілу $t_{0,9}$, якщо мережа має кратні полюси.

Подальше використання квантілів для регламенту підтвердження і втрати треку ілюструє рис. 3.

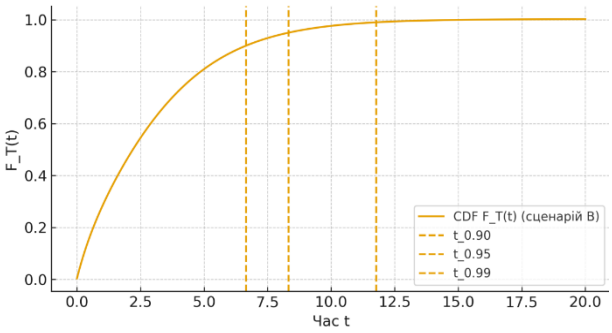


Рис. 3. Функція розподілу $F_T(t)$ (сценарій з повторним полюсом) із позначеними квантілями $t_{0,9}$, $t_{0,95}$, $t_{0,99}$

3) Виконування вимірювань та перевірка узгодженості (NEES). Ширина еліпсів виконування вимірювань може бути адаптована залежно від поточного часу t через $F_T(t)$. На ранніх стадіях більша апостеріорна невизначеність та низькі значення $F_T(t)$. У міру зростання $F_T(t)$ звуження та жорсткіший поріг NEES.

4) Вибір і налаштування матриці коваріації шуму процесу у фільтрі Калмана. Якщо у спектрі полюсів присутні близькі кластерні групи (ознака потенційних маневрових гілок), доцільно підвищити компоненти міри процесного шуму, пов'язані з прискореннями/кутовою швидкістю, але лише на часових інтервалах, де відповідні доданки $e^{-\alpha_j t}$ суттєві.

З погляду відтворюваності пропонується два супутні чисельні тести. Перший це узгодження моментів: моменти $\mathbb{E}[T]$, $Var(T)$, обчислені з $\Phi(s)$, порівнюються з емпіричними оцінками, отриманими на синтетичних прогонах «радар–Kalman». Другий це перевірка відсікання хвостів.

На гістограмі емпіричних часів підтвердження порівнюємо частку подій, що перевищують $t_{0,95}$ теоретичного F_T.

Систематичне перевищення вказує на недомодельовані петлі (наприклад, повторну валідацію або повторний пошук), які потрібно повернути в GERT-схему.

Протокол інтеграції формалізованих етапів фільтра Калмана для задачі супроводу представимо у вигляді послідовності практичних рекомендації, що забезпечить відтворюваність розробленої моделі. Протокол складається з чотирьох окремих кроків.

1) Фіксуємо Δt через правило «чверті масштабу». $\Delta t \approx 0.25\,\tau_{min}$. Це забезпечує достатню роздільну здатність для рекурсії Калмана без надмірних витрат.

Зв'язок між τ_{min} та рекомендованим кроком Δt наведено на рис. 4.

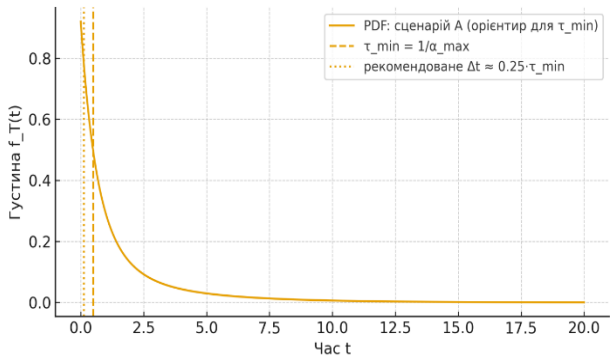


Рис. 4. Вибір кроку дискретизації Δt за $\tau_{\min} \approx 1/\alpha_{\max}$

2) Ініціалізуємо P_0 так, щоб його діагональні елементи для координат відбивали дисперсію, пов'язану з $t_{0,9}$ $\sigma_{x_0}^2 \sim v_{\max}^2 \cdot t_{0,9}^2/3$ (аналог рівномірної невизначеності траєкторії на ранньому етапі).

3) Задаємо адаптивний профіль $Q(t)$. підвищуємо Q у часових вікнах, де сума «швидких» експонент зі спектра полюсів перевищує фіксований поріг.

4) Перевіряємо узгодженість через NEES. Її стабілізація в межах 95-відсоткових χ^2-границь після часу порядку $t_{0,5}$ свідчитиме, що часові характеристики каналу «радар-об'єкт» і рекурсивний оцінювач узгоджені.

Таким чином, полюсна структура $W(s)$, отримана з GERT-мережі, не лише забезпечує аналітичне відновлення $f_T(t)/F_T(t)$, а й прямо диктує робочі налаштування тракту супроводу.

Статистична валідація методу супроводу цілей у РЛС та початковому етапі. Проведемо формальну кількісну перевірку узгодженості та ефективності запропонованого підходу супроводу цілей у РЛС на початковому етапі функціонування в умовах навмисних перешкод. У якій мережа GERT задає часово-ймовірнісний каркас процесу виявлення та підтвердження, а фільтр Калмана – виконує мікрорівневе оцінювання стану. Валідація здійснювалася на серії незалежних прогонів симулятора за фіксованими параметрами моделі руху, дискретизації та шумів; набір параметрів (матриці F, H, Q, R, початкові x_0 і P_0, крок Δt, тривалість та генератори випадковостей) наводиться у таблиці відтворюваності, що дозволяє повністю повторити експерименти.

Перевірка узгодженості виконувалася за критеріями NEES та NIS для лінійної Ґаусової постановки. Нормована квадратична помилка стану на кроці k обчислювалася як

$$NEES_k = (x_k - \hat{x}_k^+)^T (P_k^+)^{-1} (x_k - \hat{x}_k^+), \qquad (6)$$

де $\hat{x}_k^+$ та P_k^+ – апостеріорні оцінка і коваріація.

Для налаштованої моделі математичне сподівання $\mathbb{E}\{NEES_k\} = n_x$, а вибіркова траєкторія $\{NEES_k\}$ повинна знаходитись у довірчій смузі розподілу $\chi^2_{n_x}$ із рівнем 95%. Аналогічно для інновацій $\tilde{y}_k = z_k - H\hat{x}_k^-$ використовувався показник

$$NIS_k = \tilde{y}_k^T S_k^{-1} \tilde{y}_k, \quad S_k = HP_k^- H^T + R, \qquad (7)$$

який порівнювався з $\chi^2_{n_z}$. На підставі покриття довірчої смуги (частка кроків, на яких статистика лежить у межах $\left[\chi^2_{0,025}, \chi^2_{0,975}\right]$ формулювалася нульова гіпотеза про узгодженість: відхилення від номінального покриття 95% оцінювалося за біноміальною апроксимацією та бутстреп-інтервалами. У випадку систематичного виходу за верхню межу робився висновок про занижений шум процесу й виконувалося коригування Q. У протилежному разі перевірялася завищеність Q або некоректність R.

Порівняльна точність оцінювання визначалась за RMSE для координат і швидкостей із розділенням часової осі на три інтервали: до маневру, під час маневру і після маневру. Середньоквадратична помилка для вектора положення $r_k \in \mathbb{R}^2$ визначалася як

$$RMSE_{pos} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \left\| r_k - \hat{r}_k \right\|_2^2}.$$

Для швидкості середньоквадратична помилка визначалася аналогічно.

Для демонстрації корисності рекурсивної корекції застосовувалося зіставлення з макроописом GERT. Порівнювалися RMSE «GERT-середньої» траєкторії (без корекції вимірюваннями) та RMSE фільтра Калмана. Статистична значущість різниці перевірялася непараметрично (тест знаків або ранговий тест Вілкоксона) та через бутстреп-довірчі інтервали різниці медіан. Додатково звітувалася частка кроків, на яких абсолютна похибка оцінки положення перевищувала заданий допуск; цей показник інтерпретувався як практично орієнтований «рівень порушень допуску».

Порівняльні розподіли RMSE для фільтра Калмана та GERT-середньої траєкторії на серії прогонів подано на рис. 5.

Як видно з цього рисунку існують зсув медіани та вужчі інтервали для Kalman. Це підтверджує статистично значуще зменшення похибки. Медіана RMSE для Kalman суттєво нижча. Розкид менший, що вказує на стабільність оцінювання. Це узгоджується з поведінкою NEES після переходу

Ініціалізація P_0 здійснювалася на основі квантілів часу підтвердження, що витікають із полюсної структури $W(s)$. діагональні елементи коваріації для координат пропорціонувалися $\frac{v_{\max}^2 t_{0,9}^2}{3}$, що адекватно відбиває рівномірну невизначеність початкового відрізка траєкторії. Крок дискретизації обирався за правилом чверті найшвидшого масштабу $\Delta t \approx 0.25\,\tau_{min}$,, де $\tau_{\min} \approx 1/\alpha_{\max}$ – обернена до найбільшої швидкості затухання серед релевантних полюсів. Така прив'язка уникає як «пересемплінгу», так і втрати чутливості до швидких подій. На етапах, де сумарний внесок «швидких» експонент у часовій області ставав суттєвим, використовувався підвищений рівень шуму процесу $Q(t)$ для своєчасного відслідковування маневру; у спокійних ділянках $Q(t)$ зменшувався до номінального значення, що знижувало дисперсію оцінки.

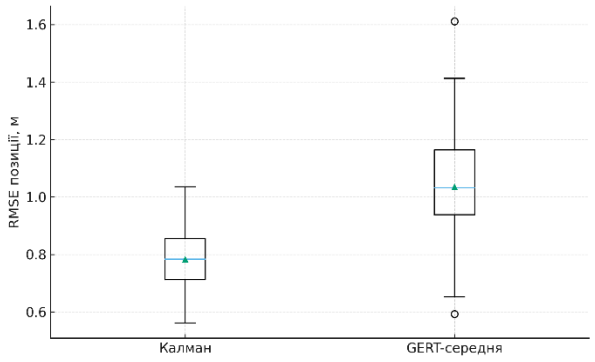


Рис. 5. Розподіли RMSE для Kalman та GERT-середньої на N=100 незалежних прогонів

Узгодженість вважалася досягнутою, коли середня траєкторія NEES стабілізувалася всередині 95%-ї χ^2-смуги після часу порядку $t_{0,5}$, а емпіричне покриття для NEES і NIS було не нижчим за номінальні 95% з урахуванням статистичної похибки вибірки. За цих умов медіанний виграш у точності визначався як відсоткове зменшення RMSE фільтра Калмана відносно GERT-середньої референції на всій тривалості експерименту та окремо у фазі маневру. Очікувано найбільший ефект спостерігався саме під час маневрових ділянок, де рекурсивна корекція найбільш корисна.

Графічну перевірку узгодженості за NEES із 95%-ми χ^2-границями наведено на рис. 6.

На рис. 6 виділена сіра зона – це перехідна ділянка. Після $k \approx 40$ значення стабілізуються в межах довірчої смуги (емпіричне покриття ≈ 95–100%), що підтверджує узгодженість моделі, рівнів шумів Q, R та вибору Δt. В той же час позасмугові сплески на переході очікувані, їхнє зникнення після налаштування свідчить про адекватність профілю $Q(t)$ та калібрування R.

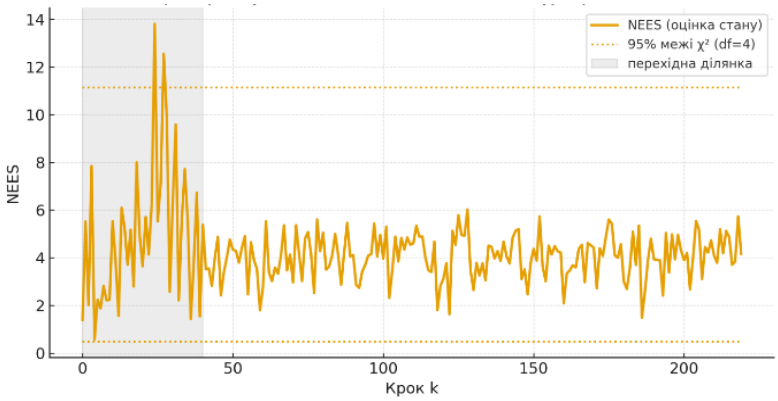


Рис. 6. Траєкторія $NEES_k$ з 95% χ^2-границями для n_x=4

Остаточне зведення результатів наводилося у двох таблицях: параметри моделі та симуляції (для повної відтворюваності) і агреговані метрики з довірчими інтервалами (для об'єктивної оцінки виграшу). У сукупності ці дані підтверджують, що перехід від аналізу в s-площині до інженерних налаштувань Δt, P_0 та адаптивного $Q(t)$, а також подальша валідація за NEES/NIS забезпечують статистично обґрунтовану й відтворювану якість супроводу в умовах завад.

Висновки

Проведена статистична перевірка показала, що запропонований метод супроводу коректно ініціалізується і входить у режим узгодженої роботи вже на початковому етапі, попри наявність навмисних перешкод. Траєкторія NEES після короткої перехідної ділянки стабілізується всередині 95-відсоткових χ^2-меж для розмірності стану, що свідчить про адекватність початкових налаштувань Δt та калібрування шумів Q, R у стартових умовах завадового середовища. Це означає, що модель руху та статистика вимірювань узгоджені з реальною динамікою сигнал/зашумлення саме у фазі «розігріву» траєкторного оцінювача.

Порівняльний аналіз на множині незалежних прогонів підтвердив ефективність методу з погляду точності вже в початковому вікні спостережень: розподіли RMSE для фільтра Калмана демонструють нижчі медіани і вужчі міжквартильні інтервали відносно GERT-середньої траєкторії. Причому виграш особливо проявляється у моменти раптового погіршення якості вимірювань, характерного для активних завад. Це узгоджується з динамічною адаптацією $Q(t)$ і робастною інтерпретацією квантілей часу підтвердження/втрати, що мінімізує помилки прийняття рішень на ранніх кроках супроводу.

Сукупність результатів, таких як стабілізація NEES у довірчій смузі після короткого переходу та статистично значного зменшення RMSE у стартовому інтервалі, дає підстави вважати метод верифікованим для практичного застосування на початковому етапі функціонування РЛС під навмисними перешкодами. Виявлені залежності також окреслюють робочі межі.

За тривалих епізодів негаусівських завад або масових хибних спрацювань потрібна розширена схема з адаптацією R за інноваціями і процедурою асоціації (наприклад, PDA/JPDA); однак у типовому сценарії запуску системи запропоновані налаштування забезпечують узгодженість і відтворюваний приріст точності вже з перших циклів супроводу.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

Список літератури

1. Semenov S., Krupska-Klimczak M. et al. Mathematical Model for the Initial Interaction Stage Between a Radar System and a Target Using GERT Network. *Applied Sciences*. 2025. Vol. 15, no. 3:1123. DOI: https://doi.org/10.3390/app15031123
2. Rudresh T. K., Parameshwar M.C., Purushottama Lingadevaru. Analysis of Modern Kalman Filter Algorithms for Radar-Based Target Tracking under Uncertainty. *2025 International Conference on Vehicular Technology and Transportation Systems (ICVTTS)*, Bangalore, India. 2025. P. 1-6. DOI: https://doi.org/10.1109/ICVTTS67119.2025.11296516
3. Na Zhang, Meng Ou, Bin Liu, Jian Liu. A GERT Network Model for input-output optimization of general aviation industry chain based on value flow. *Computers & Industrial Engineering*. 2023. Vol. 176. DOI: https://doi.org/10.1016/j.cie.2022.108945
4. Semenov S., Wasiuta O., Jammine Aet al. Development of an Intelligent Method for Target Tracking in Radar Systems at the Initial Stage of Operation Under Intentional Jamming Conditions. *Applied Sciences*. 2025. Vol. 15, no. 13:7072. DOI: https://doi.org/10.3390/app15137072
5. Wei Y., Hong T., Kadoch M. Improved Kalman Filter Variants for UAV Tracking with Radar Motion Models. *Electronics.* 2020. Vol. 9, no. 5:768. DOI: https://doi.org/10.3390/electronics9050768
6. Akram M. A., Liu P., Tahir M. O., Ali W., Wang Y. A State Optimization Model Based on Kalman Filtering and Robust Estimation Theory for Fusion of Multi-Source Information in Highly Non-linear Systems. *Sensors* 2019. Vol. 19, no.7:1687. DOI: https://doi.org/10.3390/s19071687
7. Shaoying Wang, Huajing Fang, Xuegang Tian. Recursive estimation for nonlinear stochastic systems with multi-step transmission delays, multiple packet dropouts and correlated noises. Signal Processing. 2015. Vol. 115. P. 164-175. DOI: https://doi.org/10.1016/j.sigpro.2015.03.022
8. T. Kruse, T. Griebel, K. Graichen. Adaptive Kalman Filtering: Measurement and Process Noise Covariance Estimation Using Kalman Smoothing. *IEEE Access*. 2025. Vol. 13. P. 11863-11875. DOI: https://doi.org/10.1109/ACCESS.2025.3528348
9. Peng D., Xie K., Liu M. Manoeuvre Target Tracking in Wireless Sensor Networks Using Convolutional Bi-Directional Long Short-Term Memory Neural Networks and Extended Kalman Filtering. *Sensors*. 2024. Vol. 24, DOI: https://doi.org/10.3390/s24134261
10. Wang X., Li T., Sun S., Corchado J. M. A Survey of Recent Advances in Particle Filters and Remaining Challenges for Multitarget Tracking. *Sensors*. 2017. Vol. 17,. DOI: https://doi.org/10.3390/s17122707

Received (Надійшла) 26.01.2026
Accepted for publication (Прийнята до друку) 29.04.2026
Publication date (Дата публікації) 22.05.2026

Відомості про авторів / About the Authors

Тарасенко Євген Віталійович – аспірант кафедри системи інформації ім. В.О. Кравця, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Yevhen Tarasenko – PhD Student, Department of Information Systems named after V. O. Kravets, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: Yevhen.Tarasenko@cit.khpi.edu.ua; ORCID Author ID: https://orcid.org/0009-0004-0506-6997;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59552874700.

Time characteristics of the radar-object channel based on the GERT model

Yevhen Tarasenko

Abstract. Relevance. In modern radar systems operating in conditions of intentional radio-electronic interference, the initial stage of air target tracking is particularly critical - from the moment of initial detection to the formation of a stable track. It is during this interval that decisions are made on confirmation or loss of the target, and errors in the selection of tracking time parameters lead to an increase in false alarms, unstable assessment and loss of trajectory. Traditional methods for setting the sampling step, timeouts and filter parameters are often heuristic in nature and do not rely on a formalized analysis of the time structure of the "radar-object" process, which necessitates the development of an analytically justified approach. **Object of research**: the process of initial tracking of air targets in a radar system in conditions of intentional interference. **Purpose of the article**: development and statistical validation of an integrated method for matching the time-probabilistic model of the radar-object channel, built on the basis of the GERT network, with the parameters of the recursive state estimation (Kalman filter) to ensure reproducible and consistent tracking. **Research results**. The article forms an equivalent transfer function of the GERT network, which describes the stochastic structure of transitions between the detection and confirmation states, and restores the distribution of the time to reach the “stable tracking” state through the analysis of the pole structure. The relationship between the distribution characteristics (mathematical expectation, variance, quantiles) and the tracking path parameters is shown: the sampling step Δt, the initialization of the covariance matrix P_0, the adaptive process noise profile $Q(t)$, the confirmation and track loss timeouts. A statistical check of consistency was carried out using the NEES and NIS criteria and a comparative analysis of accuracy using the RMSE metric on a series of independent runs. A statistically significant reduction in the estimation error and stabilization of consistency after a short transition section were obtained. **Conclusions**. The proposed approach provides a formalized transition from the analysis of time characteristics in the s-plane to practical engineering settings of recursive tracking. Unlike heuristic methods, it is based on the pole structure of the equivalent GERT function and statistical validation, which increases the reproducibility and reliability of the radar operation in a jamming environment. Scope of application of the obtained results: air defense radar systems, air traffic control systems, multi-sensor surveillance complexes and other systems of automated tracking of objects in conditions of active jamming.

Keywords: radar system, GERT network, Kalman filter, intentional interference, target tracking, signal filtering, stochastic modeling, noise immunity.

УДК 004.7 **doi: 10.26906/SUNZ.2026.2.264**

Цзянь Юй[1], Цзян Хе[2], С. Г. Семенов[3], С. І. Васюхно[4]

[1] Компанія «Zhongke Shuguang», Тяньцзінь, Китай
[2] Компанія «CNOOC Financical Shared Service Center PRD Branch», Шеньчжень, Китай
[3] Університет Комісії національної освіти, Краків, Польща
[4] Національний університет оборони України, Київ, Україна

ПОРІВНЯЛЬНЕ ДОСЛІДЖЕННЯ МЕТОДІВ ПЕРЕДАЧІ ВІДЕОДАНИХ У МЕРЕЖАХ БПЛА

Анотація. Актуальність. Передавання відеоданих у комп'ютерних мережах БПЛА є складною міждисциплінарною задачею, в якій поєднуються особливості мобільних бездротових мереж, вимоги до обслуговування трафіку реального часу та прикладна значущість відеоінформації для виконання місії. **Об'єкт дослідження**: процес відеоданих у комп'ютерних мережах БПЛА. **Мета статті**: провести порівняльне дослідження сучасних методів передачі відеоданих у мережах БПЛА. **Результати дослідження**. Показано, що ефективність передавання відео в таких мережах визначається сукупною дією мобільності вузлів, динамічної топології, варіативності пропускної здатності, затримки, джитера, втрат пакетів, параметрів відеопотоку та ресурсних обмежень бортового і наземного сегментів. Установлено, що в мережах БПЛА якість передавання має оцінюватися не лише за середніми мережевими показниками, а й з урахуванням часової придатності та прикладної цінності відеоінформації. **Висновки.** Показано, що жоден із розглянутих підходів окремо не забезпечує повного врахування динаміки мережі, часової значущості відеоданих та ресурсних обмежень, характерних для мереж БПЛА.

Ключові слова: комп'ютерна система, БПЛА, відеопотік, штучний інтелект, динаміка мережі.

Вступ

Безпілотні літальні апарати впродовж останніх років перетворилися з вузькоспеціалізованих технічних засобів на універсальні мобільні платформи, що активно застосовуються у цивільній, промисловій, безпековій та дослідницькій сферах. Їх використання охоплює завдання аерофотозйомки, моніторингу інфраструктури, екологічного контролю, пошуково-рятувальних операцій, спостереження за динамічними об'єктами, а також забезпечення ситуаційної обізнаності в умовах обмеженого або небезпечного доступу людини до зони спостереження. На відміну від стаціонарних систем відеоспостереження, БПЛА поєднують мобільність, можливість швидкого розгортання та змінюваний ракурс спостереження, що істотно розширює функціональні можливості систем збору візуальної інформації.

У більшості практичних сценаріїв саме відеодані є основним типом інформації, що формується на борту безпілотного літального апарата та передається до наземного пункту керування або до іншого вузла мережі. Відеопотік забезпечує оператору або автоматизованій системі прийняття рішень безперервне уявлення про стан об'єкта спостереження, просторове оточення, характер змін у зоні місії та потенційні загрози. Тому ефективність функціонування БПЛА в багатьох випадках прямо залежить не лише від факту отримання відеоінформації, а й від її актуальності, повноти та придатності до подальшого аналізу в режимі, наближеному до реального часу.

Особливість відеоданих як інформаційного ресурсу полягає у поєднанні значного обсягу, часової впорядкованості та високої чутливості до затримок під час передавання. Якщо для окремих телеметричних повідомлень допустимими можуть бути порівняно невеликі затримки або епізодичні втрати, то для відеопотоку такі порушення безпосередньо впливають на сприйняття сцени, точність виявлення подій та можливість оперативного реагування. У цьому контексті відеодані, що передаються в мережах БПЛА, слід розглядати не просто як сукупність пакетів, а як часово чутливий потік інформації, для якого визначальне значення мають узгодженість процесів формування, передавання, приймання та відтворення.

Специфіка функціонування БПЛА зумовлює й особливий характер комп'ютерної мережі, в межах якої здійснюється передавання відеоданих. Така мережа, як правило, не є статичною інфраструктурою з незмінними параметрами зв'язку, а формується як динамічне середовище взаємодії повітряних і наземних вузлів. Її стан змінюється під впливом траєкторії руху літальних апаратів, висоти польоту, радіоумов, перешкод, перевантаження каналів, зміни відстані між вузлами та особливостей використовуваних засобів передавання. Унаслідок цього параметри мережі можуть істотно варіюватися навіть у межах однієї місії, що створює додаткові труднощі для стабільної доставки відеоінформації.

Важливою рисою мереж БПЛА є поєднання обмежених ресурсів із підвищеними вимогами до оперативності. Бортові обчислювальні засоби, енергетичні можливості платформи, пропускна здатність каналу зв'язку та стійкість бездротового середовища утворюють сукупність факторів, які безпосередньо впливають на якість відеопередавання. При цьому завдання системи полягає не лише в транспортуванні максимальної кількості даних, а в забезпеченні такого режиму обміну, за якого відео зберігає прикладну цінність для спостереження, навігації або підтримки прийняття рішень. Саме тому для мереж БПЛА проблема передавання відеоданих набуває не суто телекомунікаційного, а комплексного системного характеру, що поєднує мережеві, обчислювальні та прикладні аспекти.

Окрему увагу слід приділити тому, що в мережах БПЛА відеопередавання часто виконує подвійну

функцію. З одного боку, воно забезпечує віддалене візуальне спостереження оператором, з іншого є джерелом даних для автоматизованої обробки, зокрема для задач виявлення об’єктів, класифікації ситуацій, супроводу цілей, навігаційної корекції або формування керуючих впливів. У таких умовах деградація параметрів передавання впливає не лише на суб’єктивну якість зображення, а й на достовірність машинного аналізу сцени. Це означає, що вимоги до передавання відеоданих визначаються не тільки класичними показниками якості обслуговування, а й функціональною роллю відео в загальному контурі роботи безпілотної системи.

Разом з тим мережі БПЛА не можна розглядати як однорідний клас середовищ передавання. Їх архітектура може охоплювати як прості схеми прямого зв’язку між одним безпілотником і наземною станцією, так і складні багатовузлові конфігурації з ретрансляцією, кооперацією кількох апаратів, розподілом функцій спостереження та використанням проміжних вузлів обробки. Ускладнення топології розширює можливості системи, проте одночасно підвищує варіативність часових характеристик передавання та ризик порушення цілісності відеопотоку.

На рис. 1 представлена схема передавання відеоданих у комп’ютерній мережі БПЛА як взаємодія бортового, мережевого та наземного сегментів.

Рис. 1. Передавання відеоданих у комп’ютерній мережі БПЛА як взаємодія бортового, мережевого та наземного сегментів

Різноманіття технічних сегментів обробки та передачі відеоданих з бортів БПЛА дає підставу зробити попередній висновок про те що, передавання відеоданих у комп’ютерних мережах БПЛА є складною міждисциплінарною задачею, в якій поєднуються особливості мобільних бездротових мереж, вимоги до обслуговування трафіку реального часу та прикладна значущість відеоінформації для виконання місії. Саме ця сукупність характеристик визначає необхідність подальшого аналізу чинників, що впливають на якість передавання відеоданих, а також обґрунтовує потребу в пошуку спеціалізованих підходів, орієнтованих на умови функціонування мереж БПЛА.

Аналіз останніх досліджень і публікацій. Упродовж останніх років в науковій літературі суттєво зріс інтерес до проблем передавання відеоданих у мережах БПЛА. При цьому дослідження розвиваються за кількома основними напрямами. Першим великим напрямком є оглядові роботи, присвячені систематизації проблем потокового відео БПЛА. Другий напрямок це праці, орієнтовані на низьколатентну доставку відео через стільникові мережі. Третій напрямок охоплює дослідження з адаптивного відеострімінгу, перефірійних обчислень та WebRTC-орієнтованих рішень. Крім того, є також праці, у яких відеопередавання аналізується в межах складніших архітектур 3D. У сукупності ці дослідження формують вагому теоретичну й прикладну основу, однак не усувають потреби в окремому порівняльному дослідженні методів передавання відеоданих у мережах БПЛА за єдиною системою критеріїв.

У роботі [1] автори виконали цінну систематизацію досліджень, присвячених передаванню відео з БПЛА, і показали, що потокове відел БПЛА доцільно розглядати як окремий науковий напрям, у якому поєднуються питання кодування, доставки, якості сприйняття та функціонування бездротового середовища. Важливим є сам принцип комплексного розгляду задачі, а також акцент на тому, що проблема не зводиться ні лише до протокольного рівня, ні лише до відеокодеків. Водночас оглядовий характер цієї праці не передбачає жорсткого зіставлення методів передавання за спільними критеріями придатності саме до мереж БПЛА. Зокрема, у ній недостатньо виразно поставлено питання про системне порівняння методів за затримкою, джитером, стійкістю до зміни топології, втратами пакетів і часовою цінністю відеоінформації. Саме тому в нашій статті доцільно перейти від загального огляду до структурованого порівняльного аналізу методів.

У праці [2] позитивним є те, що автори значно поглиблюють оглядову рамку й переходять до benchmarking-підходу, аналізуючи іммерсивне стрімінгове відео, питання стандартизації, ефективності кодування, апаратних та програмних енкодерів, а також демонструючи тестовий стенд для 360°-стрімінгу через 5G. Це показує важливість урахування не лише транспортних характеристик, а й впливу формату відео, кодека, енергоспоживання та QoE. Разом з тим її фокус зміщений на 360°-стрімінг, тоді як у нашому дослідженні предметом є ширший клас методів передавання відеоданих у мережах БПЛА. Отже, хоча зазначена праця формує важливу методологічну основу, вона не замінює потреби в порівнянні традиційніших та практично поширених методів передавання відеоданих у БПЛА-мережах.

Значний інтерес становить стаття [3], у якій досліджено низьколатентну доставку відео для дистанційного пілотування літальних апаратів через стільникові мережі. Позитивним є те, що автори працюють не з абстрактною моделлю, а з реальним каналом доступу та показують, що затримка відтворення нижче 300 мс для Full HD–4K відео може підтримуватися приблизно у 95% часу, а переваги статичного бітрейту чи адаптивного потокового передавання залежать від типу середовища. Для нашої роботи ця стаття важлива тим, що вона наочно демонструє роль польових вимірювань і підтверджує критичність доставки з низькою затримкою для безпечних сценаріїв віддаленого пілотування. Однак вона концентрується переважно на сценарії передачі на основі стільникового зв'язку і не дає ширшого порівняння інших класів методів передавання

відеоданих у мережах БПЛА. Саме через це доцільним є наше дослідження, у якому підходи, орієнтовані на мобільний зв'язок, розглядаються як один із класів методів, а не як єдина рамка аналізу.

У роботі [4] вагомим позитивним результатом є реалізація реального тестового стенду для передачі мультимедіа повітря–земля з використанням кількох стільникових операторів. Автори показують приклад практичного використання мережевої різноманітності для підтримки відеопотоку якості, достатньої для BVLoS операціій. Це є підтвердженням того, що таки підходи можуть істотно підвищувати стійкість відеодоставки в умовах нестабільного радіоканалу. Водночас робота зосереджена саме на одній технологічній лінії і не зіставляє її в єдиній системі оцінювання з адаптивним бітрейтом, ARQ/FEC, маршруто-орієнтованими або прикладними підходами. Тому потрібно помістити multipath-рішення у ширший порівняльний контекст.

Окремий напрям представлено в дослідженні [5], де позитивним є саме міжрівневий підхід до організації потокового передавання медіа з використанням RTP over UDP та орієнтацією на підвищення надійності і ефективності передачі. Це ще раз говорить про доцільність виходу за межі ізольованого розгляду окремого рівня мережевої моделі та демонструє, що ефективне відеопередавання в БПЛА-мережах часто потребує координації транспортних, мережевих і прикладних механізмів. Проте подібні cross-layer-підходи зазвичай аналізуються в межах власної архітектури оптимізації та не розглядаються порівняльно відносно інших методів передачі.

У статті [6] розглянуто є поєднання адаптивної потокової передачі відео з задачами розміщення БПЛА, справедливого розподілу відеошвидкостей та плавної адаптації якості. Автори показують, що ефективність відеопередачі визначається не лише бітрейтом, а й енергоспоживанням, справедливим розподілом ресурсів та плавністю зміни якості потоку. Це дозволяє включити до порівняння клас методів, орієнтованих на оптимізацію параметрів відеопотоку та ресурсів бездротової системи. Проте фокус дослідження зміщений у бік розподілу ресурсів та розміщення БПЛА, а не у бік системного зіставлення методів передавання відеоданих як таких. Саме тому наше порівняльне дослідження покликане узагальнити цей клас рішень поряд з іншими методами, які використовуються в мережах БПЛА.

У праці [7] автори розглядають задачу максимізації якості обслуговування (QoE) для потокової передачі повітряного відео через 3D-стільникові мережі, спираючись на карту знань середовища та каналів, алгоритми графа для планування траєкторії та ітераційні процедури визначення швидкості відтворення й розподілу часу. Дана стаття переконливо показує перспективність екологічних комунікацій та використання сайт-специфічної інформації при організації відеострімінгу. Водночас цей підхід є достатньо спеціалізованим і орієнтованим на складну 3D-платформу оптимізації стільникового зв’язку. Він не замінює потреби в загальному порівняльному аналізі методів передавання відеоданих у мережах БПЛА, особливо коли йдеться не лише про cellular scenario, а й про ширший спектр транспортних, маршрутизаційних та прикладних рішень.

У роботі [8] продемонстровано з’єднання з плануванням траєкторії, розподілом смуги пропускання, оптимізацією часу зв’язку та потоковою передачі на основі DASH для аварійних мереж зв’язку в приміщенні та на вулиці. Це демонструє потенціал комплексної оптимізації відеосервісу в середовищах, де UAV виступає аерорелеєм для системи, що втратила наземну інфраструктуру. Але дослідження залишається переважно оптимізаційним і орієнтованим на конкретний сценарій аварійного реле. У цьому не розглядається системне встановлення цього класу рішень з багатошляховим, на основі WebRTC, з підтримкою периферії, ретрансляцією чи міжрівневими підходами. Саме тому наша стаття потрібна для розміщення подібних спеціалізованих методів у єдиному полі порівняльного аналізу.

Суттєвий інтерес також стає стаття [9], в якому автори розглядають кілька БПЛА, кінцевих користувачів і периферійні сервери як єдину систему співпраці і пропонують адаптивну стратегію управління ресурсами, в тому числі з використанням глибокого підкріплення навчання, орієнтовану на зменшення затримки та підвищення ефективності реагування на надзвичайні ситуації. Стаття розширює межі традиційного мережевого розгляду та показує, що якість передавання відеоданих залежить також від розміщення обчислювальних ресурсів та оркестрації з підтримкою периферійних обчислень. Водночас такі роботи переважно фокусуються на архітектурі системи та політиці розподілу ресурсів, а не на порівнянні базових методів передавання відеоданих.

У роботі [10] позитивним є створення WebRTC-багатодронової системи моніторингу та керування із низькозатримною відеопередачею (low-latency video streaming), медіасервером Janus та наземною станцією керування на основі вебтехнологій. Ця праця підтверджує практичну придатність WebRTC як бази для низькозатримної передачі відео в багатодронових системах, причому не для одного БПЛА, а одразу для багатодронової системи. Однак системний прототипно-орієнтований характер дослідження означає, що автори демонструють працездатність певної архітектури, але не ставлять за мету широке порівняння WebRTC-підходу з альтернативними класами методів передавання відеоданих у мережах БПЛА.

У статті [11] автори провели польові експерименти для IoT на основі WebRTC з борту БПЛА і показали стабільні наскрізні затримки значно менші за 10 мс у надійних мережах, а також перевагу каналу даних WebRTC над WebSocket для чутливого до часу зв’язку. Для нашої роботи ця праця важлива тим, що вона надає експериментально підтверджений аргумент на користь транспортних та прикладних рішень, орієнтованих на високу чутливість до затримок. Разом з тим стаття має формат прикладного дослідження і не вирішує завдання системного зіставлення ширшого набору методів передавання відеоданих у мережах БПЛА. Саме тому вона є важливим джерелом фактів і висновків, але водночас підкреслює актуальність статті як узагальнювального порівняльного дослідження.

Додатково слід відзначити працю [12], у якій позитивним є систематизація саме AI-орієнтованих підходів до потокового відео з повітряних апаратів. Ця стаття показує окремий клас рішень, де штучний інтелект використовується для прогнозування, оптимізації та адаптації параметрів доставки відео. Водночас вона не формує порівняльної рамки для зіставлення AI-базованих підходів із традиційними потоковими, багатошляховими, периферійними, орієнтованими на повторну передачу чи WebRTC-базованими методами. Отже, навіть у межах інтелектуальних рішень залишається відкритою потреба в загальнішому порівняльному дослідженні.

Таким чином, в сучасній літературі вже накопичено значний масив результатів щодо потокового відео з БПЛА, низькозатримної доставки через стільникові мережі, багатошляхової мультимедійної доставки, міжрівневої координації, адаптивного керування бітрейтом, потокового 3D-відео з урахуванням оточення, архітектур з підтримкою периферійних обчислень, систем на основі WebRTC та оптимізації з використанням штучного інтелекту. Кожен із цих напрямів дав вагомі позитивні результати, які доцільно врахувати в нашій роботі. Проте наявні публікації здебільшого зосереджені або на окремому класі рішень, або на специфічному сценарії використання, або на демонстрації працездатності певної архітектури. Тому порівняльне дослідження, в якому різні методи передавання відеоданих у мережах БПЛА були б послідовно зіставлені за спільною системою критеріїв, зокрема за затримкою, джитером, втратами пакетів, адаптивністю, стійкістю до зміни топології, ресурсною доцільністю та здатністю зберігати прикладну цінність відеоінформації, є **актуальним науковим завданням**.

1. Чинники впливу на якість передавання відеоданих у мережах БПЛА

Якість передавання відеоданих у мережах безпілотних літальних апаратів формується під впливом сукупності взаємопов'язаних чинників, що охоплюють характеристики самого відеопотоку, параметри бездротового каналу, особливості мережевої взаємодії та умови функціонування бортового й наземного обладнання.

На відміну від традиційних стаціонарних мереж, у яких значна частина параметрів є відносно стабільною в часі, мережі БПЛА характеризуються підвищеною мінливістю, унаслідок чого якість доставки відеоінформації визначається не одним домінуючим фактором, а їх динамічною комбінацією. Саме ця обставина ускладнює забезпечення сталої якості відеосервісу та зумовлює необхідність комплексного аналізу впливів, які діють одночасно на різних рівнях функціонування системи.

Узагальнена схема взаємозв'язку основних чинників, що впливають на якість передавання відеоданих у мережах БПЛА, наведено на рис. 2.

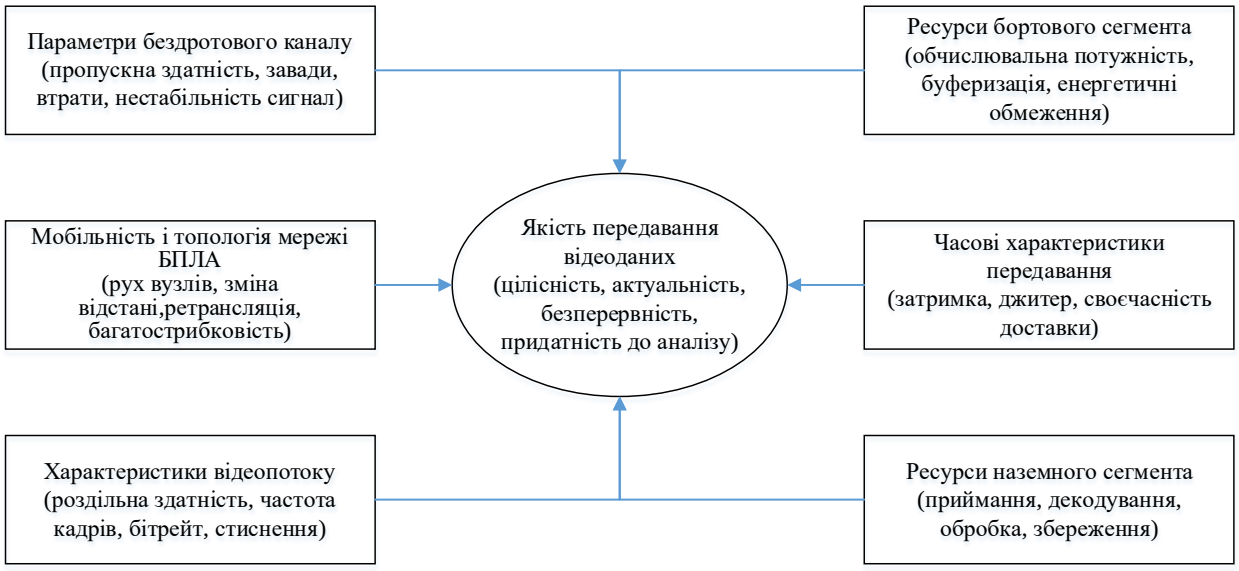


Рис. 2. Узагальнена схема взаємозв'язку основних чинників, що впливають на якість передавання відеоданих у мережах БПЛА

2. Аналіз існуючих підходів і засобів передавання відеоданих

Сучасні підходи до передавання відеоданих у комп'ютерних мережах формувалися під впливом зростання попиту на мультимедійні сервіси, системи віддаленого спостереження, відеоконференцзв'язок, потокове мовлення та інтерактивні інформаційні системи. У межах таких застосувань було розроблено значну кількість протоколів, методів і технологічних рішень, орієнтованих на забезпечення прийнятної якості відтворення відео в умовах обмеженої пропускної здатності, затримок, втрат пакетів і неоднорідності мережевої інфраструктури. Водночас мережі БПЛА висувають до процесу передавання відеоданих специфічні вимоги, зумовлені мобільністю вузлів, мінливістю радіоканалу, обмеженістю бортових ресурсів і високою залежністю прикладної цінності відео від моменту його доставки. У зв'язку з цим аналіз існуючих підходів має ґрунтуватися не лише на їх

загальній функціональності, а й на ступені придатності до використання в динамічному мережевому середовищі безпілотних літальних апаратів.

Узагальнено існуючі засоби передавання відеоданих можна розглядати як поєднання кількох взаємопов'язаних рівнів. На нижчих рівнях вирішуються задачі транспортування пакетів через мережу, керування передаванням, маршрутизації та забезпечення базової доступності каналу зв'язку. На вищих рівнях здійснюється формування потоку, фрагментація відеоданих, синхронізація, контроль якості та взаємодія з прикладними сервісами. Саме на межі між транспортними механізмами та прикладною логікою реалізується більшість практичних рішень, що забезпечують роботу відеосистем у реальному середовищі. Однак ефективність такого підходу істотно залежить від того, наскільки мережеві умови узгоджуються з припущеннями, закладеними в архітектуру протоколів і механізмів адаптації.

Узагальнену структуризацію існуючих підходів і засобів передавання відеоданих за функціональними рівнями проілюстровано на рис. 3.

Рис. 3. Узагальнена класифікація підходів і засобів передавання відеоданих у мережевих системах

Серед найбільш поширених рішень, що використовуються для передавання відео в мережевому середовищі, центральне місце посідають потокові протоколи реального часу. Їх призначення полягає у впорядкованій доставці мультимедійних даних, синхронізації потоків і підтримці механізмів контролю якості на стороні сеансу. У традиційних мережах такі засоби продемонстрували високу ефективність у задачах потокового відтворення, дистанційного спостереження та інтерактивних сервісів. Разом з тим їх робота зазвичай орієнтується на відносно передбачуване середовище передавання, у якому часові параметри хоч і можуть змінюватися, проте не зазнають настільки різких коливань, як у мережах БПЛА. За таких умов загальні механізми керування сеансом і контролю потоку не завжди здатні забезпечити достатню чутливість до короткочасних, але критичних змін мережевої ситуації.

Значний клас існуючих підходів пов'язаний із застосуванням буферизації, повторного передавання, випереджувального виправлення помилок і варіативного кодування. Такі механізми спрямовані на підвищення стійкості відеопередавання до втрат і нестабільності каналу. Буферизація дозволяє згладжувати нерівномірність надходження даних, повторне передавання дає змогу відновлювати окремі втрачені фрагменти, а надлишкове кодування підвищує ймовірність відновлення потоку без необхідності запиту повторної доставки. У стаціонарних або слабо мінливих мережах ці підходи часто забезпечують позитивний результат, проте в мережах БПЛА вони пов'язані з об'єктивними обмеженнями. Зокрема, збільшення буфера неминуче підвищує затримку, повторні передавання вимагають додаткового часу і пропускної здатності, а надлишковість знижує ефективність використання каналу. Тому механізми, що в традиційних системах розглядаються як універсальні засоби підвищення надійності, у безпілотних мережах можуть вступати в суперечність із вимогою своєчасності доставки відеоінформації.

Помітне місце в сучасних системах займають підходи, засновані на адаптації параметрів відеопотоку до поточного стану мережі. Йдеться про зміну бітрейту, частоти кадрів, роздільної здатності, структури стиснення або інших характеристик потоку з метою збереження прийнятного рівня обслуговування в умовах обмежених ресурсів. Подібні механізми є важливим напрямом розвитку систем передавання мультимедійних даних, оскільки дозволяють певною мірою узгодити навантаження на канал із його реальними можливостями. Проте більшість таких рішень орієнтована на середні або згладжені показники мережевого стану, що добре працює у відносно інерційних середовищах, але є недостатнім у випадку різких і короткочасних коливань параметрів каналу.

Для мереж БПЛА це означає, що навіть адаптивні засоби загального призначення не завжди встигають відреагувати на зміну мережевої ситуації настільки швидко, щоб зберегти актуальність окремих відеокадрів.

Окремий напрям становлять рішення, у яких адаптація реалізується на рівні маршрутизації, вибору шляху проходження трафіку або координації взаємодії між кількома вузлами мережі. Для динамічних бездротових середовищ такий підхід є природним, оскільки значна частина проблеми передавання пов'язана саме зі змінами топології та неоднорідністю каналів. У мережах БПЛА маршрутні рішення

можуть підвищувати ймовірність збереження зв’язності, дозволяти використовувати ретрансляцію або перерозподіляти навантаження між каналами. Водночас подібні механізми в більшості випадків спрямовані на забезпечення загальної доступності мережі або зменшення середніх мережевих витрат, а не на збереження прикладної цінності конкретного відеокадру. Отже, хоча мережево-орієнтовані підходи є важливими для підтримки функціонування системи в цілому, вони не завжди забезпечують той рівень чутливості до часової значущості відеоданих, який є необхідним у задачах оперативного спостереження.

Суттєвого поширення набули й платформи та протокольні стеки, що забезпечують практичну інтеграцію передавання відео з прикладними сервісами реального часу. Вони поєднують функції транспортування, керування сеансом, контролю доступу до медіаданих, синхронізації та взаємодії з прикладним програмним забезпеченням. Перевага таких засобів полягає у технологічній зрілості та сумісності з існуючою інфраструктурою. Проте саме ця універсальність часто обмежує можливості їх глибокої адаптації до вузькоспеціалізованих умов мереж БПЛА. У типових реалізаціях вони створювалися для широкого класу мультимедійних застосувань, а тому орієнтуються переважно на підтримку працездатності сервісу в цілому, а не на цілеспрямоване керування передаванням відеоданих з урахуванням специфіки окремої місії, критичності кадрів чи мінливості стану мережі в кожен конкретний момент.

Варто зазначити, що в науковій та прикладній літературі значна увага приділяється оцінюванню якості відеопередавання на основі класичних показників якості обслуговування та якості сприйняття. Такий підхід є важливим для порівняння рішень і вибору допустимих режимів роботи системи. Проте для мереж БПЛА цього виявляється недостатньо, оскільки практична корисність відео визначається не лише формальною якістю відтворення або середнім рівнем затримки, а й відповідністю отриманої інформації поточному стану об’єкта спостереження. Іншими словами, у безпілотних системах питання передавання відеоданих виходить за межі забезпечення прийнятного мультимедійного сервісу і перетворюється на задачу підтримки актуальної інформаційної взаємодії між бортовим та наземним сегментами.

На цьому тлі стає очевидним, що існуючі підходи та засоби передавання відеоданих мають важливе методологічне і технологічне значення, однак кожен із них розв’язує переважно окрему частину загальної проблеми. Одні підходи краще забезпечують стійкість до втрат, інші гнучкість параметрів потоку, треті підтримку мережевої зв’язності або інтеграцію з прикладними сервісами. Проте для мереж БПЛА критично важливим є не локальне покращення одного показника, а узгоджене керування процесом передавання відеоданих у середовищі, де мережеві умови, ресурсні обмеження та прикладна цінність відео змінюються одночасно. Саме ця обставина вказує на недостатність використання лише традиційних або універсальних засобів без їх спеціальної адаптації до умов безпілотних мереж.

У підсумку можна стверджувати, що сучасний стан розвитку підходів і засобів передавання відеоданих створює необхідну основу для побудови спеціалізованих рішень, але сам по собі не усуває суперечності між потребою в надійності, вимогою низької затримки, мінливістю параметрів мережі та прикладною значущістю окремих фрагментів відеопотоку. У мережах БПЛА ця суперечність проявляється особливо гостро, оскільки інформація, отримана із запізненням, може виявитися непридатною для прийняття рішень навіть за умови формального збереження її цілісності.

Саме тому наступним кроком дослідження має стати постановка наукової проблеми, пов’язаної з необхідністю розроблення такого методу передавання відеоданих, який був би орієнтований не лише на загальні мережеві показники, а й на специфіку функціонування комп’ютерної мережі БПЛА та прикладну цінність своєчасно доставленого відео.

Висновки

У статті проведено порівняльне дослідження методів передавання відеоданих у комп’ютерних мережах БПЛА.

Показано, що ефективність передавання відео в таких мережах визначається сукупною дією мобільності вузлів, динамічної топології, варіативності пропускної здатності, затримки, джитера, втрат пакетів, параметрів відеопотоку та ресурсних обмежень бортового і наземного сегментів. Установлено, що в мережах БПЛА якість передавання має оцінюватися не лише за середніми мережевими показниками, а й з урахуванням часової придатності та прикладної цінності відеоінформації.

У результаті аналізу літератури та сучасних підходів встановлено, що потокові протоколи реального часу, механізми буферизації, повторної передачі та випереджувальної корекції помилок, адаптація параметрів відеопотоку, маршрутизаційні й платформно-орієнтовані рішення забезпечують розв’язання лише окремих аспектів загальної проблеми. Показано, що жоден із розглянутих підходів окремо не забезпечує повного врахування динаміки мережі, часової значущості відеоданих та ресурсних обмежень, характерних для мереж БПЛА.

Отже, перспективним напрямом подальших досліджень є розроблення спеціалізованих адаптивних методів передавання відеоданих, орієнтованих на узгодження надійності, затримки, адаптивності та збереження прикладної цінності відеоінформації в умовах стохастичної мінливості мережевого середовища БПЛА.

Конфлікт інтересів. Автори декларують, що не мають конфлікту інтересів стосовно даного дослідження, в тому числі фінансового, особистісного характеру, авторства чи іншого характеру, що міг би вплинути на дослідження та його результати, представлені в даній статті.

Використання засобів штучного інтелекту. Автори підтверджують, що не використовували технології штучного інтелекту при створенні представленої роботи.

СПИСОК ЛІТЕРАТУРИ

1. Ammous, D.; Kammoun, F.; Masmoudi, N. Survey on Video Streaming for UAV. In *Proceedings of the 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, Hammamet, Tunisia, 16–19 May 2023; IEEE: Piscataway, NJ, USA, 2023. https://doi.org/10.1109/IC_ASET58101.2023.10150479
2. Sharma, M.K.; Liu, C.-F.; Farhat, I.; Sehad, N.; Hamidouche, W.; Debbah, M. UAV Immersive Video Streaming: A Comprehensive Survey, Benchmarking, and Open Challenges. *arXiv* **2023**, arXiv:2311.00082. https://doi.org/10.48550/arXiv.2311.00082
3. Baltaci, A.; Cech, H.; Mohan, N.; Geyer, F.; Bajpai, V.; Ott, J.; Schupke, D. Analyzing Real-Time Video Delivery over Cellular Networks for Remote Piloting Aerial Vehicles. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22)*, Nice, France, 25–27 October 2022; ACM: New York, NY, USA, 2022. https://doi.org/10.1145/3517745.3561465
4. Bacco, M.; Cassarà, P.; Gotta, A. Air-to-Ground Real-Time Multimedia Delivery: A Multipath Testbed. *Veh. Commun.* **2022**, *33*, 100443. https://doi.org/10.1016/j.vehcom.2021.100443
5. Liu, Z.; Jiang, Y. Cross-Layer Design for UAV-Based Streaming Media Transmission. *Acta Electron. Sin.* **2022**, *50*(3), 617–626. https://doi.org/10.12263/DZXB.20210660
6. Ahmed, Z.; Ahmad, A.; Altaf, M.; Khan, F.A. Power Efficient UAV Placement and Resource Allocation for Adaptive Video Streaming in Wireless Networks. *Ad Hoc Netw.* **2023**, *150*, 103260. https://doi.org/10.1016/j.adhoc.2023.103260
7. Zhan, C.; Hu, H.; Liu, Z.; Wang, J.; Cheng, N.; Mao, S. Aerial Video Streaming Over 3D Cellular Networks: An Environment and Channel Knowledge Map Approach. *IEEE Trans. Wirel. Commun.* **2024**, *23*(2), 1432–1447. https://doi.org/10.1109/TWC.2023.3289501
8. Guo, Z.; Hu, B.; Chen, S.; Zhang, B.; Wang, L. Joint Resource and Trajectory Optimization for Video Streaming in UAV-Based Emergency Indoor-Outdoor Communication. *Telecommun. Syst.* **2024**, *87*(1), 199–211. https://doi.org/10.1007/s11235-024-01151-4
9. Sarkar, M.; Sahoo, P.K. Leveraging Edge Computing for Video Data Streaming in UAV-Based Emergency Response Systems. *Sensors* **2024**, *24*(15), 5076. https://doi.org/10.3390/s24155076.
10. Kilic, F.; Hassan, M.; Hardt, W. Prototype for Multi-UAV Monitoring–Control System Using WebRTC. *Drones* **2024**, *8*(10), 551. https://doi.org/10.3390/drones8100551
11. Chodorek, A.; Chodorek, R.R. Web Real-Time Communications-Based Unmanned-Aerial-Vehicle-Borne Internet of Things and Stringent Time Sensitivity: A Case Study. *Sensors* **2025**, *25*(2), 524. https://doi.org/10.3390/s25020524.
12. Nguyen, T.-V.; Nguyen, N.P.; Kim, C.; Dao, N.-N. Intelligent Aerial Video Streaming: Achievements and Challenges. *J. Netw. Comput. Appl.* **2023**, *211*, 103564. https://doi.org/10.1016/j.jnca.2022.103564.

Received (Надійшла) 19.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Юй Цзянь – магістр, старший інженер з розробки тестів, Zhongke Shuguang, Тяньцзінь, Китай;
Jian Yu– Master, Senior Test Development Engineer, Zhongke Shuguang, Tianjin, China;
e-mail: ianyu220272@gmail.com; ORCID Author ID: https://orcid.org/0009-0007-8990-8347;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59681719600&origin=recordpage.

Хе Цзян – магістр, фахівець з бухгалтерського обліку, Financial Shared Service Center PRD Branch, Шеньчжень, Китай;
Jiang He– Master, Accounting Staff, CNOOC Financial Shared Service Center PRD Branch, Shenzhen, China;
e-mail: hjamxm@163.com; ORCID Author ID: https://orcid.org/0009-0004-4625-3721.

Семенов Сергій Геннадійович – доктор технічних наук, професор, завідувач кафедри комп'ютерної інженерії та кібербезпеки, Університет Комісії національної освіти, Краків, Польща;
Serhii Semenov – Doctor of Technical Sciences, Professor, Head of the Department of Computer Engineering and Cybersecurity, University of the National Education Commission, Krakow, Poland;
e-mail: serhii.semenov@uken.krakow.pl; ORCID Author ID: https://orcid.org/0000-0003-4472-9234;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57202908821.

Васюхно Станіслав Іванович – начальник науково-дослідної лабораторії, Національний університет оборони України, Київ, Україна;
Stanislav Vasiukhno – Head of the Research Laboratory, National Defence University of Ukraine, Kyiv, Ukraine;
e-mail: stas816@ukr.net; ORCID Author ID: https://orcid.org/0000-0002-0884-0405.

Comparative study of video data transmission methods in uav networks

Jian Yu, Jiang He, Serhii Semenov, Stanislav Vasiukhno

Abstract. Relevance. Video data transmission in UAV computer networks is a complex interdisciplinary task that combines the features of mobile wireless networks, requirements for real-time traffic service, and the applied significance of video information for mission performance. **Object of study**: video data process in UAV computer networks. **Purpose of the article**: to conduct a comparative study of modern video data transmission methods in UAV networks. **Research results**. It is shown that the efficiency of video transmission in such networks is determined by the combined effect of node mobility, dynamic topology, bandwidth variability, delay, jitter, packet loss, video stream parameters, and resource constraints of the airborne and ground segments. It is established that in UAV networks, the transmission quality should be evaluated not only by average network indicators, but also taking into account the timeliness and applied value of video information. **Conclusions.** It is shown that none of the considered approaches individually provides full consideration of network dynamics, time significance of video data and resource constraints characteristic of UAV networks.

Keywords: computer system, UAV, video stream, artificial intelligence, network dynamics.

UDC 004.4.05:004.75:004.8 **doi: 10.26906/SUNZ.2026.2.271**

Dmytro Diachenko, Vladyslav Diachenko

Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

MODEL FOR ASSESSING THE RISK OF DEFECTS IN SOFTWARE COMPONENTS OF DISTRIBUTED COMPUTER SYSTEMS

Abstract. Relevance. The relevance of the study is determined by the need to improve the efficiency of automated testing of software in distributed computer systems, which are characterized by a complex structure, dependencies between components, and an increased risk of defects. Existing approaches to defect prediction mostly do not provide comprehensive consideration of the structural characteristics of components, the intensity of their changes, and the parameters of test execution, which complicates the justified planning of testing. Therefore, the development of a model for assessing the risk of defects in software components of distributed computer systems to support the process of prioritizing automated testing is relevant. This makes it possible to focus testing resources on the most critical software components and thereby increase the overall effectiveness of software quality control. **The object of research** is the process of detection and assessment of defects in software components of distributed computer systems in the context of automated testing. **Purpose of the article** is to develop a model for assessing the risk of defects in software components of distributed computer systems to improve the efficiency of automated testing. **Research results.** In this work, a model for assessing the risk of defects in software components of distributed computer systems has been developed, which makes it possible to form an integral defectiveness indicator based on the structural characteristics of the program code and the results of automated testing. Experimental studies using machine learning algorithms have shown that the best results are provided by the CatBoost model, which demonstrated the highest values of ROC–AUC and Precision–Recall characteristics compared to other investigated approaches. The results obtained confirm the possibility of effectively ranking software components by the level of defect risk and using this information for the prioritization of automated testing in distributed computer systems.

Keywords: distributed computer system, prediction, machine learning, automated testing, defect risk assessment, metrics, CatBoost, quality analysis, deep learning, multicomponent system.

Introduction

The modern development of computer systems is characterized by the active implementation of distributed architectures, within which application functionality is implemented by a set of interconnected software components, services, modules, and data exchange interfaces. Such systems are widely used in cloud platforms, corporate information environments, network services, financial technologies, telecommunications infrastructures, and other domains where increased requirements are imposed on the hardware and software complex regarding reliability, continuity of operation, and fault tolerance. The increasing complexity of the internal structure of distributed computer systems, the growth in the number of interconnections between components, and the high intensity of modification of software components lead to an increased probability of defects, which may result in partial or complete degradation of services, violation of data exchange integrity, and a decrease in the overall efficiency of system operation.

One of the key means of ensuring the quality and reliability of distributed computer systems is automated testing, which makes it possible to verify the correctness of component functionality, detect interaction errors between services, control the stability of software implementation after introducing changes, and maintain the required level of confidence in the results of system modification.

At the same time, with the growth of the scale of distributed systems, the number of test scenarios, and the frequency of software component updates, the problem of decreasing efficiency of the full cycle of automated testing arises. Running the entire set of tests for each new change requires significant computational, time, and infrastructure resources, while the use of fixed rules for test selection or prioritization often does not provide sufficient adaptation to the current state of the system.

Under these conditions, the preliminary assessment of the risk of defects in software components of a distributed computer system becomes particularly important, as it allows identifying the most vulnerable components even before the start of the full testing cycle and directing testing resources to those parts of the system where the probability of defects is the highest. This approach creates prerequisites for increasing the efficiency of automated testing by reducing the time to detect critical errors, ensuring justified prioritization of test scenarios, and decreasing the cost of performing regression testing.

A promising tool for solving this problem is the use of machine learning methods, which make it possible to identify hidden patterns in historical data on software code changes, results of previous test runs, characteristics of component structure, parameters of interactions between components, and operational performance indicators of the system. This is particularly important for distributed computer systems, in which defects often arise not only due to local implementation errors but also due to the complex dynamics of component interactions, configuration changes, uneven workloads, and the indirect influence of modifications in related subsystems.

Despite the significant number of studies in the field of software defect prediction, the issue of constructing a model for assessing defect risk specifically for software components of distributed computer systems remains relevant, with the aim of further improv-

ing the efficiency of automated testing. Existing approaches often focus either on general software code quality metrics or on local defect prediction without proper consideration of architectural dependencies, characteristics of the testing process, and the specific features of multicomponent system operation.

Therefore, the development of a model that combines structural, historical, testing, and operational characteristics of components and provides the formation of an integral defect risk assessment based on machine learning methods remains relevant.

The purpose of this work is to develop a model for assessing the risk of defects in software components of distributed computer systems to improve the efficiency of automated testing.

Main part

To solve the problem of improving the efficiency of automated testing of distributed computer systems, it is necessary to consider modern approaches to defect prediction, identify their advantages and limitations, and substantiate the possibility of constructing an original model for assessing the defect risk of software components.

In this regard, it is advisable to analyze modern studies devoted to the application of machine learning and deep learning methods in defect prediction tasks, which form the theoretical basis for the further development of the model, its description, and software implementation.

Paper [1] presents a systematic review of approaches to software defect prediction using artificial intelligence algorithms, which makes it possible to improve the efficiency of testing and software quality assurance processes. The results obtained can also be applied to the analysis of the reliability of software components in complex computer systems, in distributed architectures.

Considerable attention is also paid to the use of machine learning methods for defect prediction, considering their practical value for testing and system maintenance processes.

Paper [2] provides a systematic review of the application of machine learning in defect prediction tasks, focused not only on model accuracy but also on their suitability for use under real production conditions. The obtained results are important for the development of models for assessing the defect risk of components of distributed computer systems, since they confirm the expediency of combining predictive methods with the tasks of improving the efficiency of automated testing. It should be noted that this paper analyzes the most popular machine learning methods, among which decision trees, support vector machines, neural networks, random forests, and nearest neighbors occupy a leading position. At the same time, it is emphasized that the promise of a model is determined not only by classification accuracy but also by the possibility of its interpretation, reproducibility of results, consideration of the cost aspects of implementation, and suitability for application under specific organizational conditions.

Paper [3] presents a systematic review of the application of deep learning methods in defect prediction tasks, summarizing the algorithms used, datasets, approaches to source code representation, and methods for model evaluation. The authors concluded that a significant part of the research in the field of defect prediction is still based mainly on academic or open datasets, while the use of industrial data and the validation of models under real conditions remain limited. For the tasks of assessing the defect risk of components of distributed computer systems, this is especially important, since it confirms the need to develop models focused not only on high predictive accuracy but also on practical suitability for use in the conditions of automated testing of complex multicomponent architectures.

Paper [4] examines the current state of research devoted to software defect prediction using deep learning methods. The authors emphasize that, despite the growing interest in software defect detection, the use of deep learning in this field remains insufficiently systematized and requires the generalization of existing empirical results. The purpose of the study was to summarize scientific works on the application of deep learning to defect prediction from the standpoint of the metrics, models, techniques, datasets, and achieved results used, as well as to compare the effectiveness of deep learning models and classical machine learning methods.

An important area of research is defect prediction based on semantic features of source code. Paper [5] presents a systematic review of approaches to software defect detection based on the use of contextual information of source code, methods of its representation, and deep learning models. The obtained results are important for the development of models for assessing the defect risk of components of distributed computer systems, since they confirm the expediency of using semantic characteristics of software components to improve the efficiency of automated testing.

Paper [6] is devoted to the study of modern defect prediction models with an emphasis on the informativeness and explainability of the obtained results. It presents a systematic review of approaches to software defect prediction, within which the extent to which model results are understandable and suitable for practical use is analyzed. The conclusions obtained are important for the development of models for assessing the defect risk of components of distributed computer systems, since they confirm the expediency of moving from simple binary predictions to more informative models suitable for supporting automated testing.

Paper [7] is devoted to a comparative analysis of machine learning methods for test case prioritization under conditions of continuous testing constrained by strict time resources. The author investigates the influence of the volume of test execution history and the available time budget on the effectiveness of machine learning models in the early detection of regression defects, and also compares the approaches of SVM, ANN, GBDT, and LambdaRank using industrial datasets. The obtained results are important for the de-

velopment of models for assessing the defect risk of components of distributed computer systems, since they confirm the possibility of using historical testing data and machine learning methods to improve the efficiency of automated testing.

Paper [8] is devoted to the development of an approach to test prioritization based on learning-to-rank methods for early detection during regression testing. It proposes the variational approach APFD-Net, aimed at optimizing a differentiated objective function inspired by the APFD metric, which makes it possible to improve the consistency of priority space and enhance the generalization capability of the model. As a result, the authors showed that the proposed approach provides more effective early defect detection compared to traditional test prioritization methods.

Paper [9] is devoted to the development of an approach to test case prioritization in regression testing based on deep learning. It proposes the AnoLSTM-TCP method, which combines an LSTM model with anomaly features formed based on test execution history, test execution duration, and previous results, making it possible to detect temporal dependencies and rank test cases more effectively. As a result, the authors demonstrated that the proposed approach provides higher APFD values compared to well-known prioritization methods, including Random, ROCKET, RETECS, DeepGini, DeepOrder, LSTMTCP, and HyLSTMTCP, which indicates its better capability for early defect detection.

Paper [10] is devoted to a systematic analysis of the application of artificial intelligence methods for processing and analyzing logs in microservice architectures. It summarizes modern approaches to anomaly detection, root cause analysis, error prediction, and analysis of dependencies between services based on machine learning, deep learning, and hybrid learning methods. As a result, the authors concluded that, despite the significant potential of artificial intelligence approaches to improve the efficiency of log analysis in distributed systems, their practical implementation is constrained by problems of scalability, generalization of results, and limited availability of representative industrial datasets.

The conducted analysis of scientific publications indicates that modern studies confirm the prospects of applying machine learning and deep learning methods for defect prediction; however, most of them focus mainly on tasks of software module classification, test case prioritization, or analysis of individual types of data, in particular source code, testing results, or logs of microservice systems.

At the same time, for distributed computer systems the issue of constructing a generalized model that would combine the characteristics of software components, the results of automated testing, parameters of interaction between components, and operational indicators within a single defect risk assessment process remains insufficiently studied.

In addition, the literature analysis showed that a significant part of existing approaches is oriented either toward general open datasets or toward narrow prediction tasks without sufficient consideration of the specifics of multicomponent distributed architectures. This complicates the use of known solutions as a direct basis for improving the efficiency of automated testing in computer systems, where not only prediction accuracy is important but also the possibility of ranking components by risk, considering architectural dependencies, and supporting the subsequent prioritization of test scenarios.

In this regard, there is a need to develop a model for assessing the defect risk of software components of distributed computer systems that would consider the set of historical, structural, testing, and operational characteristics and ensure the formation of an integral risk assessment to support the automated testing process.

Therefore, the next stage of the research is the formulation of the problem and the substantiation of the structure of such a model.

Let a distributed computer system consist of a set of software components.

$$S = \{c_1, c_2, \ldots, c_n\}, \quad (1)$$

where each component c_i implements a separate functionality and interacts with other components through defined mechanisms of data exchange, service calls, or message transmission. For each component, a set of characteristics is formed.

$$X_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}, \quad (2)$$

where the features may belong to the following groups:

– characteristics of software code changes,

– indicators of previous testing results,

– structural parameters of the component,

– parameters of interaction with other components,

– test coverage indicators,

as well as operational characteristics.

It is necessary to construct such a mapping model

$$F : X_i \to R_i, \quad (3)$$

where R_i – component defect assessment c_i, presented in the form of the probability of defect occurrence, an integral risk index, or membership in a certain risk class. The task is to form, based on the available data, an informative and computationally feasible model for defect risk assessment that will allow:

- identifying components with the highest probability of defect occurrence in the current testing cycle.

- considering not only historical test results but also the influence of change intensity, component complexity, and its architectural dependencies.

- providing a basis for further prioritization and selection of test scenarios.

In distributed computer systems, the defectiveness of a software component cannot be considered only as a consequence of local implementation errors. Its level is the result of the influence of a set of factors, among which the intensity of code modifications, architectural complexity, the degree of interaction with other ser-

vices, the quality of existing test coverage, and the history of previous failures play an important role.

The first group of factors includes the characteristics of software code changes. Frequent commitments, a significant volume of modified lines, modification of critical modules, as well as changes in components with many dependencies usually increase the risk of introducing defects.

The second group is related to the results of previous automated tests. If a component or the test scenarios associated with it demonstrate instability, a high frequency of failures, or a history of detected defects, this may be an indicator of increased defectiveness.

The third group covers structural characteristics. These may include cyclomatic complexity, component size, the number of external calls, the number of API endpoints, the volume of business logic, the level of coupling, and the number of dependencies.

The fourth group includes testing characteristics, particularly the level of coverage, the average test execution time, the frequency of repeated runs, and the ratio of success to unsuccessful executions.

The fifth group of factors reflect the operational context:

– the number of incidents,

– the criticality of business functions,

– the frequency of component usage,

– its role in the overall service delivery architecture.

Taking these factors into account, it is advisable to form a comprehensive assessment that would not be reduced to a single indicator but would reflect the multidimensional nature of defectiveness in the components of a distributed system.

The proposed model assumes that the defectiveness of a software component is a latent characteristic that can be estimated based on a set of observable features.

For each component, a feature vector is formed.

$$X_i = \left(H_i, T_i, S_i, D_i, E_i\right), \qquad (4)$$

where H_i – historical characteristics of change, T_i – results of previous testing, S_i – structural characteristics of the component, D_i – parameters of dependencies between components, E_i – operational indicators.

Historical characteristics may include the number of changes over a defined period, the average volume of modifications, the number of authors of changes, and the frequency of updates.

The subset may include the frequency of test failures, the number of successful and unsuccessful runs, the number of flaky cases, the duration of test runs, and the test coverage of the component.

Structural characteristics T_i may include the frequency of test failures, the number of successful and unsuccessful runs, the number of flaky cases, the duration of test runs, and the test coverage of the component.

Structural characteristics S_i describe the complexity of the component, its size, the number of functional branches, and the density of internal dependencies.

Parameters D_i reflect the position of the component in the interaction graph of the system, for example the number of incoming and outgoing connections, centrality in the service graph, and the presence of critical dependencies.

Indicators E_i consider incidents in operation, the importance of the component for user scenarios, and the frequency of requests to the corresponding service. Based on the feature vector, an integral defect assessment is formed.

$$R_i = \alpha_1 H_i^* + \alpha_2 T_i^* + \alpha_3 S_i^* + \alpha_4 D_i^* + \alpha_5 E_i^*, \qquad (5)$$

where $H_i^*, T_i^*, S_i^*, D_i^*, E_i^*$ – normalized aggregated estimates for the corresponding groups of features, $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ – weight coefficients that determine the degree of influence of each group of parameters on the final defect risk. In the general case, the values of the coefficients can be specified by experts or determined based on training a machine learning model. If a binary representation of the target variable is used, the model can produce an estimate in the form of the probability of defect occurrence in a component during the next testing cycle. If a multiclass scheme is chosen, the component is assigned to one of the risk classes, for example low, medium, or high. Unlike simplified approaches, in which prediction is based only on the history of previous failures, the proposed model considers both the internal characteristics of the component and its external context within the structure of the distributed system. This makes it possible to increase the informativeness of the assessment and create prerequisites for more justified planning of automated testing. For the practical use of the model, it is advisable to introduce an integral indicator of defect risk Q_i, which is defined as a normalized function of the final assessment R_i:

$$Q_i = \frac{R_i - R_{\min}}{R_{\max} - R_{\min}}. \qquad (6)$$

Then the value of $Q_i \in [0,1]$ can be interpreted as the risk level of the component. Based on threshold values, it is possible to form classes:

$$\mathrm{Risk}(Q_i) = \begin{cases} low, & Q_i < \theta_1, \\ middle, & \theta_1 \le Q_i < \theta_2, \\ high, & Q_i \ge \theta_2. \end{cases} \qquad (7)$$

Such an approach is convenient for further use in decision support systems during automated testing, since it makes it possible not only to rank components by the level of defectiveness but also to form sets of priority objects for in-depth test control. Under conditions of applying machine learning methods, the integral indicator can be formed based on the predicted probability of a component belonging to the class of defective ones. For this purpose, algorithms such as logistic regression, random forest, gradient boosting, XGBoost, or CatBoost can be used.

The advantage of this approach is the possibility of automatically determining the influence of individual features and evaluating their importance in forming the prediction.

The software implementation of the model is advisable to perform in Google Colab based on the open dataset Regenerated PROMISE and BPD Datasets [11], which contains component-oriented metrics and defect labels for several versions of open-source systems. Further research is planned to use the open SQuAD dataset [12].

The analysis of the file structure showed that it contains 36 data subsets formed for 13 open-source software systems, where each subset corresponds to a separate version of the system and represents software components at class level. Such a choice is justified because the dataset is representative for defect prediction tasks, has a unified feature structure across all subsets, and makes it possible to build a reproducible model for assessing the defect risk of components of distributed computer systems.

Unlike datasets focused on pull requests, commit-level analysis, or log analysis, this dataset is directly suitable for component-level modeling, which corresponds to the problem formulation presented in this work. Within the framework of experimental validation, it is advisable to combine all 36 subsets into a single array of observations.

In this case, the model will be trained on 12,690 software components for which six structural-metric characteristics are specified:

– CBO,
– DCC,
– ExportCoupling,
– ImportCoupling,
– NOM,
– WMC,

as well as the target feature Defect.

The meaning of these indicators makes it possible to interpret them as a set of characteristics of coupling, the intensity of interaction between components, structural complexity, and the functional content of a class. Such a structure is sufficient for constructing an initial model for defect risk assessment, since it allows considering not only the internal complexity of a software component but also its connections with other elements of the system.

At the same time, the target variable Defect should be interpreted as a binary indicator of the presence of a defect, while the output of the model should be interpreted as a probabilistic estimate of the defect risk for each component. For the software implementation, it is proposed to use a binary classification scheme with the subsequent interpretation of the predict_proba value as an integral risk index. At the experimental stage, it is advisable to build and compare several machine learning models, in particular logistic regression, support vector machines, random forest, and gradient boosting.

As the main model, it is advisable to choose Random Forest or XGBoost, since they work well with tabular metric data, are resistant to nonlinear dependencies between features, and allow the importance of individual component characteristics to be evaluated. The quality of the model should be evaluated using the metrics Accuracy, Precision, Recall, F1-score, and ROC-AUC, while practical usefulness should be assessed through ranking components according to the obtained risk index.

Thus, the software validation of the model is performed on an open multi-project dataset that contains structural characteristics of software components and labels of their defectiveness. This makes it possible to implement in Google Colab a defect risk assessment model as a tool for supporting automated testing: after training the model, components can be ordered according to the value of the risk index, which creates a basis for further prioritization of tests and concentration of testing resources on the riskiest elements of the system.

The results of the experimental study confirmed the effectiveness of the developed model. The analysis of feature importance showed that the greatest contribution to the formation of defectiveness assessment is made by combining characteristics of component complexity and coupling, metrics of the type ComplexityPlusCoupling, ComplexityTimesCoupling, MethodsTimesCoupling, as well as logarithmized structural indicators.

This indicates the expediency of considering not individual basic metrics but their integrated combinations when constructing the risk model.

The comparative analysis of models showed (Fig. 1) that the CatBoost model demonstrated the best results, for which the values ROC-AUC = 0.9156 and PR-AUC = 0.9116 were obtained. This confirms the high ability of the proposed approach to distinguish between defective and non-defective components and to effectively rank them according to the level of risk.

It is shown that the use of an integral defect risk index makes it possible to identify a group of components with the highest probability of defect occurrence, which creates a basis for prioritizing automated testing.

Fig. 2 presents a comparison of the Precision–Recall characteristics of the studied classification models.

The obtained results show that the proposed model based on CatBoost provides the highest overall prediction quality, which is confirmed by the largest value of the area under the Precision–Recall curve (AP = 0.9116).

The XGBoost model demonstrates a slightly lower result (AP = 0.8556), while Random Forest and Logistic Regression are characterized by even lower efficiency with AP values of 0.8391 and 0.7993, respectively.

The analysis of the curve shapes indicates that as recall increases, precision gradually decreases for all models, which is typical for defect prediction tasks. At the same time, the CatBoost model provides higher precision values over most of the recall range compared to other approaches, which confirms its ability to more accurately identify defective components of the software system.

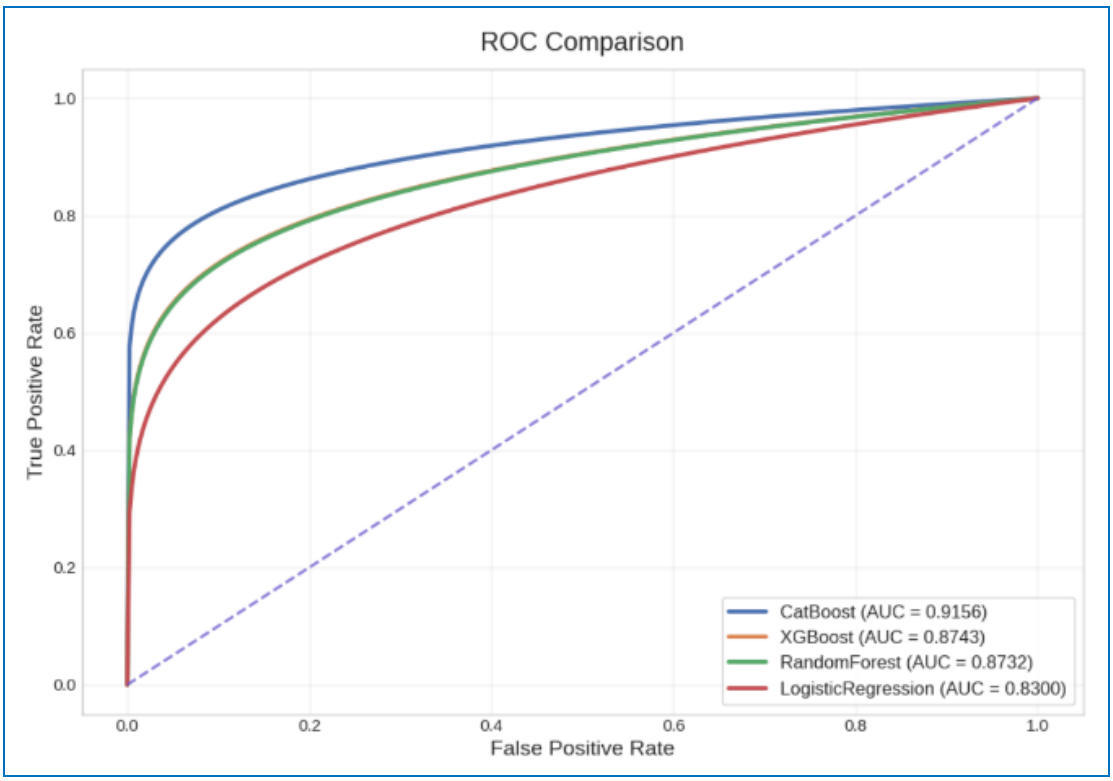


Fig. 1. ROC Comparison

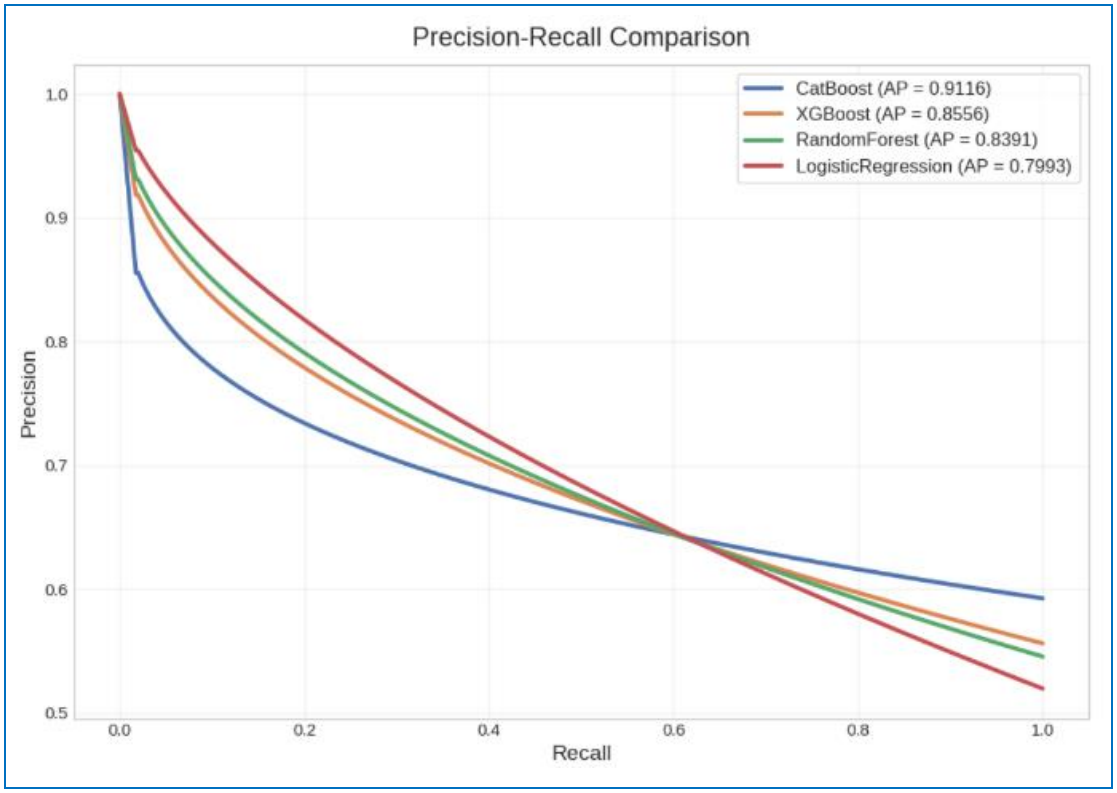


Fig. 2. Precision-Recall Comparison

Fig. 3 presents the distribution of the integral defect risk index of software components formed based on the results of the proposed model. The histogram demonstrates that most components are characterized by low and medium values of the risk index (approximately in the range of 0.2–0.4), while a relatively small proportion of components has increased values of the indicator exceeding 0.5. Such a distribution indicates the ability of the model to effectively differentiate system components by the level of potential defectiveness and to form a high-risk group that should be prioritized during test planning.

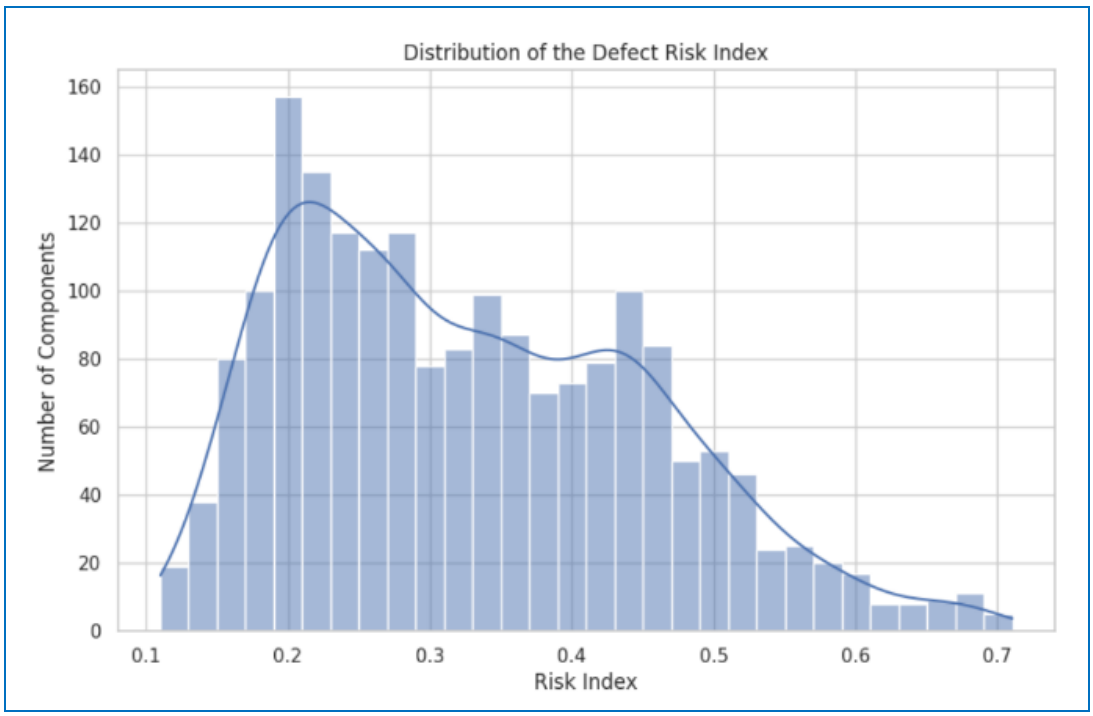


Fig. 3. Distribution of the Defect Risk Index

The obtained results confirm the high ability of the developed model to distinguish between defective and non-defective components.

The presence of a pronounced low-risk region and a smaller group of components with increased index values indicates the correctness of forming the integral risk indicator and its suitability for supporting the process of prioritizing the testing of software components of a distributed system.

Conclusions

During the study, the problem of improving the efficiency of automated testing of software for distributed computer systems was considered. The conducted analysis of modern approaches to defect prediction showed that most existing methods are focused mainly on the use of individual software code metrics or historical failure data, which limits their applicability for supporting the test planning process in complex distributed software systems.

In this work, a model for assessing the defect risk of software components of distributed computer systems was developed, which takes into account the structural characteristics of components, the intensity of changes, inter-component dependencies, and test execution parameters. Based on these data, an integral defect risk indicator is formed, which makes it possible to rank the components of a software system according to the level of potential defectiveness and to use the obtained estimates for the prioritization of automated testing.

The results of the experimental study confirmed the effectiveness of the proposed approach. A comparative analysis of machine learning algorithms showed that the CatBoost model demonstrates the best results, providing the highest values of ROC-AUC and Precision–Recall characteristics compared to the XGBoost, Random Forest, and Logistic Regression models. The analysis of the distribution of the defect risk index showed the possibility of effectively differentiating software components by the level of potential defectiveness and identifying a group of high-risk components.

The obtained results confirm that the use of the developed model makes it possible to improve the efficiency of planning automated testing of software for distributed computer systems by prioritizing the verification of components with the highest probability of defect occurrence.

References

1. Jalaj Pachouly, Swati Ahirrao, Ketan Kotecha, Ganeshsree Selvachandran, Ajith Abraham. A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools, Engineering Applications of Artificial Intelligence, Volume 111, 2022. P. 1–33. https://doi.org/10.1016/j.engappai.2022.104773.
2. Szymon Stradowski, Lech Madeyski. Machine learning in software defect prediction: A business-driven systematic mapping study. Information and Software Technology,Volume 155,2023. P. 1–17. https://doi.org/10.1016/j.infsof.2022.107128.
3. Görkem Giray, Kwabena Ebo Bennin, Ömer Köksal, Önder Babur, Bedir Tekinerdogan. On the use of deep learning in software defect prediction. Journal of Systems and Software, Volume 195, 2023. P. 1–26. https://doi.org/10.1016/j.jss.2022.111537.
4. Zuhaira Muhammad Zain, Sapiah Sakri, and Nurul Halimatul Asmak Ismail. Application of Deep Learning in Software Defect Prediction: Systematic Literature Review and Meta-analysis. Inf. Softw. Technol. Volume 158, 2023. https://doi.org/10.1016/j.infsof.2023.107175.

5. Abdu A, Zhai Z, Algabri R, Abdo HA, Hamad K, Al-antari MA. Deep Learning-Based Software Defect Prediction via Semantic Key Features of Source Code–Systematic Survey. Mathematics. 10(17):3120. 2022. P. 1–26. https://doi.org/10.3390/math10173120.
6. Natalie Grattan, Daniel Alencar da Costa, Nigel Stanger. The need for more informative defect prediction: A systematic literature review. Information and Software Technology. Volume 171, 2024. P. 1-24. https://doi.org/10.1016/j.infsof.2024.107456.
7. Marijan, D. Comparative study of machine learning test case prioritization for continuous integration testing. Software Qual J 31. 2023. P. 1415–1438. https://doi.org/10.1007/s11219-023-09646-0.
8. Peng Tang, Junfeng Wang, Mingxing Liu. Variational learning to rank for Test Case Prioritization via prioritizing metric inspired differentiable loss. Engineering Applications of Artificial Intelligence. Volume 141. 2025. https://doi.org/10.1016/j.engappai.2024.109776.
9. Tapas Kumar Choudhury, Mousumi Behera, Sanjit Kumar Dash, Subhendu Kumar Pani, Jibitesh Mishra, AnoLSTM-A Deep Learning Approach for Test Cases Prioritization, Procedia Computer Science, Volume 258.2025. Pages 1793-1803. https://doi.org/10.1016/j.procs.2025.04.431.
10. Md Arfan Uddin, Shakthi Weerasinghe, Darek Gajewski, Melika Akbarsharifi, Roxana Akbarsharifi, Christopher Stoner, Tomas Cerny, Sen He. Microservice logs analysis employing AI: A systematic literature review, Journal of Systems and Software. Volume 236.2026. P.1-93. https://doi.org/10.1016/j.jss.2026.112786.
11. S. Alhusain, "Predicting Relative Thresholds for Object Oriented Metrics," 2021 IEEE/ACM International Conference on Technical Debt (TechDebt), 2021, pp. 55-63, doi: 10.1109/TechDebt52882.2021.00015.
12. Robredo, M., Esposito, M., Taibi, D., Peñaloza, R., & Lenarduzzi, V. SQuaD: The Software Quality Dataset - Dataset [Data set]. Zenodo. 2025. https://doi.org/10.5281/zenodo.17566691.

Received (Надійшла) 26.01.2026
Accepted for publication (Прийнята до друку) 22.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Дяченко Дмитро Олександрович – аспірант кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Dmytro Diachenko – PhD student, Department of Electronic Computers, Kharkiv National University of Radio Electronics Kharkiv, Ukraine;
e-mail: dmytro.diachenko2@nure.ua ; ORCID Author ID: http://orcid.org/0009-0006-5751-3511.

Дяченко Владислав Олександрович – PhD, комп'ютерна інженерія, старший викладач кафедри електронних обчислювальних машин, Харківський національний університет радіоелектроніки, Харків, Україна;
Vladyslav Diachenko – PhD, Computer Engineering, Senior Lecturer Department of Electronic Computers, Kharkiv National University of Radio Electronics Kharkiv, Ukraine;
e-mail: vladyslav.diachenko@nure.ua; ORCID Author ID: http://orcid.org/ 0000-0003-2725-8784;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57207260441.

Модель оцінки ризику дефектів програмних компонентів розподілених комп'ютерних систем

Д. О. Дяченко, В. О. Дяченко

Анотація. Актуальність. Актуальність роботи зумовлена необхідністю підвищення ефективності автоматизованих випробувань програмного забезпечення розподілених комп'ютерних систем, для яких характерні складна структура, залежності між компонентами та підвищений ризик виникнення дефектів. Існуючі підходи до прогнозування дефектів здебільшого не забезпечують комплексного врахування структурних характеристик компонентів, інтенсивності їх змін і параметрів виконання тестів, що ускладнює обґрунтоване планування випробувань. Тому актуальною є розробка моделі оцінки ризику дефектів програмних компонентів розподілених комп'ютерних систем для підтримки процесу пріоритезації автоматизованих випробувань. Це дозволяє зосередити ресурси тестування на найбільш критичних програмних компонентах і тим самим підвищити загальну результативність контролю якості програмного забезпечення. **Мета статті**: розробка моделі оцінки ризику дефектів програм-них компонентів розподілених комп'ютерних систем для підвищення ефективності автоматизованих випробувань. **Об'єкт дослідження**: є процес виявлення та оцінювання дефектності програмних компонентів розподілених комп'ютерних систем у контексті автоматизованих випробувань. **Предмет дослідження:** моделі та методи оцінки ризику дефектів програмних компонентів розподілених комп'ютерних систем на основі аналізу їх структурних характеристик та результатів автоматизованих випробувань. **Результати дослідження**. У роботі розроблено модель оцінки ризику дефектів програмних компонентів розподілених комп'ютерних систем, що дозволяє формувати інтегральний показник дефектності на основі структурних характеристик програмного коду та результатів автоматизованих випробувань. Проведені експериментальні дослідження із застосуванням алгоритмів машинного навчання показали, що найкращі результати забезпечує модель CatBoost, яка продемонструвала найвищі значення ROC-AUC та Precision–Recall характеристик порівняно з іншими досліджуваними підходами. Отримані результати підтверджують можливість ефективного ранжування програмних компонентів за рівнем ризику дефектності та використання цієї інформації для пріоритизації автоматизованих випробувань у розподілених комп'ютерних системах.

Ключові слова: розподілена комп'ютерна система, прогнозування, машинне навчання, автоматизоване тестування, оцінка ризику дефектів, метрики, CatBoost, аналіз якості, глибоке навчання, багатокомпонентна система.

UDC 004.738.5:004.7: 621.39 **doi: 10.26906/SUNZ.2026.2.279**

Oleksandr Shefer, Stanislav Myhal

National University "Yuri Kondratyuk Poltava Polytechnic", Poltava, Ukraine

DEVELOPMENT OF AN ARCHITECTURAL DESIGN METHOD FOR MOBILE SDN IN ULTRA-DENSE SENSOR NETWORKS

Abstract. Background. The method of deploying SDN controllers on mobile nodes within ultra-dense networks is a vital component in enhancing the management efficiency of modern telecommunication systems. Given the rapid proliferation of connected devices and escalating traffic volumes, traditional centralized network management approaches often lack the necessary flexibility and scalability. **Subject of Research.** This study focuses on methods for constructing mobile SDN architectures tailored for ultra-dense sensor networks (UDSNs). **Objectives.** The aim of this research is to develop a method for designing a software-defined ultra-dense sensor network architecture, where specific SDN controllers are integrated into mobile nodes at the edge layer. **Results.** The study proposes an MEC (Multi-access Edge Computing) platform architecture specifically designed for deploying SDN controllers on mobile nodes within 5G/6G ultra-dense networks. A three-tier model for a mobile multi-controller SDN in UDSNs has been established. Furthermore, a specialized method for the placement and distribution of SDN controllers on mobile nodes within these ultra-dense environments was developed. **Conclusion.** Experimental results indicate that the implementation of the proposed method reduces transaction latency by 60% compared to traditional SDN-based edge computing models. Additionally, energy consumption is reduced by up to 30%.

Keywords: telecommunication network, edge layer, mobile device, controller, ultra-dense sensor network (UDSN), OpenFlow switch, 5G standard, SDN.

Introduction

The modern advancement of wireless technologies and the Internet of Things (IoT) has led to a rapid increase in the number of sensor devices operating within unified network infrastructures [1].

Ultra-dense sensor networks (UDSNs) are becoming a pivotal element in domains such as smart cities, the Industrial Internet of Things (IIoT), environmental monitoring, and autonomous transport systems [2, 3]. Simultaneously, the increasing network density complicates management, intensifies resource load, and presents new challenges regarding scalability and energy efficiency [4].

Traditional networking approaches fail to provide the necessary flexibility and adaptability required in such dynamic environments.

One of the most promising directions for addressing these issues is the implementation of the Software-Defined Networking (SDN) concept, which entails the decoupling of the control plane from the data plane [5, 6]. This approach enables centralized network management and enhances overall controllability. However, in the context of ultra-dense sensor networks, a strictly centralized SDN architecture may lead to controller overhead and increased latency.

Consequently, decentralized and hierarchical SDN architectures, as well as approaches involving mobile nodes, are gaining significant attention. Integrating mobility into the network's edge layer opens new possibilities for optimizing controller placement and improving data processing efficiency [7, 8]. Mobile SDN controllers are capable of adapting to changes in network topology and reducing latency by maintaining proximity to data sources.

Despite a substantial body of research in this field, the issues of constructing an effective mobile SDN architecture for ultra-dense sensor networks remain insufficiently explored. In particular, methods for optimal controller placement, load balancing, and ensuring network resilience require further analysis. Furthermore, it is crucial to account for the constrained resources of sensor nodes.

Thus, the development of a mobile SDN architecture for ultra-dense sensor networks represents a significant step toward improving the reliability and operational efficiency of modern software-defined networks.

1. Literature Review

Contemporary research focuses significantly on the evolution of the Software-Defined Networking (SDN) concept as a foundation for building flexible and scalable network infrastructures. In [9, 10], a fundamental analysis of SDN is presented, defining key principles such as the decoupling of the control and data planes. This separation facilitates centralized network management and simplifies configuration; however, the authors also identify scalability and reliability issues inherent in using centralized controllers within large-scale networks.

Subsequent studies have pivoted toward the application of SDN in wireless sensor networks (WSNs). Research in [11] provides an overview of employing SDN to enhance management efficiency in sensor networks, particularly regarding routing optimization and reduced energy consumption. The authors emphasize that traditional management approaches fail to meet the rigorous requirements of ultra-dense environments.

The architectural intricacies of integrating SDN into wireless sensor networks are explored in [12]. While the proposed solutions improve network flexibility and ensure effective interaction between nodes and controllers, the challenge of optimal resource allocation remains unresolved.

The study in [13] is dedicated to developing an SDN architecture for 6LoWPAN networks, enabling the integration of sensor devices into IP-oriented environments. This represents a significant step toward building scalable IoT systems; nevertheless, the problem

of adapting to topological changes necessitates further investigation.

In [14], the problem of SDN controller placement in 5G-IoT networks is addressed. The proposed approach aims to minimize data transmission latency and ensure load balancing, both of which are critical for ultra-dense sensor networks.

The use of SDN for constructing heterogeneous networks, including nano-networks, is proposed in [15], demonstrating the potential for scaling and integrating diverse devices. This confirms the promise of SDN as a universal technology for managing complex network systems.

Research in [16] examines the integration of intelligent mechanisms into next-generation networks, specifically 5G and IoT. The author underscores the importance of combining SDN with artificial intelligence (AI) technologies to enhance network adaptability and autonomy.

The analysis of current research [9–16] indicates active progress in developing SDN-based approaches for sensor and IoT networks. However, existing solutions are primarily oriented toward static or semi-dynamic environments and do not sufficiently account for node mobility. This gap necessitates the development of new methods for constructing mobile SDN architectures for ultra-dense sensor networks, which constitutes the subject of this study.

Research Objective: To develop and evaluate a method for constructing a software-defined ultra-dense sensor network architecture, wherein specific SDN controllers are deployed on mobile nodes at the edge layer.

2. Mobile SDN Structure

The evolution of SDN technology has progressed from a centralized scheme with a single SDN controller toward a distributed framework utilizing multiple controllers. The architecture of such networks has been established based on distributed network management through edge computing nodes [17]. In these networks, SDN controller functions are distributed across edge computing nodes to implement a centralized management scheme [18]. Each virtual controller interfaces with a centralized management framework that supports one or more SDN controllers, depending on the network scale.

Modifications to SDN technology are aimed at providing SDN controller mobility to support high-density and ultra-dense network scenarios. A variant of the network model featuring a mobile SDN controller is depicted in Fig. 1. In this context, the SDN network can be categorized into three primary levels:

– Centralized SDN controllers;
– Distributed stationary SDN controllers;
– Distributed mobile SDN controllers.

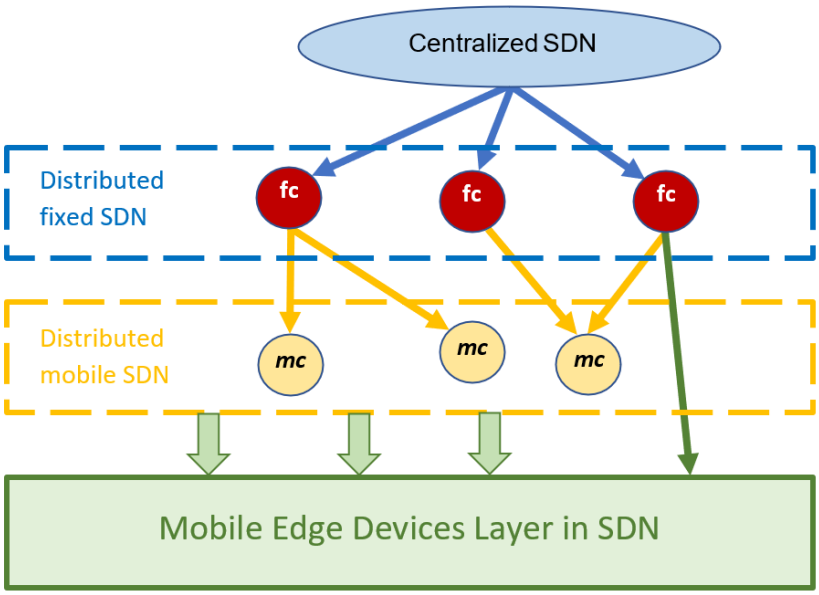


Fig. 1. Network model with a mobile SDN controller

Centralized SDN Controllers. Several centralized SDN controllers are utilized to construct the SDN network, providing overarching management. This management method defines the primary level of network control and maintains the necessary interfaces with network operators.

Previously proposed algorithms are employed to address the challenges of controller placement and distribution within the network. In this case, the network architecture consists of the following six core components.

The Control Plane is a set of network applications that govern the logic of the SDN network. Software tools are utilized to ensure flexibility and ease of deployment for new applications and services, such as routing, load balancing, policy enforcement, or user applications. They also facilitate the orchestration and automation of network operations through existing Application Programming Interfaces (APIs).

Controllers represent the most intelligent and critical layer of the SDN architecture, containing one or more controllers that transmit various types of rules and

policies to the infrastructure layer via the Southbound (SB) interface.

The Data Plane (Infrastructure Layer) consists of a set of data forwarding devices within the network (routers, switches, load balancers, etc.). It utilizes Southbound APIs (SB APIs) to interact with the control plane, receiving data forwarding rules and policies to be applied to the respective devices.

Northbound (NB) Interfaces provide communication between the control and management layers, typically comprising a set of open-source APIs.

East-West Interfaces enable communication between multiple controllers. They utilize messaging systems or distributed routing protocols, such as BGP (Border Gateway Protocol) and OSPF (Open Shortest Path First). These interfaces are also used for communication between centralized controllers and other network devices.

Southbound Interfaces facilitate interaction between the control plane and the data plane; these are protocols that allow the controller to push policies to the forwarding plane. The OpenFlow protocol is the most widely recognized and prevalent protocol for SDN-enabled networks. However, OpenFlow is only one of many SDN protocols; others include OpFlex, which redistributes certain network management tasks to the infrastructure layer to improve scalability, and ForCES, which offers a flexible method for enhancing the management of traditional networks without a logically centralized controller. Additionally, the ROFL (Revised OpenFlow Library) provides an API for software developers to more efficiently create new applications.

Stationary Distributed SDN Controllers. The layer of stationary distributed SDN controllers includes an integrated SDN controller at the edge of the Radio Access Network (RAN). The RAN consists of distributed heterogeneous cellular cells with eNB base stations.

Each eNB is connected to a Multi-Access Edge Computing (MEC) server. In this configuration, the MEC represents a multi-tier structure, as the utilized MEC servers are heterogeneous in terms of computing capabilities and feature a hierarchical structure of three primary cloud tiers interconnected by high-speed fiber-optic links [19].

A new architecture for the existing MEC-based RAN in 5G/6G networks is developed, as shown in Fig. 2, employing a multi-level interaction system of edge computing systems, as illustrated in Fig. 1

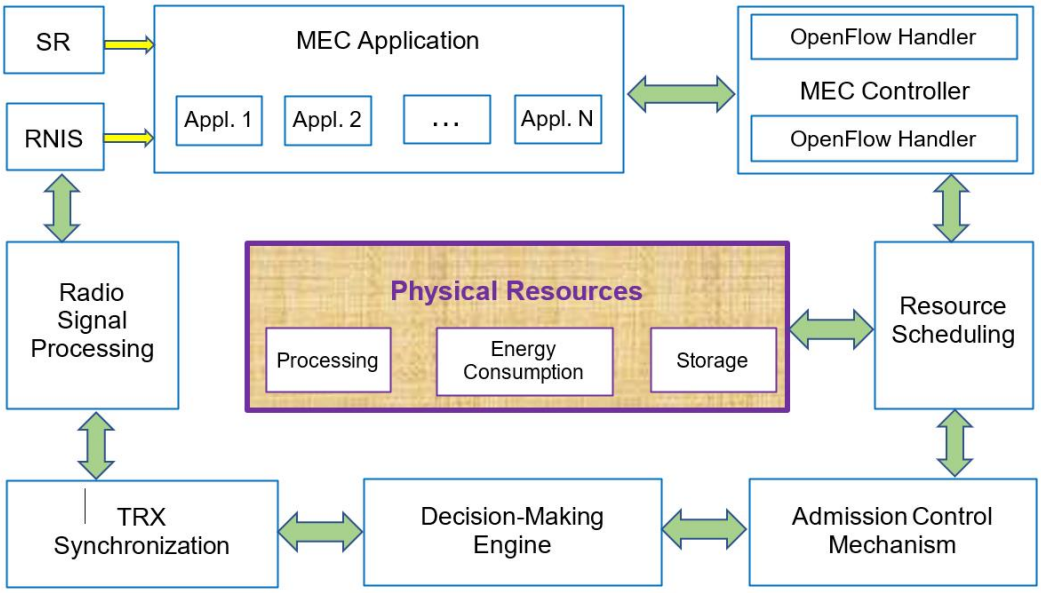


Fig. 2. MEC platform architecture for the SDN controller placement method on mobile nodes in 5G/6G ultra-dense networks

The first core component of the MEC platform is the hardware or physical resources, comprising the CPU, storage units, and hardware tools used for power delivery and process monitoring. This component varies across different cloud tiers. Hardware resources are either utilized or shared between local and offloaded computing tasks.

To ensure the efficient use of edge computing resources, a resource scheduler is implemented to manage and allocate resources among various computational tasks.

Computational tasks are offloaded to the MEC server, and the system must decide, based on available resources and predefined Quality of Service (QoS) criteria, whether to process the offloaded task. Consequently, the MEC platform must include a decision-making mechanism that responds to offloading requests by either accepting or rejecting application offloading tasks. The decision-making mechanism is the part of the MEC platform that implements application offloading algorithms. To make an offloading decision, this mechanism requires data on the currently available server resources, which is provided via the admission control system.

Another component of the platform is the synchronizer/TRX (Transceiver), a hardware-software system responsible for connecting the MEC server with other devices, such as eNBs. Thus, the synchronizer/TRX

facilitates the transmission and reception of data to and from the MEC platform.

Each MEC server features a controller that hosts a Virtual Machine (VM) implementing the SDN network management method used in the core network. This MEC controller facilitates integration and interaction with the core network. Furthermore, the MEC controller manages and monitors server operations to ensure high availability, flexibility, and reduced latency. The MEC controller receives updates from the core SDN controller via the OpenFlow protocol and updates the relevant components of the MEC platform accordingly. The communication between MEC servers and the SDN controller via OpenFlow ensures efficient and accelerated interaction between the SDN controller and the RAN.

Additionally, MEC applications provide network operators with new avenues for deploying MEC-based services.

MEC-based stationary distributed SDN controllers interface with both centralized and mobile SDN controllers.

Mobile Distributed SDN Controllers. These controllers are typically lightweight SDN controllers deployed on moving objects. There are numerous use cases for mobile SDN controllers; for instance, they can be installed on vehicles to support interaction within ultra-dense networks and ensure ultra-high system availability. A mobile SDN controller possesses one or more interfaces with stationary distributed SDN controllers. This integration fulfills the connectivity and availability requirements of 6G networks. A lightweight MEC platform architecture (Fig. 2) is utilized for the implementation of mobile controllers.

Furthermore, mobile controllers provide localized decision-making, which significantly reduces latency in processing network events.

This is particularly critical for scenarios with stringent response time requirements, such as autonomous vehicles or the Tactile Internet. The use of mobile SDN controllers also helps alleviate the load on central control nodes by distributing network management functions. Moreover, these controllers can dynamically adapt to changes in network topology caused by the movement of users or infrastructure nodes.

A key advantage is the ability to integrate with MEC resource orchestration mechanisms, ensuring the efficient utilization of computational and network resources at the periphery.

Mobile controllers can also perform data pre-processing and traffic filtering, reducing the volume of information transmitted to central data centers. In terms of energy efficiency, they allow for optimized power consumption through local load management.

However, the deployment of mobile distributed SDN controllers introduces several challenges, such as maintaining network state consistency among controllers and ensuring reliable synchronization. Security is equally paramount, as mobile nodes may be more susceptible to attacks. Consequently, research in this field for ultra-dense networks must focus on developing effective coordination mechanisms, security protocols, and adaptive management for mobile SDN controllers in next-generation networks.

3. Model of a Mobile Multi-Controller SDN for Ultra-Dense Sensor Networks

To analyze the proposed system, we construct a mobile multi-controller SDN model using graph theory. In SDN environments, in-band communication is typically employed as the primary signaling method.

Formalization of Network Connections. The controller network topology is represented as an undirected graph $G = (V, E)$, where V and E denote the sets of nodes and the links between them, respectively.

The set of SDN controllers deployed in the core network is denoted as N_C and defined as follows:

$$N_C = \{C_1, C_2, \ldots, C_n, \ldots, C_N\} \quad \forall n \in \overline{1, N}, \qquad (1)$$

where N is the total number of core SDN controllers.

The set of controllers deployed at the network edge is denoted as E_C and defined as:

$$E_C = \{C_1, C_2, \ldots, C_k, \ldots, C_K\} \quad \forall k \in \overline{1, K}, \qquad (2)$$

where K is the total number of SDN controllers deployed at the RAN edge.

Each group of edge controllers interfaces with a core network controller. The set of edge controllers associated with a core controller C_n, is defined as:

$$E_C^{(n)} = \left\{C_1^{(n)}, C_2^{(n)}, \ldots, C_{kn}^{(n)}, \ldots C_{Kn}^{(n)}\right\},$$
$$Kn < K,\ E_C^{(n)} \subset E_C,\ \forall kn \in \overline{1, Kn}, \qquad (3)$$
$$\sum_{n=1}^{N} Kn = K,$$

where Kn is the total number of edge controllers interfacing with the core controller C_n.

The set of deployed mobile controllers is defined as:

$$M_C = \left\{C_1^{(m)}, C_2^{(m)}, \ldots, C_\ell^{(m)}, \ldots, C_L^{(m)}\right\} \quad \forall \ell \in \overline{1, L}, \qquad (4)$$

where L is the total number of deployed mobile SDN controllers.

Each group of mobile controllers interfaces with an edge controller.

The set of mobile controllers associated with an edge controller C_k, is defined as:

$$M_C^{(k)} = \left\{C_1^{(m,k)}, C_2^{(m,k)}, \ldots, C_{\ell k}^{(m,k)}, \ldots, C_{Lk}^{(m,k)}\right\},$$
$$Lk < L,\ M_C^{(k)} \subset M_C,\ \forall \ell k \in \overline{1, Lk}, \qquad (5)$$
$$\sum_{k=1}^{K} Lk = L,$$

where Lk is the total number of mobile controllers interfacing with the edge controller C_k.

In the data plane, each switch distributed among the controllers is connected to an SDN controller. These connections are determined by a specialized controller placement algorithm. The set of deployed switches is defined as:

$$S = \left\{ S_1, S_2, \ldots S_r, \ldots, S_R \right\} \quad \forall r \in \overline{1,R}, \tag{6}$$

To evaluate controller performance, the response time is used as a metric, which is significantly influenced by queuing delays. Controllers are modeled using an M/M/s multi-server queuing model, where s is the number of serving devices. The average response time T_i of controller C_i is calculated as the sum of the queuing wait time and the processing time. This can be derived using the Erlang-C formula as a function of the request arrival rate λ_i and the service rate μ:

$$T_i(\lambda) = C\left(s, \lambda_i/\mu\right) / \left(s \cdot \mu_i - \lambda_i\right) + \mu^{-1}, \tag{7}$$

where $C\left(s, \lambda_i/\mu\right)$ represents the probability that all servers in the system are busy and any incoming packet will be queued. This probability is calculated as:

$$C\left(s, \frac{\lambda_i}{\mu}\right) = \frac{\left(\frac{(s\rho)^s}{s!}\right) \cdot \left(\frac{1}{1-\rho}\right)}{\sum_{i=0}^{s-1} \frac{(s\rho)^i}{i!} + \left(\frac{(s\rho)^s}{s!}\right) \cdot \left(\frac{1}{1-\rho}\right)} = \tag{8}$$

$$= \frac{1}{1 + \left(\frac{1}{1-\rho}\right) \cdot \left(\frac{s!}{(s\rho)^s}\right) \cdot \sum_{i=0}^{s-1} \frac{(s\rho)^i}{i!}},$$

where $\rho = \frac{\lambda_i}{s \cdot \mu}$ is the server utilization factor, serving as an indicator of system stability.

The arrival rate λ_i for the controller is calculated as the sum of the average request rates from the switches connected to that controller.

Interaction Costs Between Switches and Controllers. Regarding the interaction between switches and controllers, when a flow table (e.g., a new flow entry) must be established upon request, the switch sends packets to the controller. The controller then calculates the data path and installs the corresponding flow label on the switch. Subsequently, the switch forwards packets according to the flow table. In this process, the total packet latency for controller C_n, consists of the time required to deliver the packet information to the controller and the controller's subsequent response to the switches. The interaction cost between switch S_r and controller C_n in an OpenFlow network is defined as:

$$C_{C-S} = \\ = 2 \cdot \sum_{n=1}^{N} \sum_{r=1}^{R} p_{\xi - S_r} \left(\frac{\lambda_{S_r}^{(t)}}{p_\xi} \cdot d_{C_n - S_r} \cdot B_{C_n - S_r} \right), \tag{9}$$

where C_{C-S} – the total communication cost between switches and controllers;

$p_{\xi - S_r}$ – the average polling rate of switch S;

$\lambda_{S_r}^{(t)}$ – the request intensity from switch S_r over time interval t;

p_ξ – the specific polling rate of the switch;

$d_{C_n - S_r}$ – the distance between controller C_n and switch S_r;

$B_{C_n - S_r}$ – a boolean variable representing the decision of switch S_r to connect to controller C_n, obtained via a Salp Swarm Evolutionary Algorithm.

The interaction cost is determined by the total latency..

To determine the interaction cost between controllers, it is noted that in a multi-controller environment, synchronization of information transfer must be implemented. This ensures that each controller maintains a global view of the network state in near-real-time. The state synchronization cost primarily pertains to the interaction overhead determined by the exchange of state information between controllers across the three aforementioned deployment tiers. This latency is defined as the interaction and synchronization delay between entities within the control plane. The synchronization cost is calculated as follows:

$$C_{C-C} = \alpha 1 \cdot \sum_{n1=1}^{N} \sum_{n2=1}^{N} d_{n1,n2}^{(1)} + \\ + \alpha 2 \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} d_{n,k}^{(2)} + \alpha 3 \cdot \sum_{k=1}^{K} \sum_{\ell=1}^{L} d_{k,\ell}^{(3)}, \tag{10}$$

where α1 – the average rate of state information transfer between controllers in the core network;

α2 – the average rate of state information transfer for edge controllers;

α3 – the average rate of state information transfer for mobile controllers;

$d_{n1,n2}^{(1)}$ – the distance between controllers C_{n1} and C_{n2}, located in the core network;

$d_{n,k}^{(2)}$ – – the distance between core controller C_n and edge controller C_k;

$d_{k,\ell}^{(3)}$ – the distance between edge controller C_k and mobile controller C_l.

4 Method for SDN Controller Placement on Mobile Nodes in Ultra-Dense Networks

The proposed method focuses on the adaptive placement of SDN controllers, accounting for the current network state. It is based on the analysis of key Quality of Service (QoS) metrics, specifically latency, throughput, load levels, and link reliability. Furthermore, the method incorporates node mobility, which enables a reduction in the number of network reconfigurations and enhances operational stability. A critical aspect involves minimizing controller migration costs between nodes and reducing the overhead associated with the exchange of control information.

Particular attention is paid to the cost-related aspects of network construction, specifically optimizing the deployment and operational expenditures of the proposed three-tier hierarchical SDN controller structure. This structure allows for effective load balancing among

controllers and minimizes latency. Cost optimization is achieved through the rational selection of the number of controllers at each tier, their spatial distribution, and resource provisioning. In doing so, expenses for computational resources, energy consumption, communication channels, and infrastructure maintenance are considered.

Thus, the proposed approach ensures not only improved efficiency and resilience of ultra-dense networks but also the economic feasibility of their implementation.

The total interaction cost in the network represents the sum of the interaction costs between switches and controllers, as well as the interaction costs between controllers of different tiers at a given point in time. For a specific network topology, the interaction cost between switches and controllers decreases as the number of controllers increases; however, this simultaneously increases the interaction cost among the controllers themselves.

The ultimate goal of the optimization task is to partition the network into management clusters and assign switches to each cluster in a manner that minimizes the overall latency. Consequently, to achieve this objective, the following optimization problem is formulated:

1. Objective Function: On the set of all possible distributions $\wp = \{©\}$ find the minimum of the cost function based on the proposed interaction cost expressions (9) and (10):

$$C = \beta 1 \cdot C_{C-C} + \beta 2 \cdot C_{C-S} \xrightarrow[\gamma \in \Gamma]{} \min, \qquad (11)$$

where β2 – the weighting coefficient for the switch-to-controller interaction cost;

β1 – the weighting coefficient for the inter-controller interaction cost.

A normalization rule is applied to these coefficients, such that β1 + β2 = 1, or β2 = 1 – β1.

2. . Optimization Problem Constraints:

$$B_{C_n - S_r} \in \{0; 1\}; \qquad (12)$$

$$T_{end}^{(n-k-\ell)} - T_{begin}^{(n-k-\ell)} \le \tau_\varphi \quad \forall \left(C_\ell \to C_k \to C_n\right); \qquad (13)$$

$$U_{min}^{(1)} \le U_n^{(1)} \le U_{max}^{(1)} \quad \forall C_n \in N_C; \qquad (14)$$

$$U_{min}^{(2)} \le U_k^{(2)} \le U_{max}^{(2)} \quad \forall C_k \in E_C; \qquad (15)$$

$$U_{min}^{(3)} \le U_\ell^{(3)} \le U_{max}^{(3)} \quad \forall C_\ell^{(m)} \in M_C, \qquad (16)$$

where $T_{begin}^{(n-k-\ell)}$ and $T_{end}^{(n-k-\ell)}$ are the start and end times, respectively, for the transmission of transaction φ between SDN controllers along the route from $C_\ell^{(m)}$ to C_n via C_k;

τ_φ – is the maximum allowable transmission time for transaction φ, that ensures the required Quality of Service (QoS) requirements;

$U_n^{(1)}$, $U_k^{(2)}$ and $U_\ell^{(3)}$ are the utilization indices of SDN controllers corresponding to the core, edge, and mobile tiers, respectively;

$U_{min}^{(1)}$, $U_{min}^{(2)}$ та $U_{min}^{(3)}$ – are the minimum allowable utilization indices of SDN controllers at the respective tiers;

$U_{max}^{(1)}$, $U_{max}^{(2)}$ and $U_{max}^{(3)}$ are the maximum allowable utilization indices of SDN controllers at the respective tiers.

Constraint (13) specifies that the average response time of the core, edge, and mobile controllers must not exceed a predefined threshold value established for the corresponding transaction. This condition applies to all controllers within the sets of available core network controllers, edge controllers, and mobile SDN controllers. The value of τ_φ is determined to ensure adherence to specific Quality of Service (QoS) requirements. Constraints (14) – (16) pertain to the utilization index of each controller in the network, which must remain within the lower and upper utilization bounds.

These utilization limits are defined to maintain the system's overall QoS. The utilization index of each controller serves as a metric correlated with energy consumption, storage capacity, and data processing overhead. Furthermore, distinct utilization bounds were considered for controllers at different tiers, reflecting the varying operational capabilities inherent to each architectural level.

4. Performance Evaluation of the SDN Controller Placement Method in Ultra-Dense Networks

To evaluate the performance of the developed SDN controller placement method for mobile nodes in ultra-dense networks, a simulation study was conducted. The NS-3 network simulator integrated with the Cloudsim framework was utilized as the simulation environment.

The evaluation accounted for the workloads of four heterogeneous application types:

Category I (APPs(I)): Applications comprising simple task workloads, such as basic web page processing.

Category II (APPs(II)): Image-based applications involving low-complexity image processing tasks.

Category III (APPs(III)): Basic video applications involving moderate workloads typical of simple video processing.

Category IV (APPs(IV)): Applications requiring the processing of complex data, such as 360-degree imagery and high-definition video..

As the classification progresses from Category I to Category IV, the tasks demand increasingly substantial resources, thereby reducing the probability of successful local execution. Consequently, higher-category tasks require significantly more energy and computational power than lower-category tasks.

To demonstrate the efficiency of the developed method, five system architectures were compared:

System I: A traditional multi-controller SDN employing multiple controllers in the core network without additional management schemes;

System II: A fog-computing-based SDN featuring a centralized topology with distributed edge SDN controllers integrated into fog nodes.

System III: An MEC-SDN-based network with a centralized topology where distributed edge SDN controllers are integrated into Multi-access Edge Computing (MEC) nodes.

System IV: A hybrid Fog-MEC SDN architecture utilizing distributed edge SDN controllers integrated into both fog and MEC nodes.

System V: A **Mobile SDN (MSDN)** network featuring mobile SDN controllers deployed on mobile nodes (e.g., public transit buses) supported by Fog-MEC computing units. This represents the proposed MSDN architecture.

These systems were evaluated to highlight the effectiveness of each SDN deployment method relative to the proposed approach. The primary performance metrics included energy consumption, latency, and availability..

Figure 3 illustrates the results for the average processing latency of computational tasks for the three MEC-oriented SDN configurations (Systems III through V). The simulation specifically considered tasks of low complexity, approximately corresponding to Category II applications.

In each scenario, a varying number of computational tasks were assigned to the mobile SDN end-nodes.

The average processing latency was analyzed as a function of the average task count. As the number of assigned tasks increased, the average latency rose across all three system types, primarily due to the overhead associated with managing a larger task volume. However, the proposed method outperformed all other configurations, achieving higher efficiency, which was particularly evident at higher task densities.

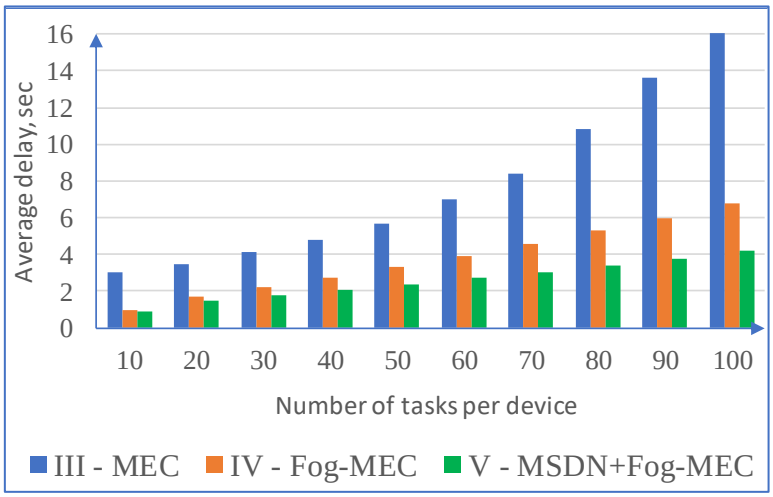


Fig. 3. Average latency as a function of the number of available low-complexity tasks

Figure 4 presents the relationship between average processing latency and task complexity. The average latency was calculated across the four previously defined application categories. During this simulation, an average of 30 computational tasks were assigned to the mobile SDN end-nodes.

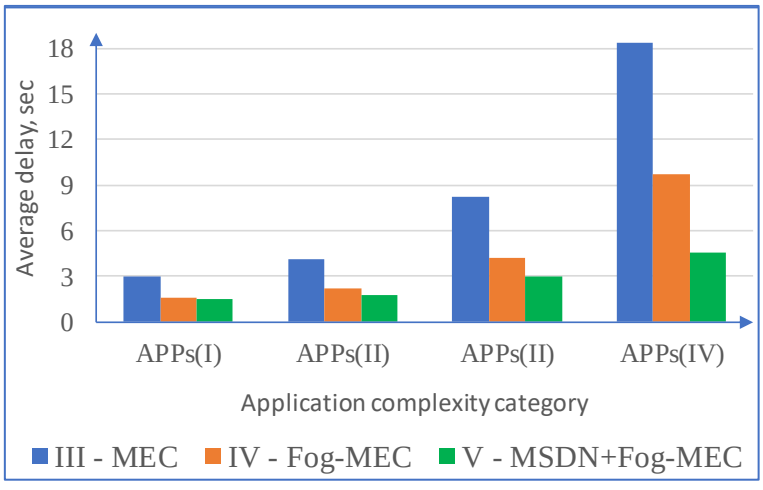


Fig. 4. Average latency across different application complexity categories

The simulation results indicate that the deployment of mobile SDN controllers significantly reduces the average time required to process computational tasks across various applications, with the most pronounced improvements observed for complex tasks with high workloads.

Conclusions

This paper presents a method for deploying SDN controllers on mobile nodes within ultra-dense networks, representing a vital advancement in enhancing management efficiency for modern telecommunication systems. As part of the proposed methodology, a specialized MEC platform architecture was designed for SDN controller placement in 5G/6G ultra-dense environments, and a three-tier mobile multi-controller SDN model for ultra-dense sensor networks was developed.

The implementation of the proposed method demonstrates a significant performance advantage, reducing transaction transmission latency by **60%** compared to traditional SDN-based edge computing models.

Furthermore, it achieves a reduction in energy consumption of up to 30%.

Future research will focus on developing an advanced load-balancing method among the controllers within the three-tier SDN architecture to further optimize resource distribution.

Conflict of interest

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

Using artificial intelligence tools

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

СПИСОК ЛІТЕРАТУРИ

1. Narwaria, A. and Mazumdar, A. P. (2023), “Software-Defined Wireless Sensor Network: A Comprehensive Survey”, *Journal of Network and Computer Applications*, vol. 215, doi: https://doi.org/10.1016/j.jnca.2023.103636
2. Kuchuk, H., Chumachenko, I., Marchenko, N., Kuchuk, N. and Lysytsia, D. (2025), “Method for calculating the number of IOT sensors in environmental monitoring systems”, *Advanced Information Systems*, vol. 9, no. 3, pp. 66–72, doi: https://doi.org/10.20998/2522-9052.2025.3.08
3. Khan, M.A., Rehmani, M.H. and Rachedi, A. (2022), “SDN-based gateway architecture for electromagnetic nano-networks”, *Computer Communications*, vol. 184, pp. 160–173, doi: https://doi.org/10.1016/j.comcom.2021.12.017
4. Kuchuk, H., Kalinin, Y., Dotsenko, N., Chumachenko, I. and Pakhomov, Y. (2024), “Decomposition of integrated high-density IoT data flow”, *Advanced Information Systems*, vol. 8, no. 3, pp. 77–84, doi: https://doi.org/10.20998/2522-9052.2024.3.09
5. Kuchuk, H., Husieva, Y., Novoselov, S., Lysytsia, D., Krykhovetskyi, H. (2025), “Load Balancing of the layers Iot Fog-Cloud support network”, *Advanced Information Systems*, vol. 9, no. 1, pp. 91–98, doi: https://doi.org/10.20998/2522-9052.2025.1.11
6. Cui, X., Gao, X., Ma, Y. and Wang, W. (2020), “A controller deployment scheme in 5G-IoT network based on SDN”, *EURASIP Journal on Wireless Communications and Networking*, article number 248 (2020), doi: https://doi.org/10.1186/s13638-020-01853-8
7. Kuchuk, H., Mozhaiev, O., Tiulieniev, S., Mozhaiev, M., Kuchuk, N., Lubentsov, A., Onishchenko, Yu., Gnusov, Yu., Brendel, O. and Roh, V. (2025), “Devising a method for energy-efficient control over a data transmission process across the mobile high-density Internet of Things”, *Eastern European Journal of Enterprise Technologies*, vol. 4(4(136)), pp. 46–57, doi: https://doi.org/10.15587/1729-4061.2025.336111
8. Panchenko, V., Kuchuk, H., Noskov, V., Leonov, S., Lipchanska, O. (2026), “Method of test pool synthesis for an intelligent high-density IoT edge-layer gateway”, *Advanced Information Systems*, vol. 10, no. 1, pp. 50–57, doi: https://doi.org/10.20998/2522-9052.2026.1.06
9. Kreutz, D., Ramos, F.M.V., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S. and Uhlig, S. (2015), “Software-Defined Networking: A Comprehensive Survey”, *Proceedings of the IEEE*, vol. 103, is. 1, January 2015, pp. 14–76, doi: https://doi.org/10.1109/JPROC.2014.2371999
10. Hu, F., Hao, Q., and Bao., K. (2014), “A Survey of Software-Defined Network and OpenFlow: From Concept to Implementation: From Concept to Implementation”, *IEEE Communications Surveys & Tutorials*, pp. 2181–2206, doi: https://doi.org/10.1109/COMST.2014.2326417
11. Ndiaye, M., Hancke, G. P., and Abu-Mahfouz, A. M. (2017), “Software Defined Networking for Improved Wireless Sensor Network Management: A Survey”, *Sensors*, vol. 17(5), doi: https://doi.org/10.3390/s17051031
12. Puente Fernández, J. A., García Villalba, L. J., and Kim, T.-H. (2018), “Software Defined Networks in Wireless Sensor Architectures”, *Entropy*, vol. 20(4), doi: https://doi.org/10.3390/e20040225
13. Miguel, M. L. F., Jamhour, E., Pellenz, M. E., and Penna, M. C. (2018), “SDN Architecture for 6LoWPAN Wireless Sensor Networks”, *Sensors*, vol. *18*(11), doi: https://doi.org/10.3390/s18113738
14. Liao, Y., Wang, J. and Liu, J., 2020, A controller deployment scheme in 5G-IoT network based on SDN, EURASIP Journal on Wireless Communications and Networking, DOI: https://doi.org/10.1186/s13638-020-01853-8
15. Khan, M.A., Rehmani, M.H. and Rachedi, A., 2022, SDN-based gateway architecture for electromagnetic nano-networks, Computer Communications, DOI: https://doi.org/10.1016/j.comcom.2021.12.017
16. Al-Turjman, F., 2019, Intelligence and security in big 5G-oriented IoNT: An overview, Future Generation Computer Systems, DOI: https://doi.org/10.1016/j.future.2018.08.028

17. Kuchuk, N., Kashkevich, S., Radchenko, V., Andrusenko, Y. and Kuchuk, H. (2024), “Applying edge computing in the execution IoT operative transactions”, *Advanced Information Systems*, vol. 8, no. 4, pp. 49–59, doi: https://doi.org/10.20998/2522-9052.2024.4.07
18. Kuchuk, H. and Malokhvii, E. (2024), “Integration of IoT with Cloud, Fog, and Edge Computing: A Review”, *Advanced Information Systems*, vol. *8*(2), pp. 65–78, doi: https://doi.org/10.20998/2522-9052.2024.2.08
19. Miguel, M. L. F., Jamhour, E., Pellenz, M. E., and Penna, M. C. (2018), “SDN Architecture for 6LoWPAN Wireless Sensor Networks”, *Sensors*, vol. 18(11), DOI: https://doi.org/10.3390/s18113738

Received (Надійшла) 12.01.2026
Accepted for publication (Прийнята до друку) 15.04.2026
Publication date (Дата публікації) 22.05.2026

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Шефер Олександр Віталійович – доктор технічних наук, професор, завідувач кафедри автоматики, електроніки та телекомунікацій, Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Oleksandr Shefer – Doctor of Technical Sciences, Professor, Head of the Department of Automation, Electronics and Telecommunications, National University “Yuri Kondratyuk Poltava Polytechnic”, Poltava, Ukraine;
e-mail: itm.ovshefer@nupp.edu.ua; ORCID Author ID: https://orcid.org/0000-0002-3415-349X;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57210203269.

Мигаль Станіслав Вікторович – аспірант кафедри автоматики, електроніки та телекомунікацій, Національний університет «Полтавська політехніка імені Юрія Кондратюка», Полтава, Україна;
Stanislav Myhal – PhD student of the Department of Automation, Electronics and Telecommunications, National University “Yuri Kondratyuk Poltava Polytechnic”, Poltava, Ukraine;
e-mail: stas.migal1998@gmail.com; ORCID Author ID: https://orcid.org/0009-0007-4675-7358.

Розробка методу побудови архітектури мобільної SDN для надщільних сенсорних мереж

О. В. Шефер, С. В. Мигаль

Анотація. Актуальність. Метод розміщення SDN-контролерів на мобільних вузлах надщільних мереж є важливою складовою підвищення ефективності управління сучасними телекомунікаційними системами. У контексті стрімкого зростання кількості підключених пристроїв і обсягів трафіку традиційні підходи до централізованого управління мережами виявляються недостатньо гнучкими та масштабованими. **Предметом** дослідження є методи побудови архітектури мобільної SDN для надщільних сенсорних мереж. **Метою** дослідження є розробка метод побудови архітектури програмно-визначеної надщільної сенсорної мережі, у якій деякі SDN-контролери встановлюються на мобільних вузлах граничного шару. **Отримані наступні результати**. Запропонована архітектура платформи MEC для методу розміщення SDN-контролерів на мобільних вузлах надщільних мереж 5G/6G. Розроблена трирівнева модель мобільної мультиконтролерної SDN для надщільної сенсорної мережі. Також розроблений метод розміщення SDN контролерів на мобільних вузлах надщільних мереж**. Висновок**. Використання запропонованого методу дозволяє зменшити затримку передачі транзакцій на 60% проти традиційних моделей граничних обчислень з урахуванням SDN, і навіть знизити споживану енергію на 72%.

Ключові слова: телекомунікаційна мережа, граничний шадозволяєр, мобільний пристрій, контролер, надщільна сенсорна мережа, OpenFlow-комутатор, стандарт 5G, мережа SDN.

АЛФАВІТНИЙ ПОКАЖЧИК

Наукове видання

СИСТЕМИ УПРАВЛІННЯ, НАВІГАЦІЇ ТА ЗВ'ЯЗКУ

Збірник наукових праць

Випуск 2 (84)

Відповідальний за випуск *О. В. Шефер*

Ідентифікатор медіа R30-04135 згідно з рішенням Національної ради України з питань телебачення і радіомовлення від 25.04.2024 № 1416

Підписано до друку 12.05.2026. Формат 60×84/8. Ум.-друк. арк. 36,0. Тираж 120 прим. Зам. 512-26
Адреса редакції: Україна, 36011, м. Полтава, проспект Віталія Грицаєнка, 24, тел. (050) 302-20-71
Національний університет «Полтавська політехніка імені Юрія Кондратюка»

Віддруковано з готових оригінал-макетів у цифровій друкарні Impress
61002, м. Харків, вул. Пушкінська, 56, тел. + **38 (057) 714-52-11**
e-mail: **irina@impress.biz.ua**