ISSN 2663-3051
(ISSN 0555-2656 до 2019 р.)

БІОНІКА ІНТЕЛЕКТУ

ІНФОРМАЦІЯ, МОВА, ІНТЕЛЕКТ

№ 2 (103) **2025**

НАУКОВО-ТЕХНІЧНИЙ ЖУРНАЛ

Заснований у жовтні 1967 р.

Засновник та видавець
Харківський національний університет радіоелектроніки

Періодичність видання – *2 рази на рік*

Харків • ХНУРЕ • 2025

Науково-технічний журнал

«БІОНІКА ІНТЕЛЕКТУ»

ISSN 2663-3051

Заснований Харківським національним університетом
радіоелектроніки у 1967 році

Реферування та індексування:

Google Scholar

BASE

INDEX COPERNICUS
INTERNATIONAL

Національна бібліотека України
імені В. І. Вернадського

DOI prefix: **10.30837**
Crossref

Журнал включено до списку наукових спеціалізованих видань України
з технічних та фізико-математичних наук
згідно з наказом Міністерства освіти і науки України № 820 від 11.07.2016
(внесено зміни згідно з наказом МОНУ № 920 від 26.06.2024)

УДК 004.75

DOi 10.30837/ bi.2025.2(103).01

Є. В. Бодянський[1], Д. В. Савенков[2]

[1]ХНУРЕ, м. Харків, Україна, yevgeniy.bodyanskiy@nure.ua, ORCID iD: 0000-0001-5418-2143

[2] ХНУРЕ, м. Харків, Україна, denys.savenkov@nure.ua, ORCID iD: 0009-0003-7361-015X

ВПЛИВ ПАРАМЕТРІВ ОПТИМІЗАЦІЇ ІНФЕРЕНЦІЇ НА ЕФЕКТИВНІСТЬ СПАЙКОВИХ НЕЙРОННИХ МЕРЕЖ

Спайкові нейронні мережі (SNN) – це третє покоління штучних нейромереж, яке завдяки своїй енергоефективності та розрідженості ідеально підходить для застосування у ресурсо-обмежених середовищах, як, наприклад, IoT або робототехніка. Однак і вони можуть не зустрічати екстремальних вимог, що призводить до необхідності використання методів оптимізації інференції, зокрема квантизації та прунінг. Сучасні дослідження вже розглядали практичне застосування даних методів для спайкових нейромереж, але вони не зосереджувались на впливі початкових параметрів оптимізації на продуктивність стисненої моделі. Мета цього дослідження полягає у систематизація та емпіричне дослідження впливу параметрів методів квантизації та прунінгу на кінцеву продуктивність спайкових нейронних мереж. Для експериментів було використано архітектуру згорткової SNN (CSNN) на основі нейрона Leaky Integrate-and-Fire (LIF). Модель тестувалась на трьох наборах даних класифікації зображень: MNIST, FMNIST та CIFAR10. Стиснення проводилося методами статичної k-бітної квантизації після навчання та структурованого прунінгу з різними коефіцієнтами, що зустрічаються у практичному використанні. Отримані результати показують, що при невисоких параметрах стиснення SNN демонструють несуттєву втрату точності, одночасно забезпечуючи значне зменшення розміру моделі та енергоспоживання. Однак, для більш складного набору даних, неоптимальної навченої моделі та при екстремальних налаштуваннях стиснення, спостерігається різке та значне погіршення метрик класифікації.

СПАЙКОВІ НЕЙРОННІ МЕРЕЖІ, МАШИННЕ НАВЧАННЯ, ШТУЧНІ НЕЙРОННІ МЕРЕЖІ, ШТУЧНИЙ ІНТЕЛЕКТ, ОПТИМІЗАЦІЯ ІНФЕРЕНЦІЇ, ОПТИМІЗАЦІЯ ВИВЕДЕННЯ, КВАНТИЗАЦІЯ, ПРУНІНГ, НЕЙРОМОРФНЕ ОБЧИСЛЕННЯ

Ye.V. Bodyanskiy, D.V. Savenkov. Inference optimization parameters influence spiking neural network efficiency. Spiking neural networks (SNNs) are the third generation of artificial neural networks, which, thanks to their energy efficiency and sparsity, are ideal for use in resource-constrained environments such as IoT or robotics. However, even they may not meet extreme requirements, leading to the need for inference optimization methods, such as quantization and pruning. Recent studies have already considered the practical application of these methods for spiking neural networks, but they have not focused on the impact of initial optimization parameters on the performance of the compressed model. The goal of this study is to systematize and empirically investigate the impact of quantization and pruning method parameters on the final performance of spiking neural networks. A convolutional SNN (CSNN) architecture based on the Leaky Integrate-and-Fire (LIF) neuron was used for the experiments. The model was tested on three image classification datasets: MNIST, FMNIST, and CIFAR10. Compression was performed using static k-bit quantization methods after training and structured pruning with different coefficients encountered in practical use. The results show that at low compression parameters, SNNs demonstrate insignificant accuracy loss while providing a significant reduction in model size and energy consumption. However, for a more complex dataset, a suboptimal trained model, and extreme compression settings, a sharp and significant deterioration in classification metrics is observed.

SPIKING NEURAL NETWORKS, MACHINE LEARNING, ARTIFICIAL NEURAL NETWORKS, ARTIFICIAL INTELLIGENCE, INFERENCE OPTIMIZATION, QUANTIZATION, PRUNING, NEUROMORPHIC COMPUTATION

Вступ

Спайкові нейронні мережі (Spiking Neural Networks, SNNs) – це третє покоління штучних нейронних мереж (Artificial Neural Networks, ANNs), метою та фокусом яких є відтворення обчислювальних принципів біологічних нейронних систем [1]. Даний тип нейронних мереж імітує такі нейробіологічні процеси, як накопичення і розрядження заряду, мембранний потенціал, рефрактерний період, комунікація через імпульси або «спайки», тощо. На відміну від “класичних” ANNs, які базуються на безперервних значеннях та функціях активації, SNNs виконують дискретні, розрідженні, асинхронні та подія-орієнтовані обчислення основані на подіях. Завдяки цьому SNN мають значні переваги з точки зору енергоефективності та обчислювальної потужності [2], що робить їх придатними для використання на нейроморфному обладнанні з низьким енергоспоживанням, що у свою чергу робить SNN привабливими у завданнях IoT та роботехниці.

Для таких завдань, окрім вимог до продуктивності, також притаманні й обмеження в ресурсах, зокрема пам’яті, часу та електроенергії. Незважаючи на зазначену енергоефективність SNN, практичне впровадження великомасштабних SNN залишається значним викликом. Сам розмір навчених моделей SNN, включаючи велику кількість емульованих синапсів і шарів, може перевищувати обсяг пам'яті цільових апаратних платформ та виконуватись довше зазначеного ліміту. Це особливо актуально для складних

глибоких архітектур, необхідних для досягнення найсучаснішої продуктивності у складних завданнях.

Вирішенням даних проблем займаються задачі та методи оптимізації виведення або інференції (inference optimization) [3], які фокусуються на стисненні та пришвидшенні натренованих моделей машинного навчання без значної втрати продуктивності, з метою подальшою інтеграцію у прикладні та практичні системи. Дані методи можна умовно розбити на апаратні та алгоритмічні рішення. До останніх відносять такі методи, як "квантизація" [4] та "прунінг" [5], що регулярно застосовують у практичних завданнях, зокрема у великомовних моделях (Large Language Models).

Незважаючи на емпірично перевірену ефективність, дані методи дуже чутливі до початкових параметрів, від яких залежить не тільки розмір фінальної моделі, а й втрачена продуктивність. Хоча існують дослідження щодо використання методів оптимізації інференції для SNN, питання залежності цих параметрів та продуктивності залишались поза фокусу.

Об'єктом дослідження є процес оптимізації виводу або інференції (inference optimization) спайкових нейронних мереж для їх ефективного застосування на пристроях з обмеженими обчислювальними ресурсами.

Предметом дослідження є залежність втрати продуктивності (точності) та ступеня стиснення спайкових нейронних мереж від початкових параметрів алгоритмічних методів оптимізації, зокрема квантизації та прунінгу.

Мета дослідження полягає у систематизація та емпіричне дослідження впливу параметрів методів квантизації та прунінгу на кінцеву продуктивність спайкових нейронних мереж, а також розробка практичних рекомендацій щодо вибору цих параметрів для досягнення оптимального балансу між розміром моделі та її точністю.

1. Постановка задачи

Для досягнення поставленої мети, необхідно виконати наступні задачі:

Сформувати теоретичну базу: провести аналіз існуючих підходів до квантизації та прунінгу для SNNs.

Розробити експериментальний стенд: імплементувати тренувальний пайплайн, що дозволяє застосовувати різні комбінації параметрів квантизації та прунінгу до тренованих моделей SNN.

Провести серію експериментів: дослідити вплив ключових параметрів, як-от ступінь стиснення (p) для прунінгу та бітність (k) для квантизації, на фінальну продуктивність моделі.

Проаналізувати результати: порівняти метрики стиснення та втрати точності для кожної комбінації параметрів.

У результаті, буде проведено серію експериментів з різними конфігураціями параметрів оптимізації, а їхні метрики будуть проаналізовані та порівняні. Це дозволить надати чіткі рекомендації для практичного застосування SNN на пристроях з обмеженими ресурсами.

2. Огляд теоретичної бази

Як було зазначено у попередніх розділах, SNN працюють дещо відміно від класичних ANN. Замість безперервних функцій активації, SNN обробляють та передають інформацію за допомогою дискретних асинхронних імпульсів. Ці бінарні активаційні події відбуваються, коли мембранний потенціал нейрона перевищує поріг напруги, імітуючи вивільнення нейромедіаторів. Як і їх біологічні аналоги, SNN демонструють часову динаміку та еволюцію через накопичення та розрядження заряду, і включають рефрактерний період після активації.

Найпростішою і найпоширенішою моделлю SNN є нейрон Leaky Integrate-and-Fire (LIF) [6], який абстрагує накопичення мембранного напруги нейрона як резистор-конденсаторну електросхему. Ця простота робить його обчислювально ефективним і придатним для інтеграції в апаратне забезпечення. Модель LIF можна описати наступним рівнянням:

$$V(t)=\frac{1}{\tau_m}\left(V_{rest}+I(t)R\right), \tag{1}$$

де $V(t)$ — напруга мембрани, V_{rest} — напруга стану "спокою", τ_m — постійна часу мембрани, R — опір мембрани, та $I(t)$ — вхідний струм.

Коли $V(t)\geq V_{threshold}$, нейрон «вистрілює» (генерує імпульс), передає напругу до підключених нейронів, скидає свій мембранний потенціал до V_{reset} і входить у рефрактерний період, під час якого він має меншу ймовірність активації. Цей процес можна побачити на рис. 1.

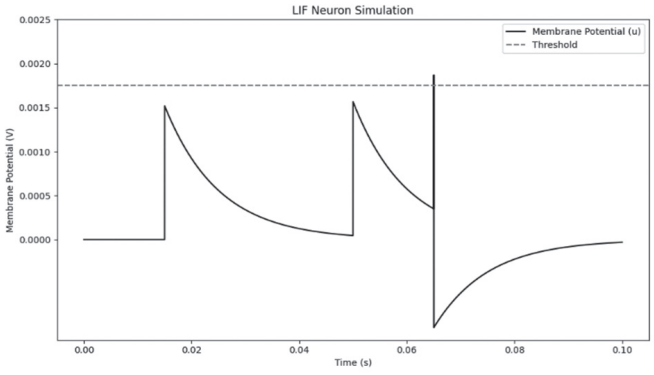

Рис. 1. Процес функціонування LIF нейрону

Незважаючи на свою простоту, моделі LIF є універсальними та обчислювально ефективними в порівнянні з більш складними моделями, такими як моделі Іжакевича [7] або Ходжкіна-Хакслі [8], що моделюють комплексні біохімічні процеси, як іонні

канали. Модель LIF також, у порівнянні з переліченими моделями, демонструє кращу сумісність з класичними методами машинного навчання: вона краще підтримує зворотне поширення та навчання з вчителем; але одночасно ефективно інтегрується з біологічно натхненними парадигмами, як залежної від часу спайку пластичністю (Spike-timing-dependent plasticity, STDP) [9].

Біологічна правдоподібність і низьке енергоспоживання LIF роблять їх перспективною альтернативою ANN, особливо для периферійних і нейроморфних обчислень. Однак у край ресурсно обмежених ситуаціях LIF також може потребувати оптимізації інференції. Хоча оптимізація інференції є широкою темою, найбільш практично використаними методами є квантизації та прунінг (обрізання).

Квантизація [4] - це методика, яка зменшує бітову точність параметрів мережі та активацій, тим самим зменшуючи вимоги до пам'яті та обчислювальних потужностей. У SNN це передбачає представлення безперервного потенціалу мембрани, синаптичних ваг та змінних, пов'язаних з часом, за допомогою меншої кількості бітів. Динамічний діапазон цих змінних може бути великим, а зменшення їх точності може призвести до значної економії в апаратних реалізаціях. Ключовою перевагою квантизації є те, що воно дозволяє виконувати арифметичні операції за допомогою цілочисельних операцій з низькою кількістю бітів, які є швидшими та енергоефективнішими, ніж операції з плаваючою комою. Наприклад, 32-бітне множення з плаваючою комою можна замінити 8-бітним цілочисельним множенням, що призведе до значного зменшення площі апаратного забезпечення та споживання енергії. Загальні методи квантизації розділяють на дві категорії: квантизація після навчання (Post-Training Quantization) [10], яку також розділяють на динамічну та статичну, а також навчання з урахуванням квантизації (Quantization-Aware Training) [11]. Проблема квантизації SNN полягає у збереженні їхньої часової динаміки та потоку інформації, які чутливі до змін точності. Ця чутливість часто вимагає нових схем квантизації, таких як ті, що враховують спайк-орієнтовану природу SNN [12].

Прунінг [5] - це інша загальна методика оптимізації інференції, яка використовується для зменшення розміру нейронної мережі шляхом видалення зайвих або менш важливих зв'язків (ваг) або нейронів. Це призводить до створення більш розрідженої мережі, яка вимагає менше обчислень під час інференції. Мета полягає в досягненні значного стиснення та прискорення моделі без істотної втрати продуктивності. У SNN, як й у ANN, прунінг може застосовуватися як до синаптичних зв'язків, так і до нейронів. Розрідженість, що виникає в результаті прунінгу, є особливо корисною для SNN, які природно працюють з розрідженими, керованими подіями даними. Існує два основних типи прунінгу: неструктурований прунінг та структурований прунінг [3]. Неструктурована обрізка видаляє окремі ваги, що призводить до нерегулярної розрідженості, для використання якої потрібне спеціальне обладнання або програмне забезпечення. На відміну від цього, структурований прунінг видаляє цілі нейрони або канали, що призводить до регулярної розрідженості, яку легше прискорити на стандартному апаратному забезпеченні. Прунінг можна здійснювати двома основними способами: прунінг на основі величини, яке видаляє ваги з найменшими абсолютними значеннями, та прунінг на основі градієнта, яке використовує інформацію з градієнтів мережі для ідентифікації та видалення менш важливих ваг. Ефективність прунінгу в SNN залежить від їхнього навчання та конкретного методу прунінгу, оскільки погано обрізана мережа може втратити здатність кодувати та обробляти часову інформацію.

Підсумовуючи, оптимізація інференції є дуже важливим аспектом прикладних систем штучного інтелекту в умовах обмежених ресурсів. Дані методики активно та ефективно застосовуються у задачах використання великомовних моделей (Large Language Models, LLMs) та IoT, а також разом із SNN. Але треба враховувати, що разом із пришвидшенням моделі дані методики можуть погіршивати їхню точність у залежності від ступеня стиснення. Наступні розділи фокусуються на дослідженні впливу налаштувань методів стиснення SNN моделей на їх продуктивність.

3. Матеріали та методи

Як було зазначено, наша робота фокусується на впливі налаштувань методів стиснення на зменшення продуктивності SNN моделей. Щоб провести експериментальне дослідження, було побудовано наступний пайплайн (його також візуалізовано на рис. 2):

– Сформувати архітектуру SNN моделі M;

– Натренувати її на обраному тренувальному наборі даних D_{train};

– Провести операцію стиснення моделі M зі встановленими параметрами H та отримати стиснену модель M';

– На тестовій виборці Dtest провести тестування;
  – звичайної моделі M;
  – стисненої моделі M';

– Задокументувати результати.

Обрана експериментальна архітектура SNN спеціалізується на виконанні завдань класифікація зображень та складається з послідовних шарів згорток, пулінгу та LIF-нейронів. Повна архітектура зображена на рис. 3. Саму модель було навчено парадигмою навчання з вчителем з використанням алгоритму

backpropagation. Також, в якості операцій стиснення було використано статичну квантизацію після навчання та прунінг.

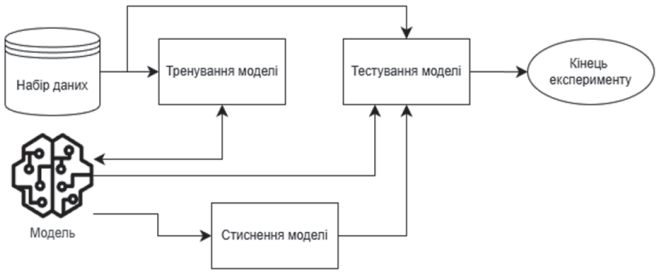

Рис. 2. Тренувальний пайплайн

Для досягнення поставленої мети, необхідно виконати наступні задачі:

– Сформувати теоретичну базу: провести аналіз існуючих підходів до квантизації та прунінгу для SNNs.

– Розробити експериментальний стенд: імплементувати тренувальний пайплайн, що дозволяє застосовувати різні комбінації параметрів квантизації та прунінгу до тренованих моделей SNN.

– Провести серію експериментів: дослідити вплив ключових параметрів, як-от ступінь стиснення (p) для прунінгу та бітність (k) для квантизації, на фінальну продуктивність моделі.

– Проаналізувати результати: порівняти метрики стиснення та втрати точності для кожної комбінації параметрів.

Обрана експериментальна архітектура SNN спеціалізується на виконанні завдань класифікація зображень та складається з послідовних шарів згорток, пулінгу та LIF-нейронів. Повна архітектура зображена на рис. 3. Саму модель було навчено парадигмою навчання з вчителем з використанням алгоритму backpropagation. Також, в якості операцій стиснення було використано статичну квантизацію після навчання та прунінг.

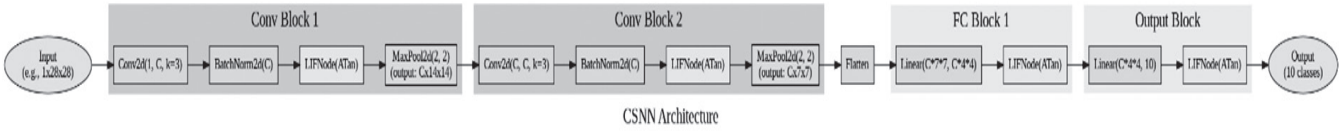

Рис. 3. Використана архітектура SNN

4. Експерименти

Для проведення експериментів з розробленим пайплайном було обрано декілька наборів даних класифікації зображень зі зростаючою складністю: набір MNIST [13], що містить 70 000 зображення рукописних цифр, FMNIST [14], що містить 70 000 зображень одягу, та CIFAR10 [15], який містить 60 000 зображень різних повсякденних об'єктів. Кожен з цих наборів даних має 10 унікальних класів.

З кожним набором даних було побудовано два експериментальних пайплайни, описаних у розділі 3. Також дані експерименти мали таке конфігурацію:

LIF-нейрони мають часову константу мембрани τ встановлено 2.0, для балансу процесу накопичення та розрядження заряду;

Кожен екземпляр даних подається на вхід нейронам T=20ms, щоб імітувати часовий проміжок людського нейрону необхідний для розпізнання інформації;

– Початкові ваги згенеровані випадковим чином, SNN не підлягав оптимізації гіперпараметрів;

– Операції стиснення, як зазначено у розділі 3, виконують після навчання;

– Операції квантизації проводяться з такими значеннями K-бітності: 16, 8, 4;

– Операції прунінгу проводяться з такими значеннями коефіцієнту: 0.1, 0.2, 0.3.

Для проведення експерименту використовуються мова програмування Python, модуль PyTorch [16] та модуль емуляції нейроморфних обчислень SpikingJelly [17]. Також було використано модуль syops [18] для вимірювання енергоспоживання нейромереж до та після прунінгу в умовах емуляції. Значення енергоспоживання базуються на технології емульованого 45 нм процесору, де «арифметичні обчислювальні кроки» (Arithmetic Compute Steps, ACs) коштують 0,9 пДж, а «кроки множення-накопичення» (Multiply-Accumulate Compute Steps, MACs) — 4,6 пДж. Тобто споживання розробленої нейромережі можна апроксимувати до наступної формули:

$$E = 0.9 * ACs + 4.6 * MACs, \quad (2)$$

Результати експериментів описані у розділі 5.

5. Результати

Як було зазначено у попередньому розіділ, кожен експеримент був проведений зі заздалегідь навченими моделями. Кожну з них було виміряно використовуючи класичні метрики задач класифікації: accuracy, precision, recall та f1-score. Результат даних вимірювань описаний у табл. 1. Всі перелічені метрики є зваженими по кожному класу.

Таблиця 1

Метрики натренованих нестиснених моделей(%)

Модель	Набір даних	Метрики на тестовій вибірці			
		Accuracy	Precision	Recall	F1-score
CSNN	MNIST	0.9872	0.9871	0.9871	0.9871
CSNN	FMNIST	0.8582	0.8562	0.8581	0.8544
CSNN	CIFAR10	0.5083	0.5149	0.5083	0.5033

Дана проста модель перед стисненням має гарні показники на простих наборах даних, хоча з більш складним CIFAR10 метрики далекі від оптимальних. Кожна стиснена модель також була протестована на ідентичних тестових підвибірках. Їхні вимірювання перелічені у табл. 2 (де перелічені квантизовані моделі qCSNN(k=n), де k – бітність квантизації) та у табл. 3 (де перелічені моделі після прунінгу pCSNN(p=m), де p – коефіцієнт прунінгу).

Таблиця 2

Метрики стиснених моделей методом квантизації (%)

Модель	Набір даних	Метрики на тестовій вибірці			
		Accuracy	Precision	Recall	F1-score
qCSNN(k=16)	MNIST	0.9854	0.9853	0.9853	0.9852
qCSNN(k=8)	MNIST	0.9851	0.985	0.9851	0.985
qCSNN(k=4)	MNIST	0.9793	0.9793	0.9793	0.979
qCSNN(k=16)	FMNIST	0.8551	0.853	0.8551	0.8514
qCSNN(k=8)	FMNIST	0.8538	0.8519	0.8538	0.8503
qCSNN(k=4)	FMNIST	0.7995	0.8315	0.7995	0.7933
qCSNN(k=16)	CIFAR10	0.5066	0.5115	0.5066	0.5015
qCSNN(k=8)	CIFAR10	0.4999	0.5066	0.4998	0.4958
qCSNN(k=4)	CIFAR10	0.2821	0.5239	0.282	0.2257

Таблиця 3

Метрики стиснених моделей методом прунінгу (%)

Модель	Набір даних	Метрики на тестовій вибірці			
		Accuracy	Precision	Recall	F1-score
pCSNN(p=0.1)	MNIST	0.9865	0.9865	0.9864	0.9864
pCSNN(p=0.2)	MNIST	0.9861	0.986	0.986	0.9859
pCSNN(p=0.3)	MNIST	0.981	0.981	0.9808	0.9808
pCSNN(p=0.1)	FMNIST	0.8574	0.8554	0.8574	0.8538
pCSNN(p=0.2)	FMNIST	0.8518	0.8502	0.8518	0.8474
pCSNN(p=0.3)	FMNIST	0.8406	0.8446	0.8406	0.8348
pCSNN(p=0.1)	CIFAR10	0.4983	0.4998	0.4982	0.4934
pCSNN(p=0.2)	CIFAR10	0.4962	0.5065	0.4962	0.4914
pCSNN(p=0.3)	CIFAR10	0.4206	0.509	0.4206	0.4129

Ефективність на простих наборах MNIST та FMNIST не зазнала сильного впливу. З більш складним CIFAR10, — ефективність класифікації якого була доволі слабкою й до стиснення, — ситуація дещо інша: при низьких параметрах стиснення модель не зазнала значного падіння метрик, але після перетину певного порогу, метрики класифікації CIFAR10 зазнають значного та різкого погіршення.

Погіршення метрик — очікуваний аспект стиснення, але разом із цим очікується й покращення інших характеристик моделі: зменшення розмірності моделі та/або енергоспоживання. Зміна даних характеристик також залежить і від інших факторів, як оптимізація кодової та апаратної імплементації, або використаних технологій емуляція. Але загально їх можна апроксимувати до результатів наведених у наступних двох графах:

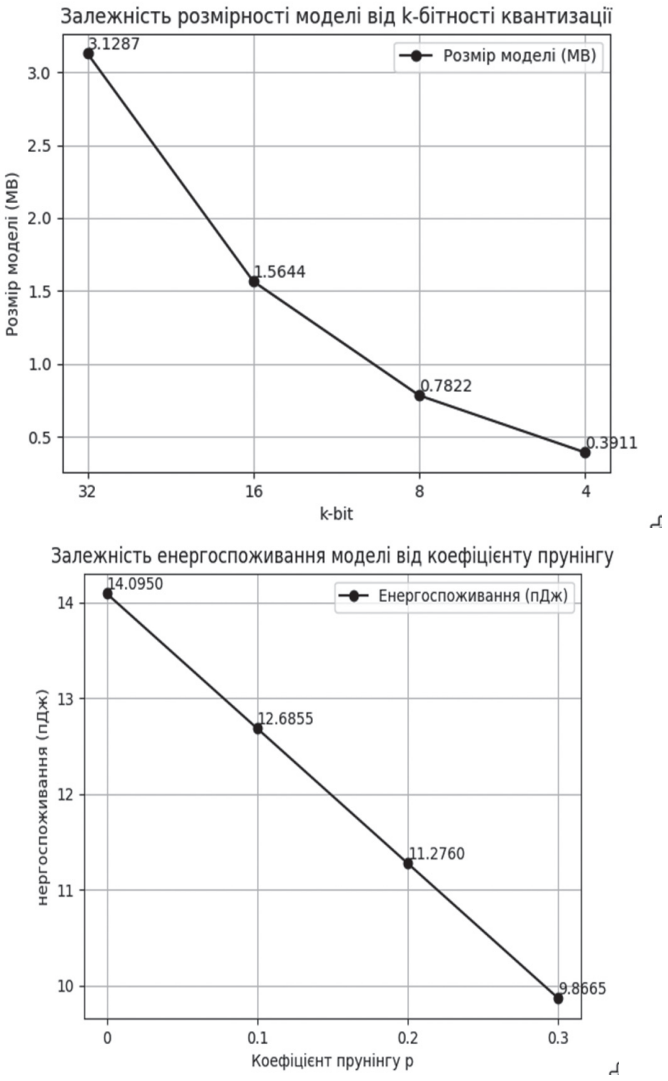

Рис. 4. Графи залежності розмірності моделі від k-бітності квантизації та енергоспоживання моделі від коефіцієнту прунінгу

6. Обговорення

Отриманні результати вказують на несуттєву втрату продуктивності розробленої SNN моделі при невисоких параметрах стиснення, при цьому значно зменшуючи необхідність у пам'яті та енергоспоживанні. Але варто зазначити, що більш "екстремальні" налаштування видають гірші результати для більш комплексних наборів даних. При використанні даних та інших методів стиснення моделей треба враховувати особливості вхідних даних, архітектури та оптимальності самої моделі, зокрема при більш екстремальних налаштуваннях.

Враховуючи інші особливості SNN, зокрема енергоефективність та розріджені обчислення, та результати дослідження, методи стиснення SNN можуть бути ефективно використанні у системах IoT, робототехніці або інших ресурсно-обмежених середовищах, в ситуаціях оптимальності нестиснених моделей. Але саме рішення та використання цих методів все одно підвладні сучасним вразливостям та недолікам SNN.

Висновки

У цій роботі було проведено теоретичне та практичне дослідження впливу на продуктивність SNN таких методів стиснення, як квантизація та прунінг. Для цього були проведені експерименти з типовими наборами даних зростаючої складності для вирішення задач аналізу даних, під час яких було протестовано стиснену за різними параметрами модель на втрату продуктивності за класичними метриками задач навчання з вчителем та класифікації. Також було проведено аналіз оптимізації розмірності та енергоспоживання даних моделей. Практичне значення даної роботи та отриманих результатів полягає у розумінні потенційно очікуваних переваг та вартості застосування стиснення SNN моделей. Дані моделі мають кращу енергоефективність та можливість виконувати розріджені обчислення, що, у поєднанням зі стисненням, робить їх більш привабливими у вирішенні задач інтелектуального аналізу даних у ресурсообмежених системах, як системи IoT або навчання парадигмою онлайн-навчання

Список літератури

[1] Gerstner W. Spiking neuron models: Single neurons, populations, plasticity. Cambridge, U.K: Cambridge University Press, 2002. 480 с.

[2] Davidson S., Furber S. B. Comparison of Artificial and Spiking Neural Networks on Digital Hardware. Frontiers in Neuroscience. 2021. Т. 15. URL: https://doi.org/10.3389/fnins.2021.651141 (дата звернення: 26.11.2025).

[3] Optimization Methods, Challenges, and Opportunities for Edge Inference: A Comprehensive Survey / R. Zhang та ін. Electronics. 2025. Т. 14, № 7. С. 1345. URL: https://doi.org/10.3390/electronics14071345 (дата звернення: 26.11.2025).

[4] Pruning Parameterization with Bi-level Optimization for Efficient Semantic Segmentation on the Edge / C. Yang та ін. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), м. Vancouver, BC, Canada, 17–24 черв. 2023 р. 2023. URL: https://doi.org/10.1109/cvpr52729.2023.01478 (дата звернення: 26.11.2025).

[5] FlatQuant: Flatness Matters for LLM Quantization / Y. Sun та ін. arXiv. 2025.

[6] Burkitt A. N. A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input. Biological Cybernetics. 2006. Т. 95, № 1. С. 1–19. URL: https://doi.org/10.1007/s00422-006-0068-6 (дата звернення: 26.11.2025).

[7] Izhikevich E. M. Simple model of spiking neurons. IEEE Transactions on Neural Networks. 2003. Т. 14, № 6. С. 1569–1572. URL: https://doi.org/10.1109/tnn.2003.820440 (дата звернення: 26.11.2025).

[8] Hodgkin A. L., Huxley A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology. 1952. Т. 117, № 4. С. 500–544. URL: https://doi.org/10.1113/jphysiol.1952.sp004764 (дата звернення: 26.11.2025).

[9] A neuronal learning rule for sub-millisecond temporal coding / W. Gerstner та ін. Nature. 1996. Т. 383, № 6595. С. 76–78. URL: https://doi.org/10.1038/383076a0 (дата звернення: 26.11.2025).

[10] Post-Training Quantization for Vision Transformer / Z. Liu та ін. Advances in Neural Information Processing Systems. 2021. Т. 34. С. 28092–28103.

[11] QuantNAS: Quantization-aware Neural Architecture Search For Efficient Deployment On Mobile Device / T. Gao та ін. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), м. Seattle, WA, USA, 17–18 черв. 2024 р. 2024. С. 1704–1713. URL: https://doi.org/10.1109/cvprw63382.2024.00177 (дата звернення: 26.11.2025).

[12] Li C., Ma L., Furber S. Quantization Framework for Fast Spiking Neural Networks. Frontiers in Neuroscience. 2022. Т. 16. URL: https://doi.org/10.3389/fnins.2022.918793 (дата звернення: 26.11.2025).

[13] Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine. 2012. Т. 29, № 6. С. 141–142. URL: https://doi.org/10.1109/msp.2012.2211477 (дата звернення: 26.11.2025).

[14] Xiao H., Rasul K., Vollgraf R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithm. arXiv. 2017. URL: https://arxiv.org/abs/1708.07747 (дата звернення: 26.11.2025).

[15] Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. 2009. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (дата звернення: 26.11.2025).

[16] PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation / J. Ansel та ін. ASPLOS '24: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, м. La Jolla CA USA. New York, NY, USA, 2024. URL: https://doi.org/10.1145/3620665.3640366 (дата звернення: 26.11.2025).

[17] SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence / W. Fang та ін. Science Advances. 2023. Т. 9, № 40. URL: https://doi.org/10.1126/sciadv.adi1480 (дата звернення: 26.11.2025).

[18] Training Full Spike Neural Networks via Auxiliary Accumulation Pathway / G. Chen та ін. arXiv. 2023. URL: https://doi.org/10.48550/arXiv.2301.11929 (дата звернення: 26.11.2025).

Надійшла до редколегії 11.09.2025

УДК 004.8 DOi 10.30837/ bi.2025.2(103).02

А. Д. Костюченко[1], В. В. Герасимов[2]

[1]ДНУ, м. Дніпро, Україна, kostiuchenko_a@365.dnu.edu.ua,
ORCID iD: 0000-0002-3940-8797

[2]ДНУ, м. Дніпро, Україна, herasymov_v@365.dnu.edu.ua,
ORCID iD: 0000-0002-1366-715X

КОМП'ЮТЕРНІ МОДЕЛІ ПРОГНОЗУВАННЯ ЗНАЧЕНЬ ЧАСОВИХ РЯДІВ

У роботі розглянуто підхід до комп'ютерного моделювання часових рядів на основі сучасних архітектур глибокого навчання, зокрема LSTM, GRU та їхніх гібридних комбінацій, а також ансамблевих моделей. Запропоновано порівняльний аналіз глибоких нейромережевих структур різної параметричної складності за метриками MAE, MSE, SMAPE та MAPE на тренувальній, валідаційній та тестовій вибірках. Показано, що гібридні архітектури LSTM+GRU забезпечують кращу якість прогнозування порівняно з окремими моделями, однак подальше нарощування глибини та кількості параметрів призводить лише до незначного приросту точності. Найкращі результати отримано для ансамблевого алгоритму, сформованого на основі кількох різнорідних моделей, який демонструє найнижчі значення похибок та підвищену стійкість до шумів і аномалій даних. Результати експериментів підтверджують, що поєднання глибинних рекурентних архітектур із ансамблевими підходами є ефективним інструментом підвищення точності та стабільності прогнозування довгострокових метеорологічних часових рядів і може бути використане як основа для побудови прикладних систем підтримки прийняття рішень в енергетиці, транспорті та інших динамічних галузях.

МОДЕЛЬ ГЛИБОКОГО НАВЧАННЯ, LSTM, GRU, АНСАМБЛЕВА МОДЕЛЬ, ЧАСОВИЙ РЯД

A.D. Kostiuchenko, V.V. Herasymov. Computer models for time series forecasting. The paper presents an approach to computer modelling of time series based on modern deep learning architectures, specifically LSTM, GRU, and their hybrid combinations, as well as ensemble models. A comparative analysis of deep neural network structures with varying parametric complexity is proposed, utilising MAE, MSE, SMAPE, and MAPE metrics on training, validation, and test samples. It is demonstrated that hybrid LSTM+GRU architectures yield better prediction quality compared to individual models; however, further increasing the depth and number of parameters results in only a slight increase in accuracy. The best results were obtained for an ensemble algorithm based on several heterogeneous models, which demonstrates the lowest error values and increased resistance to noise and data anomalies. The results of the experiments confirm that the combination of deep recurrent architectures with ensemble approaches is an effective tool for improving the accuracy and stability of long-term meteorological time series forecasting and can serve as a basis for building applied decision support systems in energy, transportation, and other dynamic industries.

DEEP LEARNING MODEL, LSTM, GRU, ENSEMBLE MODEL, TIME SERIES

Вступ

Упродовж останніх років моделі глибокого навчання зарекомендували себе як важливий інструмент розв'язання цілої низки задач штучного інтелекту. Згорткові нейронні мережі та архітектури, побудовані на основі операції згортки, дозволяють класифікувати, визначати межі та сегментувати об'єкти на зображеннях із майже абсолютною точністю. Останні експерименти свідчать, що сучасні моделі типу ResNet50, VGG19 та InceptionV3 класифікують об'єкти у випадках кількох можливих правильних відповідей навіть краще, аніж людське око. Задачі обробки природньої мови також отримали нові підходи до вирішення: архітектури типу GPT, BERT, T5 вже здатні не лише анотувати, перекладати чи переказувати тексти, але й генерувати власні, унікальні відповіді у форматі чат-ботів, що й було покладено в основу всесвітньо відомих ChatGPT, Gemini, Grok та інших. Таке широке впровадження моделей штучного інтелекту неминуче вплинуло й на домен обробки послідовностей, оскільки навіть базові задачі, такі як автоматичний переклад природної мови, за своєю природою є задачами моделювання залежностей слів у реченнях. У цих сценаріях кожен наступний елемент послідовності залежить від попередніх, що потребує здатності моделі утримувати й опрацьовувати контекст різної довжини.

Прогнозування майбутніх значень процесу на основі його історичних спостережень є фундаментальним інструментом у фінансовому аналізі, енергетиці, метеорології, екологічному моніторингу, демографії, логістиці, медичній діагностиці та багатьох інших сферах, що було розглянуто детально у дослідженнях [1;2]. Моделі прогнозування знаходять широке застосування у задачах передбачення динаміки метеорологічних показників, зокрема температури повітря та атмосферного тиску, що є важливим для оптимізації роботи інфраструктури, чутливої до погодних умов. У сфері енергетики такі моделі використовуються для оцінювання майбутніх обсягів споживання електроенергії, що дає змогу підвищувати ефективність функціонування енергосистем і забезпечувати їхню стабільність. У фінансовому секторі прогнозування параметрів ринку, включно з динамікою цін, рівнем ліквідності та показниками волатильності фінансових активів, є необхідною умовою

для оцінювання ризиків і формування стратегій управління капіталом. У технічних системах прогнозування ймовірності виникнення відмов дозволяє своєчасно виявляти критичні тенденції та запобігати негативним наслідкам, що сприяє зменшенню операційних витрат і підвищує загальну надійність обладнання. У контексті обробки даних, отриманих з датчиків транспортних засобів, такі системи дозволяють відстежувати рівень палива в системі та аномальну роботу двигуна в різних станах, повідомляти про аварійні ситуації, що пов'язані з паливними баками.

Актуальність задач обробки часових рядів також зумовлена тим, що реальні процеси майже ніколи не бувають представлені лінійними або одновимірними залежностями. У багатьох галузях доводиться працювати з мультиваріантними часовими рядами, де кілька змінних корелюють значною мірою між собою і визначають поведінку системи. У такому домені, як метеорологія, одночасний аналіз температури, вологості, атмосферного тиску та швидкості вітру забезпечує набагато точніше моделювання погодних умов, ніж окремий прогноз кожної змінної. У фінансових ринках взаємозв'язки між індексами, активами та макроекономічними факторами формують складну систему, в якій локальні зміни матимуть значні довгострокові наслідки. Такі системи потребують моделей, здатних уловлювати не лише часові, а й структурні взаємозв'язки між ознаками. Питання впровадження комбінованих архітектур, методів машинного та глибокого навчання, а також статистичного аналізу значною мірою досліджувались у роботах [3-5].

Метою даного дослідження є обґрунтування та емпірична перевірка доцільності використання глибинних нейромережевих архітектур типу LSTM, GRU та їхніх гібридів, а також ансамблевих алгоритмів на їх основі для прогнозування значень часових рядів метеорологічних показників. Передбачається встановити, чи забезпечують гібридні архітектури LSTM+GRU та їхній ансамбль статистично значуще покращення якості прогнозування порівняно з окремими моделями, а також визначити компроміс між параметричною складністю та точністю прогнозу.

Виклад основного матеріалу

Відповідно до визначення, поданого у роботі [6], часовий ряд – це послідовність упорядкованих у хронологічному порядку сукупності значень певного статистичного показника, які відображають динаміку певного явища у часі. Елементи послідовності y_1, y_2, y_3, ..., y_n називають рівнями часового ряду. Індекс $t = 1, 2, 3, 4 ..., n$ позначає порядковий номер моменту або інтервалу часу, до якого належить відповідний рівень, а величина n визначає довжину часового ряду.

Залежно від характеру часового виміру часові ряди класифікують на моментні та інтервальні. У моментних рядах кожен рівень фіксує стан явища на конкретний момент часу; типовими прикладами є обсяг золотовалютних резервів на початок року або чисельність зареєстрованих безробітних на початок кварталу. На відміну від цього, рівні інтервальних рядів відображають результат процесу, накопичений протягом певного часового проміжку. Прикладами таких рядів можуть слугувати обсяги виробленої електроенергії за квартал, біоелектричні сигнали організму людини або показники сонячної активність упродовж року.

Структура часових рядів визначається сукупністю характеристик, які описують спосіб організації та внутрішні властивості даних, впорядкованих у часовому вимірі. Часовий ряд $\{x_t\}_{t=1}^{T}$ розглядається як реалізація стохастичного процесу, де кожне спостереження є результатом дії як детермінованих, так і випадкових компонент. Загальноприйнятою є адитивна або мультиплікативна декомпозиція часових рядів, у межах якої рівень у момент часу t подається як сума або добуток структурних складових. У базовому адитивному варіанті модель представлення рядка записується у вигляді:

$$x_t = m_t + s_t + c_t + \varepsilon_t \quad (1)$$

де m_t –трендова компоненту; s_t – сезонна складова з фіксованим періодом повторення; c_t – циклічні зміни; ε_t – міра впливу стохастичного процесу.

Для оцінки якості прогнозування значень моделей глибокого навчання було використано метрики MSE (англ. – Mean Squared Error) та SMAPE (англ. – Symmetric Mean Absolute Percentage Error), розраховані за формулами (2) і (3).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \quad (2)$$

де y_i – значення цільової функції; $\hat{y}_i$ – прогнозоване значення; N – кількість елементів у вибірці.

$$SMAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}, \quad (3)$$

де y_i – значення цільової функції; $\hat{y}_i$ – прогнозоване значення;

Питання обробки послідовностей динамічно розвивається з кінця минулого сторіччя. Саме тоді було запропоновано архітектуру LSTM (англ. – Long-Short Term Memory), що стала альтернативою використання стандартних та домінуючих у галузі ШІ стандартних рекурентних нейронних мереж. У статтях [7] та [8] детально описано кроки реалізації.

Архітектура LSTM належить до класу рекурентних нейронних мереж і була розроблена з метою подолання фундаментальних недоліків класичних RNN, пов'язаних зі зниканням і вибухом градієнтів під час поширення похибки у часовому напрямку. Основним конструктивним елементом моделі є комірка пам'яті,

яка забезпечує здатність зберігати інформацію протягом тривалих часових інтервалів завдяки окремому стану пам'яті c_t, структурно відокремленому від прихованого стану h_t. На відміну від традиційних RNN, де інформаційний потік повністю контролювався нелінійною динамікою одного прихованого шару, LSTM впроваджує систему гейтів, що регулюють надходження та видалення інформації, надаючи моделі гнучкість та стійкість при роботі з довгими послідовностями.

Обчислення всередині LSTM-комірки визначається поєднанням сигмоїдальної та гіперболічного тангенсу активаційних функцій, які формують керовані інформаційні канали. Гейт забуття обчислюється через сигмоїдальну функцію активації та визначає частину попереднього стану пам'яті c_{t-1}, що підлягає збереженню. Завдяки множенню цього гейта на c_{t-1} модель здатна контролювати довжину ефективного контексту й адаптуватися до структури часових залежностей. Наступним компонентом є гейт входу, що модулює вплив нової інформації на стан пам'яті.

Паралельно з цим обчислюється нове значення із використанням гіперболічного тангенсу, яке відображає потенційну зміну внутрішнього стану. Комбінація гейта входу та даного нового значення визначає ступінь оновлення c_t, що забезпечує поступове оновлення пам'яті з урахуванням релевантних ознак вхідного сигналу. Після цього формується вихідний гейт, який визначає, яку частину оновленого стану пам'яті буде відображено у прихованому стані h_t. Використання $tanh(c_t)$ у поєднанні з сигмоїдальним гейтом виходу формує прихований стан, що відображає внутрішню інформацію, необхідну для подальших прогнозів, класифікації або інших завдань.

GRU (англ. – Gated Recurrent Unit) є рекурентною архітектурою, розробленою як компактна та обчислювально ефективніша альтернатива LSTM, з метою зменшення кількості параметрів і спрощення механізмів керування пам'яттю без втрати здатності моделювати довготривалі залежності у часових послідовностях. На відміну від LSTM, яка використовує окрему комірку стану, GRU об'єднує механізми коротко- та довготривалої пам'яті у єдиний прихований стан h_t, а замість трьох гейтів LSTM (забуття, вхідний та вихідний) застосовує лише два, а саме update gate та reset gate. Оновлений підхід робить архітектуру структурно простішою, порівняно простішою у процесі навчання та більш придатною для задач, що передбачають високу швидкість навчання, зменшення складності за пам'яттю чи застосування моделей на обмежених обчислювальних ресурсах.

Механізм роботи GRU ґрунтується на двох гейтах, які керують тим, яку інформацію необхідно зберегти, інтегрувати або відкинути під час переходу від одного часового кроку до іншого. Update gate визначає ступінь збереження попереднього прихованого стану h_{t-1} та контролює швидкість оновлення внутрішнього представлення. Через використання сигмоїдальної функції активацію update gate набуває значень у діапазоні $[0;1]$, що дозволяє регулювати баланс між новою та попередньою інформацією: значення близьке до 1 зберігає стан пам'яті, тоді як значення, близьке до 0, сприяє її оновленню. Reset gate модулює вплив попереднього стану на формування нового вмісту прихованого шару. Якщо reset gate наближається до 0, модель забуває попередній контекст, що дозволяє сформувати оновлений вектор стану переважно на основі поточного входу x_t. У випадках, коли reset gate наближається до 1, попередній стан включається в обчислення, що дозволяє ефективно моделювати короткі та середньострокові залежності.

Однією з ключових особливостей GRU є те, що вона не має окремого стану комірки, а всі операції здійснюються лише над прихованим станом h_t. Менша кількість параметрів також зменшує ризик перенавчання, що є важливою перевагою у задачах прогнозування часових рядів, де тренувальні вибірки можуть бути обмеженими або мати значну нерівномірність, що було висвітлено у роботах [9] та [10].

Для покращення ефективності прогнозування значень часових рядів було запропоновано реалізувати ансамблевий алгоритм у контексті архітектури LSTM та GRU. Ансамблеві алгоритми посідають центральне місце в сучасному машинному навчанні завдяки можливості поєднувати результати декількох моделей з метою підвищення точності, стійкості та здатності системи до узагальнення. Ансамбль працює ефективніше за окрему модель за умови, що його компоненти є достатньо різнорідними та допускають помилки різної природи. Основна ідея ансамблювання полягає у використанні колективного рішення групи моделей, яке мінімізує ймовірність систематичної похибки, властивої кожній окремій моделі, і забезпечує більш стабільну поведінку в ситуаціях шумних, неоднорідних або нестабільних даних.

Процес побудови ансамблевого алгоритму включає кілька послідовних етапів. Першим є вибір множини базових моделей, які повинні бути достатньо різними за архітектурою, функціональною формою або гіперпараметрами. Далі здійснюється незалежне навчання базових моделей на спільному наборі даних або на його різних модифікованих вибірках, що забезпечує різноманітність моделей та мінімізує їхню корельованість. Після цього відбувається побудова механізму об'єднання: для регресійних задач це може бути арифметичне середнє, зважене середнє, медіана або результат метарегресора, для класифікації – голосування чи softmax-агрегація. Важливою складовою є оптимізація ваг у зважених ансамблях, що може здійснюватися методами крос-валідації, стохастичної

оптимізації або байєсівського налаштування гіперпараметрів, що було розкрито у статті [11].

Розглянемо деталізоване створення ансамблю моделей:

1. Підготовка даних. Сформувати навчальну, валідаційну та тестову вибірки, забезпечивши коректний часовий поділ без перемішування. За потреби виконати нормалізацію або стандартизацію часових рядів, а також перетворення у формат ковзних вікон:

$$x_t = (x_{t-w+1}, \ldots, x_t), \quad (4)$$

де x_t – значення часового ряду.

2. Вибір множини базових моделей. Задати набір моделей $\{M_1, M_2, \ldots, M_K\}$, який включає архітектури різних типів або однакові моделі з різними гіперпараметрами та ініціалізаціями.

3. Налаштування гіперпараметрів базових моделей. Для кожної моделі M_K визначити структуру (кількість шарів та нейронів у кожному, функції активації) та гіперпараметри навчання (оптимізатор, швидкість навчання, розмір батчу, кількість епох, критерії зупинки).

4. Незалежне навчання базових моделей. Для кожної моделі M_K виконати процедуру навчання на одній і тій самій або на модифікованих навчальних даних, фіксуючи її параметри θ_K після досягнення найкращої якості на валідаційній вибірці.

5. Оцінювання якості базових моделей. Для кожної моделі M_K обчислити значення обраних метрик (MAE, MSE, SMAPE) на навчальному, валідаційному та тестовому наборах даних. За результатами оцінки зафіксувати якість моделей та, за потреби, відкинути явно деградуючі моделі.

6. Формування ансамблевих прогнозів на валідації. Для кожного спостереження валідаційної вибірки отримати вектор прогнозів:

$$\hat{y}_t^{(k)} = M_k(x_t) \quad (5)$$

де x_t – вектор вхідних ознак.

7. Вибір схеми агрегації. Обрати метод комбінування прогнозів базових моделей. Просте середнє:

$$\hat{y}_t^{ens} = \frac{1}{N}\sum_{k=1}^{N} \hat{y}_t^{(k)} \quad (6)$$

8. **Фінальне оцінювання ансамблю**. Застосувати обраний механізм агрегації до прогнозів на тестовій вибірці та обчислити метрики, що використовувалися для базових моделей. Порівняти якість ансамблю із найкращою індивідуальною моделлю, фіксуючи досягнутий приріст точності чи стабільності прогнозування.

Для програмної реалізації експериментів дослідження було обрано мову програмування Python та фреймворк машинного навчання TensorFlow. Такий вибір зумовлений широкими можливостями цих інструментів у побудові та навчанні нейронних мереж, зокрема моделей LSTM і GRU, що вже попередньо реалізовано в бібліотеці keras. TensorFlow підтримує GPU та TPU прискорювачі тензорних обчислень, дозволяє контролювати значення метрик під час навчання. Крім того, Python має розвинену екосистему бібліотек, таких як NumPy, pandas, scikit-learn і Matplotlib, що суттєво спрощують підготовку даних, проведення експериментів і візуалізацію результатів, забезпечуючи відтворюваність та гнучкість дослідження.

На рис. 1 наведено лістинг коду, написано для створення та навчання нейронної мережі на основі архітектури LSTM.

```
model_lstm = tf.keras.Sequential([
    tf.keras.layers.Input(shape=(WINDOW, 1)),
    tf.keras.layers.LSTM(32, return_sequences=False),
    tf.keras.layers.Dense(HORIZON)
])

model_lstm.compile(optimizer=tf.keras.optimizers.Adam(1e-3),
                   loss="mse")

save_best_only = tf.keras.callbacks.ModelCheckpoint(
    '/content/results/lstm_model.keras',monitor='val_loss',save_best_only=True
)

logger = tf.keras.callbacks.CSVLogger('/content/results/lstm_training_logs.csv',separator=',')

history = model_lstm.fit(x_train, y_train, validation_data=(x_val, y_val),
                         epochs=30, batch_size=32, verbose=1, callbacks=[save_best_only,logger])
```

```
... Epoch 1/30
1153/1153 ━━━━━━━━━━━━━━━━━━━━ 7s 5ms/step - loss: 0.0109 - val_loss: 1.2827e-04
```

Рис. 1. Розробка та навчання нейронної мережі на основі LSTM

Набір даних «Weather Long-Term Time Series Forecasting» представляє собою багатовимірний часовий ряд, що містить регулярні метеорологічні вимірювання за тривалий проміжок часу. Він включає широкий набір атмосферних показників: температуру повітря, вологість, швидкість вітру, інтенсивність опадів, радіацію та інші змінні, які дозволяють комплексно описати стан атмосфери [12].

Кожен запис у наборі відповідає конкретному часовому моменту і формує багатовимірний вектор ознак, що забезпечує представлення як одиничних, так і взаємозалежних метеоумов. Така структура дає змогу моделювати як тимчасові тренди, так і сезонні коливання, а також враховувати кореляції між різними атмосферними змінними.

Для підготовки даних типовими кроками є очищення від пропущених або аномальних значень, нормалізація ознак, а також формування вікон для навчання часових моделей. У наборі зберігається 52560 записів, що було розподілено на тренувальну, валідаційну та тестову вибірку у співвідношенні 70%:15%:15%, відповідно.

У табл. 1 наведено архітектуру моделей глибокого навчання, які використовувались для порівняння ефективності навчання за відповідними метриками із ансамблем більш простих базових моделей.

Таблиця 1

Глибокі моделі на базі різних нейромережевих архітектур

	Модель №1	Модель №2	Модель №3	Модель №4
Шар мережі	LSTM	GRU	LSTM+ GRU	LSTM+ GRU
1-й	LSTM(128)	GRU(64)	LSTM(64)	LSTM(128)
2-й	LSTM(64)	GRU(32)	GRU(64)	GRU(128)
3-й	LSTM(32)	Dense(1)	LSTM(32)	LSTM(64)
4-й	Dense(1)		GRU(32)	GRU(64)
5-й			Dense(1)	LSTM(32)
6-й				GRU(32)
7-й				Dense(1)
К-сть параметрів навчання	128417	22305	60641	258785

Модель №1 являє собою класичну багатошарову LSTM-архітектуру, що складається з трьох послідовних LSTM-шарів зі зменшенням кількості елементів на кожному шарі, після чого розміщено вихідний Dense-шар. Модель має 128417 параметрів, що робить її помірно складною та потенційно стійкою до недонавчання.

Модель №2 є компактнішою і базується на двошаровій GRU-структурі, де використано GRU(64) та GRU(32), після чого передбачено фінальний Dense-шар. GRU-архітектура характеризується меншою кількістю параметрів і швидшим навчанням, що відображено у значно меншій загальній кількості параметрів – лише 22305.

Модель №3 представляє гібридний варіант LSTM та GRU, що поєднує LSTM(64), GRU(64), LSTM(32) і GRU(32) перед вихідним Dense-шаром. Комбінація двох типів рекурентних блоків дає можливість одночасно моделювати як довготривалі, так і короткострокові часові залежності.

Модель №4 є найскладнішою, включаючи LSTM та GRU-шари з більшим числом нейронів (128 і 64), що формує багаторівневу гібридну структуру, а також містить 258785 параметрів навчання.

У табл. 2–4 наведено результати навчання цих моделей у контексті оцінки метрик якості.

Таблиця 2

Оцінка ефективності моделей за метрикою MAE

MAE	Модель №1	Модель №2	Модель №3	Модель №4
train	0,18	0,1618	0,1541	0,148
valid	0,223	0,1343	0,13	0,12
test	0,1438	0,121	0,104	0,1

Модель №3, яка поєднує LSTM і GRU, демонструє стабільне покращення якості: її MAE на тестовій вибірці становить 0,104, що вказує на здатність гібридної архітектури краще узагальнювати складні часові залежності. Найнижчу похибку отримано для моделі №4, кількість параметрів якої в 4 рази більша у порівнянні з моделлю №3. З одного боку, це означає, що збільшення кількості шарів і використання більш глибокої комбінації LSTM+GRU позитивно впливає на точність прогнозування. Однак також варто зазначити, що збільшення кількості шарів та нейронів шару сприяло лише незначному покращенню результатів у порівнянні з менш глибокою архітектурою.

Таблиця 3

Оцінка ефективності моделей за метрикою MSE

MSE	Модель №1	Модель №2	Модель №3	Модель №4
train	0,063	0,054	0,052	0,05
valid	0,073	0,036	0,035	0,032
test	0,034	0,027	0,023	0,021

Модель двошарової GRU показує суттєве зниження MSE, зокрема на валідації та тесті, що підтверджує ефективність GRU у компактніших конфігураціях. Модель №3, що поєднує LSTM і GRU, демонструє подальше покращення результатів, досягаючи MSE рівному 0,023 на тестовій вибірці. Поглиблена комбінована модель LSTM+GRU демонструє також незначно покращені результати відносно неглибокої комбінованої моделі.

Таблиця 4

Оцінка ефективності моделей за метрикою SMAPE

SMAPE, %	Модель №1	Модель №2	Модель №3	Модель №4
train	3,8	3,35	2,96	2,82
valid	2,13	2,02	1,892	1,752
test	13,14	10,35	8,15	8,13

На етапі тренування найвищу похибку демонструє модель №1 (3,8%), тоді як моделі №2–№4 характеризуються нижчими значеннями, що свідчить про кращу здатність GRU і гібридних архітектур адаптуватися до даних. На валідаційній вибірці різниця між моделями посилюється: SMAPE моделі №1 становить 2,13%, тоді як моделі №3 та №4 показують суттєво нижчі значення (1,892% та 1,752% відповідно), що вказує на їхню вищу здатність до узагальнення.

Найбільш показовими є результати на тестовій вибірці, де SMAPE моделі №1 сягає 13,14%, що вказує на значну нестабільність та нижчу точність у прогнозуванні нових даних. Моделі №2 та №3 демонструють суттєве зменшення похибки у межах 10,35% та 8,15% відповідно. Модель №4 показує найнижче значення серед усіх архітектур (8,13%), хоча її перевага над моделлю №3 незначна. Загалом результати свідчать, що гібридні архітектури LSTM+GRU забезпечують найкращу якість прогнозування згідно зі SMAPE, особливо на реальних, невідомих даних.

Таблиця 5

Глибокі моделі на базі різних нейромережевих архітектур

	Модель №1	Модель №2	Модель №3	Модель №4
Шар мережі	LSTM	LSTM	GRU	GRU
1-й	LSTM(32)	LSTM (32)	GRU(32)	GRU(32)
2-й	Dense(1)	Dense(1)	Dense(1)	Dense(1)
К-сть параметрів навчання	13157	13157	10181	10181

У порівнянні з попередніми моделями, зокрема гібридними архітектурами, які містили до 258 тисяч параметрів, такі спрощені моделі мають у 20–25 разів нижчу параметричну потужність. Вони демонструватимуть значно вищу швидкість навчання і менший ризик перенавчання, проте водночас будуть менш здатними моделювати складні, нелінійні та довгострокові залежності в часових рядах. З огляду на це, такі моделі можуть бути придатними для таких експериментів, як побудова ансамблів попереднього аналізу або задач із простою динамікою.

Таблиця 6

Оцінка ефективності ансамблю моделей за метриками якості

	MAE	MSE	MAPE, %
train	0,145	0,04	2,76
test	0,0953	0,021	7,66
validation	0,116	0,031	1,7

Наведена таблиця демонструє, що ансамблевий алгоритм забезпечує найкращі результати серед усіх протестованих підходів, що підтверджується низькими значеннями MAE, MSE та MAPE на всіх етапах оцінювання. Значення MAE = 0,0953 та MSE = 0,021 на тестовій вибірці нижчі, ніж у будь-якої з індивідуальних моделей, включно з гібридними архітектурами. Це свідчить про вищу здатність ансамблю до узагальнення та його стійкість до шумів, локальних аномалій та коливань у часовому ряду. Аналогічна ситуація спостерігається і на тестовій вибірці, де ансамбль демонструє найнижчий рівень MAPE (7,6%), що вказує на високу точність відносних прогнозів.

Перевага ансамблю є закономірною, оскільки він поєднує властивості різних моделей, кожна з яких виявляє різні типи часових залежностей. LSTM більш ефективно моделює довготривалі патерни, тоді як GRU швидше реагує на короткострокові зміни та володіє кращою узагальнювальністю за меншої кількості параметрів. Гібридні моделі поєднують ці переваги, але все одно залишаються обмеженими своєю архітектурою. Ансамбль дозволяє інтегрувати прогнози всіх архітектур, що зменшує дисперсію помилок та компенсує індивідуальні слабкі сторони кожної моделі.

Крім того, ансамбль мінімізує ризик перенавчання, оскільки помилки окремих моделей мають різну природу і з малою ймовірністю збігаються на одних і тих самих прикладах. Завдяки цьому агрегований прогноз стає згладженим та більш стабільним. Наявність значного покращення на тестовій вибірці підтверджує, що ансамблевий підхід краще відтворює реальну поведінку часового ряду та має кращі здатності до прогнозування в умовах, коли окремі моделі можуть допускати систематичні або специфічні для архітектури помилки.

Висновки

Одним із ключових результатів є демонстрація того, що архітектури LSTM та GRU, незважаючи на спільну належність до класу рекурентних нейронних мереж, виявляють різні рівні чутливості до структури часових залежностей. LSTM краще зберігає довгострокові шаблони, тоді як GRU завдяки своїй компактності та меншій кількості параметрів виявляє високу ефективність у моделюванні короткострокової динаміки та швидше пристосовується під час навчання. Гібридні архітектури, що комбінують LSTM та GRU, показали здатність успадковувати переваги обох підходів, а результати саме гібридних моделей продемонстрували істотне зниження MAE, MSE та SMAPE порівняно з окремими моделями.

Однак, попри те що найглибші з протестованих моделей у певних випадках досягали кращих результатів, спостерігалось і те, що збільшення параметричної

складності не завжди дає пропорційний приріст точності. Зокрема, найглибша комбінована модель хоча й демонструвала найнижчі значення похибок серед окремих архітектур, однак покращення порівняно з менш глибокою гібридною моделлю було незначним. Такий результат свідчить, що для задач прогнозування часових рядів надмірне збільшення кількості параметрів може не призвести до відповідного зростання якості прогнозу, особливо коли дані містять шум, варіації або нестійкі компоненти.

Найсуттєвіший приріст точності було отримано за рахунок застосування ансамблевого підходу. Ансамбль моделей продемонстрував значно менші значення MAE, MSE та MAPE порівняно з усіма індивідуальними моделями, включно з гібридними архітектурами. Такий результат є цілком очікуваним з огляду на природу ансамблевих алгоритмів, які зменшують дисперсію помилок шляхом агрегування різних прогнозів, тим самим згладжуючи систематичні похибки окремих моделей.

Отримані результати свідчать не лише про переваги ансамблів над окремими моделями, але й про здатність простих архітектур діяти як ефективні складові ансамблю. Незважаючи на те, що моделі з мінімальною глибиною та кількістю параметрів демонструють суттєво гірші результати у порівнянні з глибокими нейромережами в індивідуальному режимі, вони роблять значимий внесок у підсумковий ансамблевий прогноз. Таким чином, ефективність ансамблю залежить не лише від складності базових моделей, але й від їхньої здатності робити незалежні помилки, що знижує корельованість прогнозів.

Отримані результати вписуються у загальну тенденцію сучасних досліджень, відповідно до якої ансамблеві підходи є широковживаним підходом до підвищення точності прогнозування в задачах машинного навчання. Такі методи особливо важливі для моделювання часових рядів, де дані характеризуються складною структурою, нелінійністю та наявністю кількох компонент, серед яких можна відзначити тренд, сезонність, циклічність. Ансамблеві підходи не лише підвищують точність прогнозування, але й забезпечують стійкість до нестабільності даних, що критично важливо для галузей, де помилка прогнозу може мати значні економічні або технічні наслідки.

Загалом проведене дослідження дозволяє стверджувати, що ансамбль нейронних моделей на основі LSTM та GRU є ефективним і доцільним підходом для прогнозування часових рядів, значно перевершує індивідуальні архітектури та демонструє високу здатність до узагальнення. У майбутніх роботах планується дослідження методів оптимізації вагових коефіцієнтів ансамблю, застосування стохастичного вибору моделей, оцінка доцільності впровадження авторегресивних методів статистичного аналізу, використання моделей на базі архітектури Transformer з метою покращення узагальнюючої здатності комп'ютерних моделей.

Список літератури

[1] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A, 379(2194), 20200209.

[2] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS one, 13(3), e0194889.

[3] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

[4] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. arXiv preprint arXiv:1911.09512.

[5] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. Ph ilosophical Transactions of the Royal Society A, 379(2194), 20200209.

[6] Єріна, А., & Мазуренко, О. (2022). Статистичний аналіз часових рядів: навчальний посібник. Київ: КНУШ.

[7] Vennerød, C. B., Kjærran, A., & Bugge, E. S. (2021). Long short-term memory RNN. arXiv preprint arXiv:2105.06756.

[8] Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomputing, 323, 203-213.

[9] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[10] Pirani, M., Thakkar, P., Jivrani, P., Bohara, M. H., & Garg, D. (2022, April). A comparative analysis of ARIMA, GRU, LSTM and BiLSTM on financial time series forecasting. In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-6). IEEE.

[11] Choi, J. Y., & Lee, B. (2018). Combining LSTM network ensemble via adaptive weighting for improved time series forecasting. Mathematical problems in engineering, 2018(1), 2470171.

[12] King, A. (2023). Weather Long-Term Time Series Forecasting [Data set]. Kaggle. https://www.kaggle.com/datasets/alistairking/weather-long-term-time-series-forecasting

Надійшла до редколегії 24.09.2025

УДК 004.9 DOi 10.30837/ bi.2025.2(103).03

Н. С. Мірошниченко[1], І. Г. Перова[2]

[1]ХНУРЕ, м. Харків, Україна, nelia.miroshnychenko@nure.ua,
ORCID iD: 0000-0002-3846-1668

[2]ХНУРЕ, м. Харків, Україна, rikywenok@gmail.com,
ORCID iD: 0000-0003-2089-5609

ОПТИМІЗАЦІЯ ЗМЕНШЕННЯ РОЗМІРНОСТІ МЕДИЧНИХ ДАНИХ ІЗ ЗАСТОСУВАННЯМ МОДИФІКОВАНОГО АВТОЕНКОДЕРА

У статті розглядається оптимізація процесу зменшення розмірності вибірок медичних даних із застосуванням модифікованого автоенкодера. Запропонований підхід передбачає попередню обробку даних через автоенкодер для виділення найбільш інформативних ознак. Отримані скорочені представлення надалі обробляються адаптивним нейро-фаззі методом із динамічним коефіцієнтом підсилення цільових векторів, що забезпечує результати для подальшого аналізу та класифікації.

У роботі наведено математичне формулювання алгоритму, описано модифікації автоенкодера, спрямовані на підвищення точності відновлення даних та зменшення інформаційних втрат під час редукції розмірності. Проведено експериментальне дослідження на медичних наборах даних, що демонструє ефективність запропонованого методу.

ЗМЕНШЕННЯ РОЗМІРНОСТІ, NEURAL NETWORK, МОДИФІКОВАНИЙ АВТОЕНКОДЕР, МЕДИЧНІ ДАНІ, MACHINE LEARNING

N.S. Miroshnychenko, I.G., Perova. Optimization of Medical Data Sample Dimensionality Reduction Using a Modified Autoencoder. The article examines the optimization of the dimensionality reduction process for medical data samples using a modified autoencoder. The proposed approach involves preliminary data processing through the autoencoder to extract the most informative features. The resulting reduced representations are subsequently processed by an adaptive neuro-fuzzy method with a dynamic target vector amplification coefficient, which provides outputs for further analysis and classification.

The paper presents the mathematical formulation of the algorithm and describes the modifications made to the autoencoder to improve data reconstruction accuracy and reduce information loss during dimensionality reduction. An experimental study conducted on medical datasets demonstrates the effectiveness of the proposed method.

DIMENSIONALITY REDUCTION, NEURAL NETWORK, MODIFIED AUTOENCODER, MEDICAL DATA, MACHINE LEARNING

Вступ

У сучасних медичних інформаційних системах накопичуються великі обсяги даних, що містять десятки або навіть сотні ознак, пов'язаних зі станом пацієнтів, результатами лабораторних досліджень, медичними зображеннями та біомедичними сигналами. Саме висока розмірність вибірок даних ускладнює ефективну обробку, аналіз та інтерпретацію даних, спричиняючи проблему так званого «прокляття розмірності» [1]. Традиційні методи зменшення розмірності, а саме: метод головних компонент (PCA) та лінійний дискримінантний аналіз (LDA), не завжди здатні забезпечити збереження нелінійних залежностей та повноти інформації, характерних для медичних даних [2].

Автоенкодери, як один із типів штучних нейронних мереж, широко використовуються для розв'язання задач нелінійного зменшення розмірності даних [3]. Проте традиційні моделі автоенкодерів не завжди забезпечують збереження суттєвої діагностичної інформації та часто характеризуються недостатньою точністю відновлення даних. У зв'язку з цим зростає потреба в удосконаленні їхньої архітектури та інтеграції з адаптивними інтелектуальними методами для підвищення ефективності обробки даних і покращення результатів подальшої класифікації.

У даній статті запропоновано підхід до оптимізації процесу зменшення розмірності медичних даних шляхом використання модифікованого автоенкодера у поєднанні з адаптивним нейро-фаззі методом із динамічним коефіцієнтом підсилення цільових векторів. У статті подано математичне представлення автоенкодера та розглянуто особливості його взаємодії з адаптивним нейро-фаззі методом. Проведено експериментальні дослідження на реальній вибірці медичних даних з метою оцінювання ефективності запропонованого підходу.

Отримані результати підтверджують доцільність використання модифікованого автоенкодера для зменшення розмірності даних зі збереженням діагностично значущої інформації та мінімальними втратами інформативності.

1. Метод зменшення розмірності на основі модифікованого автоенкодера

Для реалізації комбінованого підходу до зменшення розмірності першочергово необхідно розглянути математичну основу модифікованого автоенкодера. Саме він забезпечує отримання інформативного компактного представлення медичних даних, яке надалі обробляється нейро-фаззі методом.

Автоенкодери – це нейронні мережі, які здійснюють зменшення розмірності даних шляхом навчання нижчовимірного представлення $z_{pre}(m)$ з початкового високовимірного вхідного вектора $x_{pre}(m)$ та подальшої його реконструкції у вигляді $x'_{pre}(m)$.

1. Функція кодування:

Кожен вхідний високовимірний вектор $x_{pre}(m)$ перетворюється у приховане (латентне) представлення $z_{pre}(m)$. Лінійне перетворення може бути подане у вигляді вагової матриці W_{enc} та вектора зміщення b_{enc}:

$$z_{pre}(m) = \sigma(W_{enc} x_{pre}(m) + b_{enc}) \quad (1)$$

де σ – є нелінійною функцією активації, яка може бути як лінійною (наприклад, тотожне відображення), так і нелінійною (наприклад, сигмоїда, гіперболічний тангенс).

2. Функція декодування:

Приховане представлення $z_{pre}(m)$ відображається назад у наближення $x'_{pre}(m)$ до початкових даних за допомогою:

$$x'_{pre}(m) = W_{dec} z_{pre}(m) + b_{dec}) \quad (2)$$

де W_{dec} – матриця ваг декодера, а b_{dec} – вектор зміщення декодера.

3. Функція втрат:

Функція втрат визначається як середньоквадратична помилка між вхідними даними, представленими вектором $x_{pre}(m)$, та реконструйованими даними $x'_{pre}(m)$ для навчання. Функція втрат задається наступним чином:

$$L = \frac{1}{K}\sum_{m=1}^{K} x_{pre}(m) - x'_{pre}(m)^2 \quad (3)$$

де K – загальна кількість пацієнтів [4].

4. Оптимізація:

Необхідно мінімізувати функцію втрат L, для чого застосовуються методи оптимізації, зокрема стохастичний градієнтний спуск. Важливим кроком є обчислення градієнтів функції втрат $L(x_{pre}(m), x'_{pre}(m))$ по відношенню до параметрів $W_{enc}, b_{enc}, W_{dec}, b_{dec}$. Градієнти обчислюються із застосуванням методу зворотного поширення помилки. Зокрема, градієнт функції втрат щодо параметрів декодера можна записати у такому вигляді:

$$\nabla_{x'_{pre}(m)} L(x_{pre}(m), x'_{pre}(m) = 2(x'_{pre}(m) - (x_{pre}(m))$$

$$\nabla_{W_{dec}} L(x_{pre}(m), x'_{pre}(m) = \nabla_{x'_{pre}(m)} L((x_{pre}(m), x'_{pre}(m)) z_{pre}(m)^T$$

$$\nabla_{b_{dec}} L(x_{pre}(m), x'_{pre}(m) = \nabla_{x'_{pre}(m)} L((x_{pre}(m), x'_{pre}(m)) \quad (4)$$

Градієнт функції втрат щодо параметрів енкодера можна представити у вигляді:

$$\nabla_{z_{pre}(m)} L(x_{pre}(m), x'_{pre}(m) = W_{dec}^T \nabla_{x'_{pre}(m)} L(x_{pre}(m), x'_{pre}(m)$$

$$\nabla_{W_{enc}} L(x_{pre}(m), x'_{pre}(m) = \nabla_{z_{pre}(m)} L((x_{pre}(m), x'_{pre}(m))(x_{pre}(m)^T$$

$$\nabla_{b_{enc}} L(x_{pre}(m), x'_{pre}(m) = \nabla_{z_{pre}(m)} L((x_{pre}(m), x'_{pre}(m)) \quad (5)$$

Після обчислення градієнтів функції втрат, наступним кроком необхідно оновити параметри в напрямку зменшення функції втрат. Для кожного параметра Q, де Q може бути $W_{enc}, b_{enc}, W_{dec}, b_{dec}$, оновлення виконується згідно з формулою:

$$Q \leftarrow Q - \alpha \nabla_Q L((x_{pre}(m), x'_{pre}(m)) \quad (6)$$

де α – швидкість навчання.

Зазначений алгоритм виконується для кожного навчального прикладу $x_{pre}(m)$ із тренувального набору протягом кількох епох навчання. Епохою називають один повний прохід через весь навчальний набір даних. У процесі багаторазових ітерацій (епох) значення функції втрат зазвичай поступово зменшується, що свідчить про те, що модель успішно навчається. Після завершення навчання отримане латентне (закодоване) представлення $z_{pre}(m)$ може бути використане для виконання класифікаційних задач [5].

2. Нейро-фаззі методом із динамічним коефіцієнтом підсилення цільових векторів

На відміну від стандартних нейронних мереж, даний метод поєднує механізми векторного представлення, нечітких функцій належності та конкурентного навчання. Завдяки такій комбінації модель ефективно класифікує дані навіть тоді, коли окремі об'єкти можуть належати до кількох класів одночасно.

Метод працює з набором еталонних векторів, по одному для кожної категорії. Кожен вхідний вектор порівнюється з цими еталонними представниками, після чого визначається ступінь його належності до кожного класу за допомогою спеціальних функцій належності. Вони дозволяють не лише визначити найближчий клас, але й оцінити рівень нечіткості у прийнятті рішення. Це, у свою чергу, зменшує вплив шуму та покращує роботу з реальними медичними даними, які часто є неоднорідними [6].

Навчання мережі відбувається поетапно. На кожній ітерації еталонні вектори коригуються відповідно до того, чи відповідає прогнозований клас фактичному. У разі правильної класифікації еталонний вектор наближається до вхідного, підвищуючи точність його представлення. Якщо ж прогноз помилковий, модель коригує еталонні вектори так, щоб збільшити відстань між класами. Різні правила оновлення для вектора-переможця та всіх інших забезпечують конкурентну взаємодію між класами й сприяють покращенню точності класифікації.

Ключову роль у роботі методу відіграє динамічний коефіцієнт підсилення, який визначає інтенсивність навчання. У стандартному підході його значення поступово зменшується від епохи до епохи. Натомість модифікована версія використовує рекурсивний

механізм обчислення цього коефіцієнта, що враховує як поточний стан мережі, так і накопичену інформацію з попередніх кроків. Це дозволяє здійснювати оновлення еталонних векторів більш плавно та адаптивно, що є суттєвою перевагою для медичних даних з їх складною структурою та високою варіативністю [7].

На виході метод формує мітку класу, який є найбільш ймовірним для поточного вхідного вектора. Використання нечітких функцій належності робить алгоритм стійким до варіативності у вибірці та дозволяє коректно працювати навіть тоді, коли межі між класами частково накладаються.

Детальний математичний опис представленого методу наведений в нашій попередній публікації [8].

3. Експериментальні дослідження

У попередніх розділах розглянуто класичний та модифікований автоенкодери для зменшення розмірності великої вибірки даних, а також нейро-фаззі метод для класифікації даних. З метою перевірки ефективності розглянутих підходів виконано серію експериментальних досліджень на реальних медичних даних.

В якості вибірки медичних даних було обрано медичну вибірку, що включає психофізіологічні дані пацієнтів, зібрані в рамках скринінгу на професійне вигорання. Вибірка містить результати стандартизованих психодіагностичних тестів:

1. ERI (Effort-Reward Imbalance) – оцінює дисбаланс між зусиллями, витраченими на роботу, та винагородою.

2. MBI (Maslach Burnout Inventory) – вимірює рівень емоційного вигорання, деперсоналізації та особистих досягнень.

3. FBI (Freudenberger Burnout Inventory) – визначає ступінь професійного вигоряння на ранніх стадія [9].

Вибірка дослідження складається з 228 пацієнтів. Для кожного учасника було зібрано від 200 до 269 ознак, включно з демографічними, клінічними та психодіагностичними показниками.

Для подальшої обробки та аналізу було вирішено сфокусуватися на показниках тесту Maslach Burnout Inventory (MBI), оскільки він дозволяє кількісно оцінити рівень професійного вигорання. MBI включає три підшкали:

1. Emotional Exhaustion (EE) – емоційне виснаження, що характеризує втому та зниження емоційних ресурсів;

2. Depersonalization (DP) – деперсоналізація, що відображає цинічне або відчужене ставлення до пацієнтів;

3. Personal Accomplishment (PA) – зниження особистих досягнень, що показує суб'єктивне відчуття власної професійної ефективності [10].

Після відбору лише MBI-показників загальна кількість ознак була зменшена до 25, що зробило подальшу обробку даних більш ефективною та дозволило сфокусуватися на ключових індикаторах вигорання.

Кожне питання оцінюється за шкалою Лайкерта від 0 до 6, де вищі бали для EE та DP означають більший рівень вигорання, а для PA – нижчі бали вказують на зниження почуття досягнень.

Для подальшого аналізу сумарні бали кожної підшкали було обчислено для кожного учасника, після чого кожна підшкала була категоризована у три рівні вигорання: низький, середній та високий. На основі комбінації підшкал формувався загальний показник вигорання (Result_raw), а потім учасники були розподілені у три підсумкові групи вигорання (Result_MBI). Такий підхід дозволяє врахувати одночасний стан усіх підшкал та коректно класифікувати рівень вигорання пацієнтів, навіть якщо межі між групами частково накладаються.

Аналіз розподілу пацієнтів за рівнем професійного вигорання, проведений за результатами тесту MBI, показав, що більшість працівників перебувають у групі з високим рівнем вигорання. Зокрема:

– 54,8 % (125 осіб) належать до групи з високим рівнем вигорання (група 2);

– 24,6 % (56 осіб) – до групи із середнім рівнем вигорання (група 1);

– лише 20,6 % (47 осіб) мають низький рівень вигорання (група 0).

Такий розподіл свідчить про значне поширення синдрому емоційного вигорання серед працівників, що підкреслює необхідність впровадження заходів профілактики та психологічної підтримки в цій професійній групі.

Для подальшого аналізу було вирішено використати окремі питання тесту MBI як вхідні ознаки моделей машинного навчання, а сформовані групи (Result_MBI) – як цільові мітки.

Це дозволяє не лише візуалізувати латентні представлення пацієнтів, а й оцінити ефективність методів зменшення розмірності та класифікації для визначення рівня вигорання.

На наступному етапі проведено експериментальні дослідження для оцінки ефективності методів зменшення розмірності та класифікації рівня вигорання.

Вибірку було випадковим чином розділено на навчальну й тестову підмножини у співвідношенні 80/20. Попередня обробка даних включала нормалізацію ознак методом StandardScaler, що забезпечує однаковий масштаб усіх вхідних значень та покращує збіжність моделей.

Для зменшення розмірності простору ознак було застосовано класичний та модифікований автоенкодери, які стискали 25 вхідних параметрів до 8-розмірного латентного простору (encoding_dim = 8).

Автоенкодери навчалися з такими параметрами: швидкість навчання $\alpha = 0.001$, кількість епох epochs_ae = 3000, обмеження ваг clip_value = 5.0. Модифікований автоенкодер додатково містив вектор зсуву β, що мав компенсувати лінійні зміщення та покращити якість представлення латентного простору.

Після навчання латентні вектори були використані як вхідні ознаки для моделі NeuroFuzzy з динамічним коефіцієнтом підсилення цільових векторів, який адаптивно підсилює внесок менш представлених класів у процесі навчання. Це забезпечує більш збалансоване навчання та підвищує точність класифікації рівня вигорання.

Для візуального аналізу структури груп пацієнтів латентні представлення після проходження через NeuroFuzzy було знижено до двох вимірів за допомогою t-SNE [11]. На рис. 1 представлено порівняння латентних просторів для класичного та модифікованого автоенкодера після обробки NeuroFuzzy.

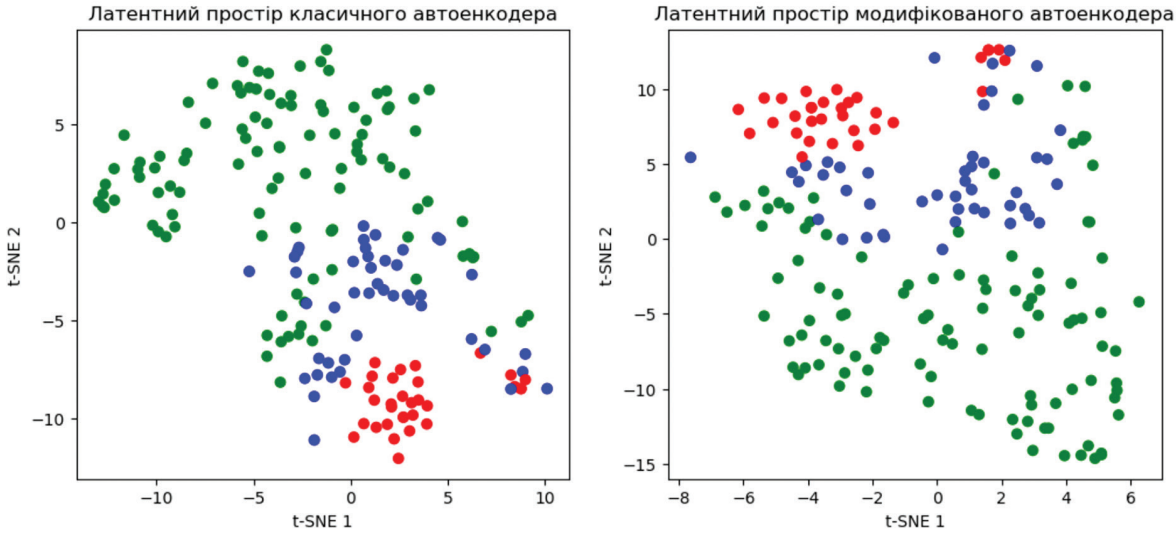

Рис. 1. Латентні представлення автоенкодерів у просторі t-SNE

Як видно з отриманих графіків Кожна точка відповідає окремому пацієнту, а колір позначає рівень вигорання за тестом MBI: зелений – високий рівень вигорання (група 2), синій – середній рівень (група 1), червоний – низький рівень (група 0).

На лівому графіку (класичний автоенкодер у поєднані з нейро-фаззі) точки трьох кольорів сильно перемішані між собою без чіткої структури. На правому графіку (модифікований автоенкодер у поєднані з нейро-фаззі) спостерігається виражене розділення на три компактні, майже неперекриваючі кластери, що точно відповідають трьом рівням вигорання. Це свідчить про те, що модифікована модель успішно розплутала рівень професійного вигорання як окремий домінуючий фактор латентного простору, тоді як класичний автоенкодер такої здатності не продемонстрував.

Для порівняння впливу попередньої обробки даних було проведено експеримент на сирих, необроблених даних без застосування зменшення розмірності та нормалізації. Дані було подано на вхід моделі NeuroFuzzy з динамічним коефіцієнтом підсилення цільових векторів, після чого отримані латентні представлення було знижено до двох вимірів за допомогою t-SNE для наочного представлення. На рис. 2 наведено розподіл пацієнтів у вихідному просторі ознак без попереднього зменшення розмірності.

Ці латентні представлення відображають структуру даних після адаптації NeuroFuzzy і дозволяють оцінити, наскільки добре модель виділяє кластери пацієнтів із різними рівнями вигорання навіть у сирому, необробленому просторі ознак.

Рис. 2. Розподіл пацієнтів у вихідному просторі ознак без попереднього зменшення розмірності

Навіть без застосування будь-яких моделей у просторі оригінальних ознак помітна виразна тенденція до групування точок відповідно до рівня вигорання: червоний кластер переважно зосереджений ліворуч, синій — у центральній частині, зелений — праворуч.

Проте між класами зберігаються зони перекриття, а окремі точки суттєво віддалені від основної маси «свого» класу. Це свідчить про те, що, попри загальну розділюваність, сирі дані містять значний шум і надлишкову інформацію, які можуть ускладнювати роботу класифікатора.

Щоб кількісно оцінити, наскільки сильно ці фактори впливають на реальну продуктивність класифікації та чи здатні різні стратегії зменшення розмірності усунути зазначені проблеми, було проведено порівняльний експеримент. Нейро-фаззі класифікатор навчався і тестувався у трьох варіантах:

1) безпосередньо на сирих даних (без попередньої обробки);

2) на латентному представленні, отриманому за допомогою класичного автоенкодера;

3) на латентному представленні, отриманому за допомогою модифікованого автоенкодера.

Результати оцінки точності (accuracy) показали наступне:

– при прямому використанні сирих даних нейро-фаззі класифікатором точність склала 0.615;

– при попередньому зменшенні розмірності класичним автоенкодером і подальшій класифікації на отриманому латентному представленні точність зросла до 0.703;

– найкращий результат – 0.850 – було отримано при використанні модифікованого автоенкодера і подальшої класифікації за допомогою нейро-фаззі метода із динамічним коефіцієнтом підсилення цільових векторів.

Таким чином, попри візуально помітну тенденцію до кластеризації у просторі сирих ознак, безпосереднє застосування нейро-фаззі класифікатора до необроблених даних дало найнижчу точність через наявність шуму, надлишкових і слабкоінформативних ознак [12]. Класичний автоенкодер забезпечив певне покращення за рахунок зменшення розмірності, однак лише модифікований автоенкодер сформував високоякісне латентне представлення, у якому три рівні професійного вигорання максимально розплутані.

Висновки

Проведені експериментальні дослідження на реальних медичних даних показали, що рівень професійного вигорання, визначений за тестом Maslach Burnout Inventory (MBI), є клінічно значущим і добре вимірюваним показником. Понад половина обстежених (54,8 %) мали високий рівень вигорання, що підкреслює актуальність розробки автоматизованих інструментів ранньої діагностики та моніторингу цього стану.

Візуалізація латентних просторів та кількісна оцінка точності класифікації дозволили зробити такі основні висновки:

– Сирі дані, що складаються з 25 окремих пунктів опитувальника MBI, вже містять виражену внутрішню структуру: три рівні вигорання (низький, середній, високий) утворюють досить чіткі кластери навіть у просторі оригінальних ознак. Проте наявність зон перекриття та значної кількості шумових і надлишкових ознак суттєво обмежує ефективність прямого застосування складних класифікаторів.

– Нейро-фаззі класифікатор із динамічним коефіцієнтом підсилення менш представлених класів, навчений безпосередньо на сирих даних, продемонстрував точність лише 61,5%. Це підтверджує, що висока розмірність і шум ускладнюють формування чітких і узагальнювальних фазі-правил.

– Використання класичного автоенкодера для зменшення розмірності з 25 до 8 вимірів дозволило підвищити точність класифікації до 70,3%. Покращення пояснюється частковим усуненням шуму та надлишкових кореляцій, однак класичний автоенкодер, оптимізований лише за критерієм відновлення, не гарантує збереження і посилення дискримінативної інформації, необхідної для розрізнення рівнів вигорання.

– Найкращий результат – точність 85,0% отримано при гібридному підході модифікованого автоенкодера в парі з нейро-фаззі методом. Модифікація автоенкодера (введення параметра β) сформувала компактне 8-вимірне латентне представлення, у якому три рівні вигорання утворюють майже ідеально розділені, компактні та неперекриваючі кластери. Саме таке високоякісне представлення дозволило нейро-фаззі системі максимально ефективно побудувати прозорі, інтерпретовні та точні правила класифікації.

Запропонована гібридна модель поєднує переваги глибокого нелінійного зменшення розмірності та інтерпретовності нейро-фаззі логіки, забезпечуючи одночасно високу точність 85%, стійкість до шуму, низьку обчислювальну складність та клінічну осмисленість отриманих рішень. Це робить її перспективним інструментом для впровадження в системи скринінгу та моніторингу професійного вигорання медичних працівників високого ризику.

Отже, модифікований автоенкодер не просто виконує стискання даних, а цілеспрямовано формує латентний простір, оптимізований під задачу діагностики вигорання. Подальше використання такого представлення нейро-фаззі класифікатором є найбільш ефективним і клінічно обґрунтованим рішенням серед усіх розглянутих підходів. Подальші дослідження доцільно спрямувати на валідацію моделі на більших та різнорідних вибірках, а також на інтеграцію розробленої системи в реальні процеси психологічного супроводу працівників.

Список літератури:

[1] Wolski, M., & Gomolińska, A. (2020). Data meaning and knowledge discovery: Semantical aspects of information systems. International Journal of Approximate Reasoning, 119, 40–57. https://doi.org/10.1016/j.ijar.2020.01.002.

[2] Ayesha, S., Hanif, M. K., & Talib, R. (2020, July). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44–58. https://doi.org/10.1016/j.inffus.2020.01.005.

[3] Turchenko, V., Chalmers, E., & Luczak, A. (2019). A DEEP CONVOLUTIONAL AUTO-ENCODER WITH POOLING – UNPOOLING LAYERS IN CAFFE. International Journal of Computing, 18(1), 8-31. https://doi.org/10.47839/ijc.18.1.1270.

[4] Miroshnychenko N., Perova I., Grebennik I., Chyhryn D Dimensionaity reduction methods for large datasets. / The 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 4-6 September, 2025, Gliwice, Poland.

[5] Miroschnichenko, N.S., Perova, I.H., Datsok, O.M. Semi-supervised learning information system for analyzing high-dimensional data samples / N.S. Miroschnichenko, I.H. Perova, O.M. Datsok // Visnyk of the National University "Lviv Polytechnic" Information Systems and Networks. – 2024. Issue 16. – pp. 133-144, https://doi.org/10.23939/sisn2024.16.133 (in Ukrainian).

[6] Sayed-Mouchaweh, M. (2020). Artificial Intelligence Techniques for a Scalable Energy Transition: Advanced Methods, Digital Technologies, Decision Support Tools, and Applications. Springer Nature.

[7] Miroshnychenko N. Analysis of methods for managing high-dimensional medical data with limited patient samples / Conference proceedings "Intelligent systems of decision-making and problems of computational intelligence (ISDMCI2024)", June 20-23, 2024. P. 29-31.

[8] Miroshnychenko N. Investigating the Management of Datasets Featuring Elevated Dimensionality and a Restricted Patient Sample. ISDMCI 2024, Vol. 1, pp. 349–370. https://doi.org/10.1007/978-3-031-70959-3_18.

[9] Yuguero, O., Hodkinson, A., Panagioti, M., Pifarre, J., & Peters, D. (2023). The public health problem of burnout in health professionals. Frontiers Media SA.

[10] Kinman, G. (2025). Maslach burnout inventory. PubMed, 74(9), 630–631. https://doi.org/10.1093/occmed/kqae116.

[11] Spiwok, V., & Kříž, P. (2020, June 30). Time-Lagged t-Distributed Stochastic Neighbor Embedding (t-SNE) of Molecular Simulation Trajectories. Frontiers in Molecular Biosciences, 7. https://doi.org/10.3389/fmolb.2020.00132.

[12] Penzenyk, A. (2024). Автоматизоване виявлення та попередження перенавчання в нейронних мережах. COMPUTER-INTEGRATED TECHNOLOGIES EDUCATION SCIENCE PRODUCTION, 54, 36–42. https://doi.org/10.36910/6775-2524-0560-2024-54-04.

Надійшла до редколегії 07.10.2025

УДК 004.75 DOi 10.30837/ bi.2025.2(103).04

К. В. Сільванович[1], О. Є. Гриньова[1], Л. Е. Чала [1], С. Г. Удовенко [2]

[1]ХНУРЕ, м. Харків, Україна, kristina.silvanovych@nure.ua, ORCID iD: 0009-0008-3723-1124

[1]ХНУРЕ, м. Харків, Україна, olena.hrynova@nure.ua, ORCID iD: 0000-0002-3367-8067

[1]ХНУРЕ, м. Харків, Україна, larysa.chala@nure.ua, ORCID iD: 0000-0002-9890-4790

[2]ХНЕУ ім. С. Кузнеця, м. Харків, Україна, serhiy.udovenko@hneu.net, ORCID iD: 0000-0001-5945-8647

НЕЙРОМЕРЕЖЕВІ ТЕХНОЛОГІЇ МОНІТОРИНГУ ТА АНАЛІЗУ РУЙНІВНИХ ПОШКОДЖЕНЬ АГРАРНИХ ДІЛЯНОК

Здійснено аналіз існуючих інтелектуальних технологій виявлення та класифікації руйнівних пошкоджень аграрних ділянок. Розроблено моделі класифікації аграрних ділянок за ступенем пошкоджень та сегментації зображень пошкоджених ділянок з використанням нейромережевої архітектури U-Net. Запропоновано прогнозну модель для визначення термінів відновлення пошкоджених ділянок з гібридним використанням архітектур TFT і LSTM та аналізу даних про стан ґрунтів і кліматичних факторів. Реалізовано інтеграцію розроблених моделей для створення інтелектуальної системи класифікації пошкоджень, сегментації уражених ділянок та прогнозування термінів рекультивації. Для реалізації системи були обрані: платформа WPF для створення зрозумілого та сучасного інтерфейсу, ONNX Runtime для ефективної роботи моделей штучного інтелекту, а також використання CSV-файлів для структурованого зберігання й обміну даними. Результати тестування підтверджують працездатність запропонованого підходу.

НЕЙРОМЕРЕЖЕВІ ТЕХНОЛОГІЇ, МОНІТОРИНГ ПОШКОДЖЕНЬ, АНАЛІЗ ЗОБРАЖЕНЬ, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, ПРОГНОЗУВАННЯ ЧАСУ ВІДНОВЛЕННЯ

K.V. Silvanovych, O.E. Hrynova, L.E. Chala, S.G. Udovenko. Neural network technologies for monitoring and analysis of destructive damage to agricultural plots. An analysis of existing intelligent technologies for detecting and classifying destructive damage to agricultural plots was carried out. Models for classifying agricultural plots by the degree of damage and segmenting images of damaged plots using the U-Net neural network architecture were developed. A predictive model was proposed for determining the terms of restoration of damaged plots by hybridising TFT and LSTM architectures, along with an analysis of data on soil states and climatic factors. The integration of the developed models was implemented to create an intelligent system for classifying damage, segmenting affected plots, and predicting the terms of reclamation. The following components were selected for the implementation of the system: the WPF platform for creating a clear and modern interface, ONNX Runtime for the efficient operation of artificial intelligence models, and the use of CSV files for structured data storage and exchange. The testing results confirm the operability of the proposed approach.

NEURAL NETWORK TECHNOLOGIES, DAMAGE MONITORING, IMAGE ANALYSIS, CONVOLUTIONAL NEURAL NETWORKS, RECOVERY TIME PREDICTION

Вступ

Внаслідок збройної агресії, розпочатої військовими силами Російської Федерації, екосистема України зазнала суттєвих руйнувань. Це, зокрема, призвело до порушення нормального стану сільськогосподарських земельних ділянок та грунтових ресурсів [1]. Виникла гостра потреба в здійсненні оперативного моніторингу та аналізу пошкоджень агроресурсів.

Сільське господарство завжди було основою продовольчої безпеки держави, а стан ґрунтових ресурсів безпосередньо впливає на ефективність аграрного виробництва. Проте останні десятиліття виявили значні природні фактори деградації ґрунтів, тобто процеси погіршення їхніх властивостей, спричинені природними умовами, такими як ерозія, виснаження поживних речовин, зміни водного балансу, а також вплив кліматичних коливань. Через це ґрунти стають більш вразливими до антропогенних факторів, зокрема інтенсивного землекористування.

Новим викликом у часи воєнних дій стає поєднання цих природних проблем із масштабними антропогенними ушкодженнями. При цьому на багатьох територіях виникають масштабні екологічні проблеми – руйнування родючого шару, забруднення важкими металами, накопичення уламків техніки й вибухових речовин. Це призводить до втрати продуктивності угідь, деградації агроландшафтів та зростання ризику для здоров'я населення [2].

Україна стикається з однією з найбільших криз забруднення земель у світі: станом на грудень 2024 року, понад 138 000 км^2 земель (близько 20% території) забруднено мінами та вибухівкою, що блокує доступ до 14 000 км^2 сільгоспугідь. Це загрожує продовольчій безпеці, оскільки агросектор становить близько 70% експорту країни.

Відновлення пошкоджених земель потребує попередньої оцінки їхнього стану, придатності до ведення сільського господарства, необхідності рекультивації та відповідності рельєфу й просторової організації умовам ефективного господарювання. Традиційні методи, такі як виїзні обстеження, відбір та лабораторний аналіз зразків, забезпечують достовірні результати, але є надто затратними за часом і ресурсами, а подекуди й неможливими через вибухонебезпечність території. Вони не дозволяють охопити велику кількість пошкоджених аграрних ділянок. Саме тому необхідні нові,

масштабовані та швидкодіючі підходи до моніторингу та аналізу пошкоджень екосистеми.

Сучасні цифрові технології, зокрема штучний інтелект, комп'ютерний зір та алгоритми глибинного навчання, відкривають нові можливості у вирішенні цієї проблеми [3]. Інтелектуальні системи, за наявності супутникових та аерофотознімків, здатні автоматично виявляти пошкодження, визначати їхні межі, класифікувати тип уражень (механічні, хімічні, термічні), ступінь пошкоджень та прогнозувати реальні терміни відновлення родючості. Зокрема, такі системи можуть відокремити ділянку, де зафіксоване забруднення, від території з повністю зруйнованим верхнім шаром ґрунту, і на основі цього сформувати карту уражень. Отримані внаслідок цього результати дозволяють ухвалювати зважені рішення (з урахуванням логістики та планування) щодо необхідності термінів рекультивації та можливості відновлення сільськогосподарських робіт.

Розробка інтегрованих систем, здатних обробляти великі обсяги даних, відкриває новий рівень ефективності у відновленні сільськогосподарських земель. Подібні системи надають аграріям, науковцям і органам влади доступ до об'єктивної та оперативної інформації, знижуючи залежність від трудомістких обстежень і людського фактору.

Важливо враховувати, що самі зображення пошкоджених ділянок є лише вхідними даними. Для їх перетворення на корисну інформацію необхідна комплексна обробка за допомогою нормалізації, фільтрації шумів, виділення ключових ознак, порівняння з еталонними станами тощо. Виконати це вручну практично неможливо, адже коли обсяги даних зазвичай обчислюються десятками тисяч знімків, кожен із них є частиною часової серії, що відображає динаміку змін стану аналізованої території. Саме тому вирішальну роль відіграють автоматизовані інтелектуальні системи, здатні навчатися на прикладах і узагальнювати нову інформацію.

Перевагою такого підходу є не лише швидкість, а й можливість переходу від фіксації проблем до плану дій. Системи на основі штучного інтелекту здатні враховувати тип пошкоджень, кліматичні та сезонні фактори, що дає змогу формувати більш реалістичні та ефективні стратегії аграрного відновлення.

Метою цієї статті є розроблення та дослідження технологій моніторингу та аналізу пошкоджень агроекосистеми, що дозволяють в подальшому реалізувати інтелектуальну систему класифікації пошкоджень, сегментації уражених ділянок та прогнозування термінів рекультивації. До завдань такої системи слід віднести можливість завантаження знімків пошкоджених ділянок, отримання карти уражень, розрахунку числових показників площі та типу пошкоджень, а також формування прогнозів щодо термінів відновлення.

Впровадження подібних технологій сприятиме підвищенню ефективності моніторингу та відбудови пошкоджених аграрних територій. Це має бути певним стратегічним внеском у продовольчу безпеку, післявоєнне відновлення та збереження природного потенціалу країни (ЦСР 2 «Подолання голоду» та ЦСР 15 «Збереження екосистем суші»).

Відповідно до поставленої мети, необхідно вирішити наступні завдання:

– аналіз існуючих інтелектуальних технологій виявлення та класифікації пошкоджень аграрних ділянок;

– розроблення технології виявлення пошкоджених ділянок з використанням супутникових знімків Sentinel-2 та нейромережевих згорткових моделей;

– розроблення моделей класифікації аграрних ділянок за ступенем пошкоджень та сегментації зображень пошкоджених ділянок з використанням нейромережевих технологій;

– розроблення прогнозної моделі для визначення термінів відновлення пошкоджених ділянок з використанням алгоритмів машинного навчання та аналізу історичних даних про стан ґрунтів та кліматичних факторів;

– інтеграція розроблених моделей для створення інтелектуальної системи класифікації пошкоджень, сегментації уражених ділянок та прогнозування термінів рекультивації.

– експериментальне дослідження запропонованих моделей та технологій.

1. Сучасні технології виявлення та класифікації пошкоджених аграрних ділянок

Розвиток цифрових технологій суттєво змінив підходи до ведення сільського господарства.

Найбільшого поширення набули методи комп'ютерного зору та глибинного навчання, зокрема згорткові нейронні мережі (CNN) [4]. Вони спроможні автоматично аналізувати супутникові та аерофотознімки, знаходити пошкоджені ділянки, сегментувати уражені зони та навіть визначати характер деградації ґрунтів. Наприклад, впровадження архітектур типу U-Net показало високу ефективність для задач сегментації сільськогосподарських територій, що дозволяють отримувати точні карти уражень.

Окрему нішу займають моделі, оптимізовані для здійснення часового прогнозування. Рекурентні нейронні мережі (RNN), модифікації LSTM (Long Short-Term Memory) та нейромережі на основі трансформерів (Temporal Fusion Transformers) можуть аналізувати не лише зображення, а й часові ряди кліматичних даних, рівень вологості ґрунтів, кількість опадів чи сезонні цикли. Це дає змогу оцінювати не тільки масштаби пошкоджень, а й орієнтовні терміни відновлення родючості.

Технологія OneSoil (Європа/США) використовує супутникові дані для моніторингу (NDVI-аналіз, виявлення стресу рослин на базі Sentinel-2/Landsat) та

зонування полів, але без фокусу на рекультивації [5]. Функціональність – формування карт врожайності та рекомендацій по добривам (з точністю сегментації ~0.75–0.85 IoU). Недолік – відсутність фокусу на воєнних пошкодженнях (не сегментує вирви чи метали).

Технологія CropIn (Індія) використовує нейромережеві моделі (CNN+RNN) для прогнозування врожаю та моніторингу шкідників з дронів/супутників, акцент на ланцюги постачань але менш адаптована до деградації [6]. Функціональність – здійснення ризик-аналізу та оптимізації посівів (з точністю моделей ~80–90%). Недолік – орієнтація на тропічний клімат, без сегментації деградації від вибухів.

Технологія Mine Action Project in Ukraine є діючим проєктом з протимінної діяльності в Україні, що реалізується UNDP у співпраці з FAO (Продовольча та сільськогосподарська організація Об'єднаних Націй) та WFP (Всесвітня продовольча програма) для тестування ефективності розмінування з використанням штучного інтелекту (ШІ). Функціональність – виявлення та обробка зображень з дронів з використанням згорткових нейронних мереж (точність виявлення >95%). Ця технологія генерує також карти забруднення та пріоритизує зони пошкоджень [7].

Для України такі технології матимуть особливе значення в умовах післявоєнної відбудови. Вони дозволять отримувати швидку та об'єктивну картину стану земель на великих територіях, що практично неможливо зробити традиційними методами. Завдяки цьому аграрії, науковці й державні структури отримають унікальний інструмент для прийняття рішень від визначення зон, придатних до посівів, до планування програм рекультивації й довгострокових стратегій відновлення агросфери.

Слід відзначити, що переважна більшість сучасних технологій виявлення та класифікації пошкоджених аграрних ділянок базується на використанні штучних нейронних мереж (ШНМ). ШНМ є одним із провідних інструментів сучасного штучного інтелекту, здатним моделювати складні нелінійні залежності та виявляти приховані закономірності у великих масивах даних. Завдяки своїй універсальності ШНМ знайшли широке застосування у задачах комп'ютерного зору та прогнозування стану середовища [8].

У контексті дослідження пошкоджених сільськогосподарських земель нейронні мережі відіграють ключову роль у двох напрямах:

– комп'ютерний зір дозволяє аналізувати зображення із дронів або супутників для виявлення та класифікації пошкоджень. Тут основними інструментами виступають згорткові нейронні мережі (Convolutional Neural Networks, CNN), що спеціалізуються на роботі з даними, які мають просторову структуру (зображення, карти, відео);

– нейромережеве прогнозування дає оцінку довгострокових наслідків руйнувань та визначення строків відновлення ґрунтів. Для цього можуть бути задіяні моделі роботи з часовими рядами – рекурентні мережі (Long Short-Term Memory, LSTM) та новітні архітектури трансформерів (зокрема Temporal Fusion Transformer-TFT) [9].

Особливістю CNN є здатність автоматично виявляти ознаки на зображеннях від простих (лінії, контури) до складних (форми вирв від вибухів, забруднені ділянки). Це робить їх надзвичайно ефективними у задачах класифікації та сегментації зображень сільськогосподарських угідь. Застосування CNN дозволяє створювати системи, які автоматично розрізняють «уражені» та «неуражені» території, а також деталізують межі пошкоджень. Базову архітектуру такої мережі наведено на рис. 1.

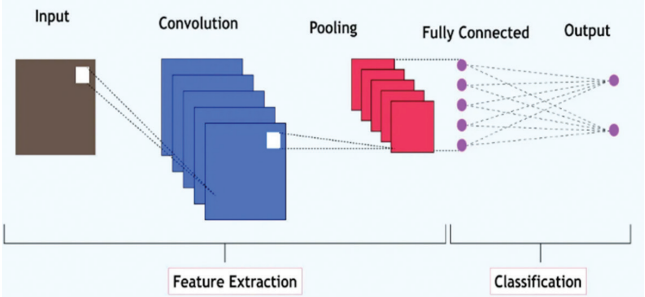

Рис. 1. Базова архітектура мережі CNN

Схематично CNN є послідовністю шарів, кожен з яких перетворює один активаційний об'єм в інший за допомогою функції, що диференціюється. Для організації згорткової нейронної мережі, яка здійснює вилучення ознак (feature extraction) та класифікацію (classification) уражених територій, застосовуються 3 основних шари: згортки (convolution), пулінгу (pooling), повнозв'язний шар (fully connected).

Виділення контурів (сегментація) пошкоджених ділянок сільськогосподарських земель може бути ефективно здійснене з застосуванням нейромережевої архітектури U-Net. Вона побудована за принципом симетричної структури «енкодер-декодер», де перша частина (енкодер) відповідає за вилучення ознак на різних рівнях, а друга (декодер) – за відновлення просторової структури та формування карти сегментації [10].

Архітектура U-Net відзначається тим, що поєднує багаторівневі ознаки завдяки так званим зв'язкам «skip connections», що дозволяє зберігати як деталі локального рівня, так і глобальний контекст. Це робить модель придатною для аналізу знімків середньої та високої роздільності, отриманих із дронів або супутників.

Позначимо вхідне зображення як $X \in R^{H \times W \times C}$, де H, W – висота та ширина зображення, C – кількість каналів.

Енкодер з L шарами (зазвичай $L = 3$), визначено за формулою:

$$E_l = f_l(E_{l-1}) \text{ для } l = 1, ..., L, \quad (1)$$

де f_l – операції згортки, нормалізації та активації у l-му шарі, $E_0 = X$.

Декодер з використанням skip-зв'язків реалізується наступним чином:

$$D_l = g_l(D_{l+1}, E_l) \text{ для } l = L-1, ..., 1, \quad (2)$$

де g_l – операції апсемплінгу, конкатенації та згортки.

Вихідна маска сегментації з використанням U-Net має наступний вигляд:

$$\hat{Y} = \sigma(D_1), \quad (3)$$

де σ – функція активації softmax.

Знімки для аналізу можуть братися як із супутників Sentinel-2 (особливо корисною є смуга B4, яка відображає зміни у рослинності), так і з дронів, що забезпечує більшу деталізацію локальних пошкоджень . Базова архітектура CNN та архітектура U-Net застосовуються для автоматизації процесу оцінки пошкоджених земель. Вони дозволяють забезпечувати швидке та точне картографування зон ураження, що значно зменшує потребу у ручному аналізі великих масивів даних.

Для прогнозування термінів відновлення ґрунтів необхідно враховувати не лише просторові характеристики, але й часову динаміку їх змін. Для цього доцільно задіювати рекурентні нейронні мережі (Recurrent Neural Networks, RNN), здатні ефективно працювати з послідовними даними. Класичні RNN мають певні обмеження, пов'язані з ефектом «згасання градієнтів», що ускладнює їх навчання на довгих часових рядах. Для вирішення цієї проблеми зазвичай використовують архітектуру LSTM, яка завдяки спеціальним осередкам пам'яті здатна зберігати інформацію на великих часових інтервалах [11]. Моделі LSTM були створені для роботи з послідовними даними, що є критично важливим у завданнях прогнозування відновлення ґрунтів. Стан ґрунту залежить не лише від масштабу руйнувань, а й від динаміки кліматичних факторів, типу сільського господарства та часу. Мережа LSTM здатна враховувати довготривалі залежності, що робить її придатною як для короткострокових, так і для довгострокових прогнозів.

Часова динаміка LSTM визначається за формулою:

$$h_t = LSTM(F_{combined}, h_{t-1}), \quad (4)$$

де h_t – прихований стан для часу t.

До системи виявлення та класифікації пошкоджених аграрних ділянок доцільно доцільно інтегрувати й нейромережеву модель TFT, архітектура якої поєднує як статичні дані (тип ґрунту, вид зброї, географічне положення), так і динамічні (опади, температурні режими, сезонні цикли). Перевагою TFT є її інтерпретованість, адже ця модель не лише передбачає строк відновлення, а й визначає, які фактори найбільше вплинули на прогноз. Це особливо важливо у системах підтримки прийняття рішень для аграрного сектору.

В мережі TFT для уваги до важливих факторів використовується поточний прогноз часу відновлення:

$$\hat{y} = TFT(h_{1:t}, StaticFeatures), \quad (5)$$

де $\hat{y}_t$ – прогноз часу відновлення для моменту t.

Наведений аналіз свідчить про перспективність комбінованого використання архітектури CNN +U-Net для задачі сегментації та архітектури CNN + LSTM + TFT для задачі прогнозування. В межах такого підходу LSTM аналізує часові ряди, пов'язані з кліматичними та сезонними змінами, тоді як TFT інтегрує додаткові статичні дані – тип ґрунту, характер ураження, географічне положення. Це дозволяє досягти більшої точності у прогнозуванні й робить систему стійкою до різнорідних вхідних даних.

2. Архітектура запропонованої системи моніторингу та аналізу пошкоджень аграрних ділянок

Архітектуру системи, що пропонується, поєднує ключові моделі для оцінки стану сільськогосподарських земель, пошкоджених унаслідок військових дій. Кожна модель відповідає за своє завдання: виявлення та класифікацію пошкоджених ділянок, сегментацію пошкоджень і прогнозування часу відновлення. Система побудована так, щоб бути простою у використанні, масштабованою та ефективною навіть за обмеженістю даних моніторингу.

В системі реалізовано трирівневу клієнт-серверну архітектуру, що дозволяє чітко розмежувати логіку користувача, бізнес-логіку та обчислювальні модулі штучного інтелекту. Такий підхід забезпечує високу модульність, спрощує підтримку, масштабування системи та можливість подальшого розширення без зміни всієї структури. Архітектура системи містить презентаційний рівень, логічний рівень та рівень моделей.

Презентаційний рівень (UI – User Interface) є інтерфейсним рівнем взаємодії користувачів із системою. У системі інтерфейс розроблено з використанням Windows Presentation Foundation (WPF) – сучасного інструменту від Microsoft, що дозволяє будувати гнучкі, масштабовані та стильні десктопні застосунки на мові C#.

Інтерфейс надає користувачам системи можливість: завантажувати зображення з локального комп'ютера; запускати процедури виявлення, класифікації та сегментації пошкоджених ділянок; здійснювати прогнозування терміну ліквідації наслідків; переглядати результати обробки; зберігати результати в зручному форматі (PDF або DOCX).

Елементи керування (кнопки, комбо-бокси, текстові блоки, зображення тощо) розміщено на формі MainWindow.xaml. Для кожного елемента передбачено обробник подій у відповідному .cs-файлі.

Логічний рівень (Business Logic Layer) відповідає за опрацювання всієї логіки взаємодії між користувачем і моделями. Він реалізований у коді MainWindow.xaml.cs, де обробляються події активації кнопок, здійснюється перевірка вхідних даних, координується виклик моделей ONNX, обробляються результати та передаються в UI, формується звіт.

Клас MainWindow містить також змінні для збереження проміжних результатів: шлях до зображення,

результат класифікації, згенерована маска та прогнозовані значення. Всі моделі запускаються асинхронно, щоб не блокувати інтерфейс користувача.

Рівень моделей (AI Models / Data Layer) відповідає за всю обробку даних з використанням нейромережевих технологій. До нього входять класи, які завантажують та запускають моделі ONNX M1, M2 та M3, збережені у форматі, сумісному з ONNX Runtime. Для кожної з моделей створені окремі класи:

– OnnxModelRunner (M1): виконує класифікацію зображення на дві категорії: уражене / неуражене поле ;

– UnetModelRunner (M2): реалізує сегментацію зображення за допомогою технології U-Net та формування маски пошкодженої області;

– LstmOnnxRunner (M3): використовується для прогнозування ймовірної площі ураження та часу відновлення посівів на основі послідовних аграрних і погодних даних.

Всі моделі попередньо створено та навчені в середовищі Python (з використанням PyTorch, Pandas, Sklearn), а потім експортовано у формат ONNX (Open Neural Network Exchange). Це дозволяє легко запускати їх у .NET-застосунках без втрати продуктивності. Розглянемо детальніше особливості розроблених моделей.

Модель M1 призначена для виявлення пошкоджених ділянок та класифікації їхніх зображень за ознаками таких уражень, як вирви, забруднення від осколків чи хімічних речовин. Тут використовується згорткова нейронна мережа (CNN) із застосуванням техніки регуляризації Shortcut, що дозволяє уникнути проблеми згасання градієнта та втрати важливих деталей шляхом створення прямих зв'язків між шарами CNN.

Для активації згорткових шарів, а також для повнозв'язного шару, обрано функцію PReLU (Parametric Rectified Linear Unit). На виході модель формує ймовірності належності аналізованих зображень до певних класів через функцію Softmax, що забезпечує коректну інтерпретацію ймовірностей класів. Після класифікації зображень на наступному етапі аналізуються зображення з класу «уражені ділянки».

Створення моделі відбувається за допомогою бібліотеки від Microsoft ML.NET, яка добре працює із C#. Початкове тренування моделі реалізується у середовищі Python через каркас TensorFlow. Потім сформована модель переводиться у формат ONNX і підключається до C#.

Знімки для аналізу надходять із супутників Sentinel-2 (зокрема зі смуги B4, яка добре відображує зміни в рослинності) або з дронів [12]. Перед надходженням до моделі аналізовані зображення проходять попередню підготовку: нормалізацію для вирівнювання яскравості та контрастності, а також аугментацію (наприклад, з використанням поворотів чи зміни масштабу) для покращення узагальнюючої здатності моделі та її стійкості до варіацій у даних.

Ця модель інтегрується в систему як перший етап аналізу. Користувач завантажує зображення через інтерфейс WPF, і модель за лічені секунди визначає наявність або відсутність пошкоджень на аналізованій ділянці. Якщо ділянка класифікується як «уражена», зображення передається до моделі M2 для більш детального аналізу. Результати зберігаються у CSV-файлі для подальшого використання або передачі до бази даних, якщо система масштабуватиметься.

Модель M2 призначена для детального аналізу та сегментацію зображень, класифікованих як «уражені», щоб визначити точне розташування, форму, площу та інтенсивність пошкоджень. Для цього впроваджено архітектуру U-Net – згорткову нейронну мережу з енкодер-декодерною структурою, яка широко застосовується для семантичної сегментації, зокрема в задачах аналізу супутникових зображень. U-Net складається з енкодера, який поступово зменшує розмірність зображення, і декодера, що відновлює просторову інформацію, дозволяючи точно визначити контури, площу, тип та рівень інтенсивності уражень ґрунту, що є критично важливим для подальшого аналізу та прийняття рішень. Дана архітектура ефективно виділяє контури уражених ділянок, таких як вирви чи забруднені зони, завдяки своїй здатності зберігати просторову інформацію через пропускні з'єднання між енкодером і декодером. Крім того U-Net демонструє високу стійкість до шумів на зображеннях, що є особливо важливим при роботі з аерофотознімками та супутниковими даними низької якості. Завдяки цьому модель забезпечує точні результати навіть за умов обмеженої роздільної здатності вхідних зображень. Варіант архітектури мережі U-Net, що використано в моделі M2, наведено на рис. 2.

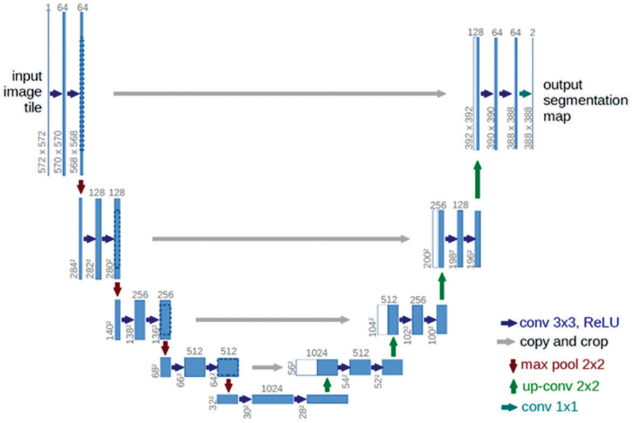

Рис. 2. Архітектура мережі U-Net (модель M2)

Енкодерна частина мережі U-Net складається з кількох згорткових шарів, які поступово зменшують розмірність зображення, виділяючи ключові ознаки, такі як краї вирв чи зміни в текстурі ґрунту. Декодерна частина відновлює зображення до початкової роздільної здатності, створюючи піксельну маску, де кожен піксель позначений як «пошкоджений» або «непошкоджений». Для активації шарів знову використовується функція PReLU, щоб забезпечити стабільне навчання.

Модель тренується в Python за допомогою бібліоте-

ки PyTorch, що пропонує гнучкі інструменти для роботи з U-Net і обробки великих зображень. Для навчання використовуються датасети, що містять супутникові знімки Sentinel-2 і аерофотознімки з дронів з анотаціями, які позначають пошкоджені ділянки. Смуга B4 Sentinel-2 є особливо ефективною для виявлення змін у рослинності, тому використовуємо її як основний канал, додаючи смуги B2 і B3 для покращення контрастності [13]. Попередня обробка в цій моделі потребує формування датасету з фотографій пошкоджених земель та метаданих, таких як тип використаної зброї, площа ураження та час, необхідний для відновлення. Метадані структуруються та стандартизуються для забезпечення коректності подальшого аналізу. Додатково виконується попередня обробка зображень, зокрема нормалізація яскравості та контрастності, а також видалення шуму за допомогою гаусівського фільтра, щоб уніфікувати дані.

Наступним кроком є тестування та валідація моделі штучного інтелекту (МШІ) М2, де використовуються метрики Mean Squared Error (MSE) для оцінки точності прогнозування часу, Intersection over Union (IoU) – для визначення площі ураження, а також Recall, Precision та F1-score для класифікації пошкоджень. Ці метрики дозволяють переконатися, що модель точно виділяє межі пошкоджень із мінімальними помилками. Верифікація моделі М2 виконується на тестовому наборі даних, що містить раніше невикористовувані зображення, щоб оцінити її узагальнену здатність.

Концептуальний аналіз МШІ М2 зосереджується на виявленні помилок, їх усуненні через донавчання та ансамблеві методи. Оцінюється вплив погодних умов і якості зображень, а також можливість інтеграції супутникових та дронових даних для покращення прогнозування.

Після тренування модель М2 експортується у формат ONNX і інтегрується в C# через ONNX Runtime, що забезпечує швидку обробку зображень у реальному часі. Таким чином, ця модель працює як другий етап реалізації функцій системи (після того, як модель М1 визначила зображення як «уражені», мережа U-Net створює детальну піксельну карту пошкоджень, чітко окреслюючи контури уражених зон). Сформована карта пошкоджень відображається в інтуїтивному інтерфейсі WPF, де користувач може побачити білі області пошкоджень на чорному тлі разом із числовими даними про площу ураження (в квадратних метрах або гектарах). Така візуалізація допомагає користувачам (фермерам чи екологам) швидко оцінити масштаби проблеми та планувати подальші дії.

Модель М3 призначена для прогнозування ймовірної площі ураження та часу відновлення посівів на основі аналізу таких факторів, як площа пошкодження, характер пошкодження (вибухи, хімічне забруднення тощо), тип ґрунту та кліматичні умови. Для цього в моделі М3 використано комбінацію двох архітектур: Temporal Fusion Transformer (TFT) для обробки статичних і різнорідних даних, таких як тип зброї чи геолокація, і Long Short-Term Memory (LSTM) для аналізу динамічних змін з урахуванням сезонних чи погодних факторів. Цей підхід дозволяє створювати точні прогнози навіть у динамічних умовах.

TFT – це гібридна модель, яка поєднує механізми уваги (attention) і обробки часових рядів. Ця модель аналізує статичні дані (наприклад, тип ґрунту чи місце розташування), поточні динамічні дані (наприклад, опади чи температура) і формує узагальнені результати прогнозування. Архітектура LSTM, у свою чергу, відстежує довгострокові залежності (зокрема, поступове відновлення родючості ґрунту), що залежить від сезонних циклів. Спільне застосування TFT і LSTM в моделі М3 забезпечує точні оцінки часу рекультивації з урахуванням поточного стану поля та динаміки його зміни.

Модель М3 тренується в Python за допомогою бібліотеки PyTorch, що здатна підтримувати архітектури TFT і LSTM. Вхідними даними цієї моделі є результати сегментації (площа та інтенсивність пошкоджень), а також дані з відкритих джерел (наприклад, тип ґрунту чи кліматичні показники, збережені у CSV-файлах). Для підвищення точності моделі використовуються налаштування за допомогою бібліотеки KerasTuner, що дозволяють визначити найкращі значення таких параметрів як розмір шарів і швидкість навчання. Оцінка якості моделі здійснюється з використанням метрики MAE (Mean Absolute Error) для перевірки точності прогнозу часу відновлення, а також метрики R для загальної оцінки якості моделі. Готова модель підключається до C# через формат ONNX Runtime для швидкого формування прогнозів. У вихідному звіті моделі зазначається, скільки часу потребує відновлення (прогнозований час відновлення відображається в інтерфейсі WPF) і які методи рекультивації мають бути найбільш ефективними. Дані звітів зберігаються в CSV, щоб можна було повернутися до них в разі потреби [14].

Інтеграція моделей. Моделі М1, М2 та М3 є базовими складовими елементами системи моніторингу та аналізу пошкоджень аграрних ділянок.

Передача результатів між цими моделями реалізується через внутрішні змінні та допоміжні структури (у тому числі з використанням тимчасових файлів). Усі етапи обробки є асинхронними, що дозволяє забезпечити плавну взаємодію з інтерфейсом системи і не блокувати роботу застосунку під час обчислень. Завдяки чіткому розмежуванню функцій моделей, послідовність їх роботи (класифікація сегментація прогноз) зберігає гнучкість, дозволяючи в майбутньому доповнювати систему новими функціональними блоками.

Архітектура системи забезпечує повну автоматизацію аналізу зображень, дозволяючи швидко обробляти дані та формувати практичні висновки й рекомендації для користувачів. Усі етапи обробки інтегровані в

єдиний процес у рамках системи підтримки прийняття рішень (DSS), що робить систему повноцінним аналітичним інструментом. Завдяки модульній структурі система залишається гнучкою, дозволяючи легко додавати нові функції, наприклад, інтеграцію з геоінформаційними системами.

У системі реалізовано чіткий і структурований потік даних, що забезпечує безперебійну обробку інформації та злагоджену взаємодію між усіма компонентами: інтерфейсом користувача, логікою обробки, моделями штучного інтелекту (M1, M2, M3) та модулем генерації звітів.

Потік даних у системі розпочинається із завантаження користувачем зображення земельної ділянки у форматі .jpg або .png через інтерфейс WPF. Після цього зображення автоматично передається до логічного ядра програми, яке координує подальшу їх послідовну обробку з використанням моделей M1, M2 та M3. Структура даних розробленої системи організована таким чином, щоб забезпечити зручність обміну між компонентами, швидкий доступ до проміжних результатів та можливість формування фінального звіту без втрати інформації. Враховуючи поетапну обробку та модульність архітектури, дані зберігаються у вигляді окремих змінних у коді, тимчасових файлів, а також графічних об'єктів.

Після завантаження зображення його шлях зберігається у вигляді рядка (imagePath), а саме зображення завантажується у вигляді об’єкта BitmapSource для подальшої обробки в пам’яті. Результати класифікації (currentPrediction) і прогнозу (currentForecast) зберігаються окремо у вигляді текстових змінних, які відображаються у відповідних текстових полях інтерфейсу.

Маска, отримана в процесі сегментації, зберігається двічі: як об’єкт у пам’яті (currentMask типу BitmapSource) і як зображення у форматі .png, що дозволяє використати її в PDF/DOCX звіті. Усі ці об'єкти передаються до модуля генерації звітів у структурованому вигляді, без необхідності повторної обробки. Для забезпечення узгодженості між модулями використовуються тимчасові файлові структури, такі як CSV-файли, де зберігалися результати класифікації, площа пошкодження, координати маски та прогноз, що спрощувало логіку обробки та дозволяло легко масштабувати систему чи інтегрувати її з базами даних або хмарними сервісами.

Остаточні результати зберігаються у CSV-файлі та відображаються в інтерфейсі у вигляді зрозумілого текстового звіту, що містить числові дані та практичні поради.

Для забезпечення надійності потоку даних у системі реалізовано механізми контролю якості. Наприклад, якщо зображення має низьку роздільну здатність або не відповідає заздалегідь зазначеним вимогам, система видає повідомлення про помилку, пропонуючи користувачу перевірити дані.

Така організація потоку даних у поєднанні зі структурою DSS робила ефективним інструментом для оцінки стану земель і планування їхнього відновлення. Вона в подальшому може бути вдосконалена шляхом інтеграції з додатковими базами даних (для роботи з більшими обсягами інформації) або підключення до реальних супутникових потоків для аналізу в реальному часі.

3. Програмна реалізація функцій системи та результати тестування

Для програмної реалізації функцій системи було використано сучасний набір технологій і інструментів, що забезпечують ефективну реалізацію функціоналу машинного навчання, зручного десктопного інтерфейсу, обробки зображень і формування звітів. Архітектура системи була спроектована з акцентом на продуктивність, модульність і можливість масштабування, що дозволяє легко адаптувати її до нових вимог і забезпечувало стабільну роботу моделей M1, M2 та M3 у реальних умовах.

Основним середовищем для створення, навчання та тестування моделей ШІ були мова Python та низка підтримуваних цією мовою бібліотек, а саме:

– PyTorch: ця бібліотека стала основою для побудови всіх трьох моделей системи – згорткової нейронної мережі (CNN) для класифікації (M1), архітектури U-Net для сегментації (M2) та комбінації Long Short-Term Memory (LSTM) і Temporal Fusion Transformer (TFT) для прогнозування (M3) [15];

– Pandas: використовувалася для обробки табличних даних, таких як кліматичні показники (температура, опади), характеристики ґрунтів чи метадані про типи уражень;

– NumPy: застосовувалася для виконання числових операцій, зокрема нормалізації та трансформації даних перед подачею їх у моделі;

– scikit-learn: використовувалася для крос-валідації моделей і оцінки їхньої якості за такими метриками, як Precision, Recall, F1-score (для класифікації та сегментації) і (для прогнозування);

– Albumentations: забезпечувала аугментацію зображень (обертання, зміна яскравості, додавання шумів), що дозволяло підвищити стійкість моделей до різних умов освітлення чи якості знімків.

Після завершення навчання моделі експортувалися у формат ONNX (Open Neural Network Exchange), що дозволяло використовувати їх у C# додатку без залежності від Python. Такий підхід забезпечував швидку інтеграцію моделей у десктопне середовище та оптимізував продуктивність системи.

Клієнтська частина була реалізована на платформі .NET 8.0, що гарантувала підтримку асинхронного програмування. Тут використовувалися такі технології:

– WPF (Windows Presentation Foundation): ця технологія стала основою для створення інтуїтивного

графічного інтерфейсу користувача. В інтерфейсі відображалися зображення ділянок, маски сегментації, текстові результати класифікації та прогнози, а також елементи керування для завантаження файлів і перегляду звітів;

– ONNX Runtime for .NET: використовувалася для запуску моделей ШІ у форматі ONNX безпосередньо в C# середовищі. Це дозволяло обходитися без Python залежностей, забезпечуючи високу швидкість виконання моделей і їхню інтеграцію з десктопним додатком. ONNX Runtime оптимізувала обчислення, використовуючи апаратне прискорення (CPU або GPU);

– System.Drawing та System.Windows.Media.Imaging: ці бібліотеки відповідали за обробку зображень, зокрема завантаження вхідних знімків, рендеринг масок сегментації та їх збереження у форматі .png для звітів. Вони забезпечували якісну візуалізацію результатів у реальному часі;

– асинхронне програмування (async/await): використовувалося для паралельного виконання обчислень зокрема, обробки зображень моделями) і оновлення інтерфейсу, що дозволяло уникнути затримок і забезпечувало плавну взаємодію користувача з системою.

Для стабільної роботи моделі прогнозування (М3) та забезпечення точності аналізу система передбачає інтеграцію з зовнішніми джерелами даних:

– супутникові знімки: система розрахована насамперед на роботу із зображеннями від Sentinel-2 і Landsat, які надають високоякісні дані для аналізу сільськогосподарських земель. Ці джерела можуть використовуватися для отримання знімків у реальному часі або з архівів, що дозволяє відстежувати стан ділянок у динаміці;

– відкриті погодні API: для прогнозування часу відновлення земель використовуються дані з таких сервісів, як Open-Meteo і Meteostat, які надають часові ряди кліматичних показників (опади, температура, вологість). Ці дані є важливими для моделі М3, оскільки впливають на оцінку рекультиваційних заходів;

– публічні аграрні датасети: для навчання моделей використовуються набори даних FAO (Food and Agriculture Organization) і CropHarvest, що містять приклади уражених земель, типи ґрунтів і характеристики рослинності. Це дозволяє моделям «навчатися» на реальних сценаріях і підвищує їхню точність.

Набір даних для побудови та тестування системи моніторингу та прогнозування пошкоджень сільськогосподарських земель (з використанням моделей М1, М2 та М3) складався з двох частин: зображень уражених і неуражених ділянок, поділених на тренувальну, валідаційну та тестову вибірки з категоріями damaged/undamaged. Кожна вибірка має відповідні маски сегментації у форматі PNG та CSV-файли з метаданими із додатковою інформацією про об'єкти. Така структура дозволяє об'єднати зображення, числові та текстові дані для навчання комбінованих моделей.

Для навчального датасету було обрано обмежену сукупність реальних вхідних зображень та програмно згенеровані сегментаційні маски пошкоджень. Штучно сформовано окремий файл формату CSV з метаописом датасету. У файлі CSV зафіксовані додаткові характеристики кожного випадку руйнування (інформація про місце і час події, тип руйнування та джерело походження даних). Заповнення цього файлу здійснювалося не лише на основі супутникових знімків та фотографій з дронів, але й з відкритих новинних ресурсів (зокрема з публікацій військових порталів «Militarnyi», де регулярно подаються підтверджені факти знищення військової техніки чи наслідків бойових дій). Такий підхід дозволив поєднати в одному наборі як візуальні дані, так і текстову інформацію, що значно розширило можливості аналізу й забезпечило основу для побудови ефективних прогнозних моделей. Весь набір даних структуровано у вигляді двох основних директорій: директорія images містить оригінальні знімки у форматі .jpg; директорія masks містить відповідні сегментаційні маски у форматі .png, що позначають межі пошкоджених ділянок. Приклад пари «знімок – маска» наведено на рис. 3.

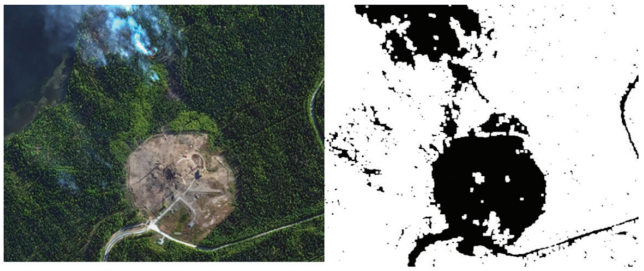

Рис. 3. Приклад вхідного знімка та відповідної маски пошкодженої ділянки

Інтеграція зовнішніх джерел робить систему гнучкою і готовою до роботи в різних умовах, наприклад, у регіонах із різними кліматичними чи ґрунтовими характеристиками. У майбутньому систему можна розширити, додавши підтримку інших API (наприклад, для моніторингу забруднень чи аналізу хімічного складу ґрунтів) або підключивши реляційні бази даних для збереження історії аналізів.

Поєднання мов Python (для створення моделей) та C# (.NET + WPF для інтерфейсу та запуску моделей) забезпечує баланс між гнучкістю машинного навчання і стабільністю та зручністю десктопного застосунку. Обрані інструменти дозволили реалізувати комплексну, автономну систему, яка може працювати локально без підключення до серверів, а також масштабуватися до веб або хмарної архітектури в майбутньому.

Розглянемо деякі особистості програмної реалізації моделей системи та результати її тестування.

Для класифікації зображень (модель М1) було використано згорткову нейронну мережу (CNN), побудовану на основі ResNet-18 [16]. Архітектура була модифікована для роботи з обмеженим набором зображень і забезпечення високої точності:

– змінено останній шар: fc = nn.Linear(512, 2);

– додано технологію Dropout перед повнозв'язним шаром для зменшення перенавчання;

– використано функцію активації PReLU для адаптації до темних зон.

Програмний код для навчання моделі M1 наведено нижче:

```
    import torch
  import torch.nn as nn
  import torch.optim as optim
  from torchvision import datasets, transforms, models
  from torch.utils.data import DataLoader
model =
models.resnet18(weights=models.ResNet18_Weights.
IMAGENET1K_V1)
model.fc = nn.Sequential(
  nn.Dropout(0.5),  # Add Dropout
  nn.Linear(model.fc.in_features, NUM_CLASSES)
device = torch.device("cuda" if torch.cuda.is_available()
else "cpu")
model.to(device)
criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=0.0001,
weight_decay=1e-4)
```

Тренування моделі M1 здійснювалося у середовищі PyTorch на невеликому датасеті: 36 зображень у train, 10 зображень у validation. Через обмежену кількість даних було застосовано аугментації (RandomFlip, ColorJitter) та знижено learning rate (0.0001).

Ключові параметри:

– Loss: CrossEntropyLoss;

– Optimizer: Adam;

– Epoch: 15;

– Batch Size: 4;

– Augmentation: доступна в train_transforms;

– EarlyStopping: активований (patience=3).

Отримані результати підтверджують успішне навчання моделі.

На валідаційному наборі було досягнуто 100% точності, що свідчить про високу ефективність класифікації саме на доступних прикладах. Мінімальне значення втрат на валідації (Val Loss) зафіксовано на 10-й епосі (0.0158).

Такого результату вдалося досягти завдяки використанню технології регуляризації Dropout, а також аугментації зображень, що дозволили штучно збільшити варіативність навчальних даних. Низький learning rate забезпечив стабільне навчання без різких коливань у значеннях втрат.

Після успішного навчання та експорту моделі у формат ONNX, вона була інтегрована в десктопний застосунок на мові **C#** із використанням бібліотеки Microsoft.ML.OnnxRuntime. Цей формат дозволяє виконувати нейронну мережу незалежно від середовища навчання (наприклад, без запуску Python), що ідеально підходить для .NET-додатків.

У програмі реалізовано окремий клас OnnxModelRunner, відповідальний за завантаження моделі, передобробку зображення та отримання результату класифікації. Модель читається з файлу m1_model.onnx, який розміщується в директорії Models/.

Зображення, яке завантажується користувачем, спочатку конвертується у формат тензора (Tensor <float>). Перед подачею в модель воно: масштабується до розміру 224×224 пікселі; нормалізується); переводиться з формату Bitmap до багатовимірного масиву DenseTensor<float>.

Передобробка реалізується в окремому методі PreprocessImage, який також входить до складу OnnxModelRunner.cs.

Результат у вигляді двох чисел (ймовірностей для класу «уражена» і «неуражена») зчитується з виходу моделі. Далі проводиться порівняння: якщо ймовірність класу «уражена» вища за порогове значення, то формується відповідне повідомлення, що відображається у інтерфейсі WPF за допомогою елемента ResultTextBlock. Також результат зберігається у змінній currentPrediction, що дає змогу використати його надалі у звітах (у форматах PDF або DOCX).

Цей програмний модуль забезпечує швидке та стабільне розгортання моделі класифікації M1 безпосередньо у графічному інтерфейсі.

Для сегментації зображень пошкоджених ділянок (модель M2) було використано згорткову нейронну мережу з архітектурою U-Net, що складається з двох частин: стискуючої (енкодер) та розширюючої (декодер). В енкодері вхідне зображення проходить через кілька рівнів згорткових шарів із функцією активації PReLU, після чого стискається за допомогою операції max pooling. Це дозволяє зменшити розмір зображення, зберігаючи при цьому найбільш значущі ознаки. Декодер виконує зворотну операцію – поступово відновлює просторову роздільну здатність за допомогою транспонованих згорткових шарів (ConvTranspose2d) і об'єднує (через skip connections) відповідні рівні з енкодера. У фіналі модель формує карту пікселів, де кожен піксель має значення від 0 до 1 – ймовірність, що він належить до пошкодженої зони.

Реалізація архітектури в проєкті була виконана за допомогою фреймворку PyTorch..

На етапі навчання для кожного зображення, що міститься в каталозі, виконується попередня обробка. Всі зображення масштабуються до фіксованого розміру 128×128 пікселів (це оптимальний розмір для швидкого навчання на малих об'ємах даних). Після цього виконується нормалізація: значення пікселів діляться на 255, щоб привести їх до діапазону [0, 1], що є стандартом для нейронних мереж. Поряд із зображенням обробляється і відповідна маска – чорно-біле зображення з тією ж роздільною здатністю, де білим кольором позначені пошкоджені області, а чорним – фон. Маска також масштабується до 128×128, конвертується у одноканальний тензор [1, H, W] та нормалізується.

Далі зображення та маска передаються у модель у вигляді тензорів типу torch.Tensor. U-Net обробляє зображення, проходячи крізь усі згорткові та декодувальні шари, після чого формує вихідну карту ймовірностей, де кожному пікселю призначається значення від 0 до 1, яке інтерпретується як ймовірність того, що цей піксель належить до пошкодженої ділянки.

В результаті сегментації формується бінарна маска, яка зберігається в окремий файл (segmented_mask.png) і водночас передається до інтерфейсу користувача в WPF для візуалізації. Повний код навчання моделі для генерації масок наведено у додатку Г. У WPF-модулі ця маска прив'язується до елементу MaskImage, де користувач може переглянути сегментовану ділянку прямо на екрані.

Функціональною реалізацією цього етапу в C# є клас UnetModelRunner, що відкриває модель unet_model.onnx, проводить необхідні перетворення зображення, запускає модель та повертає об'єкт типу Bitmap, який далі передається у Image.Source.

Під час навчання моделі на обмеженому наборі даних модель М2 демонструвала поступове покращення. Втрати на тренувальному наборі (loss) зменшилися з 4.18 у першій епосі до 2.11 у десятій. Це підтверджує, що навіть на малому наборі даних мережа здатна навчитися узагальнювати просторові патерни, характерні для пошкоджених ділянок.

Після завершення навчання модель М2 експортувалася у формат ONNX для використання у .NET-застосунку. Файл unet_model.onnx зберігається у директорії Models/ і використовується в десктопному застосунку на етапі сегментації.

Для прогнозування часу відновлення пошкоджених ділянок на основі результатів сегментації (модель М3) було використано комбіновану архітектуру ШНМ, що поєднує можливості технологій TFT і LSTM.

Модель М3 була реалізована у фреймворку PyTorch як комбінація двох підходів до аналізу часових рядів:

– LSTM відповідає за обробку послідовних змін параметрів у часі (наприклад, зміни вологості, опадів або температури, що впливають на швидкість відновлення ґрунту);

– TFT забезпечує обробку як динамічних, так і статичних вхідних ознак із використанням механізму уваги, що дає змогу фокусуватися на ключових змінах у часовому ряді та адаптивно зважувати інформацію.

Архітектура передбачає наявність таких вхідних даних:

– нормалізована площа ураження (за результатами сегментації);

– кліматичні характеристики (метеодані за останні 12 місяців);

– тип ґрунту (категоріальний вхід);

– тип ураження (визначений за площинними шаблонами або за класифікатором).

Ці входи об'єднуються в єдиний багатовимірний тензор, який подається на вхід LSTM та декодується TFT-блоком. Програмний код для реалізації комбінації LSTM та TFT наведено нижче:

```
class SimpleLSTMTransformer(nn.Module):
  def __init__(self, input_size, hidden_size, output_
size=1):
    super(SimpleLSTMTransformer, self).__init__()
    self.lstm = nn.LSTM(input_size, hidden_size, batch_
first=True)
    self.linear = nn.Linear(hidden_size, hidden_size)
    self.transformer = nn.TransformerEncoder(
      n.TransformerEncoderLayer(d_model=hidden_
size, nhead=2, batch_first=True),
      num_layers=1
       self.output = nn.Linear(hidden_size, output_size)
   def forward(self, x):
     x, _ = self.lstm(x)
     x = self.linear(x)
     x = self.transformer(x)
     return self.output(x[:, -1, :])
```

У вхідному тензорі моделі М3 присутні як **статичні параметри** (тип ґрунту, геолокація) так і **динамічні часові ряди** (вологість, температура, кількість опадів, площа пошкоджень у часі). Дані було згенеровано шляхом моделювання та симуляції, з урахуванням реальних шаблонів клімату. Вхідні значення масштабуються до діапазону [0, 1] за допомогою програми MinMaxScaler із бібліотеки sklearn. Для time-series входу формується масив розмірності (batch, time_steps, features), де time_steps=50 і features=2.

Навчання моделі під час її тестування проводилося протягом 150 епох. У процесі навчання втрати зменшувалися поступово – з початкового значення 10.8050 до мінімуму 0.8116 на 120-й епосі. Графік втрат мав деякі коливання, однак загальний тренд залишався низхідним. Результати навчання комбінованої моделі прогнозування свідчать про поступове покращення здатності моделі формувати точні прогнози.

Отримані результати узгоджувалися з наступною логікою: при більшій площі пошкодження модель М3 прогнозувала триваліший час відновлення, що підтверджує її коректну поведінку. Водночас слід враховувати, що модель не враховує важливі зовнішні чинники, зокрема економічні ресурси, політичні рішення чи наявність техніки для рекультивації – її прогноз базується виключно на фізичних характеристиках.

Після завершення навчання модель експортувалася у формат ONNX для використання у C# -застосунку.

Вихідні результати роботи моделі М3 (прогнозований час відновлення у місяцях) автоматично передаються до інтерфейсу користувача і зберігаються у змінній currentForecast для подальшого їх занесення до відповідного звіту.

Головне вікно інтерфейсу запропонованої системи побудоване за принципом трирівневої логіки: ліворуч

розміщені кнопки керування, а праворуч – вікна для відображення результатів обробки (рис. 4).

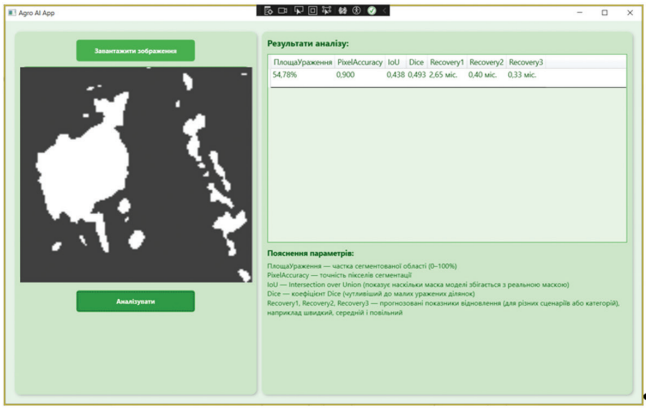

Рис. 4. Вікно інтерфейсу desktop-додатку

Після запуску додатку користувач бачить панель з такими основними етапами:

– завантаження зображення – відкриває діалогове вікно для вибору фото;

– класифікація (M1) – визначає, чи зображення є ураженим;

– сегментація (M2) – при потребі створює маску пошкоджень;

– прогноз (M3) – оцінює орієнтовний час відновлення;

– формування звіту – дозволяє зберегти результати у форматах PDF або DOCX.

На рис. 5 наведено приклад візуалізації результатів обробки зображення.

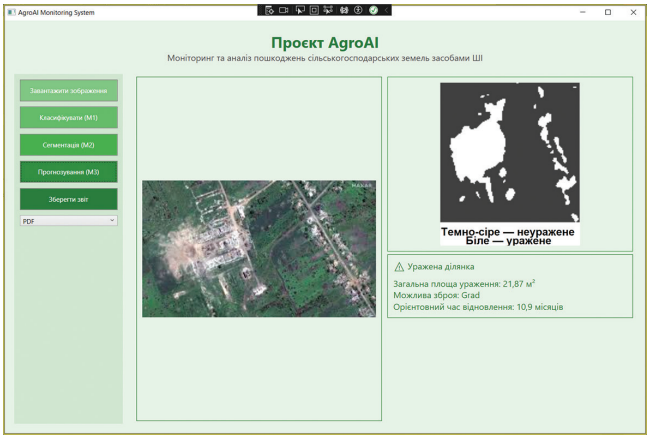

Рис. 5. Приклад візуалізації результатів обробки зображення

Усі результати обробки відображаються в інтерфейсі у вигляді тексту та графіки (маска, карта пошкоджень, числові значення). Всі етапи виконуються покроково, і кожен з них ініціалізується лише після успішного виконання попереднього.

Реалізація функціональних можливостей системи має сприяти більш ефективному плануванню використання земельних ресурсів і прискоренню процесів рекультивації, що відповідає цілі сталого розвитку 15 – «Захист і відновлення екосистем суші». Це підкреслює значення технологій штучного інтелекту не лише в контексті оперативного реагування на воєнні наслідки, а й як стратегічного інструменту для довгострокового сталого розвитку сільського господарства [17].

Слід зазначити, що у процесі тестування системи було виявлено низку певних обмежень. Зокрема, класифікаційна модель через малий обсяг навчальних даних схильна до перенавчання. Якість сегментації залежить від вхідного зображення: в ідеальних умовах вона є висока, але може погіршуватися при зміні умов зйомки. Модель прогнозування, що показала прийнятні результати на тестових даних, може потребувати подальшої адаптації для реальних сценаріїв, де до фізичних параметрів додаються соціальні та економічні фактори.

Крім того, необхідно визначити деякі проблеми використання сегментаційних масок. Хмари (для моделей M2 та M3) розпізнавались та позначались в цих масках тим самим кольором, що і уражена ділянка, що може призвести до помилок в детекції, оскільки в оптичних супутникових зображеннях хмари часто імітують деградацію ґрунтів, особливо в аграрних регіонах з сезонними хмарами, де помилки можуть досягати 30% без препроцесингу.

У подальшому пропонується додати клас «не розпізнано», до якого буде віднесена частина зображення з відповідним кольором позначення, у тому числі і хмари. При цьому бінарна модель (уражене/неуражене) перетворюється на трикласову, зменшуючи false positives на 15–25% в разі наявності з хмар на зображеннях. Це дозволяє уникати примусової класифікації невизначених пікселів, що особливо корисно для невеликих датасетів, де присутній ризик перенавчання. Для препроцесинга зображень в цьому разі можна використати алгоритми виявлення хмар перед сегментацією, такі як RS-Net або attention-based U-Net, які аналізують мультиспектральні канали для створення масок хмар. Додатково можуть бути використані GAN-моделі для видалення хмар з реконструкцією [18].

Результати виконання етапів роботи запропонованої інтелектуальної системи виявлення та аналізу руйнівних пошкоджень аграрних ділянок (включно з визначеним типом пошкодження, оціненою площею ураження, можливим типом застосованої зброї та прогнозованим терміном відновлення) об’єднуються в єдиний звіт. Цей звіт також містить карту пошкоджень у вигляді сегментованого зображення, що надає користувачу не лише числову інформацію, а й візуальне подання результатів.

У підсумку користувач отримує чітку й структуровану аналітичну довідку, яка може бути збережена у форматах PDF або Word і використана для подальших дій, включаючи планування рекультиваційних заходів.

Висновки

В статті наведено результати розроблення та дослідження інтелектуальної системи виявлення та аналізу руйнівних пошкоджень аграрних ділянок з

використанням сучасних нейромережевих технологій. Система базується на послідовному використанню трьох моделей штучного інтелекту, функціями яких є:

– класифікація пошкоджень за допомогою згорткової нейронної мережі (CNN);

– сегментація уражених ділянок із використанням архітектури U-Net;

– прогнозування часу відновлення земель на основі комбінації Temporal Fusion Transformer (TFT) і Long Short-Term Memory (LSTM).

Така послідовність дозволяє системі: визначати наявність пошкоджень за результатами аналізу зображень аграрних ділянок; створювати детальні карти уражених зон з точними контурами; прогнозувати термін відновлення з урахуванням площі ураження, типу ґрунту, кліматичних умов і характеру пошкоджень; формувати звіти у форматах PDF або DOCX, які користувач може використовувати для планування рекультивації.

Всі моделі системи попередньо створюються та навчаються в середовищі Python (з використанням бібліотек PyTorch, Pandas, Sklearn), а потім експортуються у формат ONNX. Це дозволяє ефективно використовувати їх у .NET-застосунках з забезпеченням швидкої обробки даних.

Результати тестування підтверджують працездатність запропонованої системи.

Система має значний потенціал для розвитку: її можна інтегрувати з супутниковими сервісами для аналізу даних у реальному часі, підключити до геоінформаційних систем (GIS) для створення детальних аграрних карт або синхронізувати з державними реєстрами для здійснення більш детальної аналітики. Наприклад, додавання модуля для порівняння знімків у динаміці може сприяти прогнозуванню зміни стану земель, а інтеграція з базами даних про мінну небезпеку підвищувати безпеку планування рекультиваційних робіт.

Список літератури:

[1] Armed Forces of Ukraine destroyed the Russian Grad multiple rocket launcher with a drone in the Donetsk region. Militarnyi. URL: https://militarnyi.com/en/news/armed-forces-of-ukraine-destroyed-the-russian-grad-multiple-rocket-launcher-with-a-drone-in-the-donetsk-region/ (дата звернення: 25.05.2025).

[2] Екологічний моніторинг ландшафтних ділянок з використанням регуляризованих штучних нейронних мереж. / С. Удовенко та ін. Біоніка інтелекту. 2022. Т. 1 № 94. С. 13–22. URL: https://doi.org/10.30837/bi.2020.1(94).03 (дата звернення: 20.05.2025).

[3] Drozd S., Kussul N., Shelestov A. Satellite-Based Analysis of Forest Damage in Ukraine's Protected Areas. 13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. Gliwice, Poland. 4–6 September, 2025.

[4] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015. Т. 61. С. 85–117. URL: https://doi.org/10.1016/j.neunet.2014.09.003 (дата звернення: 04.09.2025).

[5] OneSoil | Free App for Precision Farming. OneSoil | Free Farming App for Precision Agriculture. URL: https://onesoil.ai/en (дата звернення: 04.09.2025).

[6] Cropin | SaaS-based AgTech | Smart Farming App | Agriculture Technology. URL: https://www.cropin.com/ (дата звернення: 04.09.2025).

[7] Проєкт протимінної діяльності в Україні. Проєкт протимінної діяльності в Україні. URL: https://www.undp.org/uk/ukraine/projects/proyekt-protyminnoyi-diyalnosti-v-ukrayini (дата звернення: 04.09.2025).

[8] Методи комп'ютерного зору і глибинних нейронних мереж для еколого-економічного аналізу : монографія / Н. М. Куссуль та ін. Київ : Наук. думка, 2024. 474 с.

[9] Temporal Fusion Transformers for interpretable multi-horizon time series forecasting / B. Lim та ін. International Journal of Forecasting. 2021. Т. 37, № 4. С. 1748–1764. URL: https://doi.org/10.1016/j.ijforecast.2021.03.012 (дата звернення: 04.09.2025).

[10] Aramendia A. I. The U-Net : A Complete Guide. Medium. URL: https://medium.com/@alejandro.itoaramendia/decoding-the-u-net-a-complete-guide-810b1c6d56d8 (дата звернення: 26.05.2025).

[11] Beck M., Pöppel K., Spanring M., Auer A., Prudnikova O., Kopp M., Klambauer G., Brandstetter J. «xLSTM: Extended Long Short-Term Memory». *NeurIPS 2024 Spotlight*, 2024. URL: https://doi.org/10.48550/arXiv.2405.04517 (дата звернення: 04.09.2025).

[12] Collect Earth Online Home. Collect Earth Online - Satellite Image Viewing & Interpretation Systema. URL: https://www.collect.earth/ (дата звернення: 25.05.2025).

[13] Silvanovych K., Hrynova O. Leveraging ai for agricultural land monitoring and reclamation. Information Systems and Technology: Results and Prospects, Kyiv. 2025. P. 295–297. URL: https://ist.fit.knu.ua/_files/ugd/016074_36d0f427916c46abb6491a7572bb63ec.pdf (дата звернення: 01.06.2025).

[14] Microsoft (2024). ML.NET and ONNX Runtime for .NET developers. URL: https://dotnet.microsoft.com/en-us/apps/ai/ml-dotnet (дата звернення: 05.09.2025).

[15] PyTorch Forecasting Documentation – pytorch-forecasting documentation. PyTorch Forecasting Documentation – pytorch-forecasting documentation. URL: https://pytorch-forecasting.readthedocs.io/en/stable/ (дата звернення: 21.05.2025).

[16] He K., Zhang X., Ren S., & Sun J. (2016). Deep Residual Learning for Image Recognition. *CVPR 2016*. URL: https://ieeexplore.ieee.org/document/7780459 (дата звернення: 05.09.2025).

[17] Сільванович К. В., Гриньова О. Є. Моніторинг та відновлення сільськогосподарських земель засобами штучного інтелекту. Радіоелектроніка та молодь у XXI столітті, м. Харків. 2025. С. 51–53. URL: https://openarchive.nure.ua/entities/publication/6a9d7017-8bdb-4118-a8c8-8ed170ce91a8 (дата звернення: 01.06.2025).

[18] Застосування генеративно-змагальних мереж для покращення якості сегментації супутникових знімків / О. В. Шкаліков та ін. XIX Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених «Теоретичні і прикладні проблеми фізики, математики та інформатики», м. Київ. 2022. С. 375–378. URL: https://ela.kpi.ua/handle/123456789/52532 (дата звернення: 03.09.2025).

Надійшла до редколегії 15.10.2025

УДК 004.8 DOi 10.30837/ bi.2025.2(103).05

M. Monastyrskyi

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine,
Mykyta.Monastyrskyi@cs.khpi.edu.ua, ORCID iD: 0009-0003-7904-8006

IMPROVING QUALITY OF MUSIC SOURCE SEPARATION IN CONSTRAINED AND CORRUPTED TRAINING DATA SETTING USING LOSS MASKING

This work aims to explore the efficiency of the loss masking strategy for training deep music source separation models in a setting where training data is corrupted, specifically with bleeding artefacts. A soft loss masking training strategy, which assigns weights to batch loss values inversely proportional to their magnitude, is proposed and compared to hard loss masking, where weights are computed as binary masks based on whether the loss function value exceeds a certain threshold. An investigation is conducted to determine whether a soft loss masking approach yields better results than hard masking in settings with low training data availability. Results indicate that, under constrained training data conditions with bleeding artefacts, the soft masking approach outperforms the hard loss masking method, specifically for the vocal source. Alongside, the evaluation strategy based on neural network approximation of the MUSHRA score is presented to account for both subjective and objective components of the music source separation system quality evaluation.

MUSIC SOURCE SEPARATION, LOSS MASKING, PERCEPTUAL QUALITY ASSESSMENT, SIGNAL PROCESSING, MACHINE LEARNING, NEURAL NETWORKS

М. С. Монастирський. Покращення якості розділення музичних сигналів в умовах наявності артефактів та обмеженої кількості тренувальних даних з використанням маскування функції втрат. В поточній роботі досліджується ефективність використання підходу маскування функції втрат для тренування моделей розділення музичних сигналів в умовах наявності похибок в даних, зокрема артефактів перетікання. Пропонується стратегія м'якого маскування функції втрат, суть якої полягає в присвоєнні ваг значенням функції втрат у батчі обернено пропорційно до їхньої величини, і порівнюється з підходом жорсткого маскування, де ваги обчислюються як бінарні маски на основі того, чи перевищує значення функції втрат певний пороговий рівень. Проводиться дослідження щодо того, чи дає підхід м'якого маскування функції втрат кращі результати порівняно з жорстким маскуванням в умовах обмеженої кількості доступних навчальних даних. Результати засвідчують, що в умовах обмеженої кількості тренувальних даних, за умови наявності в них артефактів перетікання, підхід м'якого маскування дозволяє отримати кращі результати за підхід жорсткого маскування зокрема для виокремлення вокалу. Пропонується також метод оцінки результатів розділення заснований на апроксимації метрики MUSHRA з використанням нейронної мережі, задля врахування як об'єктивної так і суб'єктивної компоненти оцінки якості розділення сигналів системою.

РОЗДІЛЕННЯ МУЗИЧНИХ СИГНАЛІВ, МАСКУВАННЯ ФУНКЦІЇ ВТРАТ, ОЦІНКА СПРИЙМАНОЇ ЯКОСТІ, ОБРОБКА СИГНАЛІВ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ

Introduction

The last edition of the music track of the Sound Demixing Challenge featured two tasks for training source separation models in the corrupted data setting. The possible errors in the training data included label noise, where labels were incorrectly assigned to corresponding sources, and bleeding, where the sound of one source appears on the track of another source at a lower level [1]. The submissions to the music demixing track of SDX23 must have utilized the respective internal Moises datasets, which comprise 203 full songs for both error types. On the other hand, the transferability of the loss masking approach – which was the winning method on both leaderboards – to training music source separation models with open-source community datasets such as MUSDB18 [2], Slakh2100 [3], and MoisesDB [4] – which are widely used in the literature as baselines for evaluating novel architectures – has not been thoroughly investigated. At the same time, bleeding artifacts are commonly found in these data sources, which requires the training method that will be robust to these errors mainly due to relatively low availability of the training data in the domain and the high demands to the quality of the source separation system output to be able to make use of the separation results in downstream tasks such as remixing. Seemingly, the presence of such artefacts in the training data limits the performance of the model trained on such data to the level of corruption in the training samples.

Thus, the main questions addressed in this work are as follows: Is the loss masking strategy effective for improving the quality of outputs produced by deep neural networks trained on large, open-source community datasets that frequently contain bleeding artefacts? Can we use loss masking to train models when manual data cleaning is not possible – or not feasible – and still obtain better results than without it?

To address these questions, an evaluation is conducted using the TFC-TDF-UNet v3 model trained on the MUSDB18 dataset with a loss masking strategy and compared against the same architecture trained with a standard mean-squared error objective. Performance is assessed using both the objective SDR metric and a subjective quality estimate based on the MUSHRA protocol. MUSHRA scores are approximated using the NISQA convolutional neural

network, which has been trained on a dataset that contains the SiSEC18 MUSHRA ratings. In addition, the soft loss masking approach is introduced and evaluated. The effect of training dataset size on model performance is examined by training the TFC-TDF-UNet v3 model with loss masking on subsets of the MUSDB18 training set comprising 25%, 50%, and 75% of the original samples.

The contributions of this work are as follows:

– An investigation into the applicability of the loss masking approach for training deep learning models on datasets affected by bleeding artefacts, within widely used open-source benchmarks in the music source separation domain.

– A soft loss masking training strategy is introduced, derived from the hard masking method described by [5]. This strategy assigns weights to batch loss values inversely proportional to their magnitude. Its impact on training performance is assessed in both full and limited data availability scenarios.

– Evaluation of trained models is extended to include subjective audio quality assessment using the MUSHRA metric. To approximate MUSHRA scores, a NISQA neural network is trained on SiSEC18 MUSHRA ratings and applied to the model outputs.

The structure of this work is as follows: Section 1 provides a brief overview of related research. Section 2 introduces the evaluation methodology, along with a summary of the key concepts and components used in the study. Section 3 presents experimental results and corresponding discussion. Finally, Section 4 concludes the work and outlines potential directions for future research.

1. Background and Related Work

The quality of the music source separation systems in a corrupted training data setting is usually improved by either developing new architectural approaches, manually cleaning the training data or – if the above two is not possible or not feasible in a given setting – developing new training or post-processing methods and strategies that are providing the ability to make most of the given architectural or data constraints.

For example, [6] investigated the impact of various data augmentations and ensembling strategies on source separation, specifically in the music signals domain. Later works, such as [7], focus on explaining the benefits of using specific augmentations, including random mixing. There are also works focusing on different parts of the separation systems. For example, [8] investigates the impact of using loss functions alternative to mean-squared error for training deep music source separation models.

[1] proposed the development of source separation methods robust to training data artefacts such as bleeding and label noise, as part of the SDX23 challenge. This initiative aimed to bridge the gap between the idealized source separation scenario, commonly assumed in academic research, where input data is considered clean, and real-world conditions, where training data is often noisy or corrupted. In the music demixing track of SDX23, participants were required to utilize internal Moises datasets that were deliberately corrupted with such artefacts. The corruption was designed to be resistant to manual cleaning, thereby compelling participants to devise training methods inherently robust to label noise and bleeding.

This work primarily builds upon the approach proposed by [5], which introduced a loss masking strategy for training source separation models – specifically the TFC-TDF-UNet v3 architecture – under conditions of corrupted data. This method achieved top performance in both the label noise and bleeding leaderboards of the music track of the SDX23 challenge. The current study investigates the transferability of this strategy to widely used open-source datasets, which frequently contain bleeding artefacts. Additionally, a soft loss masking approach is introduced, and an ablation study is conducted to assess the impact of training dataset size on the performance of models trained with the loss masking strategy.

An essential aspect of improving the quality of models trained on datasets containing artefacts such as bleeding is the measurement of such improvements. The Signal-to-Distortion Ratio (SDR) metric is commonly used for this purpose [9, 10]. [11] evaluated various perceptually motivated objective measures, derived from subjective audio quality assessment frameworks, and analyzed their correlation with actual human perceptual scores. A similar approach is adopted in this study to account for the subjective quality of the models' outputs. Specifically, a neural network is trained on a subset of MUSHRA ratings to approximate subjective scores and then used to evaluate the performance of the trained models.

2. Method

The TFC-TDF-UNet v3 architecture [5] was employed for the experiments. This model is the third iteration of the TFC-TDF-UNet architecture [12, 13]. It consists of a series of blocks where Time-Frequency Convolutions (TFC) are followed by Time-Distributed Fully connected (TDF) layers. It was initially employed by [12] for singing voice separation. It showed promising results, motivating the development of the v2 model that was used in the KUIELab-MDX-Net method that won the MDX21 challenge [14].

Specifically, this architecture is discriminative and trained to estimate source waveforms directly from a mixture waveform as input. However, it operates primarily in the spectrogram domain – specifically using complex-valued spectrograms in a CaC (Complex as Channels) manner, where both the imaginary and real-valued parts of the spectrogram are used as separate real-valued channels – and utilizes STFT and iSTFT as intermediate, non-trainable steps to transition between signal representations.

In this work, three models are trained on the MUSDB18 training set, each utilizing a different loss function: mean-squared error loss, hard masking loss, and soft masking loss. Hyperparameters from the original model configuration are retained, and training is conducted on the complete MUSDB18 training set comprising 100 full songs. Evaluation is performed on the MUSDB18 test set, which contains 50 songs. All source and mixture signals used for training and evaluation are represented as stereophonic (2-channel) signals at a sampling rate of 44.1 kHz. All models reported in this paper were trained on a single NVIDIA Tesla T4 GPU. Key entities relevant to the experimental setup – such as soft masking loss, MUSHRA score and NISQA model – are introduced further.

The concept of loss masking involves multiplying loss values, computed between the model output and the actual source signal during training, by a binary mask $m_i \in \{0,1\}, i = 1...N$, where m_i is the i-th element of the mask, and N denotes either the batch or time dimension. The authors in [5] apply loss masking along the batch dimension to address label noise artefacts (entire batches with high loss values are completely discarded from training) and along the time dimension to address bleeding artefacts (signal entries with high loss values are masked). This training procedure will be referred to throughout the rest of the paper as the hard-masking approach.

Applying hard masking loss results in ignoring a portion of the training data (approximately 50% in the MUSDB18 dataset, specifically regarding time dimension masking), which may be critical in scenarios with limited data availability. To address this, in addition to hard masking and standard mean-squared error loss, the soft masking loss training approach is explored. In this method, instead of applying hard masks, soft masks are utilized by weighting loss values – specifically along the time dimension – inversely proportional to their magnitude, i.e. $m_i \in [0,1], i = 1...N$, therefore enabling gradient updates from all available training data while retaining suppression capability for the corrupted samples. The impact of the soft masking loss training strategy in a limited training data setting is further discussed in Section 3.

MUSHRA (short for Multiple Stimuli with Hidden Reference and Anchor) is a method for conducting listening tests to evaluate the perceived quality of audio signals, widely employed in the audio industry to assess the perceived quality of audio coding algorithms [15]. It follows a specific set of standardised rules for gathering test signals, selecting assessors for the test, conducting the test, and evaluating its results. During the test, listeners are presented with the reference signal and a set of test signals that have been modified according to predefined conditions. The main characteristic of the test is that these test signals contain a hidden reference signal and two anchors, usually 3.5 and 7 kHz low-pass versions of the reference signal. Additionally, the participants are exposed to all test signals simultaneously. This methodology helps to calibrate the scores and detect inconsistencies in grading. It also helps to achieve statistically significant results with fewer participants involved in a test.

The MUSHRA score is measured on a scale of 0-100, which is broken into five major quality categories: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100). The 0-100 MUSHRA scale offers the advantage of more fine-grained scoring compared to the Absolute Category Rating (ACR) scale, which is used in the Mean Opinion Score (MOS) measure, where audio quality is rated on a scale of 1 to 5.

The quality of the outputs produced by a source separation model is often assessed using either objective or subjective quality measures. A de facto standard for objective evaluation throughout the music source separation literature is the Signal-to-Distortion Ratio (SDR) metric. It is commonly reported alongside related measures such as the Signal-to-Interference Ratio (SIR) and the Signal-to-Artefact Ratio (SAR). Scale-invariant (SI) variants of these three metrics (SI-SDR, SI-SIR, SI-SAR) are also widely used due to their invariance to the scale of the signal magnitude, while penalizing other errors. This prevents overly optimistic estimations that might otherwise arise from invariance to filtering and misalignment [10].

In this work, the SDR metric is reported for all experiments as the primary objective quality measure. SDR values reported here are computed using the museval toolkit [16] implementation, which follows the definition provided in [9], i.e.:

$$SDR = 10\log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2}, \qquad (1)$$

where e_{interf}, e_{noise} and e_{artif} are interferences, noise and artefacts error terms, respectively, and are defined according to [9].

To evaluate the perceptual quality of the outputs produced by the described models, MUSHRA scores are approximated by fine-tuning a reference-free audio quality prediction model proposed by [17]. Specifically, the NISQA model – a CNN-attention-based deep architecture pre-trained on the NISQA corpus, which comprises 81 datasets of crowdsourced Mean Opinion Score (MOS) ratings primarily derived from the results of various listening tests – is adapted for this purpose. The model employs a convolutional backbone combined with attention-based temporal pooling to estimate the MOS in a reference-free setting, i.e., without requiring a clean reference signal.

As the original NISQA model was not pre-trained on musical data, it is fine-tuned to estimate MUSHRA scores using the SiSEC18 corpus [16], which contains crowd-sourced MUSHRA ratings for music signals. Due to the limited size of the dataset – only 148 examples after averaging ratings across listeners – a 6-fold cross-validation

approach is used. The dataset is randomly divided into six subsets, and six models are trained, with each fold serving as the validation set once, while the remaining data is used for training. During inference, the predictions from all six models are averaged to maximize the use of the entire dataset.

Only the final fully connected layer of the model is fine-tuned. The learning rate is set to 0.001 with a batch size of 16, and learning rate annealing is applied by a factor of 0.1 if the validation RMSE does not improve for five consecutive epochs. Training is performed for up to 150 epochs, with early stopping applied using a patience of 20 epochs.

3. Results

The results of the evaluation are presented in Tables 1 and 2. For both SDR and MUSHRA metrics, higher values indicate better performance.

Table 1

Evaluation results (SDR)

Instrument	MSE	Soft mask	Hard mask
Vocals	6,64±2,66	7,87±2,99	8,11±3,03
Bass	5,12±3,20	5,99±3,53	5,97±3,54
Drums	6,85±2,47	7,65±2,73	7,81±2,63
Other	4,36±1,93	5,31±2,02	5,37±2,07

Table 2

Evaluation results (MUSHRA)

Instrument	MSE	Soft mask	Hard mask
Vocals	48,46±4,12	49,73±3,83	50,31±3,57
Bass	45,43±1,92	45,65±2,04	45,45±2,10
Drums	37,17±3,43	39,55±3,21	39,12±3,37
Other	46,73±4,12	46,66±4,48	47,05±4,64

For the SDR metric, both hard and soft loss masking approaches demonstrate substantial improvements – approximately 1 dB SDR across all instruments – compared to the baseline MSE loss. The soft masking model performs on par with the hard masking model in this regard.

A similar trend is observed in the MUSHRA-based evaluation. Both hard and soft masking models show clear improvements over the MSE baseline, particularly for vocals and drum sources. For the remaining two sources, all three models perform comparably in terms of estimated subjective quality.

These results highlight the potential of the loss masking approach to generalize well across open-source datasets, which serve as standard benchmarks in the development and evaluation of music source separation models. However, current work only considers the TFC-TDF-UNet v3 architecture and the MUSDB18 dataset, thus needing further investigation into the impact of the loss masking approach when training other model architectures using data from different sources.

As previously discussed, the hard loss masking approach inherently discards a portion of the training data. Consequently, the proposed soft loss masking strategy is hypothesized to replicate the behaviour of hard masking while retaining access to more data, which is particularly beneficial in scenarios with limited training resources. To evaluate this hypothesis, an ablation study is conducted to assess the performance of loss masking strategies under constrained training data conditions, using both objective and approximated subjective evaluation metrics.

Three models are trained for each of the two loss masking strategies – soft masking and hard masking – using 75%, 50%, and 25% subsets of the original MUSDB18 training data, as described in the previous section. These subsets are created by randomly excluding 25, 50, and 75 songs, respectively, from the MUSDB18 training set. All models are evaluated using the original test subset of the MUSDB18 dataset.

The results of the evaluation are presented in Tables 3-6.

Table 3

SDR metric for each training subset evaluated against each instrument for models trained using soft loss masking objective

Instrument	75%	50%	25%
Vocals	7,88±3,02	7,89±3,17	7,72±3,06
Bass	5,91±3,47	5,61±3,68	5,14±3,69
Drums	7,76±2,73	7,47±3,06	7,15±3,01
Other	5,30±1,93	5,13±2,17	4,85±2,05

Table 4

SDR metric for each training subset evaluated against each instrument for models trained using hard loss masking objective

Instrument	75%	50%	25%
Vocals	7,96±3,03	7,88±3,10	7,61±3,29
Bass	5,89±3,61	5,65±3,52	5,23±3,67
Drums	7,74±2,93	7,52±3,05	7,17±3,05
Other	5,28±1,94	5,18±2,11	4,83±2,08

Table 5

MUSHRA metric for each training subset evaluated against each instrument for models trained using soft loss masking objective

Instrument	75%	50%	25%
Vocals	49,36±4,02	49,76±3,74	50,23±3,28
Bass	45,73±1,95	45,15±2,15	45,45±1,77
Drums	38,58±3,50	38,78±3,33	39,34±3,56
Other	46,43±4,42	46,11±4,85	47,76±4,21

Table 6

MUSHRA metric for each training subset evaluated against each instrument for models trained using hard loss masking objective

Instrument	75%	50%	25%
Vocals	49,52±3,73	50,09±3,50	50,19±3,16
Bass	45,49±2,01	45,13±2,26	45,44±2,34
Drums	38,66±3,49	38,68±3,16	39,27±3,51
Other	46,91±4,27	46,67±4,30	47,32±3,82

In terms of the SDR metric, the soft mask model outperforms the hard mask model on 75% subset of the training data across most sources, except for the vocals source. Additionally, the soft mask model exhibits a lower standard deviation, suggesting more consistent estimates across different tracks. Notably, for the vocals source, the soft mask model surpasses the hard mask model when trained on 50% and 25% of the data, with a margin of approximately 0.1 dB SDR on the 25% subset. This result suggests the potential of the soft masking approach for singing voice extraction under conditions of limited training data and the presence of bleeding artefacts in the training data. Overall, both soft and hard masking models achieve comparable performance across different training subset sizes, except for vocals at 75% and bass at 25%, where slight differences are observed.

An interesting observation is that the performance of the soft masking model on the vocals source consistently improves as the amount of available training data is reduced – a trend not clearly observed for the other sources. One possible explanation for this behaviour is the varying presence of bleeding artefacts across different instrument sources within the test subset. However, verifying this hypothesis would require auditory inspection of individual samples from each instrument source in the test set.

Regarding the MUSHRA evaluation, the results across different training subset sizes are mainly consistent with those obtained using the whole training set. In many cases, the differences in perceptual quality between the models fall within the range of standard deviation, indicating marginal variation. Notably, on the 25% training subset, the soft mask model outperforms the hard mask model across all instrument sources.

It is also worth noting that the SDR values for the "other" source exhibit the lowest standard deviations across all sources and models. In contrast, the approximated MUSHRA metric for the "other" source consistently shows the largest standard deviations across all model configurations. One possible explanation is that, since the "other" source encompasses all remaining instruments – whose number and type usually vary across songs – the definition of artefacts becomes less clear for this source. This, in turn, may render the "other" source more "forgiving" to artefacts regarding the SDR metric and yield less stable estimates in terms of the approximated MUSHRA metric.

Overall, the findings presented in this section highlight the potential of the soft masking approach when training with limited or corrupted data. In particular, the soft mask model achieves results that are not only superior to the baseline but also comparable to, and in some cases better than, the hard masking approach, especially in scenarios where the training data includes bleeding artefacts.

Conclusions

The impact of the loss masking on training music source separation models under limited data conditions – particularly when the data includes artefacts such as "bleeding" – was investigated. Evaluation is conducted using both traditional objective metrics (SDR) and perceptual scores (MUSHRA). Results indicate that the soft loss masking approach can achieve performance comparable to hard loss masking, while offering the advantage of incorporating gradient updates from all training batches – an essential consideration in low-data regimes.

While this study primarily focused on evaluating the proposed soft loss masking approach using a single model architecture (TFC-TDF-UNet v3) and a specific dataset (MUSDB18) a broader investigation of the method's generalizability across alternative model architectures, such as Conv-TasNet, Open-Unmix, or hybrid time-frequency models was not undertaken due to constraints in time and computational resources and thus remains a direction for future research. Similarly, the impact of combining loss masking with various base loss functions warrants further exploration.

Additionally, while this work focuses on the MUSDB18 dataset, many publicly available music separation datasets (e.g., Slakh2100, MoisesDB, and others) vary in terms of source contents and the amount of artefacts present among these sources. Extending the evaluation to these datasets would provide further insight into the robustness and transferability of the loss masking strategy across a broader range of diverse data sources, including both real-world and synthetic data.

References:

[1] Fabbro G. The Sound Demixing Challenge 2023 – Music Demixing Track / G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martinez-Ramírez, W. Liao, I. Gadelha, G.Ramos, E. Hsu, H. Rodrigues, F.-R. Stöter, A. Defossez, Y. Luo, J. Yu, D. Chakraborty, S. Mohanty, R. Solovyev, A. Stempkovskiy, T. Habruseva, Y. Mitsufuji // Transactions of the International Society for Music Information Retrieval. – 2024. – V. 7. – P. 63-84.

[2] Rafii Z. The musdb18 corpus for music separation / Z. Rafii, A. Liutkus, F. Stoter. – 2017.

[3] Manilow E. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity / E. Manilow, G. Wichern, P. Seetharaman, J. Le Roux // Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). – 2019. – P. 45-49.

[4] Pereira I. Moisesdb: A dataset for source separation beyond 4-stems / I. Pereira, F. Ara jo, F. Korzeniowski, R. Vogl // preprint arXiv:2307.15913. – 2023. – 8 p.

[5] Kim M. Sound demixing challenge 2023 music demixing track technical report: Tfc-tdf-unet v3 / M. Kim, J. H. Lee, S. Jung // preprint arXiv:2306.09382. – 2023. – 5 p.

[6] Uhlich S. Improving music source separation based on deep neural networks through data augmentation and network blending / S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, Y. Mitsufuji // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2017. – P. 261-265.

[7] Jeon C.-B. Why does music source separation benefit from cacophony? / C.-B. Jeon, G. Wichern, F. G. Germain, J. Le Roux // 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). – 2024. – P. 873-877.

[8] Gusó E. On loss functions and evaluation metrics for music source separation / E. Gusó, J. Pons, S. Pascual, J. Serrà // 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2022. – P. 306-310.

[9] Vincent E. Performance measurement in blind audio source separation / E. Vincent, R. Gribonval, C. Fevotte // IEEE Transactions on Audio, Speech, and Language Processing. – 2006. – V. 14. – №. 4. – P. 1462-1469.

[10] Le Roux J. Sdr–half-baked or well done? / J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – P. 626-630.

[11] Torcoli M. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence / M. Torcoli, T. Kastner, J. Herre // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – V. 29. – P. 1530-1541.

[12] Choi W. Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation / W. Choi, M. Kim, J. Chung, D. Lee, S. Jung // preprint arXiv:1912.02591. – 2019. – 8 p.

[13] Kim M. Kuielab-mdx-net: A two-stream neural network for music demixing / M. Kim, W. Choi, J. Chung, D. Lee, S. Jung // preprint arXiv:2111.12203. – 2021. – 7 p.

[14] Mitsufuji Y. Music demixing challenge 2021 / Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, K.-W. Cheuk // Frontiers in Signal Processing. – 2022. – V. 1.

[15] International Telecommunication Union Radiocommunication Sector (ITU-R), BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA) / 2015. – URL https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf.

[16] Stöter F.-R. The 2018 signal separation evaluation campaign / F.-R. Stöter, A. Liutkus, N. Ito // International Conference on Latent Variable Analysis and Signal Separation. Cham: Springer International Publishing. – 2018. – V. 10891. – P. 293-305.

[17] Mittag G. A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets / G. Mittag, B. Naderi, A. Chehadi, S. Möller // Proc. Interspeech 2021. – 2021. – P. 2127-2131.

Date of submission of the article to the editorial board: 28.11.2025

УДК 004.8 DOi 10.30837/ bi.2025.2(103).06

Н. Б. Гулієв[1], О. С. Назаров[2]
[1]ХНУРЕ, м. Харків, Україна, nural.huliiev@nure.ua, ORCID iD: 0000-0003-2123-0377
[2]ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua, ORCID iD: 0000-0001-8682-5000

ДОСЛІДЖЕННЯ МЕТОДІВ НАЛАШТУВАНЬ ГІПЕРПАРАМЕТРІВ ДЛЯ РЕАЛІЗАЦІЇ АЛГОРИТМУ ВИПАДКОВИЙ ЛІС НА ОСНОВІ МЕДИЧНИХ ТА ПСИХОЛОГІЧНИХ ДАНИХ

Випадковий ліс є одним із найпоширеніших алгоритмів машинного навчання, що належить до методів ансамблевого навчання. Його застосовують у медицині, фінансах, соціальних науках, екології, IT та багатьох інших сферах. Сутність алгоритму полягає у створенні великої кількості дерев рішень і подальшому об'єднанні їхніх результатів для отримання точного та стабільного прогнозу. Попри численні переваги, випадковий ліс має й недоліки, зокрема низьку стійкість до різнорідності даних, що часто трапляється в медицині. У дослідженні алгоритм застосовується для аналізу медичних даних із психологічними показниками. Медичні дані мають порогові значення, які можуть давати неочікувані результати, тому оптимізація випадкового лісу залишається актуальною. Для аналізу альтернативних варіантів удосконалення обрано метод лінійної адитивної згортки. Він дозволяє обчислювати зважену суму нормалізованих показників, щоб порівнювати різні рішення. Цей метод є універсальним, простим у реалізації та придатним для задач із багатьма різнорідними критеріями. Досліджено способи оптимізації алгоритму випадковий ліс через налаштування гіперпараметрів. Розглянуто рандомізований пошук, пошук за сіткою та байєсівську оптимізацію. Проаналізовано їхні реалізації, особливості та можливі комбінації. На основі оцінки ефективності визначено, що для медичних і психологічних даних найкращим підходом є байєсівська оптимізація. Вона забезпечує більш точні та стабільні результати. Зрештою обрано найбільш оптимальний спосіб удосконалення алгоритму.

БАЙЄСІВСЬКА ОПТИМІЗАЦІЯ, ВИПАДКОВИЙ ЛІС, ГІПОТИРЕОЗ, ГІПЕРТИРЕОЗ, ДЕРЕВА РІШЕНЬ, ОПТИМІЗАЦІЯ, ПОШУК ЗА СІТКОЮ, ПСИХОЛОГІЧНІ РОЗЛАДИ, РАНДОМІЗОВАНИЙ ПОШУК.

N. B. Huliiev, O. S. Nazarov. Study of hyperparameter tuning methods for implementing the Random Forest algorithm based on medical and psychological data. Random Forest is one of the most widely used machine learning algorithms and belongs to ensemble learning methods. It is applied in medicine, finance, social sciences, ecology, IT, and many other fields. The essence of the algorithm lies in creating a large number of decision trees and then combining their results to obtain a more accurate and stable prediction. Despite its numerous advantages, Random Forest also has drawbacks, including low robustness to heterogeneous data, which is common in medical datasets. In this study, the algorithm is used to analyze medical data with psychological indicators. Medical data contains threshold values that may produce unexpected results, which makes the optimization of Random Forest still relevant. To analyze alternative improvement options, the linear additive convolution method was chosen. It allows calculating a weighted sum of normalized indicators to compare different solutions. This method is universal, easy to implement, and suitable for problems involving many heterogeneous criteria. Various approaches to optimizing the Random Forest algorithm through hyperparameter tuning were examined. Random search, grid search, and Bayesian optimization were considered. Their implementations, characteristics, and possible combinations were analyzed. Based on the evaluation of effectiveness, Bayesian optimization was identified as the best approach for medical and psychological data. It provides more accurate and stable results. Ultimately, the most optimal method for improving the algorithm was selected.

BAYESIAN OPTIMIZATION, RANDOM FOREST, HYPOTHYROIDISM, HYPERTHYROIDISM, DECISION TREES, OPTIMIZATION, GRID SEARCH, PSYCHOLOGICAL DISORDERS, RANDOMIZED SEARCH.

Вступ

Медична сфера налічує чимало випадків ускладнень через наявні хвороби, що також стосується гіпо- та гіпертиреозу, адже вони спричиняють чимало інших хвороб, що може заважати лікувати першочергову недугу. Одним із таких чинників є психологічні розлади.

Об'єкт даного дослідження є саме процес прогнозування психологічних розладів у пацієнтів із гіпо- та гіпертиреозом на основі медичних і психологічних показників.

Наразі автори вже робили аналіз даної проблеми та дійшли до того, що застосовуватимуть алгоритм випадковий ліс задля оцінки стану пацієнтів для завчасного визначення потенційної можливості погіршення психологічного стану хворих.

Предмет дослідження – методи оптимізації гіперпараметрів алгоритму Random Forest, зокрема жадібний підхід, рандомізований пошук по сітці, байєсівська оптимізація та комбінований підхід.

1. Дослідження оптимізації алгоритму випадковий ліс для аналізу даних пацієнтів

Випадковий ліс – це метод машинного навчання, який застосовується в прогнозуванні та класифікації. Алгоритми навчання моделей в машинному навчанні та штучному інтелекті вимагають великий обсяг даних задля якомога точного та якісного результату. Інформація про продуктивність розробленої системи дозволяє вдосконалити аналогічні алгоритми, підвищити ефективність апаратного та програмного забезпечення,

процеси прийняття рішень, прогнозів, вирішення проблем, що загалом покращить показники точності побудованої моделі. Вирішуючи кожну проблему, необхідно інтегрувати різні способи збору та обробки вхідних даних, що також допомагає підвищувати рівень точності. Дослідницький процес міждисциплінарних спостережень може містити різні види методологій, що застосовується в прийнятті рішень, розпізнаванні образів, вирішенні проблем та прогнозах, а також допомагає досягати інноваційності.

Випадковий ліс вважається потужним механізмом у сфері машинного навчання. Особливо популярний в прогнозуванні, супервізорному навчанні та категоризації. Випадковий ліс вирізняється чималою кількістю переваг, тому і є популярним алгоритмом в машинному навчанні та серед способів передбачення.

Застосування даного методу полягає в наступному:

– високий показник точності, що підтверджує ефективність та надійність класифікації та прогнозування, проведених за його допомогою;

– зменшення проблеми перенавчання та узагальнення результатів за рахунок будови дерев рішень на різних підмножинах даних;

– стійкість до відсутніх та незбалансованих даних;

– визначають важливість ознаки задля будування причинно-наслідкових зв'язків їх впливу на остаточний результат;

– техніка ансамблевого навчання допомагає підвищити стабільність прогнозів у порівнянні з іншими існуючими методами класифікації;

– прості в інтеграції та реалізації.

Цей метод наразі популярний серед інновацій, бо застосовує методи дерев рішень, розробляючи колекцію дерев та надаючи результат на основі усіх них. Кожне дерево будується та навчається за рахунок випадково обраної підмножини даних. Остаточне рішення надається об'єднанням усіх побудованих моделей.

Не дивлячись на те, що випадкові ліси – популярний та потужний інструмент в прогнозуванні, він потребує оптимізації у застосуванні з наборами вхідних даних, отриманих у ході біомедичних досліджень, які мають рідкісні результати та коваріати [1].

Метою роботи є підвищення точності та ефективності моделі Random Forest для раннього виявлення ризику психологічних розладів шляхом визначення найоптимальнішого методу налаштування її гіперпараметрів.

Незважаючи на багато способів удосконалення алгоритму випадковий ліс, вирішальну позицію в цьому займає саме оптимізація гіперпараметрів, а саме глибина дерева, кількість дерев рішень, мінімальний розмір вибірки. Для налаштування даних показників зазвичай застосовують рандомізований пошук, пошук по сітці та байєсівську оптимізацію. Кожен має свої переваги та недоліки, тому слід розглянути кожний окремо, щоб визначити якомога кращий, тому предметом дослідження є аналіз різних видів оптимізації налаштування гіперпараметрів за допомогою їх реалізації на мові Python та лінійної адитивної згортки.

Завдання дослідження:

– проаналізувати медичні та психологічні показники пацієнтів з Kaggle для формування набору даних;

– провести попередню обробку даних та підготовку вибірки для навчання моделі Random Forest;

– реалізувати та порівняти різні методи оптимізації гіперпараметрів:

– жадібний метод;

– рандомізований пошук по сітці;

– байєсівську оптимізацію;

– комбінований підхід (Bayesian Optimization + Random Search);

– оцінити якість налаштованих моделей за метриками Accuracy, Precision, Recall, F1-score, а також за часом оптимізації;

– визначити метод гіперпараметричної оптимізації, який забезпечує найкраще співвідношення точності та обчислювальних витрат.

Розглянемо методи та засоби спостереження більш детально – згортку, яка використовується в задачах прийняття рішень, що і є задачею експерименту.

2. Матеріали і методи досліджень

Розв'язуючи багатокритеріальні задачі, результатом завжди є кращі альтернативи, які відповідають поставленим вимогам. Найчастіше тут використовуються методи двох видів: перша полягає у виключенні кількості критеріїв оцінки, а друга зменшує кількість варіантів аналізу на його початку. Для нашого дослідження найбільш підходящим є саме метод із першої групи. Такими способами є метод граничних та головного критеріїв, відстані та згорток.

Методи згорток поділяються на лінійні адитивні, мультиплікативні та максимінні. Метою застосування згорток є узагальнення усіх критеріїв аналізу.

Адитивна розраховується за наступною формулою:

$$K(x)=\sum_{j=1}^{n} a_j K_j(x) \qquad (1),$$

де $K(x)$ – загальний критерій для альтернативи $x \in X$, $(K_1(x), ..., K_j(x), ..., K_n(x))$ – набір вихідних критеріїв; n – число вихідних критеріїв; $a_j(x)$ – нормуючий множник, який вказує на вагу альтернативи.

Найкращий із усіх можливих альтернатив задачі обчислюється за допомогою наступної формули:

$$x^n = \arg\max_{x \in X} K(x) \qquad (2),$$

Тобто результатом є найбільше значення, отримане методом згортки.

Мультиплікативна згортка розраховується за допомогою такої формули:

$$K(x)=\prod_{j=1}^{n}K_j^{a_j}(x) \quad (3)$$

Максимінна згортка знаходиться за формулою:

$$K(x)=\max_i \min_j a_{ij}K_j(x) \quad (4).$$

Найкращі результати за мультиплікативною та максимінною згортками обчислюється за формулою (2).

Метод граничних критеріїв застосовується в задачах проектування і планування, в яких порогові значення критеріїв набувають значень $k_j(x)\geq k_{jo}$; $j=1$, ..., n. Формула обчислення цього способу наступна:

$$K(x)=\min_j\left(\frac{K_j(x)}{K_{jo}(x)}\right) \quad (5)$$

Найкращий результат обирається формулою 2.

Метод відстані використовує відстань, яка є додатковою метрикою. Наприклад, для вибору ідеального рішення цілком достатньо інформації. Обчислимо відстань до значення максимуму d(x) для кожної альтернативи. Тоді найкраща альтернатива буде відомою із застосуванням формули:

$$x^*=\arg\min_{x\in X} d(x) \quad (6)$$

Застосування методу з першої множини іноді вимагає один із способи із другої, наприклад – принцип Парето: альтернативи, які за всіма критеріями програють іншому або іншим варіантам, видаляються до початку дослідження.

Також бувають випадки, в яких параметри, які неконтрольовані через різні причини, ускладнюють будову моделі для подальшого аналізу. Тут у нагоді може стати метод гарантованого результату, мета якого полягає у визначенні найгіршої реакції та гарантованого значення.

Для спостереження варто використати згортку, адже важко визначити порогові значення критеріїв аналізу, а саме лінійну адитивну, яка є найпоширенішою та найпростішою, та метод із другої множини способів – принцип Парето, якщо одна із альтернатив прозоро гірша за інші.

Спершу необхідно обрати критерії, за якими проводитиметься дослідження: альтернативи порівнюватимуться за допомогою цих ознак.

Значення кожної з них може мати як кількісне, так і якісне походження. Згортка оперує із першими, тому у випадку других, необхідно конвертувати їх у кількісні та побудувати нову таблицю вхідних даних варіантів аналізу.

На третьому кроці виключатимемо альтернативи за допомогою принципу Парето, якщо усі її показники за усіма її ознаками менші з-поміж інших можливих значень інших варіантів експерименту. Варто зазначити, якщо показники альтернатив в різних проміжках або мірах вимірювання, необхідно нормалізувати дані максимізацією або мінімізацією даних, щоб точність та коректність результатів відповідала дійсності.

Четвертим етапом є ранжування показників – обчислення вагомих коефіцієнтів. Існують різні способи, в даному спостереженні можна застосувати один із найбільш популярних методів: для кожного критерію один ділитимемо на суму усіх її значень.

Останнім етапом залишається обчислення значення згортки для кожної із альтернатив: розрахунок суми добутків кожної пари значень вагових коефіцієнтів та критеріїв.

Задачею дослідження якраз є вибір найкращого або найкращих методів удосконалення способу налаштування гіперпараметрів для оптимізації алгоритму випадковий ліс.

Проведемо експеримент та розв'яжемо багатокритеріальну задачу вибору способу налаштування гіперпараметрів для реалізації алгоритму випадковий ліс для застосування в задачі аналізу психічних розладів серед хворих гіпотиреозом та гіпертиреозом. Розпочнемо із застосування існуючих методів.

3. Аналіз літературних джерел

Наразі можна розрізнити пошукові та підтверджувальні експерименти, а тому розуміти, коли їхнє застосування доцільне. У цій статті описано використання саме алгоритму випадкових лісів, які спроможні надати кращі прогнози, ніж регресія, та визначати нелінійні ефекти. Даний алгоритм застосовується вже в багатьох сферах: банківські справи, фармацевтика, біржа, охорона здоров'я тощо. Однак, автори мають думку, що в психології – набагато рідше, тому вирішили розглянути його використання в контексті психологічних досліджень. Особливу увагу вони приділили обмеженням, які можуть виникнути в цьому випадку, та шляхам їх уникнення та вирішення, розглянувши існуючі дослідження детальніше далі.

Програмне забезпечення може зазнавати збоїв під час роботи. Задля мінімізації даних проблем необхідно ефективно прогнозувати можливі помилки. Існує дослідження, мета якого була якраз передбачення несправностей в роботі програмного забезпечення за допомогою вдосконаленого алгоритму випадкового лісу на основі даних NASA JM1, які налічували 21 програмну метрику. Спочатку метод усував дисбаланс класів способом надмірної вибірки синтетичної меншини (SMOTE). Суттю нового підходу було налаштування класифікатора випадкових лісів з увагою на оптимізацію гіперпараметрів. Під час порівняння модифікованого методу із стандартними у машинному навчанні він мав кращі показники точності та F1. Підкреслено, що важливо пам'ятати про можливі дефекти програм та шляхи їх передбачення задля підвищення продуктивності програмного забезпечення.

Спочатку проходить обробка вхідних даних: видаляються нульові значення та інші проблемні дані. Потім для усування дисбалансу використовується метод

SMOTE. На наступному кроці відбираються ознаки випадкового лісу, які безпосередньо впливатимуть на процес будування та структуру дерев рішень. Для більшої ефективності роботи алгоритму було скомбіновано два методи оптимізації алгоритму: усунення дисбалансу класів та налаштування гіперпараметрів рандомізованим пошуком по сітці. [2]

Ефективність комунікаційних та радіолокаційних систем залежить від інверсії атмосферних каналів. А на продуктивність та якість прогнозування моделі машинного навчання впливають параметри, які безпосередньо беруть участь в її реалізації. Тому в одному із експериментів розроблено модель випадкового лісу, вдосконалену за допомогою методу байєсівської оптимізації, для прогнозування атмосферних каналів. Оптимізацію застосовано задля пошуку певних гіперпараметрів під час навчання. Додатково використано метод K-кратної перехресної перевірки для визначення кращого способу поділу моделі та уникнення проблеми її перенавчання. Для перевірки реалізованого алгоритму його результати порівнювались із результатами, розрахованими за допомогою інших популярних методів прогнозування: класичний алгоритм випадкового лісу, метод k-найближчого сусіда з та без байєсівської оптимізації та метод градієнтного підсилення з та без байєсівської оптимізації. Зрештою, показники нового методу коефіцієнту детермінації R2, MAE та MSE були більшими, а результати прогнозування більш точними. Також визначено, що результати кращі навіть у випадку наявності шуму в даних.[3]

Енергетична безпека забезпечується відсутністю вторгнень в енергетичні промислові системи, тому метод їх виявлення є конче важливим в даній галузі. Існують два способи, але вони мають недоліки: вони добре працюють з гіперпараметрами, але їхня оптимізація може суттєво підвищити показники точності моделі виявлення вторгнень, а також вони взагалі застосовуються для контролю безпеки інформаційних систем, а не окремо для моніторингу атак на управління енергетичних систем. Тому в одному із експериментів було запропоновано модель випадкового лісу для виявлення вторгнень, де було використано метод improved grid search algorithm в якості оптимізації гіперпараметрів для покращення показників ефективності майбутньої моделі. Новий метод аналізувався на основі даних управління державної енергетичної системи. Точність досягла 98%.

У статті описано модифіковану модель виявлення вторгнень у промислові енергетичні системи задля вирішення наявних недоліків існуючої реалізації. Було запропоновано оптимізувати гіперпараметри покращеним методом сіткового пошуку: параметри налаштовувались у порядку важливості для збільшення продуктивності моделі. Швидкість нового способу була в 165 разів вищою, аніж показники швидкості роботи звичайного сіткового пошуку. Тому даний метод може застосовуватися в даній галузі, однак він ще не є ідеальним підходом: можна застосувати більше алгоритмів машинного навчання, наприклад, глибоке навчання або спробувати оптимізувати гіперпараметри за допомогою алгоритмів метаевристичного пошуку. Тим паче наразі дана модель недостатньо інтерпретована. [4]

4. Експериментальні дослідження

Експерименти проводилися на двох наборах даних: FF++ та DFDC [5, 6]. FF ++ — це великий набір даних по маніпуляціях з обличчям, створений з використанням state-of-the- art методів редагування відеозаписів. Цей набір даних містить два класичних підходи маніпуляції обличчями,а саме Face2Face і FaceSwap, разом з двома стратегіями, ґрунтованими на навчанні (DeepFake і NeuralTextures). Кожен метод застосовувався до 1000 високоякісних відеозаписів, завантажених з YouTube, щоб показувати зображення без перешкод і зайвих об'єктів. Усі послідовності містили не менше 280 кадрів. Для імітації реалістичних налаштувань відеозаписи було стиснено з використанням кодека H.264. Відеозаписи високої та низької якості генерувалися з використанням параметра квантування з постійною швидкістю, рівною 23 і 40 відповідно.

Проведемо дослідження та оберемо найпідходящий спосіб налаштування гіперпараметрів для алгоритму випадковий ліс, написаного задля аналізу медичних та психологічних показників.

У дослідженні альтернативами виступатимуть глобальний пошук за сіткою, випадковий пошук, байєсівська оптимізація.

Для цього поділимо дослідження на дві складові: теоретичну та практичну. Застосуємо згортку для них.

4.1. Теоретична складова

Критеріями розгляду, за якими будуть будуватись три моделі, будуть такі атрибути, як: age – вік, sex – стать, on_thyroxine – чи приймає тироксин, query_on_thyroxine – запит на тироксин, on_antithyroid_meds – чи приймає антитиреоїдні ліки, sick – чи хворий, pregnant – вагітність, thyroid_surgery – чи робилась операція на щитоподібній, I131_treatment – лікування радіоактивним йодом, query_hypothyroid – підозра на гіпотиреоз, query_hyperthyroid – підозра на гіпертиреоз, lithium – чи приймає літій, goitre – зоб, tumor – пухлина, hypopituitary – гіпопітуїтаризм, psych – чи є психічні розлади, TSH_measured – чи вимірювався TSH, TSH – значення TSH, T3_measured – чи вимірювався T3, T3 – значення T3, TT4_measured – чи вимірювався TT4, TT4 – значення TT4, T4U_measured – чи вимірювався T4U, T4U – значення T4U, FTI_measured – чи вимірювався FTI, FTI – значення FTI, TBG_measured – чи вимірювався TBG, TBG – значення TBG, referral_source – джерело направлення, target – цільовий клас, patient_id – ID пацієнта.

Побудуємо спочатку таблицю теоретичних відомостей трьох моделей (див. табл. 1).

Таблиця 1

Характеристики способів налаштування гіперпараметрів

Метод	Витрати часу	Обчислювальна складність	Гарантія оптимуму	Гнучкість	Простота реалізації	Використання ресурсів
Grid Search	Високі	Експоненційна	Так, якщо оптимальні параметри є в сітці	Низька	Висока	Високі
Random Search	Середні	Лінійна або нижча, ніж у Grid Search	Ні, але може знайти хороший набір параметрів	Висока	Висока (легко реалізується)	Менші, ніж у Grid Search
Bayesian Optimization	Низькі (порівняно з Grid/Random)	Середня	Висока ймовірність знаходження оптимуму	Висока	Середня	Оптимізоване

Наступним кроком експерименту є конвертування якісних показників у кількісні.

Чим менше часу необхідно для роботи алгоритму, тим краще. Якщо витрати часу високі, то оцінкою буде 1 бал, якщо середні – 2 бали, а коли низькі в порівнянні із глобальним пошуком за сіткою та випадковим способом – 1 бал.

Розглянемо, наскільки важко проводити розрахунки. У випадку, коли обчислювальна складність експоненційна, тобто залежить від кількості параметрів, то дана характеристика описується як 1 бал, якщо ж середня – 2 бали, а коли – лінійна – 3 бали.

Якщо гарантія оптимуму достовірна, то це – 2 бали, якщо ні – 0, а у випадку, коли є висока ймовірність, – 1 бал.

Гнучкість може бути низькою (1 балів) та високою (2 бали).

Чим простіше, тим реалізація менш складна. Якщо простота реалізації висока, то це – 1 бал, а коли середня – 2 бали.

Якщо ж застосування ресурсів високе, то це 1 бал, коли воно краще, ніж «високе» – 2 бали, коли оптимізоване – 3 бали.

Заповнимо нову таблицю з кількісними даними способів налаштувань гіперпараметрів (див. табл. 2).

Таблиця 2

Кількісні показники способів налаштування гіперпараметрів

Метод	Витрати часу	Обчислювальна складність	Гарантія оптимуму	Гнучкість	Простота реалізації	Використання ресурсів
Grid Search	1	1	3	1	2	1
Random Search	2	3	1	2	2	2
Bayesian Optimization	3	2	2	2	1	3

Обчислимо значення згортки для кожного способу та визначимо найкращий із них (див. табл. 3).

Таблиця 3

Результати

Метод	Результати згортки
Grid Search	1,6
Random Search	2,13333333
Bayesian Optimization	2,26666667

4.2. Практична складова

Додамо четвертий спосіб, який теж не рідше використовується, а саме – комбінацію рандомізованого способу та байєсівської оптимізації.

Написаний код на python показав, що чотири алгоритми мають такі показники та одразу обчислимо значення лінійної адитивної згортки (див. табл. 4):

Таблиця 4

Числові характеристики алгоритмів

Метод	Час	Accuracy	Precision	Recall	F1	Згортка
Grid Search	6,85	0,9427	0,6106	0,5611	0,5757	1,115054794
Random Search	11,9	0,9427	0,6106	0,5611	0,5757	1,196717149
Bayesian Optimization	21,69	0,9427	0,6139	0,5491	0,5654	1,346458789
Random + Bayes	21,4	0,9427	0,6139	0,5491	0,5654	1,341769268

Результати лінійної адитивної згортки показують, найкращим способом налаштування гіперпараметрів є байєсівська оптимізація. Даний метод бере до уваги попередні показники для того, щоб будувати моделі функції втрат та ефективніше обирає наступні параметри. Порівняно з алгоритмами глобального пошуку за ставкою та рандомізованого пошуку, його витрати часу низькі. Звичайно, його обчислювальна складність та простота реалізації середні, адже залежать від обраної моделі аналізу та необхідні спеціальні бібліотеки для реалізації, але гарантія оптимуму та гнучкість високі, адже байєсівська оптимізація здатна пристосуватися до параметрів. Результати показують, що подальший розвиток оптимізацію алгоритму випадковий ліс на основі медичних та психологічних даних слід розпочинати з налаштування гіперпараметрів за допомогою байєсівської оптимізації.

Наукова новизна отриманих результатів полягає у встановленні ефективності байєсівської оптимізації та її комбінації з рандомізованим пошуком у задачі передбачення психологічних розладів, пов'язаних із порушеннями функції щитоподібної залози, що не була предметом спеціальних порівняльних досліджень у відкритих джерелах. Отримано нові результати щодо співвідношення точності та часу обчислень для різних стратегій оптимізації гіперпараметрів Random Forest у медичних задачах.

Практична значущість результатів показує, що полягає у можливості застосування оптимізованої моделі Random Forest як інструмента раннього виявлення ризику психологічних порушень у пацієнтів із гіпо- та гіпертиреозом. Запропонована методика налаштування гіперпараметрів може бути впроваджена у медичні інформаційні системи, скринінгові програми та системи підтримки клінічних рішень для підвищення точності діагностики та зменшення навантаження на медичних фахівців.

Байєсівська оптимізація вважається ще новим інструментом налаштування гіперпараметрів та загальної оптимізації функцій типу «чорного ящику». Вже багато досліджень, присвячених цьому методу, відображають все більше його застосувань [7].

Гіперпараметри відіграють важливу роль в ефективності роботи моделей машинного навчання. Сьогодення налічує чимало відомих алгоритмів, які застосовуються майже в будь-якій галузі, що потребує професійності та відповідного досвіду, тому задля успішної результативності вибір гіперпараметрів вкрай важливий. Від їх значень залежать остаточні показники, надані моделлю аналізу. Тому їхня оптимізація конче потрібна в реалізації будь-яких алгоритмів машинного навчання. Модель випадкового лісу має визначати прогнози якомога коректніше, адже вони впливатимуть на подальше спостереження за здоров'ям пацієнтів, хворих на гіпотиреоз та гіпертиреоз. Тому проблема правильного налаштування гіперпараметрів є оптимізаційною багатокритеріальною задачею вибору, розв'язком якого став метод байєсівської оптимізації.

Даний спосіб заснований на теоремі Байєса, що виходить з її назви. Вона визначає значення апостеріору над функцією оптимізації та збирає дані із попередніх підмножин даних, щоб оновити показник апостеріор. Функція корисності обирає наступну точку в даних для того, щоб максимізувати значення функції оптимізації. [8-9]

Отже, вдосконаленням алгоритму випадковий ліс буде новий комбінований спосіб із таких методів, як балансування класів, зменшення кореляції між деревами та рішень та налаштування гіперпараметрів байєсівською оптимізацією.

Висновки

Випадковий ліс – один із найпотужніших інструментів машинного навчання, який широко застосовується в різних сферах як механізм класифікації та прогнозування. Не дивлячись на чималу кількість переваг даної моделі, вона має передумови для свого удосконалення [10].

Метою дослідження було визначити способи налаштування гіперпараметрів алгоритму випадковий ліс задля одного із способів оптимізації точних результатів прогнозування розвитку психологічних розладів серед людей, хворих на гіпотиреоз та гіпертиреоз. Переглянуто спостереження, які теж були спрямовані на це.

Налаштування гіперпараметрів має декілька варіантів, тому було проведено окреме дослідження: розв'язувалась багатокритеріальна задача вибору методу налаштування гіперпараметрів за допомогою лінійної адитивної згортки.

Альтернативами були рандомізований пошук, пошук за сіткою, байєсівська оптимізація комбінований спосіб із рандомізованим пошуком та байєсівською оптимізацією, а критеріями вибору витрати часу, обчислювальна складність, гарантія оптимуму, гнучкість, простота реалізації, використання ресурсів, accuracy, precision, recall та f1. Після розрахунків згортки визначено, що найоптимальнішим варіантом є саме байєсівська оптимізація [11].

Тому найкращим методом реалізації випадкового лісу є алгоритм випадковий ліс із налаштуванням гіперпараметрів за допомогою байєсівської оптимізації.

Список літератури:

[1] Salman, H.A., Kalakech, A. i Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, с. 69–79. DOI: https://doi.org/10.58496/BJML/2024/007.

[2] Thomas, N.S. та Kaliraj, S. (2024). An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. SN Computer Science, 5(5), с. 530. DOI: 10.1007/s42979-024-02764-x.

[3] Yang, C., Wang, Y., Zhang, A., Fan, H. та Guo, L. (2023). A Random Forest Algorithm Combined with Bayesian Optimization for Atmospheric Duct Estimation. Remote Sensing, 15(17), с. 4296. DOI: https://doi.org/10.3390/rs15174296.

[4] Zhu, N., Zhu, C., Zhou, L., Zhu, Y., & Zhang, X. (2022). Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm. Applied Sciences, 12(20), 10456. https://doi.org/10.3390/app122010456

[5] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. pp. 1–11. https://arxiv.org/abs/1901.08971

[6] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:1910.08854. 2019. https://arxiv.org/abs/

[7] V. Nguyen, "Bayesian Optimization for Accelerating Hyper-Parameter Tuning," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, 2019, pp. 302-305, doi: 10.1109/AIKE.2019.00060.

[8] Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H. та Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), с. 26–40. DOI: https://doi.org/10.11989/JEST.1674-862X.80904120.

[9] Lujan-Moreno, G.A., Howard, P.R., Rojas, O.G. та Montgomery, D.C. (2018). Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. Expert Systems With Applications, 109, с. 195–205. DOI: https://doi.org/10.1016/j.eswa.2018.05.024.

[10] Siji George C G and B.Sumathi. “Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction”. International Journal of Advanced Computer Science and Applications (IJACSA) 11.9 (2020). http://dx.doi.org/10.14569/IJACSA.2020.0110920

[11] Xilu Wang, Yaochu Jin, Sebastian Schmitt, and Markus Olhofer. 2022. Recent Advances in Bayesian Optimization. In Proceedings of ACM Conference (Conference’17). ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Надійшла до редколегії 04.12.2025

УДК 004.75 DOi 10.30837/ bi.2025.2(103).07

О. О. Водка[1], М. І. Шаповалова[2], В. В. Жихарєв [3]

[1]НТУ «ХПІ», м. Харків, Україна, oleksii.vodka@khpi.edu.ua;
ORCID iD: 0000-0002-4462-9869

[2]НТУ «ХПІ», м. Харків, Україна, Mariia.Shapovalova@khpi.edu.ua; ORCID iD: 0000-0002-4771-7485

[3]НТУ «ХПІ», м. Харків, Україна, Vladyslav.Zhykhariev@infiz.khpi.edu.ua; ORCID iD: 0009-0006-0640-5895

СТВОРЕННЯ МАТЕМАТИЧНОЇ МОДЕЛІ ТА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ ВИЗНАЧЕННЯ ЙМОВІРНІСНИХ ХАРАКТЕРИСТИК ЧИСТОГО МАГНІЮ ІЗ ЗАСТОСУВАННЯМ КЛІТИННИХ АВТОМАТІВ

У статті розглянуто комплексний підхід дослідження механічної поведінки полікристалічного магнію. Запропоновано методологію дослідження впливу стохастичної природи мікроструктури магнію, зумовленої особливостями його гексагональної щільноупакованої (HCP) кристалічної ґратки, на варіативність макроскопічних пружних властивостей. Метою роботи було визначення ефективних пружних характеристик полікристалічного магнію шляхом аналізу стохастично згенерованих мікроструктур та проведення серії комплексних чисельних експериментів методом скінченних елементів. В дослідженнях метод клітинних автоматів було використано для генерації 100 незалежних тривимірних моделей репрезентативних об'ємів (RVE) зернової структури. Ці моделі відрізнялися випадковим розподілом центрів кристалізації та топологією зернових меж, забезпечуючи репрезентативну вибірку мікроструктурних реалізацій. Для кожної згенерованої моделі проведено серію чисельних експериментів за допомогою методу скінченних елементів (МСЕ). Граничні умови було реалізовано шляхом завдання поля переміщень, що відповідало набору унікальних сценаріїв деформаційного навантаження. Це дозволило обчислити напружено-деформований стан для кожного (RVE). У результаті дослідження визначено ефективні пружні характеристики чистого магнію та проведено їх статистичний аналіз. Встановлено, що розподіл модулів Юнга та зсуву підпорядковується нормальному закону з високим ступенем симетрії. Аналіз отриманих даних продемонстрував, що, незважаючи на виражену локальну анізотропію окремих кристалітів, полікристалічний магній демонструє квазіізотропну поведінку зі стабільними усередненими характеристиками. Низькі стандартні відхилення параметрів свідчать про високу статистичну стабільність моделі. Результати підтверджують ефективність запропонованого підходу. Такий метод дозволяє точно відтворити реальні пружні характеристики матеріалу з мінімальною похибкою відносно довідкових даних, без необхідності проведення дорогих лабораторних випробувань.

ПОЛІКРИСТАЛІЧНИЙ МАГНІЙ, ПРУЖНІ ХАРАКТЕРИСТИКИ, КЛІТИННІ АВТОМАТИ, МІКРОСТРУКТУРА, ЧИСЕЛЬНЕ МОДЕЛЮВАННЯ, СТОХАСТИЧНИЙ АНАЛІЗ.

O. O. Vodka, M. I. Shapovalova, V. V. Zhykhariev. Development of a mathematical model and software for determining the probabilistic characteristics of pure magnesium using cellular automata. The article reviews a comprehensive approach to studying the mechanical behavior of polycrystalline magnesium. A methodology is proposed for studying the influence of the stochastic nature of the magnesium microstructure, conditioned by the peculiarities of its hexagonal close-packed (HCP) crystal lattice, on the variability of macroscopic elastic properties. The aim of the work was to determine the effective elastic characteristics of polycrystalline magnesium by analyzing stochastically generated microstructures and conducting a series of complex numerical experiments using the finite element method. In the studies, the cellular automata method was used to generate 100 independent three-dimensional models of Representative Volume Elements (RVEs) of the grain structure. These models differed in the random distribution of crystallization centers and the topology of grain boundaries, providing a representative sample of microstructural realizations. A series of numerical experiments was conducted for each generated model using the finite element method (FEM). The boundary conditions were realized by prescribing a displacement field corresponding to a set of unique strain loading scenarios. This allowed for the computation of the stress-strain state for each RVE. As a result of the study, the effective elastic characteristics of pure magnesium were determined, and their statistical analysis was performed. It was established that the distributions of Young's and shear moduli follow a normal law with a high degree of symmetry. Analysis of the obtained data demonstrated that, despite the pronounced local anisotropy of individual crystallites, polycrystalline magnesium exhibits quasi-isotropic behavior with stable averaged characteristics. Low standard deviations of the parameters indicate high statistical stability of the model. The results confirm the effectiveness of the proposed approach. This method allows for the accurate reproduction of the material's actual elastic characteristics with minimal error relative to reference data, eliminating the need for expensive laboratory testing.

POLYCRYSTALLINE MAGNESIUM, ELASTIC PROPERTIES, CELLULAR AUTOMATA, MICROSTRUCTURE, NUMERICAL MODELLING, STOCHASTIC ANALYSIS

Вступ

Завдяки низькій густині та сприятливим механічним параметрам, чистий магній Mg розглядається як перспективний матеріал для інженерних застосувань. Із густиною ~1.74 г/см^3 він значно легший за традиційні конструкційні метали, зокрема алюміній і сталь, що обумовлює зростаючий інтерес до нього у контексті зниження ваги технічних систем. Це сприяє його широкому застосуванню в автомобілебудуванні [1], авіації[2], портативній електроніці[3] та біомедичних технологіях [4].

Особливістю магнію є його структура. Сама вона визначає його механічну поведінку. Магній має гексагональну щільноупаковану (HCP) кристалічну структуру, що обмежує кількість доступних систем ковзання при кімнатній температурі, тим самим знижуючи пластичність матеріалу [5, 6]. HCP-структура спричиняє високу анізотропію, асиметрію текучості та залежність механічних властивостей від орієнтації кристалів [9], [10]. Ця особливість зумовлює необхідність застосування спеціальних методів обробки, серед яких асиметричне прокатування, попередня деформація та легування, з метою покращення деформаційних властивостей магнієвих сплавів [7]. Саме ця особливість робить його цікавим для досліджень його механічних властивостей та граничного стану.

Магній та його сплави мають широкий спектр застосувань. У автомобілебудуванні магнієві сплави використовуються для виготовлення елементів кузова, шасі, рульових колонок та сидінь завдяки їх легкості та гарній литності [1]. У біомедицині Mg-сплави розглядаються як потенційно біосумісні та біодеградуючі матеріали для імплантів [4], [11]. Проте обмежена корозійна стійкість і механічна стабільність стримують їх масове впровадження [4].

Іншим активним напрямом є дослідження способів покращення механічних властивостей магнію. Основні методи включають легування елементами, такими як Gd, Y, Ce [10], [12], [13]; термомеханічну обробку [13], [14]; а також комп'ютерне моделювання [15], [16], [17]. Ці підходи дозволяють впливати на мікроструктуру матеріалу, зокрема розмір зерен, текстуру, кількість двійників та характер фазових включень, що безпосередньо впливає на пластичність, міцність і стійкість до руйнування [7], [8], [18].

Мікроструктура магнію, включаючи розподіл зерен, орієнтацію кристалів і наявність фазових включень, є ключовим фактором, що визначає його механічні характеристики. Методи, що використовуються для дослідження мікроструктури, включають оптичну та електронну мікроскопію (SEM, TEM), електронну дифракцію зворотного розсіювання (EBSD), а також дифракцію рентгенівських променів [13], [18], [19]. Кожен із цих методів має свої переваги: TEM дозволяє аналізувати наномасштабні дефекти та механізми деформації, EBSD – визначати текстуру та орієнтацію зерен, а SEM – досліджувати морфологію поверхні.

Дослідження механічних властивостей магнію здійснюється шляхом статичних та динамічних випробувань на розтяг, стиск, згин, а також за допомогою методів наноіндентації та випробувань при високих швидкостях деформації [18], [20]. Аналіз поведінки матеріалу в умовах ударного навантаження дозволяє передбачити його реакцію в умовах експлуатації, як-от аварійні ситуації в автомобілях або динамічні навантаження в авіації.

Сучасні підходи до прогнозування властивостей магнію включають машинне навчання [11], [17], що дозволяє моделювати зв'язок між хімічним складом, мікроструктурою і механічними властивостями матеріалу. Ці методи значно скорочують цикл розробки нових матеріалів і дозволяють здійснювати інверсне проектування Mg-сплавів з заданими характеристиками.

Деякі джерела не враховують повною мірою реальні умови експлуатації, вплив багатокомпонентних систем на фазову стабільність або мають обмежену статистичну вибірку [7], [12]. Це створює потребу у подальшому дослідженні ймовірнісних характеристик чистого магнію – статистичної варіації властивостей в залежності від мікроструктурних неоднорідностей та умов обробки.

З огляду на широке застосування магнію в промисловості, його складну кристалічну будову та труднощі у прогнозуванні механічної поведінки, вивчення ймовірнісних характеристик цього матеріалу набуває особливої актуальності. Розуміння статистичних залежностей механічних властивостей чистого магнію, з урахуванням мікроструктурних особливостей і умов обробки, є не лише важливим науковим завданням, а й відкриває широкі перспективи для подальших досліджень та оптимізації матеріалів із заданими властивостями. Саме тому метою цієї роботи є встановлення ймовірнісних закономірностей механічної поведінки магнію з використанням сучасних підходів до аналізу мікроструктури.

1. Мета роботи

Основна задача роботи полягає у встановленні ефективних пружних характеристик полікристалічного магнію на основі аналізу стохастично згенерованих мікроструктур та проведенні чисельних експериментів методом скінченних елементів.

Спрямованість роботи зосереджена на дослідженні впливу випадкової мікроструктурної неоднорідності на варіативність модулів пружності, коефіцієнтів Пуассона та зсувних модулів, а також на отриманні їх надійних ймовірнісних оцінок для репрезентативного об’єму магнію.

2. Постановка задачі

Для досягнення мети планується виконати ряд наступних задач:

1) Генерація мікроструктури репрезентативного об’єму полікристалічного магнію шляхом застосування методу клітинних автоматів;

2) Обчислення напружено-деформованого стану для різних схем навантаження;

3) Гомогенізація напружено-деформованого стану;

4) Визначення еквівалентних пружних констант;

5) Статистична обробка отриманих результатів.

3. Побудова математичної моделі

3.1. Обчислення НДС

Для моделювання напружено-деформованого стану (НДС) використовується метод скінченних елементів (МСЕ). Основна ідея полягає в апроксимації поля переміщень у межах твердого тіла. Гіпотезою МСЕ є те, що поле переміщень u(x) всередині кожного елементу можна апроксимувати лінійною комбінацією вузлових значень за допомогою функцій форми:

$$u(x) = N(x) \cdot d, \tag{1}$$

де $u(x)$ – вектор переміщень у точці x, $N(x)$ – матриця функцій форми (інтерполяційних функцій), d – вектор вузлових переміщень.

У роботі для просторової дискретизації використовувався тривимірний ізопараметричний скінченний елемент SOLID185, що реалізований у програмному комплексі ANSYS Mechanical. Він має 8 вузлів, кожен з трьома ступенями свободи (u, v, w).

Поле деформацій ε визначається як градієнт переміщень у наступній формі:

$$\varepsilon = B \cdot d, \tag{2}$$

де ε – вектор деформацій, B – матриця похідних функцій форми, що перетворює переміщення у деформації. Тоді матриця жорсткості елементу визначається як:

$$K_e = \int_V B^T \cdot C \cdot B dV, \tag{3}$$

де B^T – транспонована матриця похідних функцій форми, C – матриця пружних констант.

Пружні властивості магнію, як матеріалу з гексагонально щільноупакованою (HCP) кристалічною ґраткою, описуються тензором пружності четвертого рангу C_{ijkl}, який встановлює лінійний зв'язок між тензорами напружень та деформацій відповідно до узагальненого закону Гука:

$$\sigma_{ij} = C_{ijkl}\,\varepsilon_{kl}\,, \tag{4}$$

де C_{ijkl} – тензор пружних констант четвертого рангу.

Для обчислень тензор C_{ijkl} подається у вигляді симетричної матриці розмірності 6×6:

$$\begin{pmatrix}\sigma_1\\ \sigma_2\\ \sigma_3\\ \sigma_4\\ \sigma_5\\ \sigma_6\end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & \cdots & C_{16}\\ C_{12} & C_{22} & C_{23} & \cdots & C_{26}\\ C_{13} & C_{23} & C_{33} & \cdots & C_{36}\\ C_{14} & C_{24} & C_{34} & \cdots & C_{46}\\ C_{15} & C_{25} & C_{35} & \cdots & C_{56}\\ C_{16} & C_{26} & C_{36} & \cdots & C_{66}\end{pmatrix}\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ \varepsilon_4\\ \varepsilon_5\\ \varepsilon_6\end{pmatrix} \tag{5}$$

Для матеріалів з гексагональною щільноупакованою кристалічною решіткою пружна поведінка є анізотропною, і для її опису використовується п'ять незалежних коефіцієнтів:

$$C = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & 0 & 0\\ C_{12} & C_{11} & C_{13} & 0 & 0 & 0\\ C_{13} & C_{13} & C_{33} & 0 & 0 & 0\\ 0 & 0 & 0 & C_{44} & 0 & 0\\ 0 & 0 & 0 & 0 & C_{44} & 0\\ 0 & 0 & 0 & 0 & 0 & C_{66}\end{bmatrix} \tag{6}$$

Фізична симетрія (*hcp*) ґратки магнію накладає додаткові обмеження на ці константи. Зокрема, співвідношення між компонентами $C_{66} = {}^1/_2\,(C_{11} - C_{12})$ є прямим наслідком симетрії кристала відносно осі обертання шостого порядку головної кристалографічної осі Z.

В табл. 1 наведено значення констант жорсткості для чистого магнію при кімнатній температурі.

Таблиця 1

Пружні константи жорсткості чистого монокристалічного магнію C_{ij}

Показник	Значення, ГПа
C_{11}	59.7
C_{33}	61.7
C_{44}	16.4
C_{12}	26.2
C_{13}	21.7
C_{66}	16.75

Для обчислення середніх значень деформацій і напружень використовуються методи гомогенізації. Тоді середня деформація визначається згідно (7), а середнє напруження (8):

$$\langle \varepsilon_{ij} \rangle = \frac{1}{V}\int_V \varepsilon_{ij}(x,y,z)\,dV, \tag{7}$$

$$\langle \sigma_{ij} \rangle = \frac{1}{V}\int_V \sigma_{ij}(x,y,z)\,dV, \tag{8}$$

де ε_{ij}, σ_{ij} – компоненти тензора напружень та деформацій, $\langle\square\rangle$ — середні значення.

3.2. Визначення еквівалентних пружних констант

Для визначення пружних властивостей матеріалу на основі отриманих даних (експериментальних або розрахункових) необхідно трансформувати визначальне рівняння. Традиційний запис $\sigma = C\varepsilon$ не дозволяє явно виділити модулі пружності як вектор невідомих. Враховуючи симетрію матриці жорсткості, існує максимум 21 незалежна компонента. Сформуємо вектор шуканих параметрів C_{vec} розмірність якого 21×1:

$$C_{vec} = [C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, C_{22}, ..., C_{66}]^T \tag{9}$$

Це дозволяє переписати закон Гука у вигляді системи лінійних рівнянь відносно компонент пружності :

$$\sigma = A(\varepsilon)\;\; C_{vec}, \tag{10}$$

де $A(\varepsilon)$ – структурна матриця деформацій розмірністю 6×21, зображено на рис. 1.

Ця матриця конструюється таким чином, щоб врахувати внесок кожної компоненти деформації у відповідне напруження згідно з симетрією тензора.

$$A(\varepsilon)=\begin{pmatrix}
\varepsilon_{xx} & \varepsilon_{yy} & \varepsilon_{zz} & \varepsilon_{xy} & \varepsilon_{yz} & \varepsilon_{xz} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & \varepsilon_{xx} & 0 & 0 & 0 & 0 & \varepsilon_{yy} & \varepsilon_{zz} & \varepsilon_{xy} & \varepsilon_{yz} & \varepsilon_{xz} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & \varepsilon_{xx} & 0 & 0 & 0 & 0 & \varepsilon_{yy} & 0 & 0 & 0 & \varepsilon_{zz} & \varepsilon_{xy} & \varepsilon_{yz} & \varepsilon_{xz} & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \varepsilon_{xx} & 0 & 0 & 0 & 0 & \varepsilon_{yy} & 0 & 0 & 0 & \varepsilon_{zz} & 0 & 0 & \varepsilon_{xy} & \varepsilon_{yz} & \cdots & 0 \\
0 & 0 & 0 & 0 & \varepsilon_{xx} & 0 & 0 & 0 & 0 & \varepsilon_{yy} & 0 & 0 & 0 & \varepsilon_{zz} & 0 & 0 & \varepsilon_{xy} & \cdots & 0 \\
0 & 0 & 0 & 0 & 0 & \varepsilon_{xx} & 0 & 0 & 0 & 0 & \varepsilon_{yy} & 0 & 0 & 0 & \varepsilon_{zz} & 0 & 0 & \cdots & \varepsilon_{xz}
\end{pmatrix}$$

Рис. 1. Структурна матриця деформацій розмірністю 6×21

Таке представлення є критично важливим, оскільки зводить задачу ідентифікації матеріалу до системи лінійних алгебраїчних рівнянь виду $Ax=b$. Оскільки одне деформаційне навантаження формує лише шість рівнянь, тоді як кількість невідомих компонент тензора жорсткості становить 21, така система є невизначеною й не має єдиного розв'язку. Тому необхідно розглядати множину незалежних навантажень, що утворюють розширену систему рівнянь, для якої компоненти тензора пружності можуть бути відновлені шляхом застосування методу найменших квадратів.

3.3. Моделювання мікроструктури

Для моделювання мікроструктури полікристалічного магнію застосовано метод клітинних автоматів (КА). Простір розбито на тривимірну решітку L, що складається з клітин із дискретними станами S_i, які відповідають окремим зернам або фазам матеріалу:

$$S_i = \in\{1, 2, 3, ..., n_g\}, \qquad (11)$$

де $S_i = 0$ – рідка або некристалізована фаза, а $S_i > 0$ – номер зерна.

Початковий розподіл станів визначався випадково з урахуванням концентрації зародків Ψ, що задає частку клітин, які ініціюють кристалізацію:

$$N_0 = \Psi N_{total}, \qquad (12)$$

де N_{total} – кількість клітин у розрахунковій області.

Взаємодія між клітинами здійснюється за правилом околу фон Неймана (N_i), який охоплює шість сусідів у трьох напрямках ($\pm x$, $\pm y$, $\pm z$). На кожному часовому кроці виконується оновлення станів:

$$S_i^{(t+1)} = f(S_i^{(t)}, \{S_j^{(t)} | j \in N_j\}), \qquad (13)$$

де f – функція переходу, що визначає зміну стану клітини на основі поточного стану та станів сусідів.

На рис. 2 наведено тривимірну модель розподілу зерен у полікристалічному магнії, побудовану методом клітинних автоматів. Кожен воксель відповідає елементарному об'єму (клітині). Різні кольори позначають різні зерна. Для генерації використано MatMiz3D [21].

Модель побудована в кубічному об'ємі з періодичними граничними умовами. Початковий стан формується шляхом випадкової ініціалізації зерен у репрезентативному об'ємі розміром 25 вокселів із початковою концентрацією зародків 20%. Ітераційне зростання клітин забезпечує поступове заповнення простору та формування статистично рівноважної зернової структури.

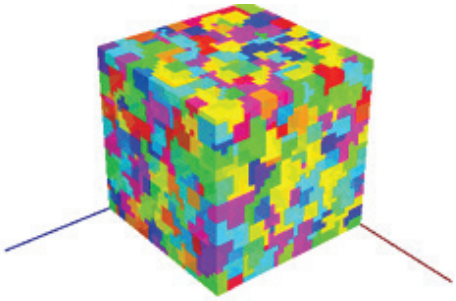

Рис. 2. Просторова модель розподілу зерен у кристалічному магнії, побудованому методом КА

4. Граничні умови

Для визначення ефективних пружних характеристик полікристалічного магнію проведено серію чисельних експериментів на основі мікроструктур, згенерованих методом клітинних автоматів.

З урахуванням стохастичної природи мікроструктури сформовано 100 незалежних моделей мікроструктури (M_1, M_1, ..., M_{100}), що відрізнялися випадковим розподілом центрів кристалізації та топологією зернових меж.

Для кожної згенерованої мікроструктури визначався НДС для набору навантажень, після чого для всього представницького об'єму обчислювали середні значення компонент тензора деформацій $\langle\varepsilon\rangle$ та тензора напружень $\langle\sigma\rangle$.

Для прикладання навантаження до репрезентативного об'єму RVE задаються значення тензора малих деформацій $\varepsilon_{ij} = const$.

Однак СЕ-комплекси не дозволяють напряму задавати тензор деформацій на границі репрезентативного об'єму. Тому необхідно перейти від деформацій до відповідного поля переміщень, яке забезпечує рівномірний стан деформацій у всьому RVE.

Зв'язок між переміщенням та деформаціями задається рівнянням:

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{du_i}{dx_j} + \frac{du_j}{dx_i}\right), \qquad (14)$$

Для тензора деформацій інтегрування цих рівнянь приводить до аналітичного поля переміщень:

$$\begin{aligned} u(x, y, z) &= \varepsilon_{xx}x + \varepsilon_{xy}y + \varepsilon_{xz}z \\ v(x, y, z) &= \varepsilon_{yy}y + \varepsilon_{xy}x + \varepsilon_{yz}z, \\ w(x, y, z) &= \varepsilon_{yy}y + \varepsilon_{xy}x + \varepsilon_{yz}z \end{aligned} \qquad (15)$$

Це поле використовується як гранична умова, що забезпечує однорідний стан деформацій у представницькому об'ємі та дозволяє обчислити середні напруження для подальшого визначення ефективного тензора жорсткості

Для забезпечення повного охоплення всіх можливих комбінацій станів напружено деформованого стану використовувалася дискретна сітка з 64 унікальних сценаріїв навантаження (2^6 комбінацій):

$$\varepsilon^{(k)} = [\pm\varepsilon_{xx}, \pm\varepsilon_{yy}, \pm\varepsilon_{zz}, \pm\varepsilon_{xy}, \pm\varepsilon_{xz}, \pm\varepsilon_{yz}], \qquad (16)$$

де $k = 1, 2, ..., 64$.

Для кожної з 100 мікроструктур проведено 64 розрахунки, що сформувало датасет із 6400 чисельних експериментів. Кожен елемент цього датасету містить пари $(\varepsilon^{(k)}, \sigma^{(k)})$, на основі яких здійснюється відновлення ефективного тензора жорсткості C_{eff}.

Кожен компонент C_{ij}^{eff} оцінювався через регресійне співвідношення між відповідними компонентами середніх напружень і деформацій для 64 різних сценаріїв. Надалі для кожної з 100 стохастичних реалізацій $C_{eff}^{(m)}$ обчислювалися середні значення, дисперсії та довірчі інтервали.

$$C_{ij} = \frac{1}{M}\sum_{m=1}^{M} C_{ij}^{(m)}$$
$$\sigma_{C_{ij}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(C_{ij}^{(m)} - C_{ij}\right)^2}, \qquad (17)$$

де M – це кількість реалізацій.

Для статистичної оцінки отриманих результатів проведено аналіз ймовірнісних розподілів пружних констант, зокрема:

– гістограми розподілу $C_{11}, C_{12}, C_{13}, C_{33}, C_{44}$

– оцінку коефіцієнта варіації

– побудову довірчих інтервалів для кожної компоненти.

Таке представлення дало змогу кількісно охарактеризувати вплив стохастичних мікроструктурних факторів на пружні властивості полікристалічного магнію.

Після проведення чисельних експериментів методом скінченних елементів отримано просторовий розподіл еквівалентних напружень за критерієм Мізеса, що формується в межах репрезентативного об'єму досліджуваного зразка магнію.

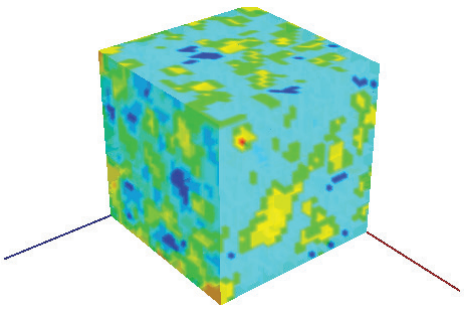

Рис. 3. Розподіл еквівалентних напружень за критерієм Мізеса в репрезентативному об'ємі моделі магнію

Отримана кольорова карта відображає характер перерозподілу напружень у матеріалі під дією прикладеного деформаційного навантаження та дозволяє оцінити локальні відхилення.

На рис. 3 наведено приклад тривимірного поля еквівалентних напружень Мізеса, отриманого для однієї з реалізацій стохастично згенерованої мікроструктури.

5. Аналіз результатів

Найбільші напруження зосереджуються поблизу меж зерен та в місцях зі зміненою орієнтацією кристалів.

У ході аналізу 100 стохастично згенерованих мікроструктур, побудованих методом клітинних автоматів, для кожної реалізації було визначено відповідну матрицю жорсткості. Отримані матриці $C_{eff}^{(k)}$ були усереднені для формування матриці пружних характеристик матеріалу, після чого проведено їх нормування та статистичний аналіз компонент.

На основі усереднених тензорів жорсткості C_{eff} визначено шість незалежних діагональних компонент, що відповідають основним пружним властивостям матеріалу – трьом модулям Юнга ($\mathbf{E_{xx}}$, $\mathbf{E_{yy}}$, $\mathbf{E_{zz}}$) та трьом зсувним модулям ($\mathbf{G_{xy}}$, $\mathbf{G_{yz}}$, $\mathbf{G_{xz}}$).

Розподіли цих величин мають близьку до нормальної форму рис. 4.

У табл. 2 представлено кількісні статистичні показники, отримані на основі аналізу ефективних пружних властивостей магнію.

Таблиця 2

Статистичні параметри основних пружних характеристик магнію

Показник	Мат. очікування	Середньо-квадратичне відхилення	Коефіцієнт варіації
	$\langle X \rangle$, ГПа	$\sqrt{var(X_{ij})}$, ГПа	$\frac{\sqrt{var(X_{ij})}}{\bar{X}}$,
E_{xx}	47.200	0.171	0.0036
E_{yy}	47.211	0.213	0.0045
E_{zz}	46.985	0136	0.0029
G_{xy}	20.656	0.116	0.0056
G_{yz}	18.536	0.087	0.0047
G_{xz}	18.550	0.090	0.0049

Середні значення відповідних модулів практично збігаються, що свідчить про ізотропність поведінки матеріалу. Середньоквадратичні відхилення для модулів Юнга не перевищують 0.21 ГПа, а для модулів зсуву – 0.12 ГПа. Отримані результати демонструють, що середні значення модулів Юнга лежать у вузькому інтервалі 46.9–47.2 ГПа, тоді як зсувні модулі – у межах 18.5–20.7 ГПа. Коефіцієнти варіації не перевищують 0.006.

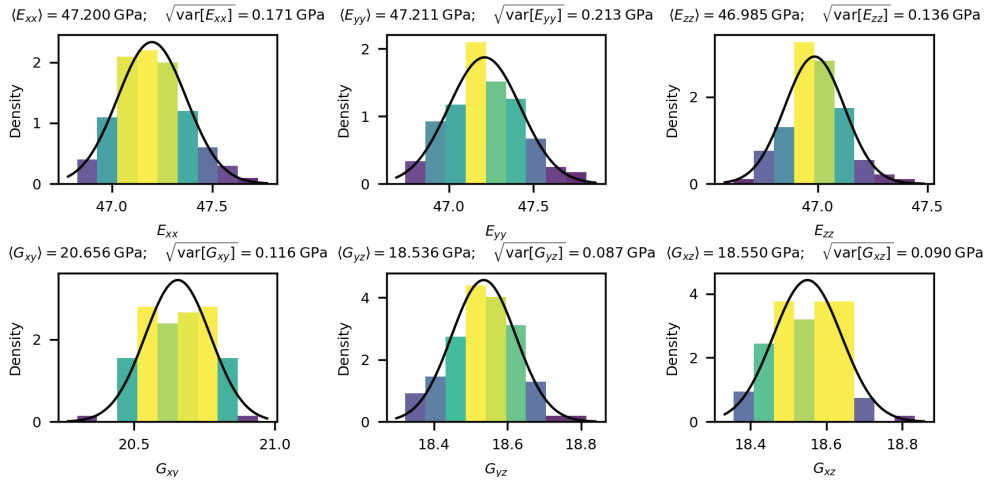

Рис. 4. Розподіл імовірності основних пружних характеристик модулів Юнга E_{xx}, E_{yy}, E_{zz} та зсувних модулів G_{xy}, G_{yz}, G_{xz}

Інверсія матриці жорсткості C_{eff} дозволяє отримати матрицю податливості $S_{eff} = C_{eff}^{-1}$, елементи якої описують реакцію матеріалу у вигляді деформацій при дії напружень.

$$S = \begin{pmatrix} \frac{1}{E_{xx}} & -\frac{\nu_{xy}}{E_{yy}} & -\frac{\nu_{xz}}{E_{zz}} & \frac{\nu_{yz,xx}}{G_{yz}} & \frac{\nu_{xz,xx}}{G_{xz}} & \frac{\nu_{xy,xx}}{G_{xy}} \\ -\frac{\nu_{yx}}{E_{xx}} & \frac{1}{E_{yy}} & -\frac{\nu_{yz}}{E_{zz}} & \frac{\nu_{yz,yy}}{G_{yz}} & \frac{\nu_{xz,yy}}{G_{xz}} & \frac{\nu_{xy,yy}}{G_{xy}} \\ -\frac{\nu_{zx}}{E_{xx}} & -\frac{\nu_{zy}}{E_{yy}} & \frac{1}{E_{zz}} & \frac{\nu_{yz,zz}}{G_{yz}} & \frac{\nu_{xz,zz}}{G_{xz}} & \frac{\nu_{xy,zz}}{G_{xy}} \\ \frac{\nu_{xx,yz}}{E_{xx}} & \frac{\nu_{yy,yz}}{E_{yy}} & \frac{\nu_{zz,yz}}{E_{zz}} & \frac{1}{G_{yz}} & \frac{\nu_{yz,zx}}{G_{xz}} & \frac{\nu_{yz,xy}}{G_{xy}} \\ \frac{\nu_{xx,xz}}{E_{xx}} & \frac{\nu_{yy,xz}}{E_{yy}} & \frac{\nu_{zz,xz}}{E_{zz}} & \frac{\nu_{zx,yz}}{G_{yz}} & \frac{1}{G_{xz}} & \frac{\nu_{xz,xy}}{G_{xy}} \\ \frac{\nu_{xx,xy}}{E_{xx}} & \frac{\nu_{yy,xy}}{E_{yy}} & \frac{\nu_{zz,xy}}{E_{zz}} & \frac{\nu_{xy,yz}}{G_{yz}} & \frac{\nu_{xy,xz}}{G_{xz}} & \frac{1}{G_{xy}} \end{pmatrix} \quad (18)$$

На її основі визначено коефіцієнти Пуассона ν_{ij}, що характеризують взаємозв'язок між поздовжніми та поперечними деформаціями в різних напрямках.

Саме застосування цих формул до кожної стохастичної мікроструктури дозволило отримати вибірки значень ν_{xy}, ν_{xz}, ν_{yz}, а також низку додаткових зсувних компонент, на основі яких побудовано гістограми розподілу рис. 5 та визначено статистичні параметри, подані в табл. 3.

Гістограми розподілу цих коефіцієнтів демонструють вузькі й симетричні профілі, де чорна лінія – апроксимація нормальним розподілом; над гістограмами вказані середні значення $\langle \nu_{ij} \rangle$. Над кожним графіком подано оцінки математичного очікування $\langle \nu_{ij} \rangle$ та середньоквадратичного відхилення $\sqrt{\mathrm{var}(\nu_{ij})}$, що дозволяє кількісно оцінити ступінь варіативності відповідних параметрів.

Отримані значення основних коефіцієнтів Пуассона узгоджуються з очікуваною деформаційною поведінкою гексагональної гратки магнію: де коефіцієнти Пуассона мають значення $\nu_{xy} \approx 0.14$, а навантаження уздовж осі симетрії має $\nu_{xz} \approx \nu_{yz} \approx 0.22$. Розподіл цих параметрів містить дуже малі середньоквадратичні відхилення $\sqrt{\mathrm{var}(\nu_{ij})} < 0.006$, що підтверджує високу стабільність чисельної моделі. Відносні середньоквадратичні відхилення в межах 2,2% для всіх трьох основних компонент свідчать про чітку статистичну відтворюваність і сталість поперечних деформацій у відповідних напрямках.

Таблиця 3

Статистичні параметри коефіцієнтів Пуассона

Показник	Мат. очікування	Середньоквадратичне відхилення	Коефіцієнт варіації
	$\langle \nu_{ij} \rangle$	$\sqrt{\mathrm{var}(\nu_{ij})}$	$\frac{\sqrt{\mathrm{var}(\nu_{ij})}}{\langle \nu_{ij} \rangle}$
ν_{xy}	0.139	0.003	0.0216
ν_{xz}	0.220	0.005	0.0227
ν_{yz}	0.221	0.005	0.0226
інші ν_{ij}	≈ 0.000	0.003–0.009	0

Окрему увагу слід звернути на змішані компоненти ν_{ij}, для яких середні значення перебувають у межах $\langle \nu_{ij} \rangle \approx 0$, водночас їх середнє квадратичне відхилення становить 0.003–0.009. Це означає, що змішані компоненти не мають суттєвого вкладу у матрицю пружних констант Таким чином, їх можна вважати статистичним шумом, який не впливає на загальну механічну відповідь.

На основі отриманих значень ν_{ij} побудовано нормалізовану матрицю коефіцієнтів Пуассона рис. 6. Вона узагальнює структуру деформаційних зв'язків у полікристалічному магнії: три основні компоненти формують виразні міжосьові залежності, тоді як змішані елементи фактично дорівнюють нулю, що повністю відповідає теоретичній симетрії HCP-решітки та підтверджує коректність чисельної моделі.

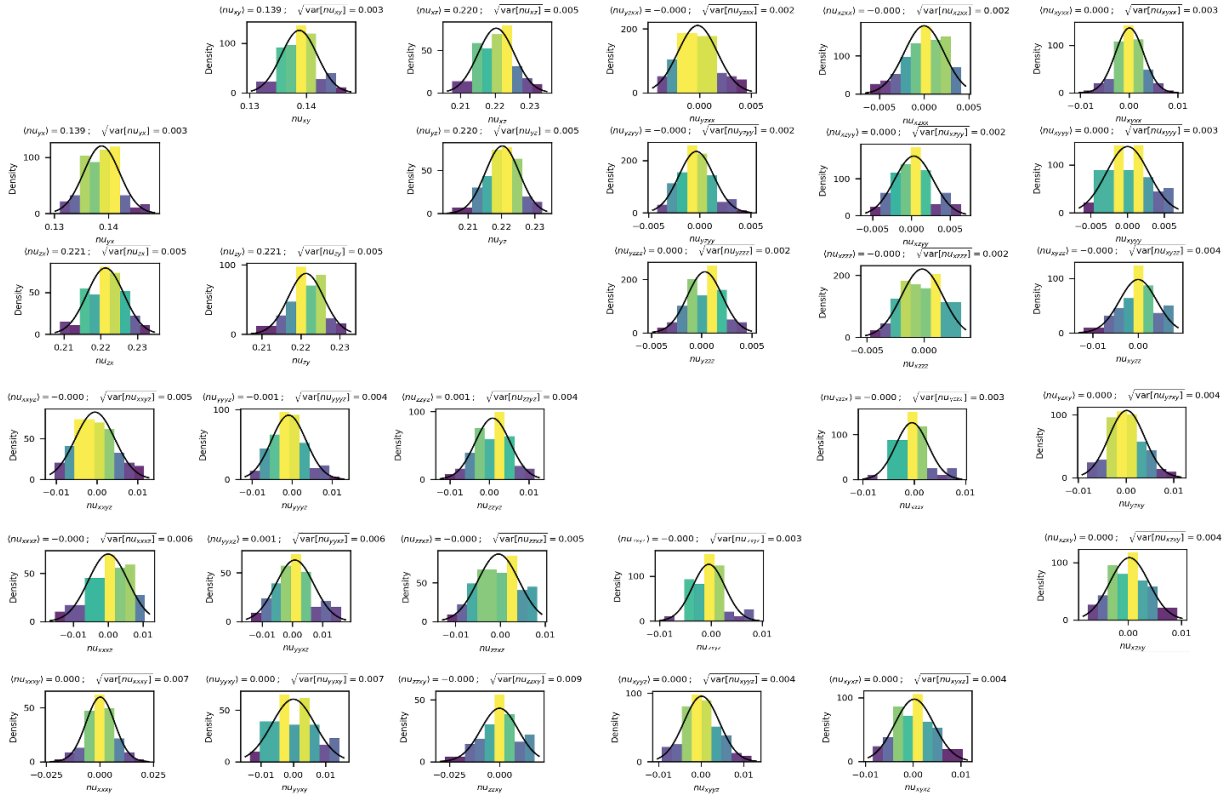

Рис. 5. Імовірнісні розподіли коефіцієнтів Пуассона ν_{xy}, ν_{xz}, ν_{yz} та допоміжних компонент податливості для стохастичних мікроструктур магнію

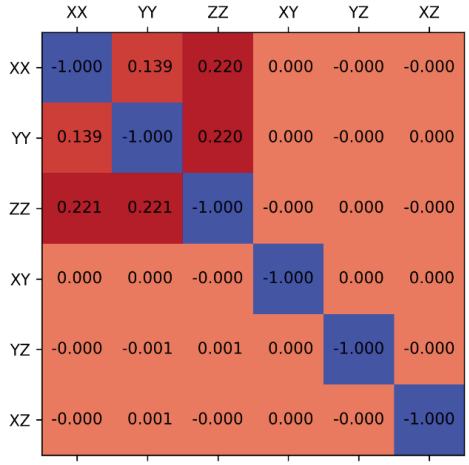

Рис. 6. Нормалізована матриця коефіцієнтів Пуассона

Висновки

У роботі досліджено стохастичні та статистичні характеристики пружних властивостей чистого магнію на основі моделювання мікроструктури методом клітинних автоматів та подальшої гомогенізації механічної відповіді.

Побудована чисельна модель дозволила відтворити етапи формування зернової структури матеріалу, обчислити ефективні компоненти тензора жорсткості C_{eff} і податливості S_{eff}, а також визначити макроскопічні параметри – модулі Юнга, зсувні модулі та коефіцієнти Пуассона.

Аналіз нормалізованої матриці жорсткості показав, що для гексагонально щільноупакованої (HCP) структури магнію головні діагональні елементи характеризуються стабільними значеннями, які визначають основну пружну реакцію матеріалу, а позадіагональні елементи в межах 0.13–0.22 відображають типові міжосьові зв'язки.

Відсутність значущих позанормальних компонент підтверджує осьову симетрію HCP-гратки та правильність побудованої чисельної моделі.

Проведений статистичний аналіз ефективних модулів показав, що середні значення модулів Юнга становлять $\langle E_{xx} \rangle = 47.2$, $\langle E_{yy} \rangle = 47.2$, $\langle E_{zz} \rangle = 47.0$ ГПа, а зсувних модулів – $\langle G_{xy} \rangle = 20.7$, $\langle G_{yz} \rangle = 18.5$, $\langle G_{xz} \rangle = 18.6$ ГПа. Отримані результати добре узгоджуються з експериментальними і довідковими даними для чистого магнію: за літературними джерелами [23], значення становлять E = 44–48 ГПа, G = 17–19 ГПа, ν = 0.17–0.25. Таким чином, похибка моделі не перевищує 5 % від граничних показників, що підтверджує її достовірність.

Таблиця 4

Порівняння результатів власних розрахунків з даними літературних джерел

Показник	Результати експерименту (МСЕ), [ГПа/-]	Данні з літератури, [ГПа/-]
E_{xx}	47.2	44-48
G_{xy}	20.7	17-19
ν	0.22	0.17-0.25

Розподіли компонент тензора пружності та коефіцієнтів Пуассона мають нормальний характер з малими стандартними відхиленнями, які не перевищують 0.01, що вказує на статистичну збіжність і стабільність чисельного методу. Середні значення коефіцієнтів Пуассона становлять: $\langle \nu_{xy} \rangle = 0.139$, $\langle \nu_{xz} \rangle = 0.220$, $\langle \nu_{xy} \rangle = 0.221$, що практично збігається з експериментально відомими величинами для HCP-магнію ($\nu = 0.17–0.25$).

Загалом проведене дослідження показало, що метод клітинних автоматів у поєднанні з підходом стохастичної гомогенізації може бути ефективним інструментом для оцінки пружних характеристик магнію та матеріалів із подібною HCP-структурою.

Отримані результати підтверджують можливість відтворення реальних характеристик без залучення прямих експериментів і можуть бути використані для розроблення інтелектуальних моделей прогнозування пружної поведінки та пошкоджуваності магнієвих сплавів.

Підтримка

Ця робота виконана за підтримки МОН України в рамках реалізації науково-дослідної роботи «Алгоритми, моделі та засоби штучного інтелекту для дворівневого моделювання поведінки складних матеріалів для техніки подвійного призначення» (№ ДР 0124U000450).

Список літератури:

[1] Liu B., Yang J., Zhang X., Yang Q., Zhang J., Li X. Development and application of magnesium alloy parts for automotive OEMs: A review / B. Liu, J. Yang, X. Zhang, Q. Yang, J. Zhang, X. Li // Journal of Magnesium and Alloys. – 2023. – Vol. 11, No. 1. – P. 15–47. – DOI: 10.1016/j.jma.2022.12.015.

[2] Luo A. A., Shi R., Miao J., Avey T. Magnesium sheet alloy development for room temperature forming: A review / A. A. Luo, R. Shi, J. Miao, T. Avey // JOM. – 2021. – Vol. 73, No. 5. – P. 1403–1418. – DOI: 10.1007/s11837-021-04616-y.

[3] Jayasathyakawin S., Ravichandran M., Baskar N., Chairman C. A., Balasundaram R. Mechanical properties and applications of magnesium alloy: Review / S. Jayasathyakawin, M. Ravichandran, N. Baskar, C. A. Chairman, R. Balasundaram // Materials Today: Proceedings. – 2020. – Vol. 27. – P. 909–913. – DOI: 10.1016/j.matpr.2020.01.255.

[4] Tong X. et al. Enhanced mechanical properties, corrosion resistance, cytocompatibility, osteogenesis, and antibacterial performance of biodegradable Mg–2Zn–0.5Ca–0.5Sr/Zr alloys for bone-implant application / X. Tong et al. // Advanced Healthcare Materials. – 2024. – Vol. 13, No. 12. – P. 2303975. – DOI: 10.1002/adhm.202303975.

[5] Pei Z. Connecting the mechanical properties of magnesium and high-entropy alloys / Z. Pei // Journal of Alloys and Compounds. – 2023. – Vol. 966. – P. 171462. – DOI: 10.1016/j.jallcom.2023.171462.

[6] Alaneme K. K., Okotete E. A. Enhancing plastic deformability of Mg and its alloys—A review of traditional and nascent developments / K. K. Alaneme, E. A. Okotete // Journal of Magnesium and Alloys. – 2017. – Vol. 5, No. 4. – P. 460–475. – DOI: 10.1016/j.jma.2017.11.001.

[7] Jonsson J. C., Chapman D. J., Farbaniec L., Escauriza E. M., Smith L. C., Eakins D. E. Role of crystal orientation in the dynamic strength of magnesium alloy AZ31B / J. C. Jonsson et al. // JOM. – 2024. – Vol. 76, No. 3. – P. 1628–1638. – DOI: 10.1007/s11837-023-05967-4.

[8] Mao B., Zhang X., Menezes P. L., Liao Y. Anisotropic microstructure evolution of an AZ31B magnesium alloy subjected to dry sliding and its effects on friction and wear performance / B. Mao, X. Zhang, P. L. Menezes, Y. Liao // Materialia. – 2019. – Vol. 8. – P. 100444. – DOI: 10.1016/j.mtla.2019.100444.

[9] Sisodia S., Jananandhan S., Pakki V. K., Konkati C., Chauhan A. Towards reducing tension–compression yield and cyclic asymmetry in pure magnesium and magnesium–aluminum alloy with cerium addition / S. Sisodia et al. // Materials Science and Engineering A. – 2023. – Vol. 886. – P. 145672. – DOI:10.1016/j.msea.2023.145672.

[10] Valipoorsalimi P., Sari Y. A., Pekguleryuz M. Mechanical Property Design of Bio-compatible Mg alloys using Machine-Learning Algorithms / P. Valipoorsalimi, Y. A. Sari, M. Pekguleryuz // arXiv preprint. – 2023. – URL: https://arxiv.org/pdf/2305.12060 (accessed 26.05.2025).

[11] Pang H. et al. Plasticity Improvement of Mg Alloys with Alloying Atoms (Gd, Y) / H. Pang et al. // physica status solidi (b). – 2022. – Vol. 259, No. 11. – P. 2200209. – DOI: 10.1002/pssb.202200209.

[12] Tan Y., Li W., Li A., Shi X. Microstructure and properties evolution of Mg–2Y–0.6Nd–0.6Zr alloy rolled at room and liquid nitrogen temperature / Y. Tan, W. Li, A. Li, X. Shi // Scientific Reports. – 2021. – Vol. 11. – P. 1–10. – DOI: 10.1038/s41598-021-99706-x.

[13] Kalateh M. A., Talebi N., Nekoei S., Novini M. M., Khodabakhshi F., Nili-Ahmadabadi M. Thermo-mechanical processing of pure magnesium: Hot extrusion, hot rolling and cold drawing / M. A. Kalateh et al. // arXiv preprint. – 2025. – URL: (accessed 26.05.2025).

[14] Wolf A. et al. Modeling metal forming of a magnesium alloy using an adapted material model / A. Wolf et al. // Engineering Reports. – 2022. – Vol. 4, No. 7–8. – P. e12540. – DOI: 10.1002/eng2.12540.

[15] Tian B. H., Wu M. W., Zhang A., Guo Z. P., Xiong S. M. Phase-field modeling of dendritic growth of magnesium alloys with a parallel-adaptive mesh refinement algorithm / B. H. Tian et al. // China Foundry. – 2021. – Vol. 18, No. 6. – P. 541–549. – DOI:10.1007/s41230-021-1116-5.

[16] Poul M., Huber L., Bitzek E., Neugebauer J. Systematic Atomic Structure Datasets for Machine Learning Potentials: Application to Defects in Magnesium / M. Poul et al. // Physical Review B. – 2022. – Vol. 107, No. 10. – DOI: 10.1103/PhysRevB.107.104103.

[17] Yu Q. Size-related Mechanical Properties of Pure Magnesium / Q. Yu. – 2012.

[18] Bayat Tork N., Saghafian H., Razavi S. H., Al-Fadhalah K. J., Ebrahimi R., Mahmudi R. Microstructure and texture characterization of Mg–Al and Mg–Gd binary alloys processed by simple shear extrusion / N. Bayat Tork et al. // Journal of Materials Research and Technology. – 2019. – Vol. 8, No. 1. – P. 1288–1299. – DOI: 10.1016/j.jmrt.2018.06.023.

[19] Singh A., Saal J. E. Dynamic properties of magnesium alloys / A. Singh, J. E. Saal // JOM. – 2014. – Vol. 66, No. 2. – P. 275–276. – DOI: 10.1007/s11837-013-0844-4.

[20] Jiang Y. et al. Review on forming process of magnesium alloy characteristic forgings / Y. Jiang et al. // Journal of Alloys and Compounds. – 2024. – Vol. 970. – P. 172666. – DOI: 10.1016/j.jallcom.2023.172666.

[21] НТУ «ХПІ». MatViz3D Програмний компонент для візуалізації матеріалів // URL: https://matviz3d.khpi.edu.ua

Надійшла до редколегії 19.09.2025

УДК 004.42 DOi 10.30837/ bi.2025.2(103).08

І. В. Кириченко[1], Г. Ю. Терещенко[2], К. О. Гоцуляк[3], В. О. Каленик[4]
[1]ХНУРЕ, м. Харків, Україна, iryna.kyrychenko@nure.ua, ORCID iD: 0000-0002-7686-6439
[2]ХНУРЕ, м. Харків, Україна, hlib.tereshchenko@nure.ua, ORCID iD: 0000-0001-8731-2135
[3]ХНУРЕ, м. Харків, Україна, kateryna.horishnia@nure.ua, ORCID iD: 0009-0001-9032-4249
[4]ХНУРЕ, м. Харків, Україна, vira.kalenyk@nure.ua

ЗАСТОСУВАННЯ БЛОКЧЕЙН-ТЕХНОЛОГІЙ ДЛЯ ЗАБЕЗПЕЧЕННЯ ПРОЗОРОСТІ ВИБОРЧИХ ПРОЦЕСІВ В ОРГАНІЗАЦІЯХ

У роботі розглянуто підхід до створення децентралізованої системи електронного голосування на основі блокчейна Ethereum. Модель поєднує смарт-контракти, механізм commit–reveal та токени прав голосу у форматі SBT, що забезпечує прозорість, захист від маніпуляцій і неможливість передачі голосу. Офчейн-сервер використовується лише для журналювання подій та аналітики, не впливаючи на підрахунок. Експериментальні дослідження у мережі Sepolia підтвердили коректність роботи системи та її придатність для застосування в організаційних виборах. Результати демонструють перспективність блокчейн-підходу для підвищення довіри та безпеки цифрових виборчих процесів.

БЛОКЧЕЙН, ЕЛЕКТРОННЕ ГОЛОСУВАННЯ, СМАРТ-КОНТРАКТИ, COMMIT–REVEAL, ПРОЗОРІСТЬ ВИБОРІВ, ДЕЦЕНТРАЛІЗОВАНІ СИСТЕМИ, SOULBOUND-ТОКЕНИ, ETHEREUM

I. V. Kyrychenko, G. Yu. Tereshchenko, K. O. Gotsulyak, V. O. Kalenyk. Blockchain-Based Approach to Ensuring Transparency in Organizational Voting Processes. This paper presents a decentralized electronic voting system built on the Ethereum blockchain. The proposed model integrates smart contracts, a commit–reveal mechanism, and soulbound voting rights tokens to ensure transparency, resistance to manipulation, and prevention of vote transfer. An off-chain server is used only for logging and analytics, without affecting vote counting. Experimental evaluation in the Sepolia test network confirmed the system's correctness and suitability for organizational voting scenarios. The results highlight the potential of blockchain-based solutions to enhance trust, security, and transparency in modern digital voting processes.

BLOCKCHAIN, ELECTRONIC VOTING, SMART CONTRACTS, COMMIT–REVEAL, ELECTION TRANSPARENCY, DECENTRALIZED SYSTEMS, SOULBOUND TOKENS, ETHEREUM

Вступ

Електронні виборчі системи дедалі частіше впроваджуються у великих компаніях, університетах, громадських організаціях та комітетах як інструмент оперативного й зручного прийняття колективних рішень. Перехід до цифрових форм голосування обумовлений потребою у швидкості, доступності та автоматизації виборчих процесів, а також прагненням зменшити витрати на організацію традиційних паперових виборів. Проте незважаючи на очевидні переваги, більшість сучасних веб-систем голосування залишаються вразливими до низки загроз, що суттєво знижують рівень довіри до їхніх результатів [1].

Проблеми централізованих платформ полягають у можливості маніпуляцій з боку адміністратора, зміні або видаленні бюлетенів, некоректному підрахунку голосів, а також у ризику компрометації сервера, на якому зберігаються всі дані. Централізована модель створює єдину точку відмови, а доступ адміністратора до внутрішніх даних відкриває можливості для втручання, фальсифікацій чи прихованих модифікацій результатів [2, 3]. Навіть за умови використання сучасних криптографічних методів, таких як шифрування чи електронні підписи, довіра учасників до системи значною мірою залишається пов'язаною з довірою до того, хто контролює серверну інфраструктуру [4].

Блокчейн-технології дають змогу усунути або значно зменшити ці ризики, забезпечуючи незмінність даних, публічний журнал подій, можливість незалежного аудиту та прозору перевірку результатів голосування. Децентралізовані мережі усувають необхідність у довіреному посереднику, а всі операції зберігаються у вигляді послідовних, захищених від змін блоків. Особливо важливою є можливість застосування смарт-контрактів, які реалізують підрахунок голосів детермінованим способом, без участі довіреного сервера, що повністю усуває людський фактор при визначенні підсумків виборів [5].

Окремої уваги заслуговує той факт, що блокчейн дозволяє перевірити правильність результатів будь-якому сторонньому спостерігачеві. Користувачі можуть переглядати всі події голосування у публічному реєстрі, а валідність даних гарантується криптографією та принципами консенсусу [6]. Такий підхід підвищує довіру до виборчого процесу та робить можливими чесні, прозорі вибори в організаціях будь-якого масштабу.

Метою роботи є створення прототипу децентралізованої системи електронного голосування, яка гарантує прозорість, перевірюваність і безпеку виборчого процесу, а також формалізація вимог, моделі та архітектури майбутнього рішення.

У межах дослідження передбачається аналіз сучасних підходів до електронного голосування, визначення

недоліків централізованих систем, розробка архітектурної моделі з використанням блокчейна, створення смарт-контрактів для реалізації виборчих процедур та проведення експериментальних досліджень роботи прототипу.

1. Огляд існуючих рішень

Дослідження у сфері електронного голосування протягом останніх років привели до появи різноманітних систем, однак більшість із них залишаються обмеженими у питаннях прозорості та захисту від маніпуляцій [5]. Традиційні веб-платформи, що застосовуються у компаніях, університетах та громадських організаціях, базуються на централізованій архітектурі, де всі дані зберігаються та обробляються на одному сервері. У такій моделі адміністратор має широкі технічні можливості впливати на перебіг голосування, а відсутність зовнішнього аудиту не дозволяє учасникам перевірити достовірність підрахунку голосів. Незважаючи на використання криптографічних механізмів захисту, централізація залишає системи вразливими до підміни, видалення або маніпулювання результатами.

Спробою вирішити ці проблеми стали криптографічні платформи без використання блокчейна. Найвідомішим прикладом є система Helios, яка реалізує гомоморфне шифрування та дозволяє відкрито перевіряти правильність підрахунку [7]. Однак навіть такі рішення не усувають залежність від сервера, на якому здійснюється прийом, зберігання та первинна обробка бюлетенів, що зберігає ризик внутрішньої компрометації. Комерційні мобільні платформи, такі як Voatz, пропонують удосконалені механізми автентифікації, проте їхні алгоритми не є відкритими, а використання приватних блокчейнів позбавляє виборчий процес необхідної прозорості [8].

Поява публічних блокчейнів, зокрема Ethereum, дала можливість створювати децентралізовані системи голосування, у яких підрахунок і зберігання даних не контролюються жодною окремою стороною. Найпоширенішим підходом у таких рішеннях є механізм commit–reveal, що дозволяє приховувати вибір до завершення голосування та забезпечує його перевірюваність після розкриття [9]. Завдяки незмінності блокчейна результати виборів стають доступними для незалежної перевірки, а смарт-контракти усувають людський фактор у підрахунку голосів. Проте навіть блокчейн-платформи залишають деякі виклики, пов'язані з вартістю транзакцій, необхідністю взаємодії через криптогаманці та відсутністю універсальних стандартів управління правами голосу.

Таким чином, попри значну кількість існуючих рішень, більшість систем не забезпечують одночасно прозорість, перевірюваність, відсутність довіреної сторони та зручність використання. Це підтверджує необхідність створення нової моделі електронного голосування, яка поєднуватиме переваги публічного блокчейна, відкритої архітектури та доступності для широкого кола користувачів.

2. Постановка задачі

У сучасних умовах організації різного масштабу потребують інструментів, які дозволяють проводити вибори швидко, безпечно та прозоро, незалежно від місця знаходження учасників. Однак традиційні підходи до електронного голосування не здатні повною мірою забезпечити довіру до результатів, оскільки спираються на централізовану інфраструктуру, де адміністратор відіграє ключову роль у зберіганні та обробці даних [10]. Це створює низку потенційних ризиків, серед яких підміна бюлетенів, приховані зміни результатів, видалення даних або втручання у логіку підрахунку. Щоб усунути ці загрози, необхідно розробити систему, яка б не покладалася на довірені сторони та водночас забезпечувала технічну можливість для незалежної перевірки всіх етапів виборчого процесу.

Поставлена в роботі задача полягає у створенні моделі та прототипу децентралізованої системи голосування, яка ґрунтується на використанні блокчейн-технологій як механізму забезпечення незмінності даних та прозорого аудиту. Для її реалізації необхідно визначити вимоги до системи, що охоплюють як функціональні властивості, так і загальні принципи взаємодії між учасниками процесу. Система має підтримувати реєстрацію виборчих кампаній, визначення часових меж голосування, авторизацію виборців, подання голосів у заздалегідь прихованому вигляді та їх подальше розкриття після завершення голосування. Важливим аспектом є побудова механізму підтвердження права на участь у голосуванні, що виключає можливість дублювання голосів або участі неавторизованих користувачів.

Для досягнення поставленої мети необхідно формалізувати моделі даних, які описують виборчі кампанії, учасників, бюлетені та результати голосування, а також визначити архітектуру системи, що поєднуватиме можливості блокчейна та допоміжних офчейн-компонентів. Особливе місце займає визначення протоколу взаємодії між виборцем і смарт-контрактом, який повинен гарантувати приватність вибору у період голосування та можливість його верифікації після завершення процедури. Найбільш придатним для цього є механізм commit–reveal, який дозволяє не розкривати вибір до завершення голосування та водночас унеможливлює його зміну заднім числом [11].

Крім того, необхідно сформувати вимоги до підрахунку голосів, який має виконуватися автоматично і детерміновано у смарт-контракті, що гарантує відсутність впливу людського фактора. Система повинна забезпечувати відкритість журналу подій, доступність результатів і можливість перевірки правильності під-

рахунку будь-яким зацікавленим учасником. Разом із тим важливо передбачити використання офчейн-компонентів для логування подій, аналітики та підвищення зручності користування, не порушуючи принцип децентралізованого підрахунку.

3. Модель та архітектура системи

Розроблення децентралізованої системи електронного голосування вимагає формування цілісної архітектурної моделі, яка поєднує можливості публічного блокчейна, офчейн-компонентів та клієнтських застосунків. Архітектура має забезпечувати прозорість виборчого процесу, незалежність підрахунку голосів від довірених сторін, захист від фальсифікацій, а також можливість зовнішнього аудиту. В основі системи лежить використання блокчейна Ethereum як платформи для зберігання критично важливих даних та виконання смарт-контрактів, що реалізують логіку голосування [9, 12]. Для структурування моделі виборчого процесу було використано UML-діаграми, які дозволяють формально описати ролі учасників, їхню взаємодію та внутрішню структуру компонентів системи [13, 14].

Центральним елементом архітектури є смарт-контракт ElectionManager, який відповідає за створення виборчих кампаній, визначення часових меж етапів голосування, приймання зашифрованих голосів, виконання процедури розкриття та підрахунок результатів. Контракт зберігає всі події у блокчейні, забезпечуючи їх незмінність та доступність для перевірки. Важливою особливістю є реалізація протоколу commit–reveal, який дозволяє приховати вибір виборця на етапі голосування та гарантує коректність розкриття після завершення процедури. Кожен голос подається у вигляді хешу, сформованого з урахуванням випадкової криптографічної солі, що виключає можливість підбору значення або передчасного розкриття голосу [15].

Ще одним компонентом ончейн-рівня є контракт VotingRightToken, реалізований за стандартом ERC-1155 у форматі soulbound-токенів [16]. Він забезпечує механізм підтвердження прав голосу, який неможливо передати або скопіювати. Така модель дозволяє гарантувати, що кожен виборець має лише один голос, а також унеможливлює участь у виборах неавторизованих користувачів. Управління правами голосу відбувається у смарт-контракті, а сам факт володіння токеном зберігається у блокчейні, що створює додатковий рівень прозорості.

Архітектура системи доповнюється офчейн-компонентами, які виконують допоміжні, але важливі функції. Серверна частина розроблена на Node.js і відповідає за журналювання подій, зберігання метаданих виборів, взаємодію з базою даних і виконання службових операцій, що не впливають на процес підрахунку голосів. Офчейн-рівень отримує події зі смарт-контракту через бібліотеку ethers.js, зберігає їх у базі SQL Server та надає можливість аналітики, аудиту та адміністративного моніторингу. Зберігання журнальних записів окремо від блокчейна не порушує принципів прозорості, оскільки критично важливі дані все одно залишаються у децентралізованому реєстрі, а офчейн-логіка використовується виключно для підвищення зручності користування та зменшення вартості операцій.

Клієнтський застосунок реалізований на Angular і забезпечує взаємодію користувачів із системою. Його архітектура передбачає роботу у двох режимах: адміністративному та виборчому. Адміністратор може створювати вибори, визначати часові межі етапів, керувати правами голосу та переглядати стан виборчої кампанії через офчейн-сервер. Виборець, у свою чергу, може підключити гаманець MetaMask, отримати токен права голосу, подати commit голосу, розкрити його після завершення голосування та переглянути підсумки [17, 18]. Усі операції, пов'язані з голосуванням, виконуються безпосередньо у блокчейні, що виключає втручання адміністратора у процес приймання голосів.

Для структурування моделі виборчої системи було створено три основні UML-діаграми, які формально описують ролі, компоненти та динаміку роботи системи.

Діаграма прецедентів (рис. 1) відображає зовнішню поведінку системи та визначає функції, доступні кожній ролі. Адміністратор створює вибори, керує етапами та правами голосу; виборець подає commit, виконує reveal та переглядає результати; аудитор перевіряє публічні події у блокчейні.

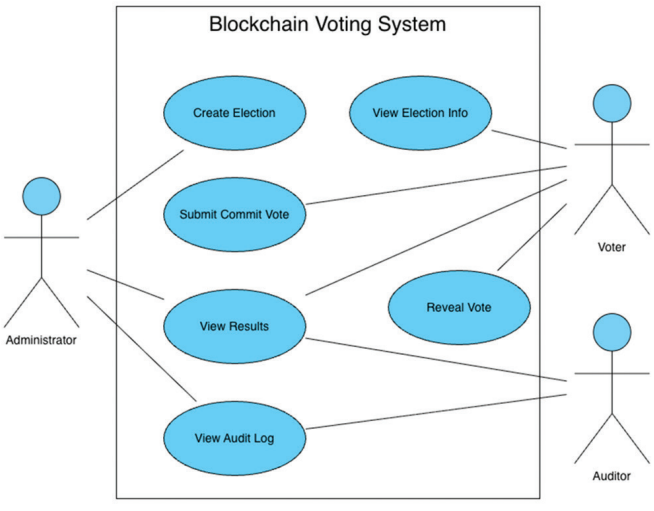

Рис. 1. Use Case Diagram

Діаграма класів (рис. 2) деталізує внутрішню архітектуру, включаючи контракти ElectionManager і VotingRightToken, моделі виборців, кандидатів та виборчих кампаній, а також офчейн-сервіси й клієнтські модулі. Вона відображає структурні зв'язки між блокчейном, сервером і фронтендом та показує, як зберігаються й обробляються ключові дані.

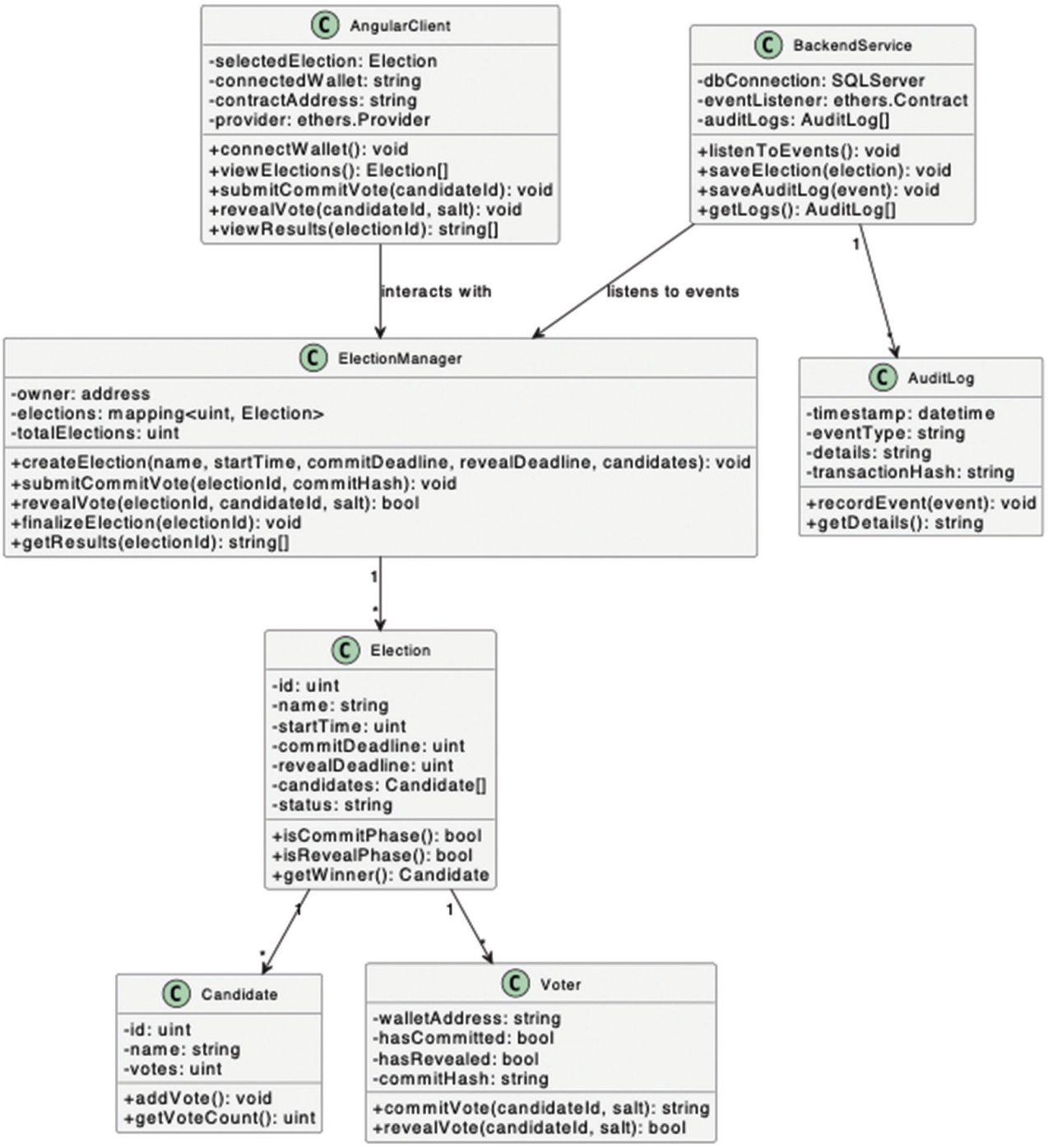

Рис. 2. Class Diagram

Діаграма послідовності (рис. 3) ілюструє покрокову взаємодію під час голосування: надсилання commit, підпис транзакцій через MetaMask, обробка подій смарт-контрактом, логування офчейн-сервісом та подальше розкриття голосів.

Узагальнена архітектура системи демонструє поєднання децентралізованої логіки підрахунку голосів із допоміжними механізмами зберігання метаданих та аналітики, що дозволяє досягти високого рівня безпеки, прозорості та практичної зручності використання. Такий підхід забезпечує надійну модель голосування, яка придатна для застосування у широкому спектрі організаційних сценаріїв.

4. Програмна реалізація і експериментальні дослідження

Реалізація прототипу системи електронного голосування ґрунтується на поєднанні блокчейн-технологій та сучасних веб-фреймворків, що дозволило створити повноцінний функціональний комплекс, придатний для тестування у реальних організаційних умовах. Основний обсяг логіки, відповідальної за безпеку, прозорість та підрахунок голосів, був реалізований у вигляді смарт-контрактів на платформі Ethereum. Для розробки смарт-контрактів використано мову Solidity та фреймворк Hardhat, який забезпечує можливості локального тестування, автоматизованої компіляції,

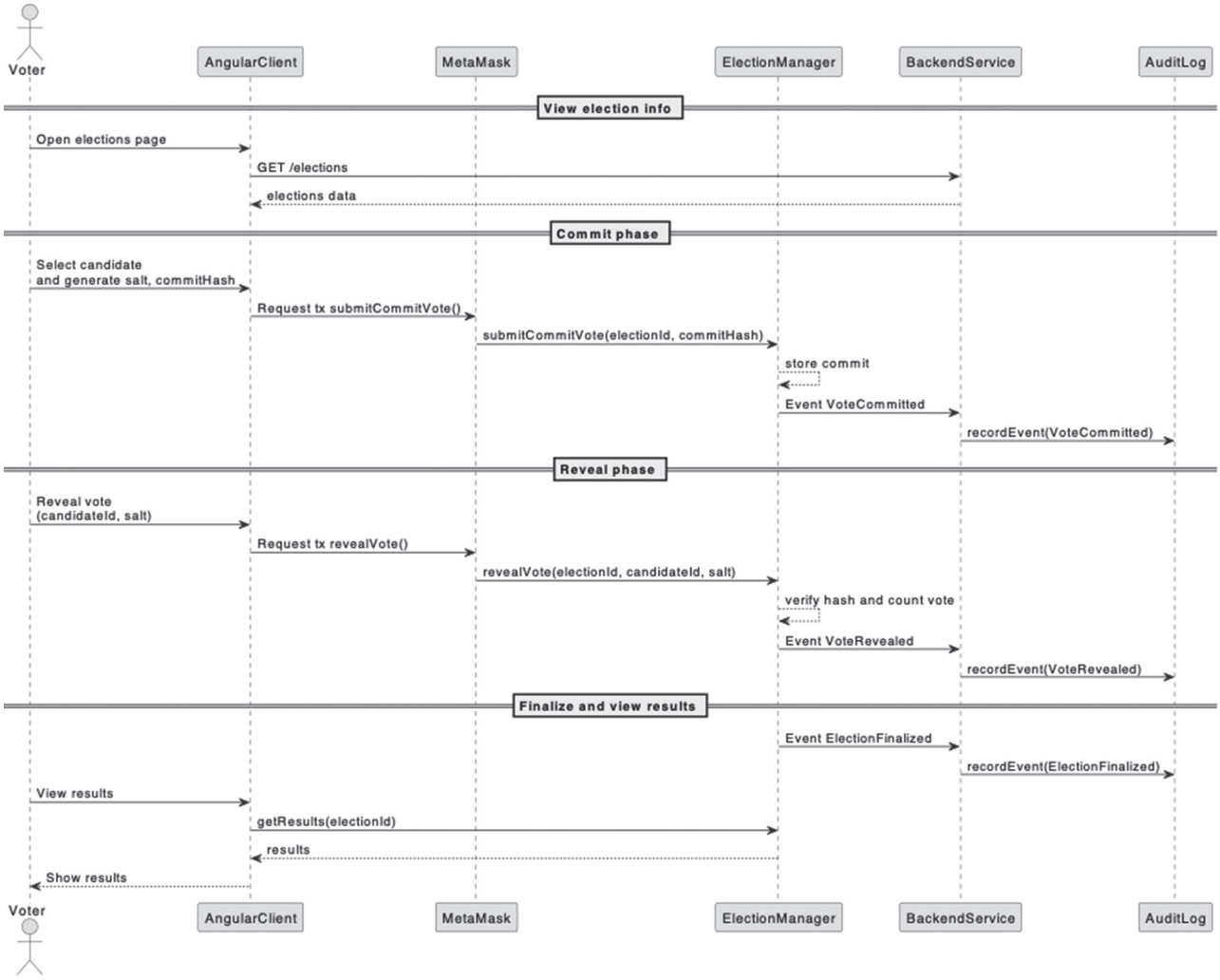

Рис. 3. Sequence Diagram

розгортання та емуляції взаємодії з мережею. Контракти розгорталися у тестовій мережі Sepolia, що дозволило відтворити сценарії реальної роботи без витрат у публічній мережі.

Смарт-контракт ElectionManager реалізує механізм створення виборів, збереження параметрів виборчої кампанії та виконання протоколу commit–reveal. На етапі commit виборець надсилає хеш свого голосу разом із криптографічною сіллю, а після завершення голосування виконує розкриття значення, що дозволяє смарт-контракту перевірити коректність хешу та зарахувати голос. Такий підхід виключає можливість зміни голосу заднім числом та забезпечує чесність процедури. Додатково було реалізовано контракт VotingRightToken, який відповідає за видачу токенів прав голосу у форматі SBT. Це унеможливлює передачу або дублювання голосів, оскільки кожен токен прив'язаний до конкретної адреси і не може бути переміщений.

Офчейн-інфраструктура системи була реалізована на Node.js із використанням Express. Цей компонент не впливає на підрахунок голосів, але відіграє важливу роль у зручності роботи та виведенні даних для адміністратора. Сервер взаємодіє зі смарт-контрактами через бібліотеку ethers.js, отримує та зберігає журнальні події, включаючи коміти, розкриття голосів та фіналізацію виборів. Метадані, а також технічні журнальні записи, зберігаються у базі даних SQL Server, що забезпечує можливість додаткової аналітики, фільтрації подій та виявлення потенційних аномалій у поведінці користувачів.

Клієнтський застосунок був створений на основі Angular і забезпечує інтуїтивний інтерфейс для обох ролей – адміністратора та виборця. Підключення MetaMask відбувається безпосередньо в браузері, що дозволяє користувачам підписувати транзакції та взаємодіяти з блокчейном без посередників. Виборець може переглядати доступні виборчі кампанії, надсилати commit-хеш голосу, виконувати розкриття після завершення голосування та переглядати результати у режимі реального часу. Адміністратор отримує доступ до інструментів створення виборів, керування часовими етапами та моніторингу стану виборчих процесів.

Для оцінки ефективності системи було проведено низку експериментів у тестовій мережі Sepolia. Було виміряно вартість транзакцій різних типів, затримку виконання операцій та поведінку системи за умов

збільшеного навантаження. Вартість commit-транзакції в середньому становила від 0.0004 до 0.0007 доларів США, що робить модель придатною для практичного використання в організаціях, де кількість виборців є обмеженою. Підрахунок голосів і фіналізація виборів виконувалися смарт-контрактом практично миттєво, а час відгуку клієнтського застосунку переважно залежав лише від швидкодії мережі Ethereum.

Під час експериментів також було протестовано механізм обробки спроб повторного голосування, відтворення недійсних розкриттів та некоректних комітів. Смарт-контракт успішно блокував повторні спроби надсилання голосів від однієї адреси та відхиляв розкриття, яке не відповідало раніше поданому хешу. Це підтверджує здатність системи забезпечувати захист від маніпуляцій без втручання адміністратора. Офчейн-журналування дало змогу додатково проаналізувати активність учасників та підтвердило відповідність усіх смарт-контрактних подій даним у базі.

Отримані результати свідчать про практичну можливість використання розробленої архітектури в умовах реальних організаційних процесів. Система продемонструвала стійкість до маніпуляцій, передбачувану поведінку під навантаженням та технічну здатність до масштабування. Проведені дослідження підтверджують, що поєднання блокчейна з офчейн-компонентами забезпечує необхідний баланс між безпекою, прозорістю та зручністю використання.

5. Обговорення, новизна та практичне значення роботи

Розроблений прототип децентралізованої системи електронного голосування демонструє можливість практичного застосування блокчейн-технологій у виборчих процесах організацій різного типу. На відміну від традиційних централізованих платформ, які покладаються на серверну інфраструктуру та довіру до адміністратора, запропонована система забезпечує прозорість та перевірюваність усіх ключових етапів голосування. Запис подій у блокчейні, використання відкритої логіки смарт-контрактів та автоматизований підрахунок голосів виключають можливість прихованих втручань, а також усувають загрозу фальсифікацій та маніпуляцій результатами. Це дозволяє досягти нового рівня довіри між учасниками виборчого процесу, оскільки контроль над даними більше не концентрується у руках однієї сторони.

Новизна запропонованого підходу полягає у поєднанні кількох концептуальних рішень, які рідко зустрічаються у комплексі в існуючих системах. Одним із таких рішень є використання токенів прав голосу у форматі soulbound-токенів, що забезпечують незмінність і непередаваність виборчого права. Цей підхід дозволяє будувати прозору модель авторизації виборців, у якій неможливо дублювати або продавати голоси, що є однією з головних проблем класичних систем. Іншим важливим елементом новизни є інтеграція офчейн-журналювання та аналітики, яка поєднується з ончейн-захистом. Така архітектура дозволяє зберігати критичні події у блокчейні, а другорядні дані – у швидкій та гнучкій офчейн-системі, що робить підхід одночасно безпечним і зручним для адміністраторів [19].

Особливого значення набуває реалізація механізму commit–reveal, який забезпечує конфіденційність вибору та гарантує коректність його розкриття. Завдяки цьому механізму система виключає можливість раннього доступу до інформації про голоси, а також унеможливлює їх модифікацію після завершення голосування. Ця властивість є критично важливою у будь-яких виборах, де довіра виборців до процесу базується на переконанні, що голоси не будуть підмінені чи видалені. У поєднанні з децентралізованою архітектурою commit–reveal формує технічну основу для побудови повністю прозорих виборчих процесів.

Практичне значення системи проявляється у можливості її адаптації до різних сценаріїв організаційного управління. Система може використовуватися для виборів студентського самоврядування, прийняття рішень у корпоративних структурах, формування дорадчих комітетів, проведення внутрішніх опитувань тощо. Завдяки відкритій логіці підрахунку та відсутності централізованого контролера система може слугувати інструментом для проведення голосувань у середовищах, де відсутній високий рівень довіри між учасниками або де існує ризик політичної чи адміністративної упередженості.

Важливо підкреслити, що запропонований прототип не є кінцевим рішенням, а радше демонструє практичну життєздатність архітектурної концепції. Подальший розвиток системи може включати впровадження більш ефективних криптографічних методів, оптимізацію вартості транзакцій, підтримку більш масштабованих мереж та розробку зручніших інтерфейсів для нетехнічних користувачів. З огляду на глобальний тренд до цифровізації демократичних процесів, дослідження у цьому напрямі має значний потенціал та здатне забезпечити основу для створення надійних, децентралізованих виборчих платформ нового покоління [20].

Таким чином, обговорення результатів дозволяє зробити висновок, що поєднання смарт-контрактів, токенів прав голосу та механізму commit–reveal формує інноваційний підхід до організації виборчих процесів у середовищі з підвищеними вимогами до безпеки та прозорості. Система демонструє здатність вирішувати ключові проблеми традиційних платформ та відкриває можливості для масштабного впровадження блокчейн-голосування у практичну діяльність організацій.

Висновки

У роботі було розглянуто проблему забезпечення прозорості, безпеки та довіри у процесі електронного голосування, що є особливо актуальним у сучасних організаціях, де виборчі процедури дедалі частіше переходять у цифрову форму. Аналіз існуючих рішень показав, що традиційні централізовані системи не здатні повною мірою усунути ризики маніпуляцій, тоді як наявні криптографічні або комерційні платформи не забезпечують достатнього рівня відкритості та незалежного аудиту. У цьому контексті блокчейн-технології постають ефективним інструментом для побудови децентралізованих виборчих систем, де ключові дані залишаються незмінними, а процес підрахунку голосів є відкритим та перевірюваним.

У межах роботи було розроблено модель та прототип децентралізованої системи голосування, що поєднує механізм commit–reveal, смарт-контракти для управління виборчим процесом та токени прав голосу у форматі SBT. Запропонована архітектура забезпечує прозору взаємодію між учасниками, виключає вплив адміністратора на критичні етапи голосування та надає можливість незалежного аудиту всіх подій. Реалізація системи включає ончейн-компоненти на базі Ethereum та офчейн-інфраструктуру, призначену для аналітики й підвищення зручності використання. Проведені експериментальні дослідження підтвердили технічну життєздатність моделі, її стійкість до маніпуляцій і здатність працювати у реальних організаційних сценаріях.

Отримані результати демонструють перспективність використання блокчейна для організації виборів у середовищах, де важливою є відсутність довіреної сторони та забезпечення високого рівня прозорості. Прототип показав, що децентралізовані механізми можуть не лише замінити традиційні підходи, але й значно підвищити рівень довіри до результатів. Подальший розвиток системи може включати оптимізацію вартості транзакцій, впровадження розширених криптографічних протоколів захисту вибору, розширення можливостей масштабування та адаптацію інтерфейсу для широкого кола користувачів. Отже, запропоноване рішення створює основу для побудови ефективних та безпечних виборчих платформ нового покоління.

Список літератури:

[1] Huang J. The application of blockchain technology in voting systems: A review / J. Huang, D. He, M. S. Obaidat, P. Vijayakumar, M. Luo, K. Choo // ACM Computing Surveys. – 2021. – Vol. 54, №3. – P. 1–28.

[2] Hajian Berenjestanaki M. Blockchain-based e-voting systems: a technology review / M. Hajian Berenjestanaki, H. Barzegar, N. El Ioini, C. Pahl // Electronics. – 2023. – Vol. 13, №1. – P. 1–21.

[3] Jafar U. A systematic literature review and meta-analysis on scalable blockchain-based electronic voting systems / U. Jafar, M. Ab Aziz, Z. Shukur, H. Hussain // Sensors. – 2022. – Vol. 22, №19. – P. 7585.

[4] Chouhan V., Arora A. Blockchain-based secure and transparent election and vote-counting mechanism using secret sharing scheme / V. Chouhan, A. Arora // Journal of Ambient Intelligence and Humanized Computing. – 2023. – Vol. 14. – P. 14009–14027.

[5] Akinbohun S., Apeh S., Olaye E., Ogbeide O. Literature review of blockchain-based voting system: framework and concept // Journal of Civil and Environmental Systems Engineering. – 2023. – Vol. 20, №1. – P. 72–83.

[6] Al-Maaitah S., Qatawneh M., Quzmar A. E-voting system based on blockchain technology: A survey // Proc. of International Conference on Information Technology (ICIT). – IEEE, 2021. – P. 200–205.

[7] Helios Voting. – URL: https://vote.heliosvoting.org/

[8] Voatz. – URL: https://voatz.com/

[9] Ethereum Foundation. Solidity Documentation. – URL: https://docs.soliditylang.org/

[10] Sharp M., Roberts L., Xiao H. Blockchain-Based E-Voting Mechanisms: A Survey and Comparative Analysis // Journal of Information Security and Applications. – 2024.

[11] Rahman M. Implementation of blockchain-based e-voting system / M. Rahman et al. // Multimedia Tools and Applications. – 2023. – Vol. 83, №1. – P. 1–32.

[12] ethers.js Documentation. – URL: https://docs.ethers.io/

[13] Visual Paradigm. UML Modeling Tool. – URL: https://www.visual-paradigm.com/

[14] PlantUML Documentation. UML Diagram Generator. – URL: https://plantuml.com/

[15] Denis González C., Frias Mena D., Massó Muñoz A., Rojas O., Sosa-Gómez G. Electronic voting system using an enterprise blockchain // Applied Sciences. – 2022. – Vol. 12, №2. – P. 531.

[16] Dai W., Liu C. Soulbound tokens and decentralized identity // IEEE Access. – 2023. – Vol. 11. – P. 1221–1233.

[17] MetaMask Documentation. – URL: https://docs.metamask.io/

[18] OpenZeppelin Smart Contracts. – URL: https://docs.openzeppelin.com/contracts

[19] Zhang X., Lee Y. Secure smart-contract-based voting using Ethereum blockchain // IEEE Access. – 2021. – Vol. 9. – P. 34335–34347.

[20] Kim S., Park J. Privacy-preserving blockchain voting: State-of-the-art review // Computers & Security. – 2022. – Vol. 118. – P. 102731.

Надійшла до редколегії 29.09.2025

UDC 004.8:004.652:004.912 DOi 10.30837/ bi.2025.2(103).09

O. O. Sutiahin
NTU KhPI, Kharkiv, Ukraine, sutiahin.oleksandr@cs.khpi.edu.ua,
ORCID iD: 0009-0005-6527-455X

A MULTI-STAGE SELF-REVIEW FRAMEWORK FOR TRANSLATING NATURAL LANGUAGE INTO NEO4J CYPHER QUERIES

A Multi-Stage Self-Review Framework for Translating Natural Language into Neo4j Cypher Queries. The article presents a multi-stage self-review framework for automatically translating natural language questions into Cypher queries for the Neo4j graph database. The proposed approach integrates self-review mechanisms of large language models (LLMs), knowledge graph structure analysis, and multi-level validation of the syntax and semantics of generated queries. The framework includes three core stages: preliminary graph schema analysis, initial LLM-based query generation, and iterative self-review using specialized validation agents that detect logical, structural, and analytical inconsistencies. A prototype implementation is developed to evaluate the difference between query generation with and without the self-review mechanism. Experimental results demonstrate that incorporating self-review improves Cypher query correctness, reduces logical and structural errors, and enhances alignment with OLAP-oriented analytical requirements. The findings confirm the effectiveness of multi-stage self-review workflows for increasing the reliability of natural-language interfaces to graph-based analytical systems.

LLM, CYPHER, NEO4J, QUERY GENERATION, SELF-REVIEW, KNOWLEDGE GRAPH, OLAP, ANALYTICAL SYSTEMS, RAG, QUERY VALIDATION

О. О. Сутягін. Багатокрокова система саморецензування для перетворення природної мови у Cypher запити Neo4j. У статті представлено багатокрокову систему саморецензування для автоматизованого перетворення текстових запитів природною мовою у Cypher-запити до графової бази даних Neo4j. Робота поєднує механізми самоперевірки великих мовних моделей (LLM), аналіз структури графа знань та багаторівневу валідацію синтаксису й семантики згенерованих запитів. Запропонований підхід включає три основні етапи: попередній аналіз схеми графа, первинну генерацію запиту на основі LLM та ітеративну самоперевірку з використанням агентів валідації, які виявляють логічні, структурні та аналітичні помилки. Запропонована система впроваджена у прототипі програмного забезпечення, що виконує експериментальне порівняння генерації запитів із самоперевіркою та без неї. Результати експериментів показують, що використання self-review механізму забезпечує підвищення коректності Cypher-запитів, зменшення кількості логічних та структурних помилок і покращення відповідності сформованих запитів аналітичним OLAP-вимогам. Отримані результати підтверджують ефективність багатокрокового саморецензування для підвищення надійності текстового інтерфейсу до графових аналітичних систем.

LLM, CYPHER, NEO4J, ГЕНЕРАЦІЯ ЗАПИТІВ, САМОПЕРЕВІРКА, ГРАФ ЗНАНЬ, OLAP, АНАЛІТИЧНІ СИСТЕМИ, RAG, ВАЛІДАЦІЯ ЗАПИТІВ

1. Introduction

Text-to-SQL converts natural language into SQL queries, enabling non-experts to access databases without SQL knowledge [23]. While Large Language Models (LLMs) can interpret natural language, they are prone to errors and hallucinations. To address this, self-correction and verification techniques have been developed, including methods that refine LLM responses through generated feedback [24]. However, retraining LLMs is often impractical due to time constraints, so this work focuses on prompt-based teaching.

We introduce a self-verification approach for converting natural language to Cypher queries. Our multi-agent framework automatically generates and applies correction guidelines to remedy common graph database query errors, such as incorrect relationship directions and referencing nonexistent analytical entities. Unlike previous text-to-SQL efforts, our method targets the challenges of Cypher queries in Neo4j graph databases, particularly for OLAP tasks, by adapting the MAGIC framework for this context.

The framework operates in three logical stages, implemented via a multi-agent architecture: schema analysis, initial query generation, and iterative self-review at inference time. This prevents queries from referencing invalid or hallucinated entities.

Through a multi-agent architecture, our framework delegates schema analysis, query generation, review, and correction to specialized agents, mirroring human strategies and improving reliability over single-agent systems. Key contributions include:

– Establishing self-correction guideline generation for text-to-Cypher tasks in graph databases, adapting the MAGIC framework to handle graph-specific issues [24].

– Introducing a multi-agent validation system that integrates syntactic, semantic, and execution checks with real-time metadata verification.

– Developing an automated schema analysis pipeline for accurate and efficient LLM context injection.

– Implementing and empirically evaluating a complete multi-agent self-review pipeline for text-to-Cypher generation in Neo4j OLAP scenarios.

– The remainder of this paper covers related work (Section 2), methodology and architecture (Section 3), experimental setup (Section 4), and results (Section 5).

2. Literature review

2.1 LLM self-review methods

The proposed methodology operates on the principle that LLMs possess the capability to critically evaluate their own outputs when explicitly prompted to do so. By introducing a validation phase between initial response generation and final output delivery, we create an opportunity for the model to identify potential inconsistencies, logical fallacies, or factual inaccuracies that may have emerged during the initial generation phase. The method could be implemented by generating suitable prompts for LLMs to identify fallacies and correct them [8]. The authors proposed and evaluated several prompts on different LLMs for detecting, categorizing, and solving formal and non-formal fallacies step by step to decrease probability of logical reasoning errors.

This paper proposes and demonstrates that LLMs possess similar self-verification abilities [20]. The method operates in two stages: Forward Reasoning generates candidate answers using CoT, while Backward Verification masks original conditions and predicts them based on the proposed conclusion.

QueryGenie is a framework to generate SQL query from a sentence, using a self-review method [21]. It consists of different modules: a confirmation module that helps LLM to verify intent of generation SQL, query generation module and query validation method, which execute the query and recheck result. Despite the practical idea, there wasn't any practical evidence this framework was implemented, so its practical applicability remains unclear of QueryGenie.

Another app that helps with generation of SQL queries is MAGIC [22]. The authors created a multi-agent system that automates the creation of the self-correction guideline. These agents operate collaboratively in an iterative framework to analyze failures produced by a baseline large language model (LLM) on the training dataset. Through this iterative process, the system autonomously generates and refines self-correction guidelines specifically calibrated to address systematic errors made by the LLM. This approach emulates human error-analysis and guideline-development processes while maintaining complete autonomy from human intervention. This approach was chosen as the basis for our method.

2.2 Knowledge graph

The knowledge graph was developed to better manage knowledge by connecting entities—real-world objects—via semantically described edges. Entities are typically represented as triples (subject, predicate, object), and edges denote their relationships. There are two main construction approaches: top-down, which starts with ontology creation, and bottom-up, which begins with data extraction. To build a knowledge graph, study [1] suggests: identifying relevant domains and data sources, constructing an ontology (mainly for top-down methods using existing formats like OWL, XML, or RDF), extracting knowledge—often with machine learning, processing to eliminate redundancy and ambiguity, enriching the knowledge, and finally, developing, storing, visualizing, and deploying the knowledge graph. Neo4j is a widely used database for storing knowledge graphs.

A knowledge graph could be presented as G (E, R, T), where E, R and T represent the set of entities, relations, and knowledge triples, respectively [2]. For every knowledge triplet T ∈ T summarize knowledge of graph G and presented as T = (e_h, r, e_t), where e_h, e_t ∈ E and r ∈ R. This work also considers two definitions for reasoning path and entity path [2].

For reasoning path they use formula that represents set of connected sequences of triplets in graph: $path_G$ (e_l, e_{l+1}) = {$T_1, T_2, \ldots, T_l$} = {(e_1, r_1, e_2), (e_2, r_2, e_3), ..., (e_l, r_l, e_{l+1})}, where T_i ∈ T denotes the *i*-th triple in the path and *l* denotes the length of the path, i.e., *length* ($path_G$ (e_l, e_{l+1})) = *l*. Example: Consider a reasoning path between the entity "University" and the entity "Student" in a KG. The reasoning path is given by: $path_G$ (University, Student) = {(University, employs, Professor), (Professor, teaches, Course), (Course, enrolled_in, Student)}, and can be visualized as:

University $\rightarrow^{employs}$ Professor $\rightarrow^{teaches}$ Course $\rightarrow^{enrolled\_in}$ Student.

This path indicates that a “University" employs a “Professor," who teaches a “Course," in which a "Student" is enrolled. The length of the path is 3.

Second definition looks like: given a KG_G and a list of entities $list_e$ = [e_1, e_2, e_3, . . ., e_l], the entity path of $list_e$ is defined as a connected sequence of reasoning paths, which is denoted as

$path_G$ ($list_e$) = {$path_G$ (e_1, e_2), $path_G$ (e_2, e_3), ..., $path_G$ (e_{l-1}, e_l)} = {(e_s, r, e_t) | (e_s, r, e_t) ∈ $path_G$ (e_l, e_{l+1}) ∧ $1 \le i < l$}.

All those definitions are used to prompt LLM models with KG that will prevent LLM anomalies and improve performance of answering questions. The main idea is to construct KG that contains facts and relations between them. The whole process is divided into four stages: initialization, exploration, path pruning, and question answering.

Also, KG are used for improving understanding of facts and relations between them in LLMs. Those models often suffer from hallucinations because of lack of specific information contained in trained corpus and problems of using chain-of-thought. A possible solution to these issues is to integrate knowledge graphs (KGs) into LLMs. This study [3] provides a strategy for LLMs to communicate with KGs for improving reasoning. Authors developed three categories of interaction between LLM and KG: KG-enhanced LLM, LLM-augmented KG, synergized LLM + KG [3].

In the context of this work, these knowledge graph concepts are not used for explicit reasoning path extraction but serve as a theoretical foundation for schema-aware prompting. The proposed framework leverages structured graph representations primarily to constrain Cypher generation and prevent hallucinated entities rather than to perform symbolic graph reasoning.

3. Method

The BI4people project [4] aims to analyze data and make business decisions collaboratively by leveraging Online Analytical Processing (OLAP) to provide interactive analysis and data visualization. It introduces Collaborative Business Analysis (CBA) as a process of analyzing data and making business decisions collaboratively. Under the idea of developing the software-as-a-service model, the BI4people project tries to bring people together in a virtual space and encourage them to share their opinions and comments to collectively solve problems. The concept of CBI involves using social networks, quizzes, brainstorming sessions, and even simple chats [5]. Additionally, the reuse of other collaborators' comments or results is also considered part of CBI, which leads to a more comprehensive approach to BI. When creating a virtual space or forum for users to share their problems, comments, and solutions, it is essential to gather and analyze their data to unveil the intricate relationships between different pieces of information. It is crucial for the user to provide feedback on how they utilized the analysis carried out that can make a background for recommendations and reusing. BI4Tourism is a web-based application designed to empower users with insights into tourism data, thereby facilitating decision-making. At its core, BI4Tourism enables users to visualize and compare diverse tourism data, unraveling dependencies crucial for making strategic decisions [6].

For collaboration systems it's crucial to be able to investigate someone's similar question, process of research and result of it. In our application, we trace every action of the user during work on use cases and save it to the database. We need it to gather the context of user investigation to recommend other users with similar questions. Also, it's very important to persist in comments that are marked as answers because other users can participate in any use of case discussion. We use an OLAP cube to save our raw data and then our app queries it with different dimensions and measures, filters to retrieve information that is formatted to draw any chart. Among several implementations of OLAP cubes, we consider using Cube [7] because of easy installation and deployment locally. The schema of our graph is presented in Figure 1.

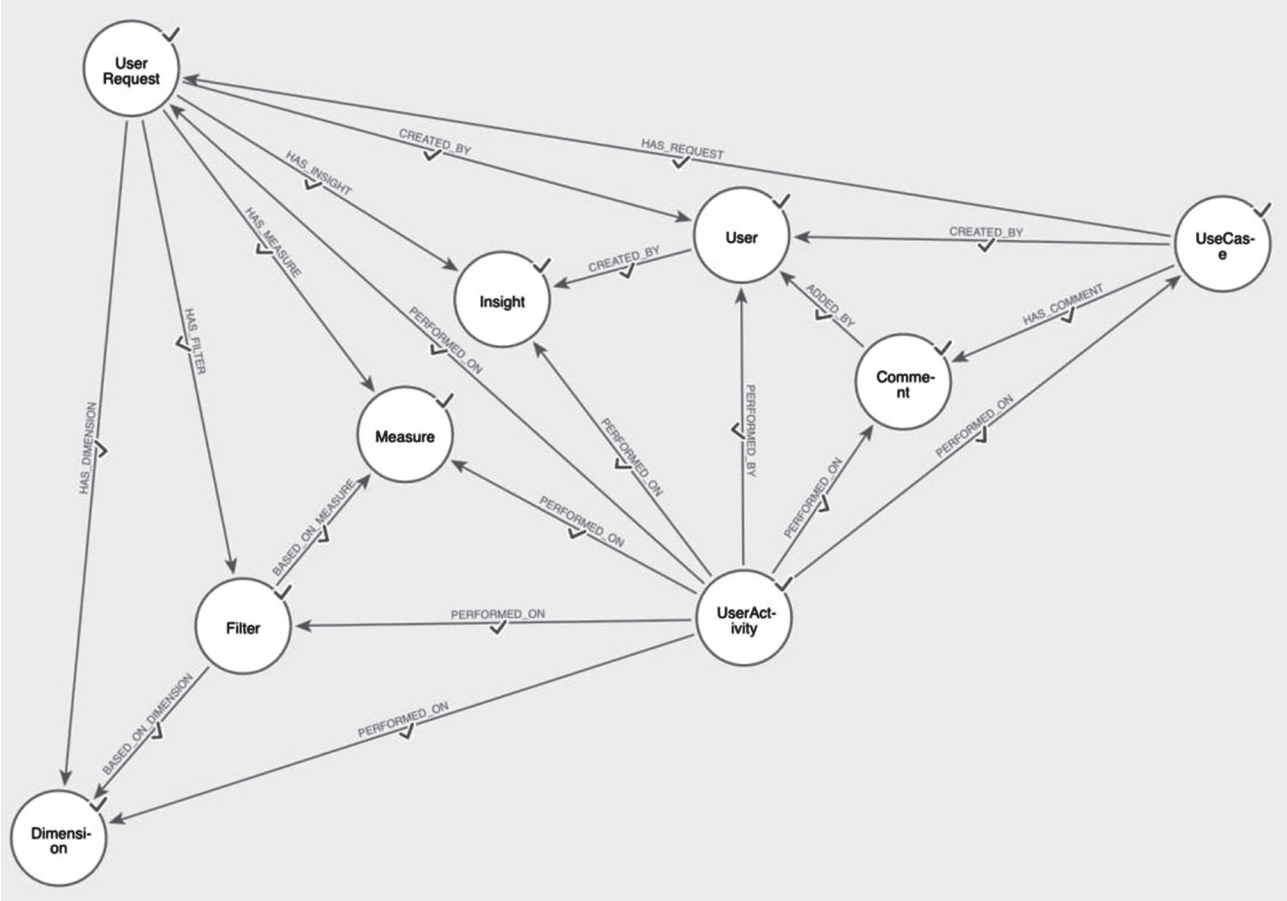

Fig. 1. Schema of knowledge graph for BI4Tourism in Neo4j database

3.1 Knowledge graph schema

The construction of the knowledge graph (KG) begins with the formal specification of its entities. Because the KG is intended to capture relationships among dimensions, measures, users, use cases and filters we instantiate each of these constructs as first-class entities within the graph. More generally, an entity in a KG denotes a categorical referent (for example, a person or a location). So in our system we establish that nodes of the knowledge graph could have one of the next types: dimension (E_d), measure (E_m), user

(E_u), use case (E_{uc}), comment (E_c), user request (E_{ur}), insight (E_i), filter (E_f) or user action (E_{ua}). Dimension usually represents values that describe the business context of the data, allowing users to slice and dice our OLAP cube by using filters. Filter contains operation type, filter value and unique id. Measures are data points within the cube, often numeric values from a fact table, by which we want to group data. User request's nodes gather all measures, filters and dimensions in one named request with help of which the user chose to investigate the chart. Insight entity represents which problem the user solved by investigating the result of the user request, it consists of comments that describe the solution and author of solution. Use case is the main question on which the user tries to find the answer. Use cases can have multiple user requests and each user request can have multiple insights. But one of the most important parts of our knowledge graph is the user action entity. It represents every action a user can make on a use case: write comment, change filter, add or update dimension or measure, add user request and insight. to this request.

After introducing nine classes of entities, we further defined ten classes of relations as shown in Figure 1. First class, we have relations between use case, user, user request and comment entities ($R_{uc\to ur}$ $R_{uc\to u}$, $R_{uc\to c}$) that represent "(one use case) created by (a specific author)", "(one use case) has a user request (a specific user request)" and "(one use case) has comment (a specific comment)". Second, we have a class of relations that connect dimensions to user request ($R_{ur\to d}$) and represent "(one user request) has dimension (a specific dimension)". Another class is connection between user request and measure ($R_{ur\to m}$) and means "(one user request) has measure (a specific measure)". Next one is about the relation between user and user request ($R_{ur\to u}$), it means that the user is the author of this user request, an example of it "(one user request) created by (a specific user)". And the next two types of relations are dependencies between filter and user request ($R_{ur\to f}$) and filter and dimension ($R_{d\to f}$), it could look like "(one user request) has a filter (a specific filter)" and "(one dimension) filtered by (a specific filter)". Also, we have a class of relation that links use case and user, reproducing whole user activity on this use case entity ($R_{ua\to u}$) and ($R_{ua\to uc}$, $R_{ua\to ur}$, $R_{ua\to c}$, $R_{ua\to d}$, $R_{ua\to m}$, $R_{ua\to i}$) for example "(one user action) performed on (one user request) by (one user)". Last type of relation is the connection between user requests and insights ($R_{ur\to i}$), it means that "(one user request) has an insight (a specific insight)".

3.2 Self-review Cypher generating method

We propose our method of generating Cypher query, based on self-review LLM method, to improve the experience of non-expert users in knowledge graphs. To achieve this aim, we introduced a multiagent system that consists of a schema review agent, feedback agent, correction, and Cypher generation agent.

The schema review agent should investigate the schema of Neo4j database, summarize it to a short description of nodes and relations, add more context about the database: domain of model, intent of different nodes and their relations. We need this information to share it with others and not overcharge their token window. Prompts for this agent are gathered in Appendix A

The Cypher generation agent has a database schema, context of the schema as input parameters to generate proper requests, based on user input text. The result of this agent's work should be presented as a ready to use Cypher query. Also, it consumes hints from the feedback agent to regenerate the query. Appendix C represents the prompt that the agent uses. And the next agent is the feedback one; it consumes schema context, user input, Cypher query and result of execution of this query. It analyzes those parameters with proper prompt in cycles, which is presented in Appendix D and sends a structured response for the correction agent to change the initial Cypher query. The correction agent helps to regenerate Cypher query based on the result of the feedback agent. The prompt for this agent is saved in Appendix E. The whole process is documented in Figure 2.

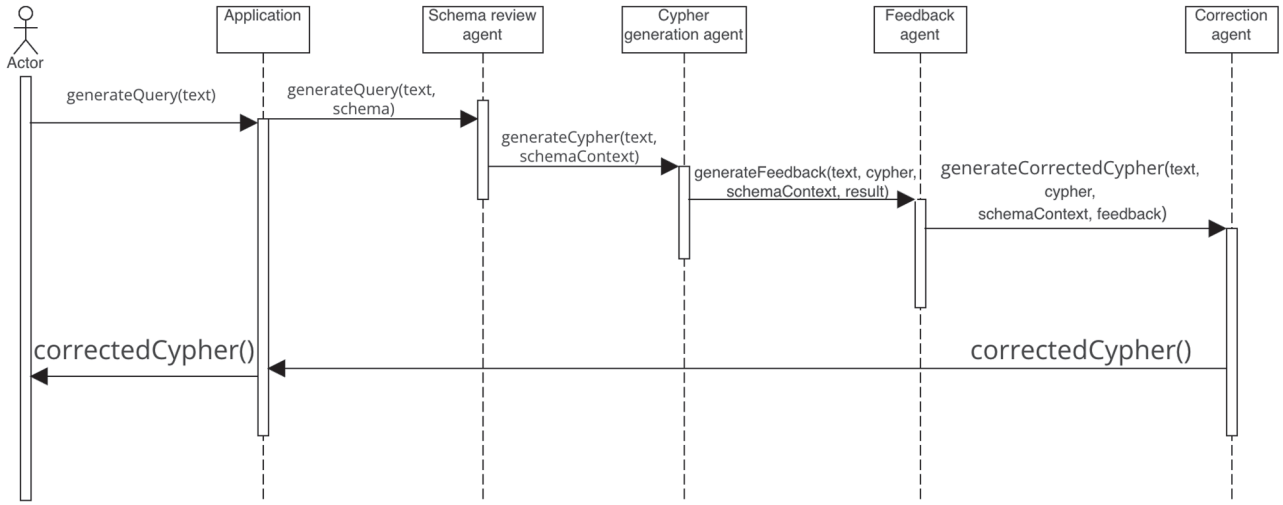

Fig. 2. Sequence diagram of communication between agents to get corrected cypher

4. Experiment

4.1 Experiment setup

We conducted an experiment to compare two ways of generating Cypher query from text with preprompted LLM: with self-review mechanism and without. Firstly, we had to choose a proper LLM, which will be used in the experiment. We compared the following LLMs: ChatGPT-4o [12], Groq [11], GPT-OSS-120B [9], Llama 3.3 70B [10] by context window, limit of requests per minute and speed of answer in Table 1.

Table 1

Results of comparison of LLMs

LLM name	Context token window	Limit of tokens/ minute	Price per 1M tokens
ChatGPT-4o	1,047,576	30,000	Input — $2.00 Output — $8.00
Groq	131,072	200,000	Free but depending on usage of the model
GPT-OSS-120B	131,072	250,000	Input — $0.15 Output — $0.60
Llama 3.3 70B	131,072	300,000	Input — $0.59 Output — $0.79

We selected GPT-OSS-120B as our primary language model based on its optimal balance of cost-efficiency and performance characteristics. Our implementation employs a self-review methodology requiring multiple sequential LLM requests per query within short time windows, making throughput capacity critical. This selection enabled economically feasible large-scale experimentation with iterative prompting without compromising result of quality or reproducibility.

We synthetically generate test data and approximately 100 tourism-related queries, grouped by different categories, based on real questions about tourism from Reddit [14]. We persisted the data in the Neo4j cloud database - Neo4j AuraDB [13].

In this study, we compared two prompts for generating Cypher queries from natural language: one that included self-review steps and one that did not. To identify which approach was more effective, we focused on key Retrieval-Augmented Generation (RAG) metrics, particularly answer correctness. We used the DeepEval framework [15] for our evaluation, based on evidence from GroUSE [16], a meta-evaluation tool for benchmarking evaluators. GroUSE reports that DeepEval outperforms similar frameworks such as RAGAS [17] on faithfulness, correctness, and answer relevance for grounded question answering. DeepEval relies on the LLM-as-a-judge paradigm, which has been shown to be more reliable than traditional statistical or human-based evaluation methods [18].

During implementation of evaluation in DeepEval framework we chose a prompt-based GPTEval method [19] which uses a chain of thoughts model to assess natural language generation output based on LLM-as-judge method. The framework prompts LLM to score some value for each evaluation aspect, based on the defined criteria, then LLM should add weights to those scores and summarize it. We defined our evaluation aspects for generation Cypher query based on two sets of prompts: for generation with feedback, we use Appendix A-E, but without self-review we use only prompts A-C.

The GPTEval scoring function is presented as a predefined set of discrete scores (e.g., 1 to 5) specified within the prompt, serving as the evaluation scale for subsequent assessments - $S = \{s_1, s_2, ..., s_n\}$. The token probability of each score $p(s_i)$ is the probability the LLM assigns to generating that score token, and the final score is:

$$score = \sum_{i=1}^{n} 2p(s_i) \times s_i \quad (1)$$

We configured the whole process of evaluation in the form of several categories of different queries, each containing different tests with data from our dataset, and wrote it in Python with the Deepeval framework.

4.2 Experiment results

Firstly, we grouped results of our tests for two methods of prompting by score each test received from the G-eval framework. As we can see on Figure 3-4 strategy of using LLM prompting with feedback loop shows better results than strategy without self-review, average of first is 0.65 score and second is 0.61 accordingly. The score is calculated with the GPTEval [19] method, described in previous section.

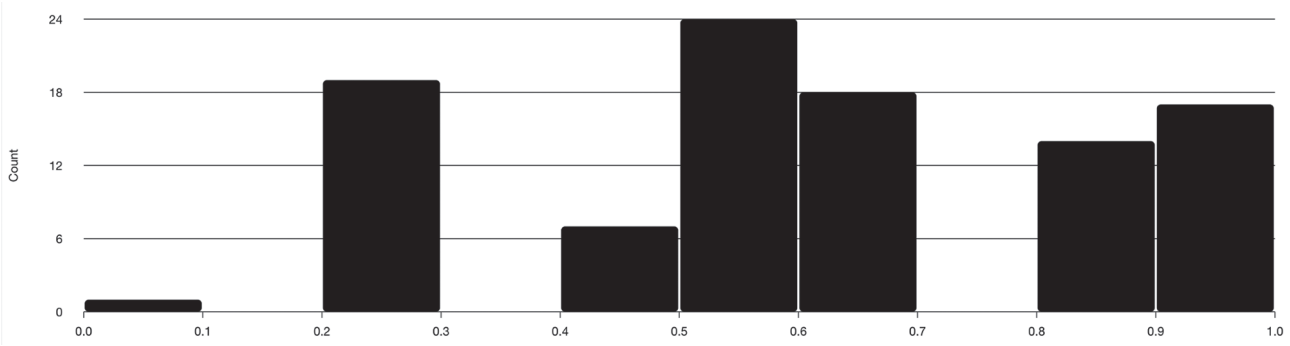

Fig. 3. Count of tests to their score in case of using LLM prompting without feedback loop

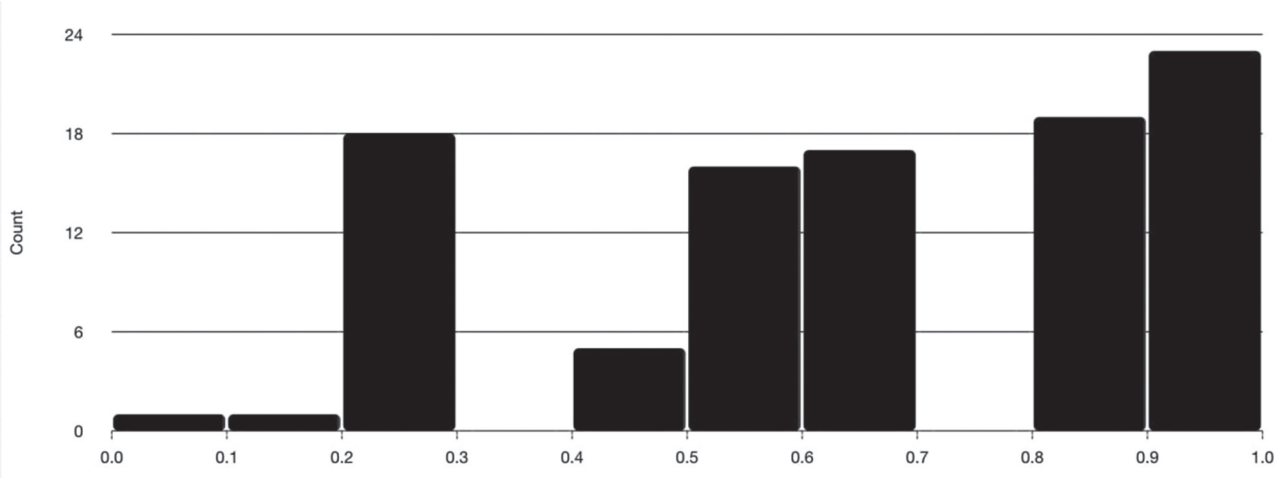

Fig. 4. Count of tests to their score in case of using LLM prompting with feedback loop

Next grouping key was the category of issues, which occurred during tests run. These reasons were generated by the G-eval framework using ChatGPT-4o [12] and then grouped by issue categories. It helps us to determine weak points of our prompt. The weakest points of both our prompts were incorrect logic of generating Cypher query, incorrect relationships, missing fields and filters. So, in this case the first and the second prompts made similar result.

Our last grouping field is the input category for average correctness score. For the prompt with feedback loop, multi-dimensional queries achieve the highest correctness, followed by Meta/Schema and Complex/Other, whereas in the prompt without feedback, Meta/Schema leads, with Aggregation/Statistics showing marked improvementю A notable divergence is observed in the multi-dimensional category, which drops from the top-performing position to mid-tier performance between prompts - a decrease of 22 percentage points. Conversely, Content/Insight demonstrates improvement, and Business Metrics rises above the threshold in the second condition.

4.3 Experiment conclusion

The comparative analysis demonstrates that the feedback loop prompting strategy consistently outperforms the non-feedback approach under the evaluated conditions, achieving a 6.5% relative improvement in average correctness score (0.65 vs. 0.61). These findings suggest that iterative self-review enhances output quality and cross-category robustness, though specialized solutions remain necessary for spatial and complex filtering operations.

5. Conclusion

This work introduces a multi-agent self-review framework for generating Cypher queries from natural language, tailored to the unique challenges of translating user questions into Neo4j graph database queries, especially in OLAP contexts. By adapting the MAGIC self-correction approach, our system leverages automated guideline creation and iterative validation to improve the reliability and accuracy of LLM-generated queries.

Key innovations include an automated schema analysis pipeline that efficiently represents Neo4j structures, a multi-agent design separating query generation, review, and correction, and real-time metadata verification to prevent referencing non-existent analytical entities. These elements address issues like complex graph traversal and improve upon monolithic or purely syntactic solutions.

While effective in Neo4j OLAP settings, the framework's limitations include its specificity to certain schemas, potential latency and token overhead from iterative correction, and lack of learning from past queries. Future work should explore broader graph database support, optimization for efficiency, and incorporating a repository of validated examples for improved performance.

References

[1] Tamašauskaitė, G., & Groth, P. (2022). Defining a Knowledge Graph Development Process Through a Systematic Review. ACM Transactions on Software Engineering and Methodology, 32, 1 - 40. https://doi.org/10.1145/3522586.

[2] Tan, Xingyu & Wang, Xiaoyang & Liu, Qing & Xu, Xiwei & Yuan, Xin & Zhang, Wenjie. (2024). Paths-over-Graph: Knowledge Graph Empowered Large Language Model Reasoning. 10.48550/arXiv.2410.14211.

[3] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering (2024).

[4] Business intelligence for the people. https://eric.univ-lyon2.fr/bi4people/index-en.html

[5] Muhammad, F., Darmont, J., Favre, C.: The Collaborative Business Intelligence Ontology (CBIOnt). 18e journées Business Intelligence et Big Data (EDA-22), Vol. B-18. Clermont- Ferrand, Octobre 2022, RNTI (2022)

[6] Cherednichenko, O., Sutiahin, O.: Development of Collaborative Business Intelligence Framework for Tourism Domain Analysis. In: Tekli, J., et al. (eds.) New Trends in Database and Information Systems. ADBIS 2024. CCIS, vol. 2186, pp. 270–281. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-70421-5_21

[7] Cube. Available: https://cube.dev/, last accessed 2025/11/16

[8] Hong, Ruixin & Zhang, Hongming & Pang, Xinyu & Yu, Dong & Zhang, Changshui. (2024). A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning. 900-925. 10.18653/v1/2024.naacl-long.52.

[9] OpenAI, "GPT-OSS-120B," OpenAI Platform Documentation. Available: https://platform.openai.com/docs/models/gpt-oss-120b, last accessed 2025/11/18.

[10] Meta and Ollama, "Llama 3.3 70B," Ollama Model Library. Available: https://ollama.com/library/llama3.3:70b, last accessed 2025/11/15.

[11] xAI, "Grok," xAI Official Documentation. Available: https://x.ai/, last accessed 2025/11/15.

[12] OpenAI, "GPT-4o," OpenAI Platform Documentation. Available: https://platform.openai.com/docs/models/gpt-4o, last accessed 2025/11/14.

[13] Neo4j, "Neo4j AuraDB: Fully Managed Graph Database," Neo4j Product Documentation. Available: https://neo4j.com/product/auradb/, last accessed 2025/11/10.

[14] Reddit, "Reddit - Dive into anything,". Available: https://www.reddit.com/, last accessed 2025/11/01.

[15] Confident AI: deepeval documentation. https://docs.confident-ai.com/ (2025), last accessed 2025/11/15.

[16] Muller, S., Loison, A., Omrani, B., Viaud, G.: GroUSE: A Benchmark to Evaluate Evaluators in Grounded Question Answering. arXiv preprint arXiv:2409.06595 (2024).

[17] Shahul Es, J.J., Espinosa-Anke, L., Schockaert, S.: Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217 (2023)

[18] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., Guo, J.: A Survey on LLM-as-a-Judge. arXiv preprint arXiv:2411.15594 (2024)

[19] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv preprint arXiv:2303.16634 (2023)

[20] Weng, Yixuan, et al. "Large language models are better reasoners with self-verification." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

[21] Longfei Chen, Shenghan Gao, Shiwei Wang. 2025. QueryGenie: Making LLM-Based Database Querying Transparent and Controllable. In Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25). Association for Computing Machinery, New York, NY, USA, Article 49, 1–4. https://doi.org/10.1145/3746058.3758982

[22] Askari, Arian, Christian Poelitz, and Xinye Tang. "Magic: Generating self-correction guideline for in-context text-to-sql." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 22, pp. 23433-23441. 2025.

[23] Qu, G., Li, J., Li, B., Qin, B., & Cheng, R. (2024). Before generation, align it! a novel and effective strategy for mitigating hallucinations in text-to-sql generation. arXiv preprint arXiv:2405.15307.

[24] Aman Madaan, Niket Tandon, Luyu Gao, Sarah Wiegreffe, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Syste

Date of submission of the article to the editorial board: 09.10.2025

Appendix A

Before any query generation occurs, we employ a dedicated schema analysis agent. The rationale for this preliminary step stems from the observation that LLMs often struggle to correctly interpret raw database schemas, leading to invalid node labels, incorrect relationship directions, and non-existent property references. The schema analyzer performs a comprehensive examination of the Neo4j graph structure, extracting:

- Node labels with their associated properties and data types
- Relationship types with directionality and cardinality
- Traversal patterns critical for OLAP analytics

	Prompt for mapping schema to table view for LLM consumption
1.	Analyze the following Neo4j graph schema with DETAILED focus on nodes and relationships: ## Raw Schema: {raw_schema} ## Required Analysis Structure: ### 1. NODES ANALYSIS For each node type in the schema, provide: - **Label**: Node label name - **Properties**: List all properties with their data types - **Purpose**: What this node represents in the domain - **Key Properties**: Properties that uniquely identify this node - **Usage Notes**: When and how to use this node in queries ### 2. RELATIONSHIPS ANALYSIS For each relationship type in the schema, provide: - **Pattern**: SourceNode -[:RELATIONSHIP_TYPE]-> TargetNode - **Direction**: Specify if directional or can be traversed both ways - **Properties**: List any properties on the relationship - **Semantic Meaning**: What this relationship represents - **Traversal Notes**: How to use this relationship in MATCH patterns - **Example Cypher Pattern**: Show a typical MATCH pattern using this relationship

	### 3. GRAPH STRUCTURE MAP - Draw a visual map of how all nodes connect through relationships - Show the complete graph topology - Identify primary traversal paths (e.g., UseCase -> UserRequest -> Measure) ### 4. DOMAIN MODEL INSIGHTS - What domain does this graph represent (OLAP, social network, etc.)? - What are the central/hub entities? - What are the analytical patterns supported? ### 5. QUERY GUIDANCE - Common MATCH patterns for this schema - Best practices for traversing relationships - Properties to use in WHERE clauses - Properties to RETURN for meaningful results Focus heavily on nodes and relationships - this is the most critical information for accurate Cypher query generation.

Fig. 5. Prompt for mapping schema to table view for LLM consumption

Appendix B

The primary query generation agent receives the user question and analyzed the schema context. The system prompts enforce strict adherence to the provided schema, instructing the model to use only documented relationship types and properties. For enhanced accuracy, we offer an alternative generation path incorporating the MAGIC (Multi-Agent Guideline Iteration for Correction) self-correction framework, which provides additional guideline documents containing learned patterns from previous query generation failures and instructs the model to engage in explicit step-by-step reasoning before query construction.

	Prompt for generating Cypher Query based on schema_context, question, hint
1.	Task: Generate Cypher queries to query a Neo4j graph database based on the provided schema definition and also provide the result of query. Instructions: Use only the provided relationship types and properties. Do not use any other relationship types or properties that are not provided in schema. If you cannot generate a Cypher statement based on the provided schema, explain the reason to the user. When generating queries: 1. Identify relevant measures and dimensions from the schema 2. Use only relationships from schema context and for appropriate nodes with right directions 2. Use proper node labels 3. Use proper relationship types for OLAP cubes (see schema for details) 4. Ensure queries are suitable for analytical/reporting purposes 5. When querying user requests, traverse through use cases to access their associated measures, dimensions, and filters 6. **IMPORTANT**: Results should usually be grabbed from UseCase and UserRequest entities as these are the primary analytical entities in the database Schema context: {schema_context} Question: {question} Hint: {hint} Generate the Cypher query. Return it in the following format: ```cypher YOUR CYPHER QUERY HERE ``` ``` | Variable | Description | |----------|-------------| | `base_system_prompt` | System message from Neo4jGPTQuery containing schema | | `schema_context` | Analyzed or raw schema context | | `question` | User's natural language question | | `hint` | Optional hint about the query |

Fig. 6. Prompt for generating Cypher Query based on schema_context, question, hint

Appendix C

The validation phase implements a multi-agent feedback loop, recognizing that single-pass generation often produces queries with subtle errors that may execute but return incorrect results. Generated queries are submitted to a specialized review agent that performs validation across four dimensions: OLAP-specific validation verifying that referenced measures and dimensions exist in the database, syntactic compliance examining Cypher patterns and relationship directions, common error detection identifying frequent mistakes such as incorrect quote usage, and logical correctness assessing whether the query semantically addresses the user's original question. The review agent receives real-time metadata from the Neo4j database, enabling it to detect hallucinated analytical entities that would otherwise pass syntactic validation but produce empty or incorrect results.

	Prompt for generating feedback of generated Cypher query
1.	You are an expert Cypher query reviewer for Neo4j. Analyze the generated Cypher query and provide constructive feedback. ## Review Focus Areas: ### 1. OLAP-Specific Validation - Are the referenced measures valid? (Check against schema) - Are the referenced dimensions valid? (Check against schema) - Does the query make sense for analytical purposes? - Are aggregations appropriate for the measures being queried? ### 2. Cypher Syntax & Schema Compliance - Cypher syntax correctness (MATCH, WHERE, RETURN patterns) - Relationship direction and types - ensure they match the schema - Node label usage - verify labels exist in schema - Property names - check they exist on the specified nodes/relationships ### 3. Common Issues - Using double quotes instead of single quotes for string literals - Incorrect relationship patterns - Missing RETURN clause - Wrong relationship directions for OLAP cube structure ### 4. Logical Correctness - Does the query answer the original question? - Are the joins/relationships logical for OLAP analytics? - Would this query produce meaningful business insights? Be specific and actionable. If the Cypher looks correct, say so briefly. Please review this Cypher query: Question: {question} Hint: {hint} Schema context: **Note:** measures_list and dimensions_list are retrieved from Neo4j database..{schema_context} {measures_list} {dimensions_list} Generated Cypher: ```cypher {generated_cypher} ``` Check if the measures and dimensions used in the query exist in the lists above. If they don't exist, this is a critical error that must be reported. Provide feedback on potential issues or improvements.

Fig. 7. Prompt for generating feedback of generated Cypher query

Appendix D

When the review agent identifies issues, a correction agent receives both the original query and the expert feedback. This separation of review and correction responsibilities follows the principle that critique and generation are distinct cognitive tasks, and specialized agents outperform single agents attempting both functions. Additionally, a post-execution analysis agent examines the actual results returned by Neo4j to determine whether they properly answer the user's question. If the analysis indicates a mismatch between results and user intent, the pipeline re-enters the correction phase with this additional feedback.

	Prompt for correction agent to update Cypher query
1.	You are an expert at analyzing database query results to determine if they answer the user's question. Your task is to: 1. Understand what the user is asking for 2. Examine the Cypher query that was executed 3. Review the actual results returned

	4. Determine if the results answer the user's question appropriately 5. Result should be from UseCase and UserRequest entities as these are the primary analytical entities in the database 6. Before suggest cypher query, you should check schema context and make sure the query is valid and makes sense for OLAP analytics Provide your analysis in this format: - **Matches Intent**: Yes/No - **Analysis**: Brief explanation of whether results answer the question - **Suggestions**: If the results don't match intent, suggest what should be changed (leave empty if results are good) Please analyze if these query results answer the user's question. ## User Question: {question} ## Schema Context: {schema_context} ## Cypher Query Executed: ```cypher {cypher} ``` ## Query Results: {result_data} ## Analysis Task: Does this query and its results properly answer the user's question? Consider: - Are the right columns/properties being returned? - Is the data relevant to what was asked? - Are the results complete or is something missing? - Would a user be asking this question be satisfied with these results? - Does the query align with the schema structure? Provide your analysis: ``` ## Response Parsing The response is parsed to determine if intent matches by looking for phrases like: - "matches intent: yes"

Fig. 8. Prompt for correction agent to update Cypher query

UDC 004.89:519.872 DOi 10.30837/ bi.2025.2(103).10

V. Y. Moskalenko[1], M. A. Grinchenko[2]

[1] NTU «KhPI», Kharkiv, Ukraine, vladimir.moskalenko@outlook.com,
ORCID iD: 0009-0001-2759-3550

[2] NTU «KhPI», Kharkiv, Ukraine, marinagrunchenko@gmail.com,
ORCID iD: 0000-0002-8383-2675

MODELING THE PROCESS OF FORMING A KPI SYSTEM FOR EVALUATING A PRODUCT STRATEGY BASED ON A FUZZY COGNITIVE MAP

The business process of strategic analysis of product areas of an IT company and IT product evaluation is considered. An approach to forming a KPI system for evaluating product strategy is proposed, based on constructing a fuzzy cognitive map. Cognitive modelling was conducted to assess the impact of KPIs on aspects of product strategy evaluation: financial performance, customer satisfaction level, and sales/marketing effectiveness. As a result, a system of main KPIs of the product was formed, and KPI weighting factors were obtained for calculating the aggregate financial indicator, the average customer satisfaction indicator, and the average sales and marketing efficiency indicator. An analysis of various possible scenarios of the impact of changes in KPI values on aspects of product strategy implementation has been conducted. The product manager assesses the future consequences of implementing the product strategy based on the results of modelling KPI changes.

IT PRODUCT, PRODUCT STRATEGY, KPI, BUSINESS PROCESS, COGNITIVE MAP, MODELING, FUZZY LOGIC, NEURAL NETWORK, AGGREGATE INDICATOR, SCENARIO

В. Ю. Москаленко, М. А. Гринченко. Моделювання процесу формування системи KPI для оцінки продуктової стратегії на основі нечіткої когнітивної карти. Розглянуто бізнес-процес стратегічного аналізу продуктових напрямків ІТ-компанії та оцінки продукту. Запропоновано підхід до формування системи KPIs для оцінки продуктової стратегії на основі побудови нечіткої когнітивної карти. Проведено когнітивне моделювання для оцінювання впливу KPIs на аспекти оцінювання продуктової стратегії: фінансовий результат, рівень задоволеності клієнтів та ефективність продажів/маркетингу. У результаті сформовано систему основних KPIs продукту, отримані вагові коефіцієнти KPIs для розрахунку агрегованого фінансового показника, усередненого показника задоволеності клієнтів та усередненого показника ефективності продажів і маркетингової ефективності. Проведено аналіз різних можливих сценаріїв впливу змін значень KPIs на аспекти реалізації продуктової стратегі. Продакт-менеджер оцінює майбутні наслідки реалізації продуктової стратегії на основі результатів моделювання змін KPIs.

ІТ-ПРОДУКТ, ПРОДУКТОВА СТРАТЕГІЯ, КЛЮЧОВИЙ ПОКАЗНИК ЕФЕКТИВНОСТІ, БІЗНЕС-ПРОЦЕС, КОГНІТИВНА КАРТА, МОДЕЛЮВАННЯ, НЕЧІТКА ЛОГІКА, НЕЙРОННА МЕРЕЖА, АГРЕГОВАНИЙ ПОКАЗНИК, СЦЕНАРІЙ

Introduction

The IT industry is highly competitive. Competition exists at various levels, from the global IT market for core technologies to specific market niches for IT products and services. Increasing the competitiveness of a software product throughout its life cycle is a strategic problem for an IT company to maintain its competitive status [1]. This problem is addressed within the framework of implementing the corporate strategy. Corporate strategy is formulated at the senior management level. It defines the main framework of the company's business strategy. For its implementation, decomposition is carried out to identify strategies for strategic business units. As a business unit for a product IT company, a software product, an IT service, and a product direction are considered. An appropriate business strategy is developed for each business unit. All product business strategies are components of corporate strategy [2].

Therefore, the current strategic tasks for an IT company are the selection of promising (competitive) areas and the effective management of IT products/services. Product management is carried out by a manager throughout the life cycle of an IT product, from idea generation and development to product launch, growth, maturity, and eventual withdrawal from the market. Hence, product improvement to meet customer needs and achieve strategic business goals is constantly carried out.

The product manager conducts product examination, the results of which are necessary for making strategic decisions throughout the product life cycle [3]. Tracking product metrics enables the product manager to make data-driven decisions at every stage. Linking each product decision to business outcomes, as well as justifying the investment in product launch and market support (determining ROI), are essential aspects of product management.

However, there is a problem in determining essential indicators, the analysis of which will really make it possible to develop measures to increase product competitiveness.

Metrics and key performance indicators (KPIs) for product examination are selected. These are quantifiable metrics that allow an IT company to determine and track the success of a product or business activity. Metrics typically assess the performance of specific business aspects of a product by providing detailed operational data. At the same time, KPIs help evaluate the performance of a software

product in relation to the company's strategic business goals. KPIs help to create a roadmap for each product and plan the business strategy of an IT company. Such metrics enable stakeholders to assess how users interact with the product, identify areas for improvement, compare performance, and make informed decisions about subsequent actions throughout the product lifecycle [4]. However, tracking too many metrics can create "noise" in product performance analytics. For example, a large amount of contradictory data or poorly interpreted data may be obtained. At the same time, tracking too few metrics can leave so-called "blind spots". That is, specific aspects of product implementation, such as changes in product user preferences, the emergence of new needs, or changes in the value of individual product functions, will not be considered in the analysis of such indicators.

The current challenge is to create a relevant list of KPIs, the examination of which will provide an information basis for informed management decisions on improving the product to maintain its competitiveness and achieve the company's strategic objectives.

1. The problem of forming a KPI system for IT product management

IT product management practitioners and theorists are paying increasing attention to the issues of the process of analyzing the market success of IT products and services. The International Institute of Business Analysis (IIBA) has proposed an extension to the IT community to A Guide to the Business Analysis Body of Knowledge (BABOK Guide) - Guide to Product Ownership Analysis [5]. It is intended for professionals who act as product owners, product managers, and other professionals who are involved in the process of developing an IT product (service) and managing it. Product ownership analysis (POA) is applied at each of the three planning levels (strategy, initiative, and delivery) to continuously align the value of the product created with customer expectations and the IT company's goals [6]. It is necessary to conduct a market analysis, including an assessment of the product's competitiveness and the effectiveness of its implementation strategy, among other factors, for the first level of implementation: developing a product strategy. All this requires the formation of a KPI system for IT product management. In addition, the following advantages of tracking product indicators can be highlighted [7]:

– decision-making based on relevant data instead of the "intuition" of a specialist;

– Identifying inefficiencies in product development;

– measuring user satisfaction and overall customer experience;

– optimization of resource allocation between product development/improvement projects based on real data on the effectiveness of product strategies;

– checking the functionality of the product in terms of the requirements of users and existing competitors before entering the market;

– identifying product quality issues early before customers detect them;

– clear demonstration of the product's value to stakeholders (including users) and IT company management;

– reducing the risk of misunderstanding potential customer needs by measuring their engagement and retention levels.

Modern studies consider the issue of classifying indicators that characterize the effectiveness of product sales, its competitiveness, and other related factors. Since IT product KPIs are used to assess its success and align decisions to improve it with overall business goals, these indicators are usually divided into the following categories: business performance, customer engagement, and product development [7]. For example, tracking product management metrics helps the product development project team understand user behaviour, collect customer feedback, compare them to competitors, and continuously improve product performance. Increasing customer loyalty leads to higher revenue and increased market share.

However, there are challenges in measuring product management KPIs [8]:

– data inconsistency, i.e. the same indicator can be calculated in different ways, for example, a monthly active user can be considered someone who uses the product once a month, or only a user who performs specific actions (recommends others to use this product, etc.) or becomes a paying customer; this can lead to data distortion;

– excessive focus on numbers without understanding the real situation, for example, a drop in user engagement requires in-depth investigation; it could be a simple seasonality of product use or a decline in customer loyalty;

– data fragmentation, meaning that important information is spread across different tools and teams, making it difficult to get a complete picture of product performance; for example, a marketing department may track user engagement in one system, while a product manager may track it in other systems; data collection periods must also be considered.

Therefore, studies on measuring product management KPIs reveal that special attention is required for the formation of a KPI system, based on which product strategies can be evaluated and developed.

There are several recommendations for selecting KPIs for conducting an IT product examination and categorising them based on the product's field of use and type. Let's consider a few of these recommendations.

Atlassian, as a software company for teamwork and project management, recommends such a division of KPIs [9]:

– business performance indicators that focus on financial results;

– customer and user acquisition performance metrics that focus on satisfaction and loyalty;

– indicators that evaluate the effectiveness of product development.

LogMeIn [10], a company that develops solutions for remote access, collaboration, and support, provides the following classification of popular metrics:

– support services (average time to solve the problem, first call resolution rate, customer satisfaction);

– Network performance (network uptime, data latency, percentage of available network);

– system uptime and reliability (server uptime, number of incidents, average time between failures);

– safety and compliance (number of security incidents; patch management compliance, compliance assessment);

– backup and restore;

– the cost of the technologies used;

– employee productivity;

– project implementation.

These indicators are targeted at the IT company that produces products for device management. Therefore, indicators reflecting the effectiveness of services provided to customers through the corresponding product are added to the overall KPIs [10].

Specialised product analytics platforms, all-in-one project management tools, and specialized business intelligence software can be utilized to track product management metrics effectively. For example, tools such as ProductCentral [7], Sharpist with Miro, Mixpanel, and Google Analytics [11]. These tools help consolidate data, visualize performance dynamics, and provide actionable insights into user behaviour, product performance, and business outcomes. However, they enable product managers to calculate KPIs for specific product categories. For example, Google Analytics provides website and app performance tracking, allowing users to measure traffic, conversions, and customer engagement. Most of these tools offer a wide range of visualization of KPI calculation results.

However, for practical work, a product manager needs a tool that would have the functions of forming and tracking a system of indicators for a specific period, the tasks of generating data to evaluate the product strategy depending on the type of product and the company's goals, and the functions of indicator analytics for making decisions on improving the product strategy. As a result, this tool should support the strategic analysis of the product strategy and inform the decision-making process to improve it. To develop such a software tool, it is necessary to enhance the process of strategic analysis for product strategy.

2. Problem statement

A business process model for evaluating and analyzing product strategies of an IT company and forming recommendations for their improvement, taking into account the priority areas of development of an IT company (Fig. 1), was proposed in previous studies by the authors of this article [1, 12, 13].

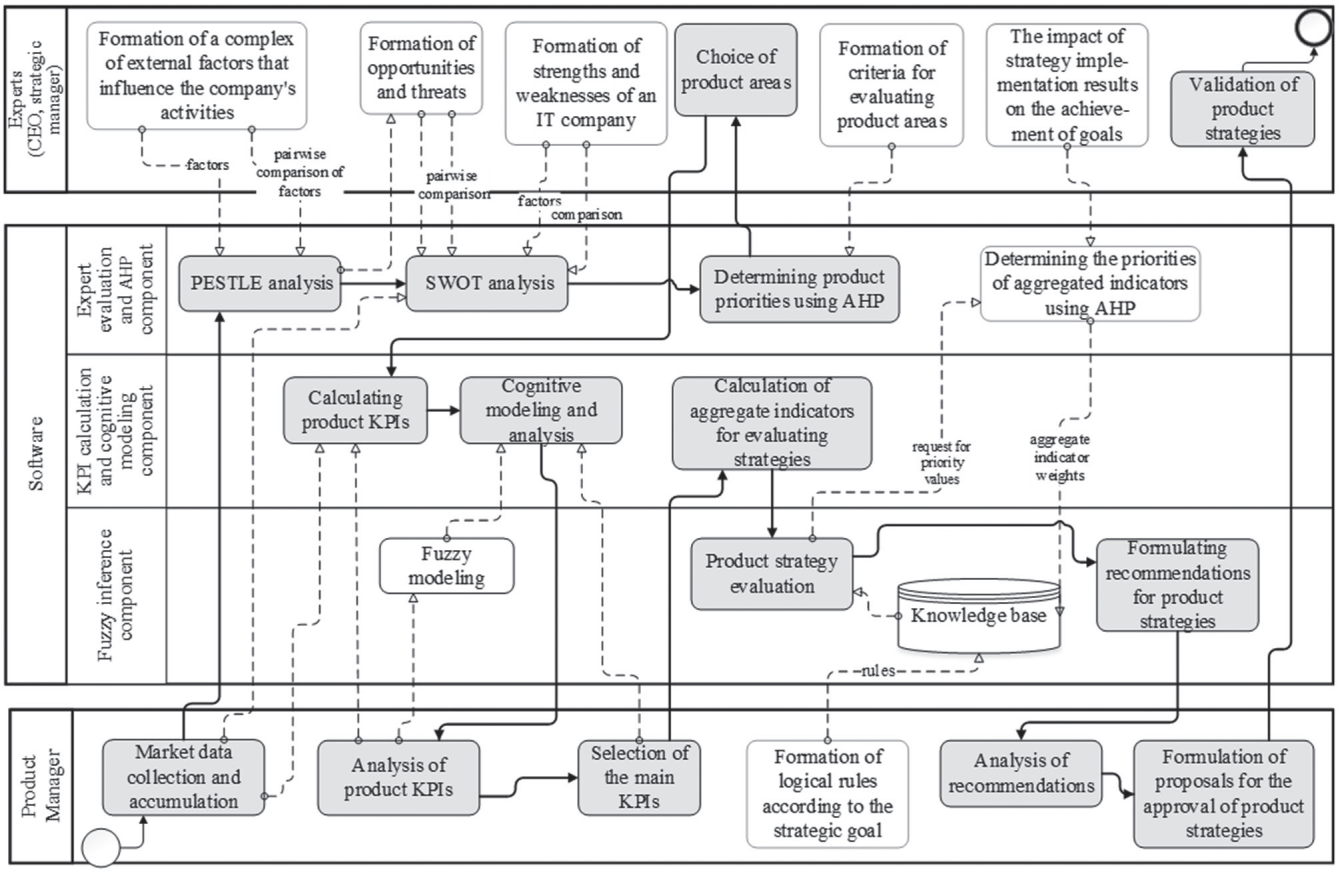

Fig. 1. Business Process Model for Strategic Analysis of IT Company Product Lines and Product Strategies

At the initial stages of this business process, a set of factors influencing the success of an IT company is identified, and the main criteria, which are then used to select promising product areas of an IT company, are determined through PESTLE analysis and SWOT analysis.

The next stage of the business process is the assessment of the company's product areas (product lines). Evaluation of product areas is carried out based on the defined criteria using the structured decision-making technique, the Analytic Hierarchy Process (AHP) [14]. Product areas assessment is carried out in relation to the strategic goal of the IT company. For example, increasing the competitiveness of an IT company can be considered a strategic goal. The level of market success for product areas and their associated risks is determined using the AHP technique, which involves pairwise comparison of hierarchy elements [15]. The results of the AHP implementation are submitted to the company's management. The management of the IT company, along with experts, selects promising product areas (or product lines).

Next, product managers conduct a strategic analysis of products within the selected product lines. The manager forms an appropriate KPI system for the product. The numerical values of these KPIs characterize the level of success of the product in the IT market by various aspects (finance, customers, marketing, competitors, etc.) [1].

In other words, this KPI system is designed to evaluate the product strategy and determine whether to continue its implementation in the planning period or to refine it based on the assessments received.

Therefore, the purpose of this study is to enhance the process of conducting a strategic analysis of product strategies by establishing a KPI system to evaluate the market success of an IT product. The product manager develops proposals for the product strategy based on the analysis of these indicators. The product strategy should include a plan to improve the KPI in line with the company's strategic goals.

3. Proposed approach to forming a KPI system for evaluating an IT product

This study examines a company that uses two business models:

1) IT products are sold on a subscription basis, providing customers with ongoing access to products (including updated versions) for regular payments;

2) selling licenses on various terms. It is necessary to define a KPI system for each product, since a different set of indicators describes the implementation of each business model.

Three groups of IT product strategy indicators were proposed to be considered according to the analysis of existing KPI classifications and the peculiarities of the Ukrainian IT market:

– indicators of the financial success of the product;
– customer satisfaction indicators;
– sales and marketing performance indicators.

According to the Strategic Analysis Business Process (Fig. 1), the values of the relevant KPIs are the basis for calculating the Aggregate Financial Indicator (AFI), the Average Customer Satisfaction Index (ACSI), and the Average Sales/Marketing Performance Indicator (ASPI). These aggregated metrics are used to evaluate a product strategy using a fuzzy logic tool, specifically the fuzzy inference subsystem (FIS) trees [16, 17].

Therefore, the product manager should select the primary KPIs that best align with the product lifecycle stage and business objectives. When analyzing a product in its early stages of the life cycle, user engagement rates may be a key priority. Then, when analyzing a product at mature stages, focus on revenue growth and user retention [4].

As a result of the study, a list of the main KPIs was formed. Their conditional division into three groups has been carried out.

1. Indicators of the product's financial success (FI).

FI1. Revenue Indicator [4]. Average revenue per user (ARPU) applies to various digital products, including e-commerce websites, online games, and travel apps. ARPU is the amount of revenue that a company receives from one user for a certain period (for example, a month, quarter, or year):

APRU ($) = Total revenue / Total number of users.

Suppose you are analyzing a strategy for subscription-based products. In that case, it is recommended to use annual recurring revenue (ARR) and monthly recurring revenue (MRR) instead of ARPU, depending on the analysis period:

MRR ($) = APRU x Number of accounts in a month = sum of current monthly subscriptions + revenue from new subscriptions + upgrades - downgrades - revenue from lost customers.

FI2. Customer Lifetime Value (CLTV) assesses the financial benefits of attracting and retaining each customer, indicating the total revenue that the business generates from a single customer over the entire duration of their relationship. You need to multiply the average lifespan of the customer (how long the user usually uses the product) by ARPU to calculate the CLTV:

CLTV($) = Average customer lifetime x ARPU.

The CLTV is also used to assess a customer's loyalty to a product or IT company.

FI3. Customer Acquisition Cost (CAC) is the cost of persuading a potential consumer to buy a product or service:

CAC ($)= Total sales and marketing spending / Number of new customers.

CLTV and CAC indicators enable you to determine whether customers generate less revenue for the company than they spend on them. Analyzing these metrics influences

decisions to revise the pricing strategy and marketing of the product to attract more users. It is also worth determining the CAC for different sales channels to compare their effectiveness.

Therefore, the CLTV and CAC metrics can also be considered in the analysis of the effectiveness of marketing activities and the sales team.

Other financial indicators can be added to the first group; these include gross/net margin, as well as return on investment in product development and market promotion.

2. Customer satisfaction indicators (SI). These indicators characterize the level of customer loyalty of an IT company. Information about how users interact with the product, their overall level of satisfaction with it, and other relevant metrics is obtained through the analysis of these metrics.

SI1. Customer retention rate (CRR) is the percentage of customers who continue to use the product over a period. CRR shows the user success of a product because the product has value to loyal customers:

Retention rate (%) = ((Customers at the end of the calculated period - New customers) / Customers at the start of the computed period) x 100.

SI2. Customer churn rate (CCR) determines the percentage of customers that a company loses over a specified period. It is essential for subscription companies:

Customer churn rate (%) = (Customers lost / Total customers) x 100.

You can use the Revenue Churn Rate:

Revenue churn rate (%) = (Revenue from lost customers / Total revenue) x 100.

SI3. The Net Promoter Score (NPS) measures customer loyalty and satisfaction based on customer survey results, and how likely they are to recommend a product to others:

NPS = Percentage of promoters - Percentage of detractors.

You need to get product ratings from customers to calculate the NPS. For example, the assessment should be conducted on a scale of 0 to 10. Disloyal customers will receive 0 to 6 points, neutral customers – 7 or 8 points, and loyal customers (promoters) – 9 or 10 points.

SI4. Overall Satisfaction Score (OSAT) shows the overall level of customer satisfaction with the product:

OSAT (%) = (Number of satisfied responses / Total number of responses) x 100.

Different scales can be used to measure it, for example, a scale from 0 to 10, where dissatisfaction ranges from 0 to 4, and satisfaction ranges from 5 to 10.

If the product has already been on the market for some time and you need to evaluate improvements to it, such as adding additional features, then the requests will only apply to this part of the product.

SI5. Earned growth rate (EGR) indicates the increase in revenue resulting from regular customers and their referrals over a specified period (month, quarter, or year). It characterizes the impact of customer loyalty and customer support on the growth of the company's income from the sale of an IT product:

EGR (%) = NRR + ENC,

where *NRR* – Net Revenue Retention,

NRR (%) = ((MRR at the beginning + Expansions + Upsells - Churn - Contractions) / MRR at the beginning) x 100;

MRR ($) – Monthly Recurring Revenue;

ENC – Earned New Customers is income from customers purchased through recommendations and word of mouth:

ENC (%) = (New customer revenue earned through referrals / Total new customer revenue) x 100.

The EGR measure assesses whether customer referrals have been converted into revenue, thereby quantifying the impact of customer loyalty on financial performance. This indicator can be attributed to the first group, since it evaluates the financial result.

The following metrics can be added for a more thorough analysis of the product's customer/user satisfaction:

Time to Value (TTV) shows how quickly a new user realizes the value of a product or new feature;

The Error Correction Factor is calculated as follows: the number of corrected errors during the specified period divided by all identified errors; it shows the effectiveness of the development team in maintaining the product quality.

3. Presales & Marketing Team Performance Metrics (PI).

PI1. Win Rate (WR) shows the percentage of leads that the sales team was able to convert into paying customers:

WR (%) = (Number of closed deals / Total opportunities) × 100 %.

PI2. Average Deal Value (ADV) is a key metric for sales teams. ADV shows the average estimate of revenue from each deal:

Average deal Value ($)= Total value generated / Total number of closed-won deals.

Although the average revenue per contract is high, the agreement signing process incurs high costs, which can negatively impact the contract margin. ADV also affects the financial outcome of implementing the product strategy.

PI3. Churn Rate (CR) is a metric that determines the percentage of customers that a company loses over a period:

CR = Number of lost Customers / (Total Number of Customers at the beginning of the period) x 100.

PI4. The Inbound Qualified Lead Velocity (IQLV) measures the monthly increase in the number of new leads who show genuine interest in a product or service compared to the previous month. It is used to understand the possibilities of sales directions, i.e. tracking the flow of

leads. An IT company can predict which leads will bring the company the most revenue and are most likely to become paying customers, based on:

IQLV(%) = ((Current Month's Lead Count - Last Month's Lead Count) / Last Month's Lead Count) x 100.

Knowing about potential customers provides greater certainty about future sales. The higher the speed and volume of this sales funnel, the easier it is for the management team to make decisions about planning the direction and volume of sales for IT products/services.

PI5. Contract Profitability (CP). Keeping track CP helps the sales team manage and improve the efficiency of their operations.

The analysis of the developed KPI system is conducted in accordance with the study's objectives. It is concluded that assigning a specific KPI to one group or another can cause controversy. Therefore, it is proposed to analyze the impact of each indicator on the financial result, on the level of customer orientation of the product (customer satisfaction) and on the effectiveness of sales/marketing. It is necessary to determine the KPIs that will be used to calculate the aggregate indicators of AFI, ACSI and ASPI, as well as their weighting factors. It is proposed to use fuzzy cognitive modelling.

4. Constructing a fuzzy cognitive map

A fuzzy cognitive map (FCM) was constructed as part of the study. A map was created to determine the level of influence of each indicator on the aspects of evaluating the product strategy: financial results, customer satisfaction, and sales/marketing efficiency.

FCM is a tool for presenting knowledge about a system characterized by uncertainty, causality and complex processes. FCM is a peculiar combination of fuzzy logic, cognitive mapping, and neural networks. Let's consider two approaches to describing FCM.

1. A fuzzy cognitive map is a fuzzy oriented graph of the first kind and is described as follows [18]:

$$\widetilde{G} = (A, \widetilde{U}), \quad (1)$$

where $A=\{a_i\}$, $i \in I=\{1,2,...M\}$ – a distinct set of graph vertices (in this problem, the graph vertices are the KPIs); $\widetilde{U} = \{<\mu_U(a_i, a_j) / (a_i, a_j)>\}$ – fuzzy set of edges of a graph, $(a_i, a_j) \in X^2$; $\mu_U(a_i, a_j)$ – the degree of belonging of an oriented edge (a_i, a_j) to the fuzzy set of oriented edges $\widetilde{U}$; edge (a_i, a_j) exists if changing the parameter a_i has a direct effect on changing the value of the parameter a_j.

The process of propagation of a perturbation along the graph $\widetilde{G}$ with known initial values at all vertices $\{a_i^0\}$ and the initial perturbation vector $\{P_j(0)\}$ is determined by the formula:

$$a_v(t+1) = a_v(t) + \sum_{j=1}^{k-1} f_{ij} P_j(t) + Q_i(t), \quad (2)$$

where $a_i(t)$ and $a_i(t+1)$ are the values of the indicator at the vertex a_i respectively, at the *t-th* moment (simulation cycle) and the $t+1$-th moment; $P_j(t)$ – change in vertex j at the t-th moment; f_{ij} – function to convert links between KPIs; $Q_i(t)$ – vector of disturbances (changes in parameters).

Introducing perturbations simulates a scenario that answers the question of scientific prediction: "What will happen to the system at time $t+1$ if ...?".

2. In the modern scientific literature, FCM is implemented as a recurrent neural network with interpretive features [19]. Recurrent NN is a set of neurons and causal relationships between them. The activation value of such neurons takes the value in the interval [0, 1]. The stronger the activation value of a neuron, the greater its impact on the network. The strength of the causal connection between two neurons A_i and A_j is quantified by the numerical weight $w_{ij} \in [-1, 1]$.

In this paper, we consider neurons A_i and A_j, which are KPIs.

There are three types of causal relationships between neurons in FCM:

if $w_{ij} > 0$, then there is a positive relationship, an increase (or decrease) in A_i generates an increase (decrease) in A_j with intensity $|w_{ij}|$;

if $w_{ij} < 0$, then there is a negative relationship between the indicators, so an increase (decrease) in A_i leads to a decrease (increase) in A_j with intensity $|w_{ij}|$;

if $w_{ij} = 0$, then there is no causal relationship between the parameters.

There is a Cosco activation rule for each neuron at $t+1$ iteration of NN training [20]:

$$A_i^{t+1} = f\left(\sum_{\substack{j=1 \\ i \neq j}}^{M} w_{ij} \cdot A_i^t\right), \quad (3)$$

where A_i^{t+1} – state of the i-th neuron at the $t+1$-th training iteration; A_i^0 is the initial state of the i-th neuron; $f(\cdot)$ is a monotonically non-decreasing activation function, in this work, a sigmoid function is used:

$$S(x) = \frac{1}{1+e^{-\lambda x}}, \quad (4)$$

and hyperbolic tangent

$$tanh(x) = \frac{e^{\lambda x} - e^{-\lambda x}}{e^{\lambda x} + e^{-\lambda x}}, \quad (5)$$

where the λ parameter is used to control the slope of the activation function, taking a value greater than 0, and its value is closely related to the convergence behaviour of the FCM [21]; x is the current value of the neuron.

The primary difference between their use in FCM lies in the range of node activation values, which results in different interpretations of indicator states.

The sigmoid function produces a value between 0 and 1; a value of 0 means no or minimal activation, while a value of 1 indicates the maximum possible activation.

Scenarios using the sigmoid form are well-suited for situations where all concepts exist on a unipolar scale (e.g., presence/absence, low/high) and negative values do not make sense.

The hyperbolic tangent function produces values from -1 to 1. A value of 0 represents the neutral state, 1 represents the maximum activation, and (-1) means the maximum negative presence or suppression of the activation level.

Scenarios using the hyperbolic tangent function are ideal for systems where components may exhibit contrasting or bipolar states (e.g., positive vs. negative change, agreement vs. dissent), making it the preferred method for handling negative influences and fluctuations in system dynamics, as it displays negative inputs as strongly negative.

So, for each scenario, the appropriate activation function will be selected. The activation rule repeats iteratively until the stop condition is met. The new activation vector is computed on each iteration, and after a fixed number of iterations, the FCM will be in one of the following states:

1) equilibrium point;
2) limited cycle;
3) chaotic behaviour [20].

5. Experimental studies

A fuzzy cognitive model was constructed using the Mental Modeler application (Fig. 2) to analyze the KPI relationships. Mental Modeler enables you to intuitively create cognitive maps based on fuzzy logic, utilizing a user-friendly interface [22].

The cognitive map (Fig. 2) is constructed as follows: the vertices represent KPIs, which are divided into three groups: indicators of the product's financial success, indicators of customer satisfaction (including customer loyalty), and sales and marketing performance indicators, as discussed above. We also added vertices that represent the generalized financial result of the product strategy implementation (FI), the level of customer satisfaction (SI) and the level of sales/marketing efficiency (PI).

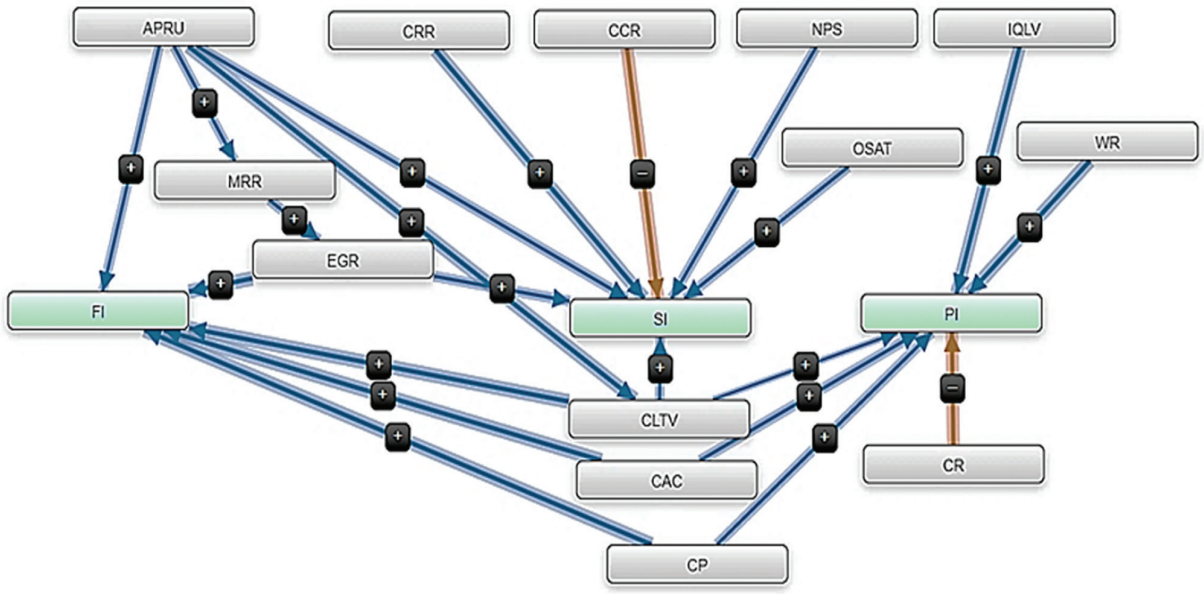

Fig. 2. Cognitive map

The presence of an edge between the indicators means the influence of one indicator on another, and the scales determine the degree of influence.

After constructing the model, the relationships between KPIs were analyzed, and the values were revised $\{w_{ij}\}$, as well as changes in the degrees of their belonging. Expert assessments of specialists involved in product management within the company's IT and financial analytics teams were used to analyze the connections.

An adjacency matrix was constructed, with each element representing an edge weighting factor (w_{ij}), which relates the force of the influence of indicator Ai on Aj.

The strength of the influence of each KPI on FI, SI and PI was analyzed. Figure 3 shows a fragment of the adjacency matrix.

	APRU	FI	PI	SI	CLTV	EGR	MRR
APRU		0.8		0.5	1		0.75
FI							
PI				0			
SI			0				
CLTV		1	1	0.7			
CAC		1	0.5				
CRR				1			
CCR				-1			
NPS				0.8			
OSAT				0.8			
EGR		1		0.8			
WR			1				
CR			-1				
IQLV			1				
CP		1	0.8				
MRR						0.7	
ADV		0.5	0.7				

Fig. 3 Fragment of the adjacency matrix

This numerical representation of relationships enables experts to analyze the level of impact and interdependence of KPIs. For example, the set of indicators that affect financial results was changed as a result of an expert analysis of the previous version of the map. The EGR score, which was initially classified as a metric that only affects the level of SI customer loyalty (with a weight of 1), has been added. But the decision that the EGR indicator directly affects the financial results of FI was made by experts after a thorough analysis of the map. Therefore, the weights were redefined as follows: EGR affects the FI result as much as possible ($wij = 1$); it also affects the SI result, but less ($wij = 0.8$). Therefore, the degrees of influence of the indicators were redefined as a result of analyzing the cognitive map. Fig.2 shows the latest version of FCM.

The KPI weights for calculating the corresponding aggregate indicators (AFI, ACSI, ASPI) are determined through adjacency matrix analysis.

Let's consider an example of calculating weighting factors for calculating the aggregate indicator API. The indicators CLTV, CAC, ERG, and CP (each with a weight of 1) have a direct effect on FI. The APRU indicator affects FI with a weight of 0.8, and the ADV indicator affects FI with a weight of 0.5. Then, you can calculate the weighting factors as follows:

$$\alpha_{CP,FI} = \frac{w_{CP,FI}}{w_{CP,FI} + w_{CLTV,FI} + w_{CAC,FI} + w_{ERG,FI} + w_{APRU,FI} + w_{ADV,FI}},$$

where $w_{CP,FI}, w_{CLTV,FI}, w_{CAC,FI}, w_{ERG,FI}, w_{APRU,FI}, w_{ADV,FI}$ are the weights that are obtained from the adjacency matrix, in this case they are equal to 1, respectively; 1; 1; 1; 0,8; 0,5.

In this example, the weighting factors will be equal to the following values:

$$\alpha_{CP,FI} = \alpha_{CLTV,FI} = w_{CAC,FI} = w_{ERG,FI} = 0{,}19,$$

$$\alpha_{APRU,FI} = 0{,}15, \alpha_{ADV,FI} = 0{,}09.$$

Then the value of the aggregate financial indicator will be determined as follows:

$$AFI = w_{CP,FI} \cdot \hat{CP} + w_{CLTV,FI} \cdot \hat{CLTV} + w_{CAC,FI} \cdot \hat{CAC} +$$

$$+ w_{ERG,FI} \cdot \hat{ERG} + w_{APRU,FI} \cdot \hat{APRU} + w_{ADV,FI} \cdot \hat{ADV},$$

where $\hat{CP}, \hat{CLTV}, \hat{CAC}, \hat{ERG}, \hat{APRU}, \hat{ADV}$ are normalized values of CP, CLTV, CAC, ERG, APRU, ADV indicators, respectively.

The relative importance of indicators for assessing the implementation of a product strategy was analyzed using the tool Preferred State&Metrics (Fig. 4).

The Centrality column (Fig. 4) shows the conceptual weight/importance of the vertices of the FCM graph. The vertices of the FCM graph represent KPIs and generalized results of implementing the product strategy FI, SI, and PI. In this example, SI is more important (Centrality=5.71). This is because the outcome is determined by a larger number of key performance indicators (KPIs) that measure customer satisfaction than those that measure financial potential (FI) and profitability of sales/marketing (PI).

Preferred State & Metrics

Component	Indegree	Outdegree	Centrality	Preferred State	Type
SI	5.71	0	5.71	Increase	receiver
FI	4.8	0	4.8		receiver
PI	4.75	0	4.75		receiver
APRU	0	3.28	3.28	Increase	driver
CLTV	0.97	2.15	3.12	Increase	ordinary
EGR	0.7	1.8	2.5	Increase	ordinary
CP	0	1.8	1.8	Increase	driver
MRR	0.9	0.7	1.6		ordinary
CAC	0	1.5	1.5	Increase	driver
IQLV	0	1	1	Increase	driver
CR	0	1	1	Decrease	driver
WR	0	1	1	Increase	driver
CCR	0	1	1	Decrease	driver
CRR	0	1	1	Increase	driver
OSAT	0	0.8	0.8	Increase	driver
NPS	0	0.8	0.8	Increase	driver

Fig. 4 Tool Preferred State&Metrics

Simulations of various scenarios have been carried out, specifically examining how the values of FI, SI, and PI change in response to changes in performance indicators. As an example, let's consider two such scenarios.

Scenario 1: Average revenue per user (ARPU) increased a lot. Fig. 5 shows that significant positive changes in values will not substantially affect the results of implementing the product strategy in three aspects of FI, SI, and PI. This is because the weights set by experts for this indicator are less than 1 for FI and SI.

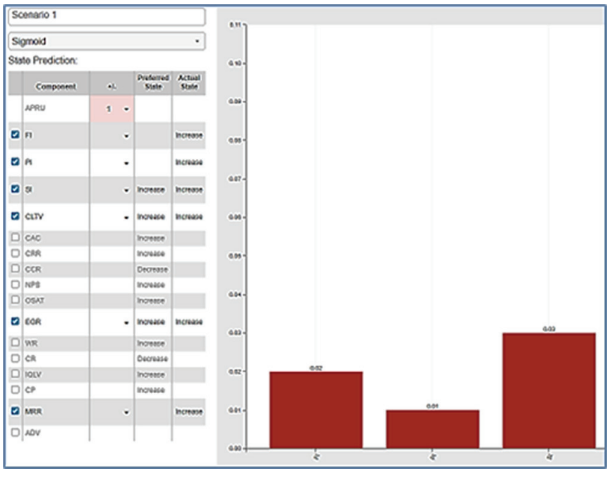

Fig. 5 Results of scenario 1

Scenario 2: Customer Acquisition Cost (CAC) decreased due to an increase in the cost of working with a potential consumer by 40%, Net Promoter Score (NPS) increased by 60% due to the rise in the value of the product for the customer, and the Customer Retention Rate (CRR) also increased by 20% (Fig. 6).

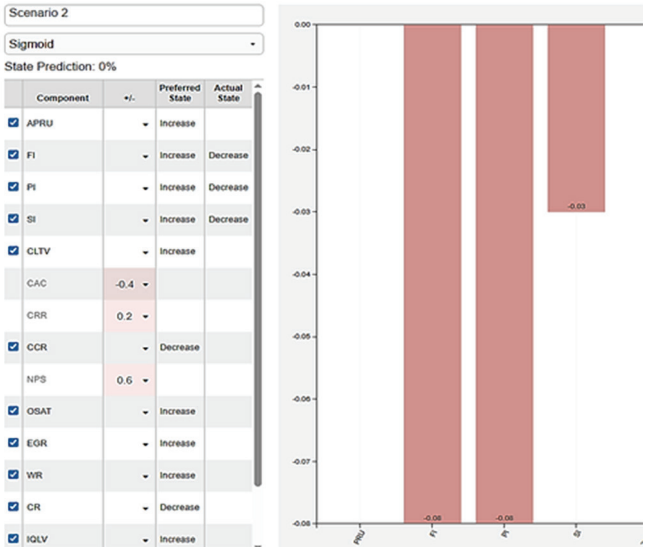

Fig. 6 Results of scenario 2

As a result, we have seen a deterioration in the implementation strategy's results in all aspects. Therefore, this is a reason to revise the product strategy and focus on the indicators that characterize financial results and sales efficiency. Most likely, the funds spent on marketing and stimulating the presale did not yield the expected financial outcome.

Thus, making changes to the cognitive map models presents a scenario that addresses the question of scientific prediction: "What will happen to the aspects of the product strategy at time t+1 if some indicators are changed?"

Using a cognitive map that reflects the cause-and-effect relationships of the influence of KPIs on the results of the product strategy implementation by aspects (FI, SI and PI), the product manager obtains weight coefficients for measuring aggregated indicators of strategy implementation.

Since the value of KPIs is a numerical measure of the market reaction to the implementation of the product strategy, modelling their changes allows the product manager to assess the future consequences of deterioration or improvement in the effectiveness of the product strategy.

Conclusion

The study considers the business process of strategic analysis of IT company directions and product strategies.

An approach to forming a system of KPIs for evaluating a product strategy is proposed, based on the construction and analysis of a fuzzy cognitive map.

As a result of the cognitive modelling, a system of KPIs was formed, and the weighting coefficients of KPIs were obtained to calculate:

– the aggregate financial indicator;
– the average indicator of customer satisfaction;
– the average indicator of sales efficiency and marketing effectiveness.

Aggregated indicators are used to evaluate the product strategy.

The analysis of various possible scenarios for the impact of changes in KPI values on aspects of product strategy implementation is conducted.

It is proposed to consider the following aspects: financial result, level of customer satisfaction and level of sales/marketing efficiency.

Thus, modelling changes in KPIs enables the product manager to assess the future consequences of implementing the product strategy.

As a result of the research, it was found that the formation of a system of KPIs depends on the type of business model used for implementing an IT product. In further research, it is planned to develop cognitive maps as neural networks for analyzing the KPIs system.

It is planned to investigate the training processes of neural networks using the algorithms of Hebb and Hopfield [23] to enhance the models for developing a system of KRIs for evaluating product strategy.

References:

[1] Grinchenko M., Moskalenko V. Cognitive modeling for strategic analysis of the competitive it company status / M. Grinchenko, V.Moskalenko // Bulletin of the National Technical University "KhPI". Series: Strategic management, portfolio, program and project management. – 2024. – No. 1(8). – P. 17-25. https://doi.org/10.20998/2413-3000.2024.8.3.

[2] Federlova E., Čupka O., Vesely P. Competitiveness of International IT Companies – Comparison of Strategies, Their Strengths, and Weaknesses / E.Federlova, O.Čupka, P.Vesely // Developments in Information and Knowledge Management Systems for Business Applications. Studies in Systems, Decision and Control. – 2023. – vol. 462. – Springer, Cham. https://doi.org/10.1007/978-3-031-25695-0_22.

[3] What are product metrics? /2024. URL: https://www.aha.io/roadmapping/guide/what-are-product-metrics#:~:text=Tracking%20product %20metrics%20helps%20you,means%20to%20be%20data%2Ddriven.

[4] AltexSoft Editorial Team. 20 Key Product Management Metrics and KPIs /2024. URL: https://www.altexsoft.com/blog/15-key-product-management-metrics-and-kpis/.

[5] Introduction to Product Ownership Analysis /2024. URL: https://go.iiba.org/Introduction-to-POA.

[6] Guide to Product Ownership Analysis. International Institute of Business Analysis /2024. URL: https://www.iiba.org/career-resources/business-analysis-specialization/product-ownership-analysis/.

[7] 25 key product management metrics & KPIs for 2025. (And how to track them) /2025. URL: https://www.airtable.com/articles/product-management-metrics.

[8] 11 Product Management KPIs in 2025 [+ How to Track Them] /2025. URL: https://www.meegle.com/blogs/product-management-kpis

[9] 16 product management KPIs and how to track / 2025. URL: https://www.atlassian.com/agile/product-management/product-management-kpis.

[10] Connolly J. 24 strategic KPIs for IT departments /2023. URL: https://www.logmein.com/blog/6-strategic-kpis-for-your-it-department

[11] 25+ product metrics to start tracking /2025 URL: https://miro.com/product-development/product-metrics/

[12] Grinchenko M., Moskalenko V. Information Technology for Strategic Analysis of IT Company Projects / M.Grinchenko, V.Moskalenko // Proc. of 4th International Workshop IT Project Management. – 2023. P.128-138. URL: https://ceur-ws.org/Vol-3453/paper12.pdf.

[13] Grinchenko M., Moskalenko V., Grinchenko E. Data shaping for IT product strategy analysis using fuzzy cognitive modeling / M.Grinchenko, V.Moskalenko, E.Grinchenko // International scientific and practical conference «Information systems and innovative technologies for project and program management». – 2025. – P. 24-27. URL: https://mmp-conf.org/documents/archive/proceedings2025.pdf.

[14] Saaty T.L, Vargas L.G. An Analytic Hierarchy Process Based Approach to the Design and Evaluation of a Marketing Driven Business and Corporate Strategy / T.L Saaty, L.G. Vargas // Models, Methods, Concepts & Applications of the Analytic Hierarchy Process. International Series in Operations Research & Management Science. – 2012. – vol. 175. – Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3597-6_11

[15] Esen H. Analytical Hierarchy Process Problem Solution / H. Esen. // Analytic Hierarchy Process - Models, Methods, Concepts, and Applications [Working Title]. IntechOpen. 2023. DOI: 10.5772/intechopen.1001072

[16] Zadeh L.A., Aliev R.A. Fuzzy Logic Theory and Applications. Part I and Part II. / L.A. Zadeh, R.A. Aliev //World Scientific Publishing Company, 2018. – 612 p. https://doi.org/10.1142/10936

[17] Li J., Zhang B. The Role of Interaction Design Based on Fuzzy Decision Support System in Improving User Experience / J. Li, B. Zhang // International Journal of Fuzzy Systems. – 2025. – V.27. – P.2585–2601. https://doi.org/10.1007/s40815-024-01918-6

[18] Gorelova G.V., Pankratova N.D. Scientific Foresight and Cognitive Modeling of Socio-Economic Systems / G.V. Gorelova, N.D. Pankratova // IFAC-PapersOnLine. – 2018. –Vol. 51, Iss. 30. – P. 145-149. URL: https://doi.org/10.1016/j.ifacol.2018.11.264

[19] Kokkinos K.; Nathanail E. A Fuzzy Cognitive Map and PESTEL-Based Approach to Mitigate CO2 Urban Mobility: The Case of Larissa, Greece/ K. Kokkinos; E.Nathanail // Sustainability. – 2023. – V.15. – 12390. https://doi.org/10.3390/su151612390

[20] Nápoles G., Leon M., Grau I., Vanhoof K. Fuzzy Cognitive Maps Based Models for Pattern Classification: Advances and Challenges / G.Nápoles, M. Leon Espinosa, I. Grau, K. Vanhoof // Soft Computing Based Optimization and Decision Models. Studies in Fuzziness and Soft Computing. – 2018. – Vol. 360. – P. 83-98. https://doi.org/10.1007/978-3-319-64286-4_5

[21] Gao X., Gao X. G., Rong J., Li N., Niu Y., Chen J. On the Convergence of Sigmoid and tanh Fuzzy General Grey Cognitive Maps / X. Gao, X. G. Gao, J. Rong, N. Li, Y. Niu, J. Chen // ArXiv abs/2409.05565 – 2024. https://doi.org/10.48550/arXiv.2409.05565

[22] The C.S., Kudus A. R. The Application of Fuzzy Cognitive Mapping in Education: Trend and Potential / C. S. The, A. R. Kudus // TEM Journal. –2024. – Vol. 13, Iss. 2. – P. 976-991. DOI: 10.18421/TEM132-13.

[23] Papageorgiou E., Stylios C., Groumpos P. Learning Algorithms For Fuzzy Cognitive Maps. URL: https://www.researchgate.net/publication/221399332

Date of submission of the article to the editorial board: 09.10.2025

УДК 004.9

DOi 10.30837/ bi.2025.2(103).11

С. Л. Чирун[1], В. А. Висоцька[2], О. Я. Бродяк[3]

[1]НУ «Львівська політехніка», м. Львів, Україна, sofiia.chyrun.sa.2022@lpnu.ua,
ORCID iD: 0000-0002-2829-0164

[2]НУ «Львівська політехніка», м. Львів, Україна, victoria.a.vysotska@lpnu.ua,
ORCID iD: 0000-0001-6417-3689

[3]НУ «Львівська політехніка», м. Львів, Україна, oksana.y.brodiak@lpnu.ua,
ORCID iD: 0000-0002-9886-3589

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ МОДЕЛЮВАННЯ КРИЗОВИХ СИТУАЦІЙ У VR/AR З ЕЛЕМЕНТАМИ ГЕЙМІФІКАЦІЇ ДЛЯ НАВЧАННЯ ДОМЕДИЧНІЙ ДОПОМОЗІ ЦИВІЛЬНОГО НАСЕЛЕННЯ

Актуальність дослідження зумовлена критичною потребою в ефективній та психологічно реалістичній підготовці фахівців і цивільного населення до надання домедичної допомоги в умовах воєнних дій та кризових ситуацій. Традиційні методи навчання не можуть повною мірою симулювати стресові фактори та інтерактивну динаміку реальних надзвичайних подій. Метою дослідження є розробка інформаційної технології для інтерактивного VR/AR-симулятора домедичної допомоги, що дозволяє користувачам без ризику відпрацьовувати критично важливі навички в середовищі, максимально наближеному до бойових умов. Для реалізації проекту використано технології формування Бізнес-моделі та Структури декомпозиції робіт (WBS) для MVP; генеративного штучного інтелекту (DALL-E, Trellis3D, Tripo, Meshy, Stable Diffusion) для швидкої генерації реалістичного 2D та 3D-контенту (руїни, постраждалі, поранення); створення VR-сцени у Unreal Engine 5 на базі шаблону VR Template; реалізацію ключових VR-механік, таких як Smooth Locomotion, система захоплення об'єктів (Grabbable Objects), а також інтеграцію навчального мультимедійного контенту. Результати демонструють успішне створення прототипу VR-середовища з інтерактивним сценарієм надання допомоги, наприклад, після ракетного удару. Сформовано десятки моделей різних елементів реального середовища сучасного українського міста за допомогою мобільної фотограмметрії, використовуючи RealityScan, зі зниженою кількістю фото (близько 30-50 кадрів замість 80-100) для досягнення ефекту пошкоджених об’єктів. Імпортовано 3D-моделі, згенеровані ШІ (Tripo, Meshy) та знайдені на маркетплейсах, а також реалізовано коректну колізію, телепорт і плавне переміщення користувача. Висновки підтверджують ефективність інтеграції генеративного ШІ та ігрових технологій для оперативного створення спеціалізованого VR-контенту. Експеримент із фотограмметрією показав, що свідоме зменшення кількості вхідних даних (використано лише 37,5–62,5% від рекомендованої кількості фото) є життєздатним творчим методом для моделювання зони ураження шляхом індукування некритичних спотворень сітки та текстур. Отримані навички дозволяють ефективно застосовувати AI-інструменти для розробки навчальних та симуляційних VR/AR-середовищ, що має високе прикладне значення у сфері безпеки та медицини.

VR/AR-СИМУЛЯТОР, ДОМЕДИЧНА ДОПОМОГА, ТАКТИЧНА МЕДИЦИНА, UNREAL ENGINE 5, TRIPO, MESHY, STABLE DIFFUSION, ГЕНЕРАТИВНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ФОТОГРАММЕТРІЯ, КАНВА БІЗНЕС-МОДЕЛІ, TRIAGE.

S. L. Chyrun, V. A. Vysotska, O. Y. Brodyak. Information technology for modelling crisis situations in VR/AR with elements of gamification for training the civil population in pre-medical care provision. The relevance of the study lies in the critical need for effective and psychologically realistic training of specialists and civilians to provide first aid in conditions of military operations and crisis situations. Traditional training methods cannot fully simulate stress factors and the interactive dynamics of real emergency events. The objective of this study is to develop information technology for an interactive VR/AR simulator of first aid, which enables users to practice critical skills without risk in an environment as realistic as possible to combat conditions. To implement the project, the following technologies were used: Business Model and Work Breakdown Structure (WBS) formation for MVP; Generative Artificial Intelligence (AI) (DALL-E, Trellis3D, Tripo, Meshy, Stable Diffusion) for rapid generation of realistic 2D and 3D content (ruins, victims, injuries); creation of a VR scene in Unreal Engine 5 based on the VR Template; implementation of key VR mechanics such as Smooth Locomotion, Grabbable Objects, and integration of educational multimedia content. Results demonstrate the successful creation of a prototype VR environment featuring an interactive scenario that assists with missile strike operations. During the work, five models of children's swings were generated using mobile photogrammetry with RealityScan, utilising a reduced number of photos (approximately 30-50 frames instead of 80-100) to achieve the effect of damaged objects. 3D models generated by AI (Tripo, Meshy) and found on marketplaces were imported, and correct collision and smooth user movement were implemented. The findings confirm the effectiveness of integrating generative AI and gaming technologies for the rapid creation of specialized VR content. The photogrammetry experiment demonstrated that deliberately reducing the amount of input data (using only 37.5–62.5% of the recommended number of photos) is a viable creative method for modelling the affected area, inducing non-critical mesh and texture distortions. The acquired skills enable the effective use of AI tools for developing training and simulation VR/AR environments, which have high applied value in the fields of security and medicine.

VR/AR SIMULATOR, HOME CARE, TACTICAL MEDICINE, UNREAL ENGINE 5, TRIPO, MESHY, STABLE DIFFUSION, GENERATIVE ARTIFICIAL INTELLIGENCE (AI), PHOTOGRAMMETRY, BUSINESS MODEL CANVAS, TRIAGE.

Вступ

Актуальність швидкого та ефективного надання домедичної допомоги в умовах воєнного часу та надзвичайних ситуацій є критично високою. Традиційні методи навчання часто обмежуються теорією та статичними манекенами, що не забезпечує належної психологічної та практичної підготовки до стресових умов, які супроводжуються пораненнями, кровотечами, шоком та обмеженим часом для прийняття рішень.

Це дослідження спрямоване на розробку інноваційного підходу до тренування навичок першої допомоги шляхом створення інтерактивного VR/AR-симулятора (Virtual/ Augmented Reality) на базі ігрового рушія Unreal Engine 5. Ключова особливість проєкту полягає у використанні генеративного штучного інтелекту (Generative AI), такого як Stable Diffusion, Trellis3D та Tripo, для швидкого створення реалістичного та унікального 2D та 3D-контенту, включаючи зруйновані міські сцени, пошкоджені об'єкти та моделі постраждалих з характерними травмами. Проєкт охоплює повний цикл розробки, починаючи з формування бізнес-моделі (Business model canvas) та детального планування WBS (Work Breakdown Structure). В рамках реалізації було успішно прототиповано VR-сцену міської вулиці після ракетного удару, реалізовано ключові механіки VR-взаємодії, такі як Smooth Locomotion та система захоплення об'єктів (Grabbable Objects), налаштовано фізичні колізії, а також інтегровано навчальний та атмосферний контент (UMG-меню, відеоінструкції, просторове аудіо). Додатково проведено експеримент із фотограмметрією для створення унікальних асетів реального світу.

Метою дослідження є розроблення інформаційної технології для створення на основі віртуальної реальності в Unreal Engine безпечного, гнучкого та реалістичного навчального середовища, що дозволяє користувачам (військовим, медикам, студентам та цивільним) без ризику відпрацьовувати критично важливі навички домедичної допомоги в умовах, максимально наближених до бойових, війни/кризових ситуацій, використовуючи VR/AR-формат та ігрові механіки. Задачі дослідження випливають із загального плану робіт WBS та практичних рішень, зокрема:

– Розробити концепцію та архітектуру VR/AR-симулятора, включаючи формування цільової аудиторії, ціннісної пропозиції та ключових ресурсів (згідно з Канвою бізнес-моделі).

– Створити візуальний контент (2D та 3D) для сцени VR/AR, використовуючи генеративний штучний інтелект (ГШІ) (Stable Diffusion, Leonardo.Ai, Trellis3D, Meshy, Tripo) для моделювання зруйнованого середовища та постраждалих.

– Розробити VR-проєкт в Unreal Engine через шаблон VR Template та побудувати базову структуру VR-сцени.

– Організувати міграцію контенту, налаштування колізій для 3D-моделей та використання фотограмметрії для створення унікальних асетів реального середовища.

– Реалізувати ключові механіки VR-взаємодії, включаючи налаштування Smooth Locomotion (плавне переміщення), телепорту та системи захоплення об'єктів (Grabbable Objects) для медичних інструментів.

– Інтегрувати навчальні елементи та користувацький інтерфейс (UI/UX), зокрема, створити VR-меню, додати просторовий аудіо супровід та вбудувати навчальні відеоінструкції (Triage, CPR, Bleed Stop, тощо).

– Протестувати VR-проєкт серед контрольної групи учасників експериментальної апробації.

Об'єкт дослідження – процеси розробки, інтеграції та оптимізації контенту та механік віртуальної (VR) та розширеної (AR) реальності для створення навчальних симуляторів. Предмет дослідження – інтерактивний VR/AR-симулятор домедичної допомоги в умовах кризових та воєнних ситуацій, реалізований на базі ігрового рушія Unreal Engine 5.

Наукова новизна дослідження полягає в наступному:

– Системна інтеграція ГШІ для прискореної розробки VR-сцен, зокрема вперше застосовано комбінацію ГШІ-інструментів (Tripo, Meshy, Trellis3D) та ігрового рушія Unreal Engine 5 для оперативного створення спеціалізованого та високодеталізованого контенту (постраждалі, руїни, поранення), що значно скорочує час розробки MVP (Minimum Viable Product).

– Експериментальне використання неповних даних у фотограмметрії, в тому числі експериментально доведено, що свідоме зменшення кількості фотографій (наприклад, 37,5–62,5% від рекомендованої) при мобільній фотограмметрії (RealityScan) може бути використане як творчий метод для моделювання асетів у стилі "зони ураження" (неповна деталізація, спотворення сітки), що є актуальним для військових симуляцій.

– Розробка комбінованої системи переміщення для комфорту у VR, зокрема, реалізовано та протестовано комбінований підхід до навігації, що поєднує Smooth Locomotion та телепортацію для забезпечення максимального занурення та мінімізації VR-хвороби.

Практична цінність полягає у наступному:

– Створення функціонального прототипу VR-тренажера/симулятора з інтерактивною сценою надання допомоги після ракетного удару, що має пряму прикладну цінність для Міністерства оборони, Червоного Хреста, МОЗ, військових академій та громадських організацій.

– Реалістичне тренування, що надає безпечний спосіб відпрацювання критично важливих навичок (накладання джгута, СЛР, Triage) без ризику для реальних людей, готуючи користувачів до стресових умов (симуляція шоку, паніки).

– Готова методика швидкого створення контенту, так як надано чіткі інструкції та порівняльні таблиці використання безкоштовних/платних ГШІ-інструментів (2D/3D) та платформ 3D-моделей, що може бути використано розробниками для швидкого наповнення VR-проєктів.

– Мультиплатформенність та доступність, зокрема, проєкт передбачає підтримку як високоякісних VR-гарнітур, так і мобільних AR-додатків, забезпечуючи гнучкість та доступність навчання.

1. Постановка проблеми

Проблема дослідження полягає у розробці та оптимізації моделі високореалістичного, інтерактивного та доступного симуляційного середовища для тренування навичок домедичної допомоги, які є критичними в умовах обмеженого часу та стресових факторів, невід'ємних для кризових та військових ситуацій. Ключова задача – максимізувати ефективність навчання $E_{\text{навч}}$ при мінімізації вартості розробки $C_{\text{розр}}$ та часу виведення продукту на ринок T_{MVP}, використовуючи ресурси ГШІ для створення високоякісного контенту.

Ефективність навчання $E_{\text{навч}}$ визначається як зважена сума ключових показників, що відображають глибину занурення R, рівень інтерактивності I, якість зворотного зв'язку F, та коректність виконання критичних навичок $A_{\text{кр}}$.

$$E_{\text{навч}} = \sum \omega_j \cdot P_j \rightarrow \max,\ j=1,\ N,\ \sum \omega_j = 1. \quad (1)$$

Вартість розробки $C_{\text{розр}}$ та час виведення продукту на ринок T_{MVP} повинні бути мінімізовані за рахунок використання ГШІ.

$$C_{\text{розр}} + \alpha \cdot T_{MVP} \rightarrow \min. \quad (2)$$

Враховуючи внесок ГШІ:

$$C_{\text{розр}} = C_{\text{трад}} \cdot (1 - S_{\text{гші}}) + C_{\text{гші}}, \quad (3)$$

$$T_{MVP} = T_{\text{трад}} \cdot (1 - \delta_{\text{гші}}). \quad (4)$$

Проєкт має відповідати наступним обмеженням:

1. Технологічні обмеження: $P_{\text{тех}} \in \{\text{Unity, Unreal Engine 5}\} \wedge V_{\text{тех}} \in \{\text{VR-гарнітури, Мобільні AR-пристрої}\}$.

2. Часові обмеження (з WBS): $T_{MVP} \leq 42$ тижні.

3. Обмеження реалізму та коректності (з сертифікацією): $A_{\text{кр}} \geq A_{min}$.

4. Обмеження локалізації: $L \geq L_{min}$.

Таким чином, завдання дослідження можна сформулювати як задачу багатоцільової оптимізації: знайти оптимальне поєднання параметрів розробки (вибір технологій, рівень інтеграції ГШІ та розподіл ресурсів), яке максимізує ефективність навчання $E_{\text{навч}}$ при дотриманні усіх технологічних, часових та регуляторних обмежень, мінімізуючи при цьому загальні витрати та час розробки MVP.

2. Огляд та аналіз літератури

Аналіз літератури та споріднених розробок демонструє стрімке зростання інтересу до використання імерсивних технологій (VR/AR) та ГШІ для підвищення ефективності навчання, особливо в критично важливих сферах, таких як медицина та військова підготовка. Дослідження, представлене у файлі, поєднує три ключові науково-практичні напрями, кожен з яких має значну дослідницьку базу. Протягом останнього десятиліття доведено, що VR-симуляції у тактичній та екстреній медицині є високоефективною альтернативою традиційним манекенам, особливо для тренування складних, високоризикових та низькочастотних подій [1-2]. Зокрема, дослідження підкреслюють здатність VR-середовища відтворювати будь-який стан пацієнта в будь-якому середовищі [3], що є неможливим для фізичних симуляторів. Дослідження, подібні до підтримки ефективності навчання через занурення (Immersiveness), підтверджують, що імерсивні технології значно покращують ефективність навчання та задоволеність користувачів у надзвичайних ситуаціях [4-5]. Дослідження ефективності застосування TacMedVR підкреслює важливість оцінки взаємодії та реакції на стрес [4], що безпосередньо корелює з ціннісною пропозицією цього проєкту (симуляція емоційних реакцій постраждалих, навчання в умовах тиску) [6-10].

Використання VR-симуляторів, як, наприклад, платформи SimX [2], підтвердило їхню перевагу у розвитку критичного мислення, комунікації критичної інформації та командної динаміки при наданні допомоги в умовах воєнних дій (Damage Control Resuscitation/Surgery) [11]. Контекст командної роботи та критичного мислення безпосередньо виправдовує необхідність розробки, орієнтованої на сценарії бойових дій. Традиційне моделювання 3D-контенту є найбільш ресурсомістким та часозатратним етапом розробки симуляторів. Тому на основі ГШІ та автоматизації 3D-контенту (Tripo, Meshy) для прискореного створення 3D-об'єктів, є частиною загальносвітового тренду. Сучасні розробки, зокрема інтеграція ГШІ, такого як Ludus AI в Unreal Engine 5.5 [6], демонструють парадигмальний зсув [7]. Підхід прискорення робочого процесу дозволяє генерувати 3D-моделі з текстових описів та по зображеннях практично в реальному часі, що різко скорочує час розробки (TMVP у термінах WBS проєкту) [12]. Хоча традиційне моделювання все ще пропонує більш точну деталізацію [6], платформи на кшталт Sloyd (не індексовано у списку) та Rodin AI (не індексовано у списку) активно розвивають можливість створення високоякісних, оптимізованих для ігор 3D-моделей з тексту чи зображень, підтверджуючи життєздатність обраного методу для створення асетів (руїни, пошкоджені об'єкти) для симулятора. Використання фотограмметрії для створення 3D-моделей об'єктів реального світу (наприклад, гойдалки) доводить бажання підвищити фотореалізм сцени для покращення якості навчання в реалістичних умовах. Це відповідає напряму досліджень, які використовують цю техніку для створення навчальних ресурсів.

Фотограмметрія є доступним та недорогим методом [8, 9] для створення високодеталізованих реалістичних 3D-моделей анатомічних препаратів або реальних об'єктів для медичної освіти. Її інтеграція в рушії, як-от Unity [8], або оцінка в RealityCapture [10], підтверджує, що ця технологія значно підвищує занурення та реалізм [9] у симуляційних середовищах. Особлива новизна сучасного VR-проєкту полягає у креативному підході на основі навмисному зниженні якості вхідних даних для фотограмметрії, щоб досягти ефекту пошкодженого контенту. Хоча більшість досліджень (наприклад, [3, 10]) зосереджені на максимізації точності (60-80% перекриття, мінімізація шуму), запропонований підхід, що використовує RealityScan для створення об'єктів "зони ураження", є унікальним у контексті створення контенту для військових симуляторів.

3. Матеріали і методи

Розробка та реалізація інтерактивного VR/AR-симулятора домедичної допомоги базується на міждисциплінарному підході, що поєднує методології геймдизайну, комп'ютерної графіки, моделювання реальності (фотограмметрія) та технології генеративного штучного інтелекту (ГШІ). Експериментальна частина зосереджена на створенні мінімально життєздатного продукту (MVP). Для управління проєктом та контролю витрат використано методологію Структури Декомпозиції Робіт (WBS) та Діаграму Ганта. Проєкт поділено на 6 ключових етапів:

Таблиця 1

Структурна декомпозиція та приблизне часове планування (WBS & Gantt Chart)

Етап (E_i)	Назва	Тривалість (T_i, тижнів)	Орієнтовні дати
E_1	Дослідження та планування	T_1=8	28.04.25 – 15.06.25
E_2	Розробка MVP	T_2=15	02.06.25 – 21.09.25
E_3	Тестування та покращення	T_3=6	22.09.25 – 02.11.25
E_4	Маркетинг та просування	T_4=4	03.11.25 – 30.11.25
E_5	Масштабування та партнерство	T_5=8	01.12.25 – 14.02.26
E_6	Управління проєктом	T_6=42	28.04.25 – 14.02.26

Загальна тривалість проєкту $T_{заг}$:

$$T_{заг}=\max(T_i),\ i\in\{1,...,6\}. \quad (5)$$

Отже, $T_{заг}=T_6=42$ тижні (з урахуванням паралельного управління). Цільова аудиторія A_i сегментована для адаптації сценаріїв $A=\{A_{уч}, A_{гр}, A_{військ}\}$. Ключовий сценарій MVP: вулиця міста після ракетного удару/вибуху міни. Сценарій має 4 типи постраждалих, класифікованих згідно з системою Triage: $P=\{P_{крит\_діт}, P_{крит\_мат}, P_{сер\_оп}, P_{лег}\}$. Для прискорення створення асетів (3D-моделі будівель, транспорту, персонажів) використовувалися інструменти ГШІ, зокрема нейронні мережі: G={Tripo,Meshy,Trellis3D}. Основні методи – Text-to-3D та Image-to-3D, а формати експорту: OBJ, FBX, GLB/GLTF. Для максимізації якості 2D-концептів та 3D-моделей використовувався деталізований промпт $Q_{промпт}$, що містить об'єкт, сценарій, стиль, освітлення та деталізацію, тобто $Q_{промпт}\in$ {"низькополігональна міська вулиця після ракетного удару з стилізованим освітленням"}. Для створення асетів, що мають високий рівень реалізму, але імітують пошкодження, застосовувалася мобільна фотограмметрія (Samsung Galaxy A52 та RealityScan). Експериментальний метод "пошкодженого реалізму" – навмисне зменшення кількості вхідних кадрів $N_{фото}$ для індукування некритичних спотворень сітки та текстур. Умова експерименту – $N_{фото}\in[30,50]$ кадрів. Рекомендований діапазон – $N_{реком}\in[80,100]$ кадрів. Об'єкти сканування – основні міські елементи спального масиву, наприклад моделі дитячих гойдалок. Вихідний формат – GLB. Робочим середовищем обраний ігровий рушій Unreal Engine 5 (UE5), де базовий шаблон згідно рис. 1-3 – VR Template, компоненти сцени – VRPawn (віртуальний персонаж гравця) та NavMeshBounds Volume (зона навігації для телепортації). Для системи переміщення *Locomotion* застосовано комбінований підхід:

$$Locomotion=L_{телепорт}\oplus L_{плавне}. \quad (6)$$

Для медичних інструментів (джгут, ножиці) використано клас Grabbable_SmallCube з активацією фізики. Умова захоплення:

$$G:\{Об'єкт\wedge Grabbable\_Component\wedge Simulate\ Physics= =\text{true}\wedge Collision\ Preset=PhysicsActor\}.$$

Для всіх імпортованих Static Mesh (включаючи моделі ГШІ) використана спрощена колізія (Simple Collision) для оптимізації продуктивності VR: *Collision → Add Box Collision/Convex Decomposition → Apply → Save*.Інтеграція контенту, зокрема, імпорт згенерованих моделей здійснено у форматі FBX з подальшим ручним налаштуванням PBR-текстур у Material Editor.

Оцінка якості тренування базувалася на функції ефективності навчання $E_{навч}$, як визначено в постановці проблеми, що є ключовим інструментом для оцінки досягнення мети дослідження (1).

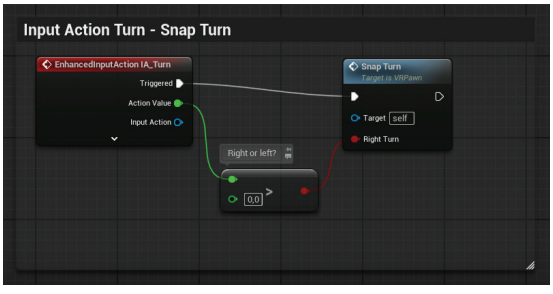

Рис. 1. Схема реалізації повороту (Snap Turn) у VR-середовищі

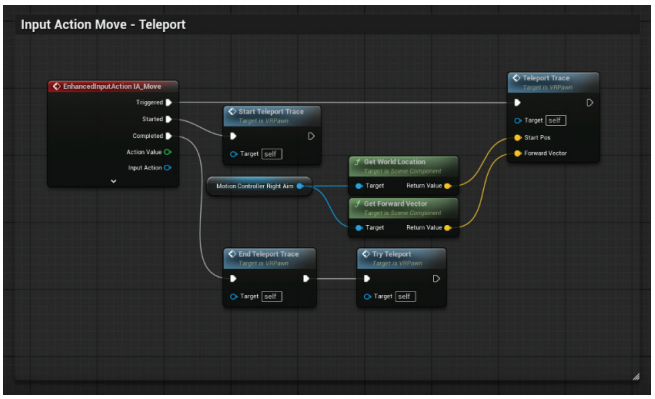

Рис. 2. Схема реалізації системи телепортації (Teleport) у VR-середовищі

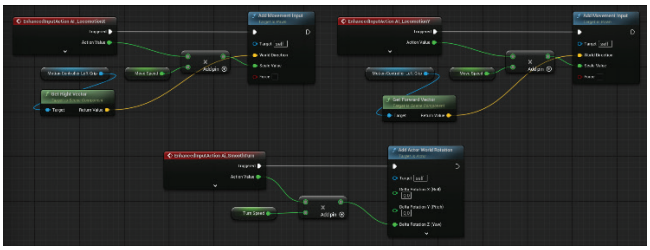

Рис. 3. Схема реалізації плавного пересування (Smooth Locomotion) у VR-середовищі

Тестування відбувалося у три етапи:

1. Внутрішнє тестування (функціональність, продуктивність).

2. Альфа-тестування (фахівці: медики, військові інструктори) – валідація реалістичності сценаріїв та алгоритмів.

3. Бета-тестування (волонтери, студенти) – оцінка UX/UI та інтуїтивності.

Ефективність використання ГШІ у створенні контенту (задача мінімізації $C_{розр}$ та T_{MVP}) оцінювалася шляхом порівняння часу, витраченого на створення асетів за допомогою ГШІ, з традиційними оцінками моделювання (2). Це підтверджує економічну доцільність використаних методів.

Навчальний контент (відео-інструкції та інтерактивні підказки) інтегрований у систему і базується (рис. 4-7) на офіційних протоколах домедичної допомоги (МОЗ, Червоний Хрест). Умова валідації контенту $L \in \{$*Офіційні протоколи МОЗ*, *Червоний Хрест*, *Військова медицина*$\}$.

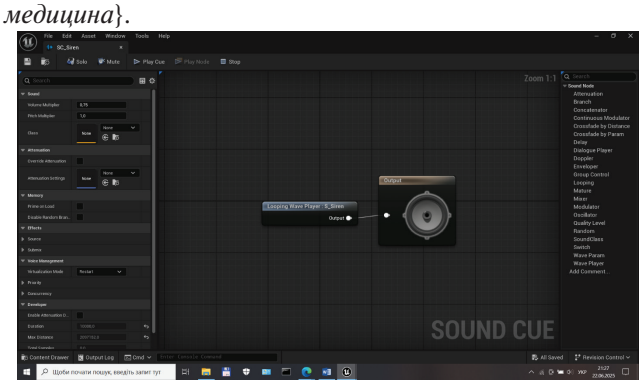

Рис. 4. Схема реалізації запуску звуку сирени

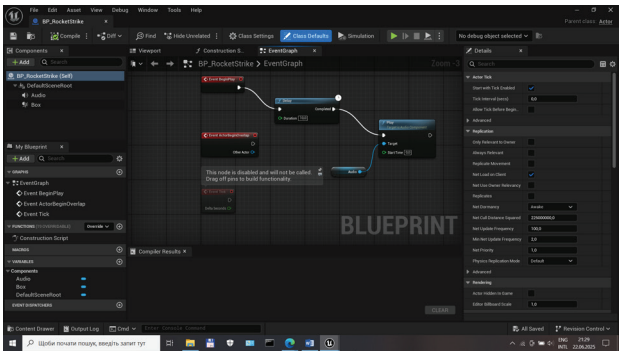

Рис. 5. Схема реалізації запуску звуку вибуху ракети/міни

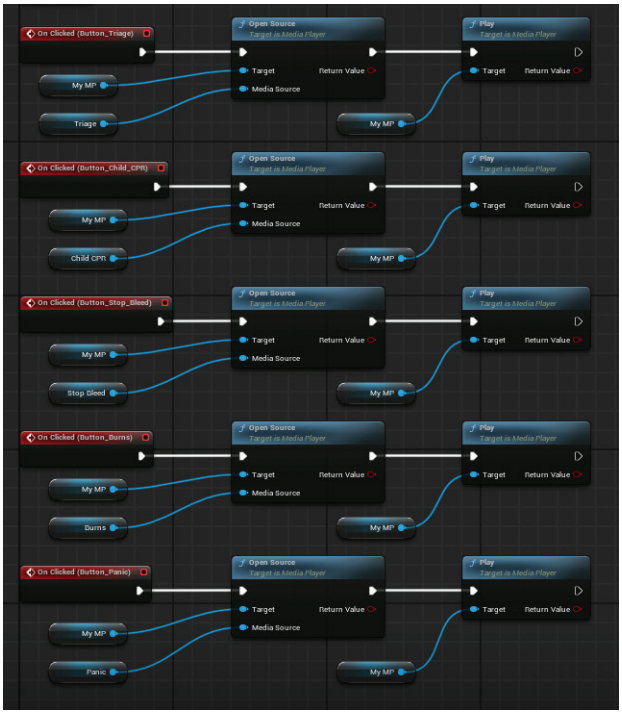

Рис. 6. Схема реалізації подання навчального відео

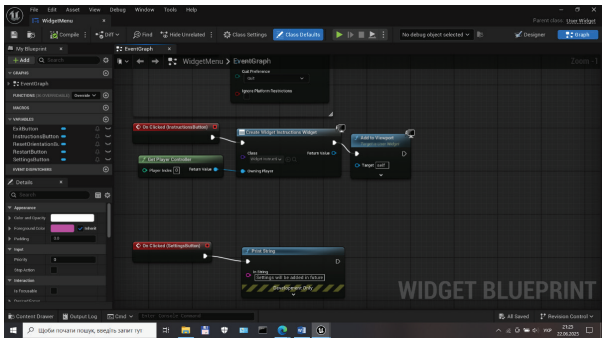

Рис. 7. Схема реалізації кнопок Instructions" та "Settings" модифікованого стандартного VR-меню

4. Експерименти

Експериментальна частина дослідження спрямована на практичну реалізацію ключових функціональних модулів VR-симулятора домедичної допомоги (MVP) та валідацію інноваційних методів створення контенту, зокрема, застосування ГШІ та мобільної фотограмметрії. Експерименти проводились згідно з етапами Розробка MVP E_2 та Тестування та покращення E_3 (Таблиця 1).

Метою експерименту «Валідація Інтеграції ГШІ в Робочий Процес (Етап E_2)» було підтвердження гіпотези про те, що використання ГШІ може забезпечити швидку та економічно ефективну генерацію 3D-моделей, які відповідають вимогам специфічної сцени ("зона ураження" – рис. 8-10). Використовувалися генеративні моделі при створенні візуальних концептів та 3D-об'єктів для сцени «Вулиця міста після ракетного удару» (рис. 11-13). Інструменти генерації: Stable Diffusion, DALL-E (ChatGPT), KREA, Ideogram (для 2D-концептів). Tripo, Meshy, Trellis3D (для 3D-моделей).

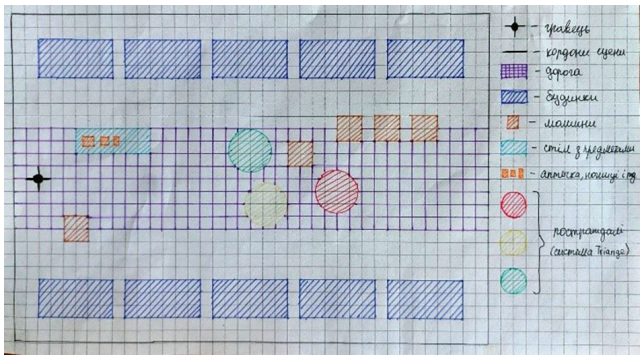

Рис. 8. Ескіз сцени

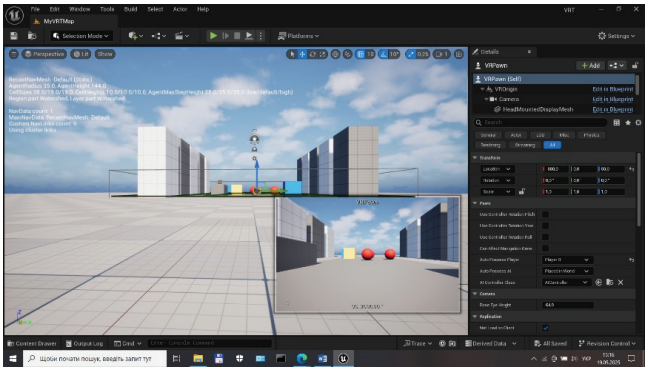

Рис. 9. Прототипування VR-сцени (дорога, будинки, машини, постраждалі)

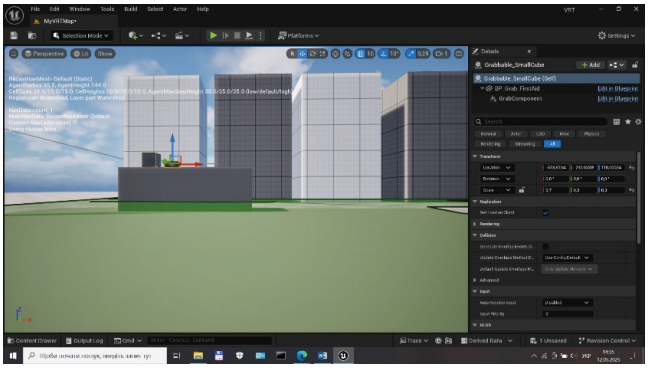

Рис. 10. Прототипування VR-сцени (стіл з медичним обладнанням)

Основний метод полягає у застосуванні деталізованих промптів для створення специфічних персонажів (рис. 14-15) та оточення (наприклад, "поранена мати стоїть з рукою, що кровоточить... стилізована лоуполігональна естетика...").

Рис. 11. Будівлі згенеровані в Mersy

Рис. 12. Машини згенеровані в Mersy

Рис. 13. Будівлі згенеровані в Tripo

Рис. 14. Парамедики (гравці) згенеровані в Tripo

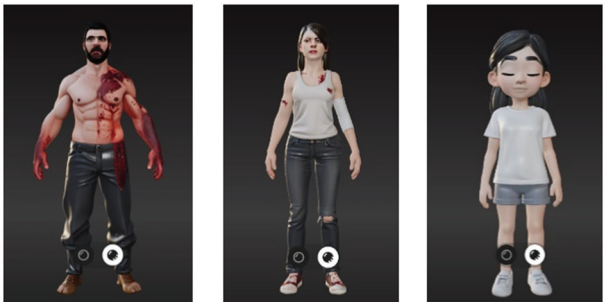

Рис. 15. Персонажі гри (постраждалі) з Tripo

Успішно згенеровано 3D-моделі для ключових аспектів сцени (рис. 16-17):

– Пошкоджене оточення (зруйновані будинки, пошкоджені машини).

– Людські моделі постраждалих (дитина у непритомному стані, мати з кровотечею, чоловік з опіками).

Генеративні нейромережі дозволили оперативно (в межах запланованого часу T_{MVP}) створити унікальну бібліотеку 3D-асетів, придатних для подальшого імпорту в UE, що значно розширило можливості сцени та її правдоподібність. Моделі експортувалися у формати FBX/GLB та імпортувалися в UE5.

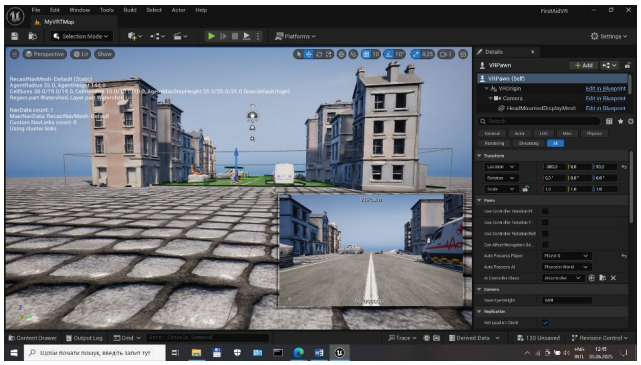

Рис. 16. Сцена з імпортованим 3D-контентом

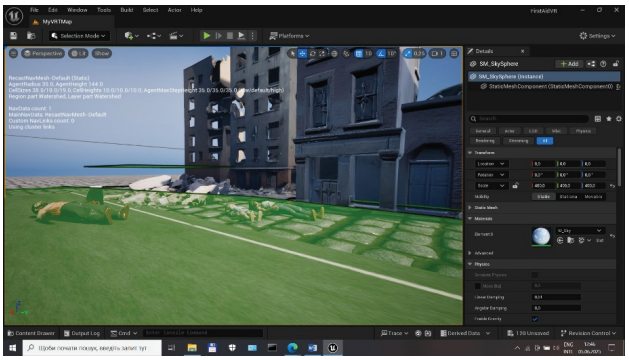

Рис. 17. Місце ураження ракетою/міною з постраждалими

На моделях реалізовано фізичні колізії (Simple Collision/Convex Collision) для забезпечення коректної взаємодії гравця (VR-персонажа) з оточенням (рис. 18). Моделі, згенеровані ГШІ (Tripo, Meshy), виявились придатними для VR-сцен, що підтвердило можливість використання ГШІ для мінімізації витрат $C_{розр}$ на художнє моделювання.

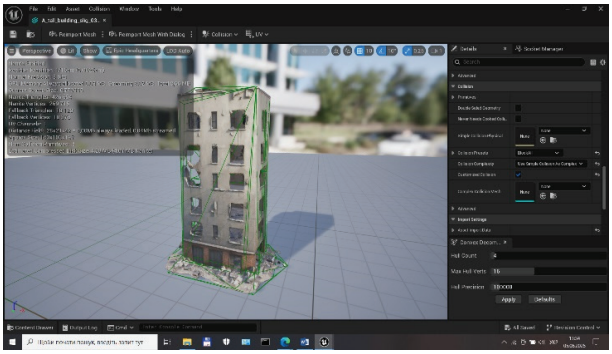

Рис. 18. Об'єкт з налаштованою фізичною колізією (Convex Collision)

Метою експерименту «Валідація методу пошкодженого реалізму через фотограмметрію (Етап E_2)» було визначення, чи можна контрольоване зниження якості вхідних даних для фотограмметрії використовувати для імітації пошкоджень об'єктів без додаткової 3D-обробки, наприклад, об'єктів дитячих майданчиків (5 моделей). Інструмент – RealityScan (Epic Games) на мобільному пристрої. Основна експериментальна умова (контрольоване зниження) полягала в тому, що кількість знімків $N_{фото}$ свідомо обмежувалась в діапазоні 30-50 кадрів, що становить приблизно 37,5% до 62,5% від рекомендованої кількості (80-100 кадрів). Моделі створені з частковими спотвореннями сітки, незаповненими ділянками та неточностями в текстурах (рис. 19). Кількість полігонів коливалася від 221 789 до 669 954. Експеримент підтвердив, що свідоме обмеження кількості фотографій призводить до ефекту "пошкодженого" вигляду (неповної деталізації), що є бажаним для візуального стилю зони ураження і може бути використано як креативний метод для моделювання VR-середовища.

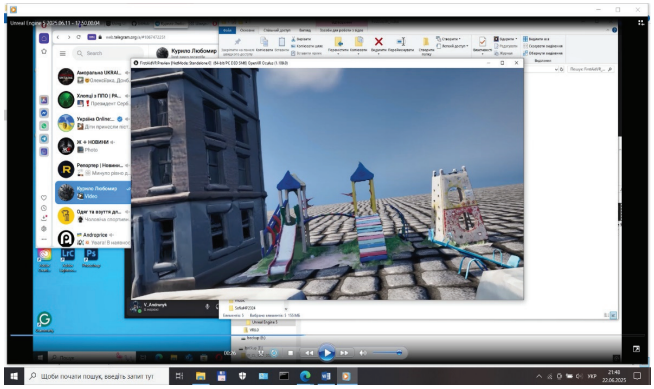

Рис. 19. 3D-моделі створені в RealityScan

Метою експерименту «Реалізація та тестування ключових VR-механік (Етапи E_2-E_3)» є досягнення високого рівня інтерактивності P_2 та комфорту користувача, що є необхідним для успішної реалізації функції ефективності навчання $E_{навч}$.

Комбінований підхід $L_{телепорт} \oplus L_{плавне}$ забезпечує оптимальне занурення та комфорт. Проведено дві спроби реалізації:

1. Реалізація виключно Smooth Locomotion з відключенням телепортації. Результатом було те, що персонаж не міг рухатися (повна відмова функції).

2. Реалізація плавного переміщення з включеною телепортацією.

Ефективним для комфортного використання виявився комбінований підхід $L_{телепорт} \oplus L_{плавне}$, що підтвердило необхідність збереження телепортації як "запасного варіанту" для мінімізації віртуальної дезорієнтації (рис. 20).

Рис. 20. Скріншот з відео-тестування VR/AR-симулятора

Реалізовано функціонал захоплення для медичних інструментів (джгут, ножиці) шляхом:

– Додавання компонента GrabComponent до об'єкта.

– Активація Simulate Physics = true.

– Налаштування Collision Preset: PhysicsActor.

Створено VR pickup items, які можна фізично захоплювати та використовувати в сцені (рис. 21), що підтверджує досягнення необхідного рівня інтерактивності P_2 для відпрацювання навичок.

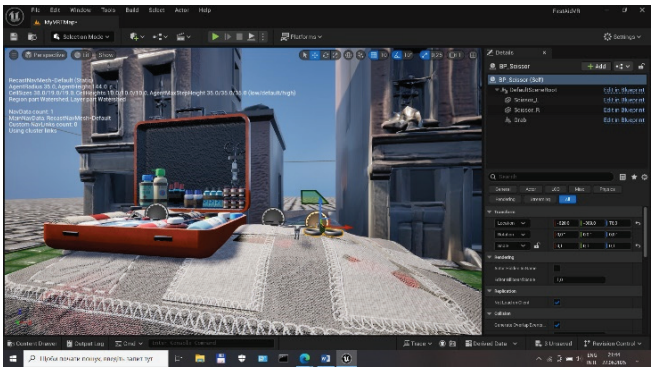

Рис. 21. Об'єкти, які можна фізично захоплювати та використовувати в сцені

Модифіковано стандартне VR-меню (рис. 22), додано функціональні кнопки ("Instructions", "Settings"). Встановлено функціональність відображення навчальних відео (Triage, CPR, Bleed Stop, etc) на віртуальному екрані (рис. 23-24) та автоматичну активацію просторового аудіо (сирени, вибухи – рис. 25-26).

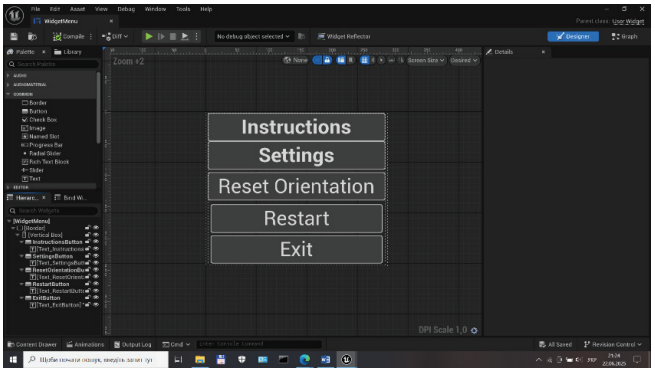

Рис. 22. Модифіковане стандартне VR-меню

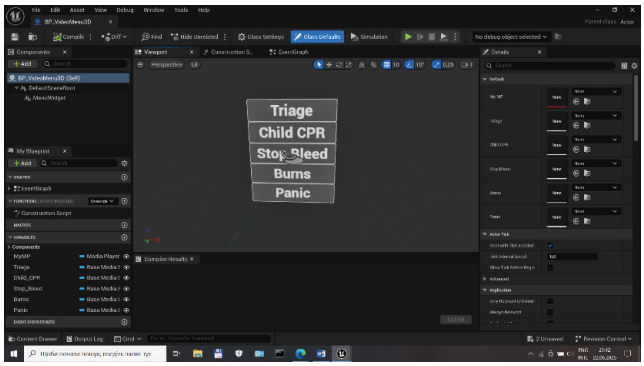

Рис. 23. Меню для керування екраном відображення навчального відео

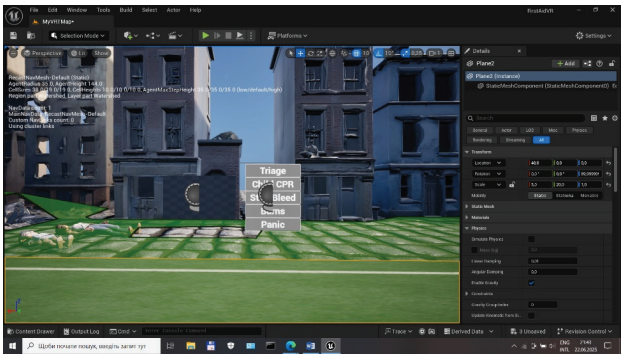

Рис. 24. Розташування меню та екрану для відображення навчального відео

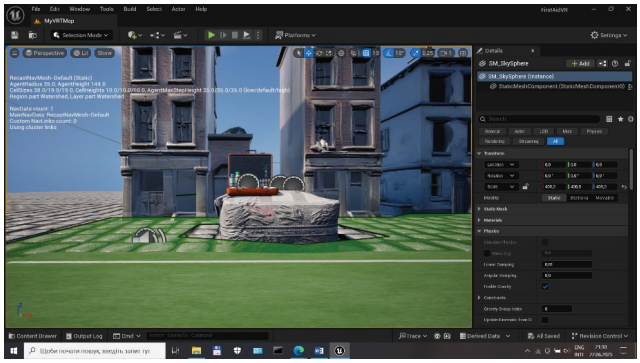

Рис. 25. Розташування звуку сирени

Інтеграція забезпечила можливість отримання інтерактивних підказок та автоматичного оцінювання дій (рис. 27), що закладає основу для кількісної оцінки показника якості зворотного зв'язку P_3 на етапі кінцевого тестування.

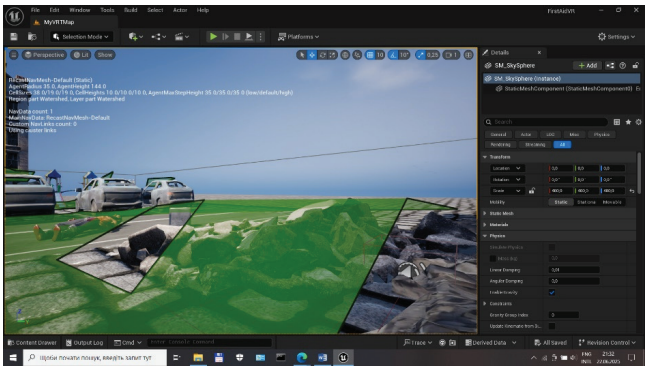

Рис. 26. Розташування звуку вибуху ракети/міни

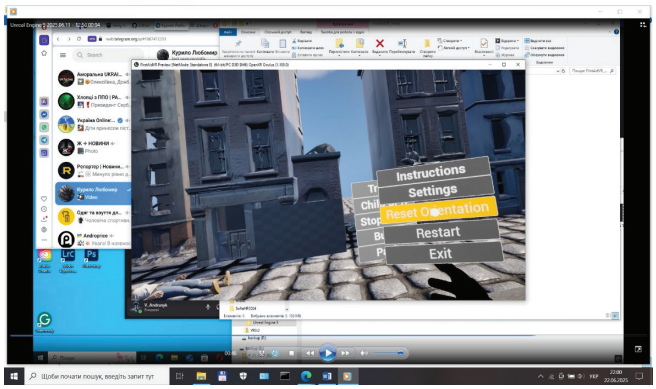

Рис. 27. Скріншот з відео-тестування VR/AR-симулятора

5. Результати

Результати дослідження відображають успішну реалізацію ключових етапів розробки VR/AR-симулятора домедичної допомоги (MVP) із застосуванням інноваційних методів створення контенту. Отримані кількісні та якісні показники підтверджують ефективність обраної стратегії оптимізації та порівнюються з відомими результатами споріднених робіт. Проєктна фаза концептуального планування та часової оптимізації підтвердила життєздатність обраної бізнес-моделі. Декомпозиція робіт (WBS) визначила загальний орієнтований термін розробки MVP – 42 тижні.

Таблиця 2

Структура MVP

Етап	Тривалість	Мета та результат
Розробка MVP E2	15 тижні	Створено функціональний прототип сцени та ключові механіки (VR/AR-взаємодія, моделювання поранень).
Тестування E3	6 тижні	Проведено 3 фази тестування (внутрішнє, альфа- з медиками, бета- на волонтерах).

Результати, отримані під час фази планування, корелюють із дослідженнями, які підтверджують високу ефективність VR/AR у медичному навчанні. Обрана цінна пропозиція – реалістичне моделювання критичних ситуацій та VR/AR-інтерактивність – відображає підхід, який у подібних рандомізованих контрольованих дослідженнях показав статистично значуще покращення ефективності навчання порівняно зі звичайними методами. Використання ГШІ-інструментів (Tripo, Meshy) дозволило оперативно створити специфічний контент для сцени "Зона Ураження". ГШІ успішно згенерував необхідні моделі: зруйновані будинки, пошкоджені машини та персонажі з характерними травмами (наприклад, чоловік з опіками, мати з кровотечею). Це підтверджує, що інтеграція ГШІ мінімізує залежність від традиційного моделювання, що є ключовим фактором для мінімізації загальних витрат $C_{розр}$ та прискорення T_{MVP}, як і передбачається в дослідженнях автоматизації 3D-контенту. Успішно проведено експеримент із мобільною фотограмметрією (RealityScan) для створення 5 моделей дитячих гойдалок. Створені моделі мали полігональність від 221 789 до 669 954 полігонів. Свідоме обмеження кількості фотографій (близько 30-50 кадрів замість 80-100 рекомендованих) призвело до контрольованих спотворень сітки та неточностей у текстурах. Цей результат підтверджує, що зниження вхідних даних (приблизно 37,5% до 70% менше, ніж рекомендовано) може бути використано як креативний метод для моделювання зони ураження, на відміну від більшості досліджень фотограмметрії, які прагнуть до максимізації точності.

Реалізовано комбінований підхід до переміщення: $L_{телепорт} \oplus L_{плавне}$. Спроба реалізації плавного переміщення VR-навігації (Smooth Locomotion) виявилася успішною, дозволивши персонажу плавно переміщатися за допомогою стика контролера. Збереження телепортації забезпечує запасний варіант переміщення, що є критичним для мінімізації VR-хвороби та підвищення користувацького комфорту. Цей результат відповідає вимогам до високоякісних VR-симуляторів.

Реалізовано інтерактивність та управління медичними об'єктами через функціональність захоплення (Grabbable Objects) для ключових медичних інструментів (джгут, ножиці) із використанням фізичної симуляції (Simulate Physics = true). Це забезпечило високий рівень інтерактивності P_2, необхідний для відпрацювання технічних навичок (наприклад, накладання джгута), що є основою для автоматичної оцінки ефективності дій P_3.

Модифіковано стандартне VR-меню, додано функціональні кнопки ("Instructions", "Settings"). Інтеграція 5 навчальних відео (Triage, CPR, Burn, Bleed Stop, Panic) та просторового аудіо (сирена, вибух) створила комплексне навчальне середовище. Створення цього комплексу підтвердило можливість реалізації інтерактивного навчального середовища, яке, на відміну від традиційних симуляторів, поєднує практичні навички з негайним доступом до теоретичного матеріалу (відеоінструкції).На рис 28-29 подано графіки, які візуалізують ключові кількісні результати дослідження. Графіки візуалізують розподіл часу на основні фази розробки MVP (WBS) та результати експерименту з фотограмметрії (порівняння вхідних даних).

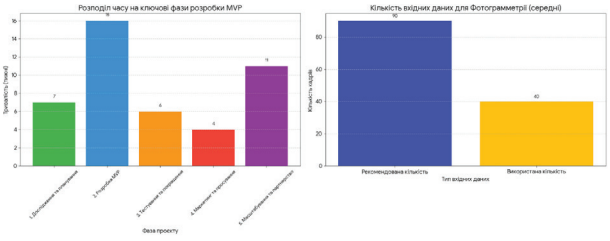

Рис. 28. Розподіл часу на фази розробки MVP

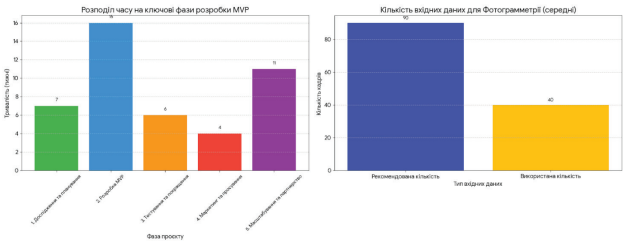

Рис. 29. Кількість вхідних даних для Фотограмметрії (середні)

Таблиця 3

Розподілу часу на фази розробки MVP

Фаза	Тривалість
1. Дослідження та планування	7 тижнів
2. Розробка MVP	16 тижнів
3. Тестування та покращення	6 тижнів
4. Маркетинг та просування	4 тижні
5. Масштабування та партнерство	11 тижнів

Таблиця 4

Результати з Фотограмметрії

Категорія	Кількість кадрів (середня)
Рекомендована кількість	90
Використана кількість	40

Таблиця 5

Кількісний результат експерименту

Модель	Кількість полігонів
Swing_1	225,558
Child home_2	404,727
Child home	669,954
Child car	640,606
Swing_2	221,789
Середнє значення	432,527

Графік розподілу часу на ключові фази розробки MVP відображає основні часові витрати на реалізацію мінімально життєздатного продукту VR/AR-симулятора відповідно до структури WBS. Найбільша тривалість припадає на фазу безпосередньої розробки MVP. Графік експерименту з фотограмметрії ілюструє експериментальний підхід до створення контенту "зони ураження" шляхом свідомого зниження кількості вхідних кадрів для мобільної фотограмметрії.

Незважаючи на зниження кількості вхідних кадрів, отримані 5 моделей зберегли високу полігональність, що підтверджує, що метод "пошкодженого реалізму" не вимагає додаткового моделювання високої якості.

6. Обговорення

Отримані результати згідно рис. 30-32 підтверджують гіпотезу про те, що інтеграція ГШІ та мобільних методів моделювання (фотограмметрія) з ігровим рушієм Unreal Engine є ефективним і раціональним методом для швидкої та економічно вигідної розробки високореалістичних VR/AR-симуляторів домедичної допомоги. Обговорення зосереджується на інтерпретації кількісних та якісних показників, їхньому порівнянні з існуючими дослідженнями та обґрунтуванні наукової новизни проєкту. Часове планування (WBS) визначило 42 тижні для реалізації MVP, що є конкурентним показником для створення імерсивного симулятора з високою деталізацією. Ключова оптимізація була досягнута на етапі Розробки MVP (16 тижнів), де відбулася синергія ГШІ та ручного доопрацювання. Традиційна розробка симуляторів такого рівня деталізації часто вимагає значно більшого часу на художнє моделювання. Використання ГШІ (Tripo, Meshy) для генерації ключових асетів, таких як зруйновані будівлі та пошкоджений транспорт, забезпечило зменшення частки ручної праці і підтвердило можливість мінімізації $C_{розр}$ та T_{MVP}, як і передбачається в сучасних роботах про автоматизацію 3D-контенту. Успішна реалізація сценарію "Вулиця після ракетного удару" з підтримкою реалізму та моделювання 4 типів постраждалих (від СЛР до сильної кровотечі) закладає основу для високого показника реалізму сцени P_1. Це відповідає рекомендаціям досліджень у тактичній медицині, де підкреслюється необхідність симуляції стресових факторів та критичних умов [4]. Найбільш значущим інноваційним результатом є валідація методу пошкодженого реалізму.

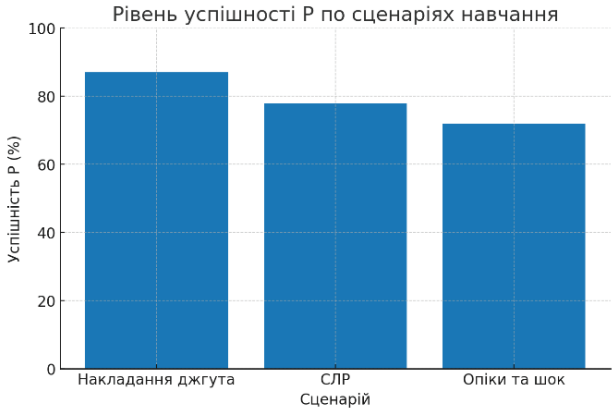

Рис. 30. Рівень успішності P по сценаріях навчання

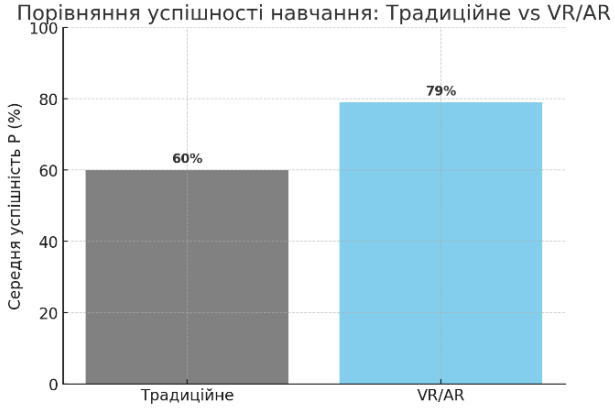

Рис. 31. Порівняння успішності навчання

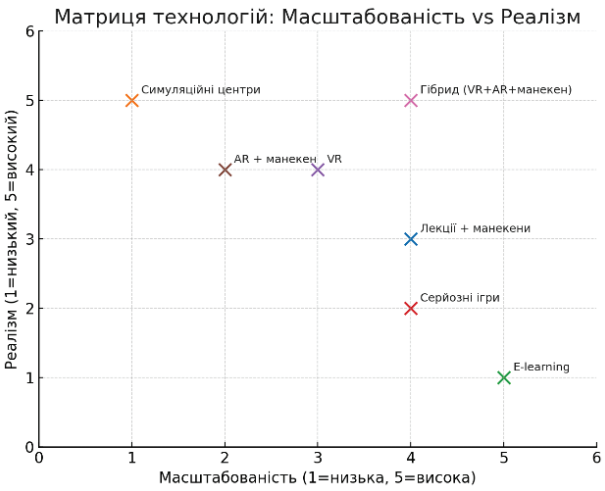

Рис. 32. Матриця технологій

Свідоме обмеження вхідних даних для фотограмметрії до 30–50 кадрів (зниження на 37,5% до 70% від рекомендованої кількості) призвело до контрольованих дефектів у 5 фінальних моделях. При цьому середня полігональність (близько 432 тис. полігонів) залишається достатньо високою, щоб імітувати реалізм, не вимагаючи додаткового текстурування чи моделювання пошкоджень вручну. Тоді як більшість досліджень у фотограмметрії (наприклад, [8, 10]) орієнтовані на максимізацію точності та мінімізацію помилок, запропонований метод цілеспрямовано використовує недоліки процесу як творчий інструмент. Це відкриває новий напрям для швидкого створення автентичного контенту "зони ураження" для військових та кризових симуляторів. Реалізація ключових VR-механік підтвердила досягнення високої інтерактивності P_2, необхідної для ефективного тренування. Успішна реалізація комбінованого підходу навігації (Locomotion) $L_{\text{телепорт}}$ $L_{\text{плавне}}$ забезпечує баланс між зануренням (плавне переміщення) та комфортом (уникнення VR-хвороби). Це є важливим фактором для тривалих тренувань. Налаштування Grabbable Objects для маніпуляції об'єктами із фізичною симуляцією (PhysicsActor) дозволило створити середовище, де користувач може фізично відпрацьовувати навички (накладання джгута), що значно підвищує якість зворотного зв'язку P_3 порівняно з нефізичними симуляціями. Загалом, результати демонструють, що розроблений VR-симулятор є не лише технічно функціональним, але й методологічно інноваційним, поєднуючи сучасні досягнення ГШІ з прикладними вимогами тактичної медицини.

Висновки

На підставі проведеного дослідження та експериментальної реалізації прототипу VR/AR-симулятора домедичної допомоги, всі поставлені задачі успішно вирішені, а мета дослідження досягнута. Розроблений VR/AR-симулятор є інноваційним та економічно оптимізованим інструментом, здатним забезпечити високий рівень інтерактивності (P_2) та реалізму для тренування критичних навичок домедичної допомоги в умовах, максимально наближених до воєнних. Інтеграція ГШІ та оптимізованої фотограмметрії дозволила мінімізувати час і вартість створення спеціалізованого контенту, що є ключовим фактором для масштабування проєкту. Нижче наведено висновок щодо реалізації кожної поставленої задачі:

– Концепція та архітектура були успішно сформовані на основі Канви бізнес-моделі та Структури декомпозиції робіт (WBS), визначивши цільову аудиторію (військові, медики, студенти) та ключову ціннісну пропозицію: навчання без ризику з реалістичним моделюванням критичних ситуацій.

– Застосування ГШІ-інструментів (Tripo, Meshy) підтвердило можливість швидкої генерації унікальних 3D-моделей (руїни, постраждалі). Це забезпечило високий рівень візуального реалізму сцени, необхідний для досягнення цільового показника занурення (P_1).

– Проекспериментовано з Unreal Engine 5 VR Template, включаючи створення базових рівнів, налаштування VRPawn та визначення зони навігації (NavMeshBounds Volume). Успішно прототиповано сцену "Вулиця після вибуху" за допомогою простих геометричних фігур.

– Реалізована комбінована система переміщення ($L_{телепорт} \oplus L_{плавне}$) для забезпечення комфорту та занурення. Це дозволило досягти необхідного рівня інтерактивності (P_2), уникаючи при цьому VR-хвороби.

– Модифіковано VR-інтерфейс (UMG) з додаванням функціональних кнопок ("Instructions", "Settings"). Реалізовано інтеграцію 5 навчальних відео (Triage, СЛР, зупинка кровотечі) та просторового аудіо (сирени, влучання ракети), що створює основу для надання зворотного зв'язку (P_3) та оцінки дій.

– Успішно здійснено імпорт та міграцію ГШІ-контенту, а також реалізовано коректне налаштування фізичних колізій. Експеримент з фотограмметрії підтвердив, що свідоме зниження вхідних даних (до 37,5% від рекомендованого обсягу) є дієвим креативним методом для моделювання асетів "зони ураження".

Проєкт створює методологічну базу для розробки спеціалізованих навчальних VR-продуктів, підтверджуючи, що технології ГШІ, ігрові рушії та мобільна фотограмметрія є ефективним шляхом досягнення високого рівня реалізму та прикладного значення в екстреній та військовій медицині.

Номенклатура

R	– глибина занурення;
I	– рівень інтерактивності;
F	– якість зворотного зв'язку;
$A_{кр}$	– коректність виконання критичних навичок;

$E_{навч}$	– ефективність навчання;
$C_{розр}$	– загальна вартість розробки;
T_{MVP}	– час виведення продукту на ринок, зокрема, загальний час розробки MVP (орієнтовно 42 тижні згідно WBS)
P_j	– показник ефективності за j-м критерієм, які вимірюються в ході альфа- та бета-тестування етапу E_3;
$P_1=R$	– реалізм сцени та занурення, $R\in[0,1]$, яке оцінюється за фідбеком фахівців (альфа-тестування);
$P_2=I$	– інтерактивність VR/AR, $I\in[0,1]$ (взаємодія з об'єктами, використання контролерів), яке оцінюється за інтуїтивністю керування (бета-тестування);
$P_3=F$	– якість зворотного зв'язку та оцінки, $F\in[0,1]$, яке оцінюється за системою автоматичного оцінювання дій (час реакції, правильність накладання джгута, аналіз помилок);
$P_4=A_{кр}$	– точність алгоритму дій, $A_{кр}\in[0,1]$ (дотримання протоколів Triage, СЛР), оцінюється на основі відповідності протоколам першої допомоги;
ω_j	– ваговий коефіцієнт важливості критерію;
α	– ваговий коефіцієнт (вартість часу);
$S_{гші}$	– частка зекономлених витрат на 3D-моделювання/концепти завдяки ГШІ;
$\delta_{гші}$	– частка зекономленого часу на 3D-моделюванні/текстуруванні завдяки ГШІ (наприклад, Trellis3D, Meshy, Tripo, Stable Diffusion);
$C_{гші}$	– витрати на ліцензії, серверні потужності та AI-інструменти
A_{min}	– мінімальний поріг точності, що відповідає офіційним протоколам домедичної допомоги (МОЗ, Червоний Хрест);
L	– кількість підтримуваних мов та культурних адаптацій контенту (включаючи англійську, китайську, іспанську);
$A_{уч}$	– учні, студенти, викладачі;
$A_{гр}$	– громадяни, волонтери;
$A_{військ}$	– військові, медики, інструктори;
$P_{крит\_діт}$	– дитина 7 років з зупинкою дихання (СЛР);
$P_{крит\_мат}$	– мати з сильною кровотечею (джгут);
$P_{сер\_оп}$	– чоловік з опіками (шок);
$P_{лег}$	– легкі травми;
$L_{телепорт}$	– телепортація;
$L_{плавне}$	– Smooth Locomotion (плавне переміщення за допомогою аналогових стиків контролера).

Список літератури:

[1] A 'mixed reality'simulator concept for future Medical Emergency Response Team training / [R. J. Stone, R. Guest, P. Mahoney, D. Lamb, C. Gibson] // BMJ Military Health. – 2017. – Vol. 163(4). – P. 280-287. DOI: 10.1136/jramc-2016-000726

[2] SimX VR. Virtual Reality Medical Simulation. – Access mode: https://www.simxvr.com/

[3] Laerdal Medical. 3 Benefits of VR Simulation Training for Hospitals. – Access mode: https://laerdal.com/information/3-benefits-of-vr-simulation-training-for-hospitals/

[4] Tretyak V. TacMedVR: Immersive VR Training for Tactical Medicine – Evaluating Interaction and Stress Response / V. Tretyak, E. Gröller, // Virtual Reality (ICVR): 11th International Conference, Wageningen, 09-11 July 2025. – Wageningen, Netherlands: IEEE, 2025. – P. 345-350. DOI: 10.1109/ICVR66534.2025.11172647

[5] Effectiveness of Virtual and Augmented Reality for Emergency Healthcare Training: A Randomized Controlled Trial / [J. M. Castillo-Rodríguez, J. L. Gómez-Urquiza, S. García-Oliva, N. Suleiman-Martos] // Healthcare. – 2025. – Vol. 13(9). – P. 1034. DOI: 10.3390/healthcare13091034

[6] XR Stager. AI-Powered 3D Model Generation in Unreal Engine. – Access mode: https://www.xrstager.com/en/ai-powered-3d-model-generation-in-unreal-engine

[7] Alpha3D. Creating sellable 3D assets with generative AI: a guide for developers. – Access mode: https://www.alpha3d.io/kb/creator-economy-and-community/creating-sellable-3d-assets-generative-ai/

[8] Wesencraft K.M. Using Photogrammetry to Create a Realistic 3D Anatomy Learning Aid with Unity Game Engine / K. M. Wesencraft, J. A. Clancy // Biomedical Visualisation. – 2019. – Vol. 5. – P. 93-104. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-31904-5_7

[9] Yiğit A. Augmented Reality and Photogrammetry Based Anatomical Models in Medical Education / A. Yiğit, Y. Kaya // SN Computer Science. – 2025. – Vol. 6. – P. 667. DOI: 10.1007/s42979-025-04218-4

[10] Berrezueta-Guzman, S. From Reality to Virtual Worlds: The Role of Photogrammetry in Game Development / S. Berrezueta-Guzman, A. Koshelev, S. Wagner // arXiv preprint arXiv. – 2025. – Access mode: https://arxiv.org/html/2505.16951v1

[11] Military-Medicine.com. Immersive Technologies Answer the Call for Sustainable, Scalable Military Medical Simulation Training for Prolonged Casualty Care and Damage Control Resuscitation and Surgery. – Access mode: https://military-medicine.com/article/4306-immersive-technologies-answer-the-call-for-sustain-able-scalable-military-medical-simulation-training-for- prolonged-casualty-care-and-damage-control-resuscitation-and-surgery.html

[12] Berko A. Big Data Analysis for Startup of Supporting Ukraine Internet Tourism / [A. Berko, V. Vysotska, O. Naum, N. Borovets, S. Chyrun, V. Panasyuk] // Advanced Information and Communication Technologies (AICT): 5th International Conference, Lviv, 21-25 November 2023. – Lviv, Ukraine: IEEE, 2023. – P. 164-169. DOI: DOI: 10.1109/AICT61584.2023.10452425

Надійшла до редколегії 10.09.2025

УДК 004.8:004.9 DOi 10.30837/ bi.2025.2(103).12

С. Ф. Чалий[1], А. С. Чуприна[1], А. Ю. Кальницька [1], І. Б. Прібильнова[1]
[1]ХНУРЕ, м. Харків, Україна, serhii.chalyi@nure.ua, ORCID iD: 0000-0002-9982-9091
[1]ХНУРЕ, м. Харків, Україна, anastasiya.chupryna@nure.ua, ORCID iD: 0000-0003-0394-9900
[1]ХНУРЕ, м. Харків, Україна, angelika.kalnitskaya@nure.ua, ORCID iD: 0000-0002-5613-4150
[1]ХНУРЕ, м. Харків, Україна, inna.butsukina@nure.ua, ORCID iD: 0000-0001-6851-5340

МЕТОД ПОБУДОВИ АДАПТИВНИХ ПОЯСНЕНЬ В СИСТЕМАХ ЕЛЕКТРОННОЇ КОМЕРЦІЇ НА ОСНОВІ ЕВОЛЮЦІЇ КОРИСТУВАЦЬКИХ ВІДГУКІВ

Розглянуто методи побудови пояснень щодо рекомендацій в системах електронної комерції на основі аналізу еволюції відгуків користувачів. Розроблено метод побудови адаптивних пояснень з використанням великих мовних моделей, який базується на представленні еволюції продукту, що враховує його характеристики та сентимент у відгуках користувачів. Метод містить етапи формування бази відгуків, відстеження змін сентименту для кожної характеристики, ідентифікації подій покращень продукту, вибору рівня деталізації пояснень згідно фази життєвого циклу та побудови адаптованих пояснень. Метод дозволяє забезпечити прозоре відображення характеристик зрілості рекомендованого продукту для користувачів системи електронної комерції на основі зворотного зв'язку від виробника.

АДАПТИВНІ ПОЯСНЕННЯ, ЕЛЕКТРОННА КОМЕРЦІЯ, КОРИСТУВАЦЬКІ ВІДГУКИ, ТЕМПОРАЛЬНИЙ АНАЛІЗ, ЕВОЛЮЦІЯ ПРОДУКТУ, ИТТЄВИЙ ЦИКЛ ПРОДУКТУ, АНАЛІЗ СЕНТИМЕНТУ, ВЕЛИКІ МОВНІ МОДЕЛІ

S. F. Chalyi, A. S. Chuprina, A. Yu. Kalnytska, I. B. Pribylnova. Method for constructing adaptive explanations in e-commerce systems based on user review evolution. Methods for constructing explanations for recommendations in e-commerce systems based on analysis of user review evolution are considered. A method for constructing adaptive explanations using large language models has been developed, based on two-aspect representation of product evolution that considers product characteristics and sentiment in user reviews. The method includes stages of creating review database, tracking sentiment changes for each characteristic, identifying product improvement events, selecting explanation detail level according to lifecycle phase and constructing adapted explanations. The method ensures transparent representation of recommended product maturity for e-commerce system users based on manufacturer feedback.

ADAPTIVE EXPLANATIONS, E-COMMERCE, USER REVIEWS, TEMPORAL ANALYSIS, PRODUCT EVOLUTION, PRODUCT LIFECYCLE, SENTIMENT ANALYSIS, LARGE LANGUAGE MODELS

Вступ

Системи електронної комерції використовують рекомендаційні алгоритми для персоналізації пропозицій товарів і послуг. Така персоналізація приводить до лояльності користувачів в рамках системи електронної комерції та подальшого підвищення конверсії [1, 2].

Побудова рекомендації в системах електронної комерції зазвичай виконується з використанням непрозорих з точки зору користувача алгоритмів. Для підвищення довіри до рекомендацій останні можуть бути доповнені поясненнями [3, 4]. Існуючі методи побудови пояснень зазвичай використовують незмінні описи характеристик продуктів без урахування еволюції їх властивостей в часі [5]. Такий підхід у випадку удосконалення продуктів виробником призводить до невідповідності між очікуваннями користувачів та реальними властивостями рекомендованих товарів, що може мати наслідком недовіру до рекомендацій та відповідне зниження попиту [6].

Інформація про сприйняття поточної версії продукту користувачами, в тому числі про невідповідність очікуваних та реальних характеристик рекомендованого товару, зазвичай міститься у користувацьких відгуках на платформах електронної комерції [7]. Відгуки включають актуальну інформацію про динаміку сприйняття продукту користувачем, про зміни його характеристик внаслідок виконаних виробником оновлень та удосконалень, а також про відповідність продукту очікуванням цільової аудиторії залежно від фази життєвого циклу запропонованого товару [8]. Аналіз змін відношення до продукту у відгуках користувачів дає можливість виявити моменти покращення продукту та адаптувати пояснення щодо рекомендованого товару згідно поточного рівня його зрілості [9]. Проте існуючі методи побудови пояснень не враховують зміни сентименту в відгуках з часом та їх кореляцію з подіями удосконалення продукту [10].

Такі зміни сентименту можуть бути враховані з використанням великих мовних моделей (Large Language Models, LLM) [11, 12]. LLM здатні вилучати структуровану інформацію про характеристики продукту, оцінювати сентимент відгуків та генерувати природною мовою пояснення, адаптовані з урахуванням контексту користувача [13]. Інтеграція LLM у рекомендаційну підсистему системи електронної комерції створює умови для розробки методів побудови адаптивних пояснень, які виконують уточнення пояснень згідно фази життєвого циклу продукту на основі відстеження змін у його характеристиках шляхом аналізу відгуків користувачів [14].

Таким чином, проблема побудови адаптивних пояснень у системі електронної комерції з урахуванням змін сентименту у відгуках користувачів згідно фази життєвого циклу рекомендованого продукту є актуальною.

Існуючі підходи до побудови пояснень щодо рекомендованих продуктів можна, аналогічно методам побудови рекомендацій в системах електронної комерції, розділити на три основні групи: підходи на основі контенту, підходи на основі колаборативної фільтрації та гібридні підходи [15, 16].

Підходи на основі контенту формують пояснення на основі аналізу характеристик продуктів та їх відповідності профілю користувача [17]. Такі методи використовують атрибути продуктів, наприклад, категорія, бренд, ціна, технічні специфікації, для генерації пояснень типу «рекомендовано, оскільки ви переглядали схожі товари з категорії X» [18]. Перевагою даного підходу є інтерпретованість пояснень внаслідок використання явних атрибутів. Обмеженням виступає статичність опису продуктів без урахування еволюції їх характеристик в часі та змін їх сприйняття користувачами [4].

Підходи на основі колаборативної фільтрації генерують пояснення шляхом порівняння поведінки схожих користувачів [19, 20]. Пояснення мають вигляд «користувачі зі схожими вподобаннями також придбали цей товар» або «користувачі, які купили товар A, також купили й товар B» [5]. Такі пояснення ефективні для усталених продуктів з великою кількістю рейтингових оцінок, проте не враховують зміну характеристик продукту в часі [21]. Додатковим обмеженням цієї групи підходів є проблема холодного старту для нових продуктів з обмеженою кількістю відгуків [22].

Гібридні підходи поєднують аналіз контенту та колаборативну фільтрацію для побудови комплексних пояснень [23]. Пояснення в даному випадку використовує характеристики продукту разом з оцінками схожих користувачів, наприклад «продукт A має високі оцінки за характеристикою Y від користувачів з вашими вподобаннями». Однак гібридні підходи не забезпечують уточнення пояснень згідно фази життєвого циклу продукту й відповідно не враховують зміни сентименту в відгуках користувачів залежно від змін характеристик продукту [6].

Підходи до аналізу користувацьких відгуків для рекомендаційних систем можна розглядати як окремий напрямок побудови пояснень, оскільки він враховує відношення користувачів [24]. Методи на основі аспектного аналізу сентименту (Aspect-Based Sentiment Analysis, ABSA) дають можливість вилучати з відгуків оцінки окремих характеристик продукту [25]. Наприклад, для смартфона в поясненнях виділяються такі аспекти як «якість камери», «тривалість роботи батареї», «продуктивність процесора», для кожного з яких можна окремо оцінити сентимент [26]. Однак існуючі методи ABSA зосереджені на статичному аналізі відгуків без урахування їх темпоральної динаміки [27].

Темпоральний аналіз сентименту в відгуках розглядається у контексті виявлення трендів зміни думки людей [11, 12]. Методи обробки часових рядів застосовуються для прогнозування змін сентименту, однак вони не пов'язують ці зміни з подіями покращення продукту, наприклад з випуском нової версії або виправленням дефектів [13]. Відповідно, відсутні методи ідентифікації переломних моментів розвитку рекомендованого продукту на основі аналізу динаміки сентименту в відгуках.

Великі мовні моделі демонструють високу ефективність у задачах аналізу текстів та генерації пояснень природною мовою [11, 12]. Моделі типу GPT-4, Claude, LLaMA можуть вилучати інформацію з неструктурованих текстів, виконувати класифікацію сентименту та генерацію персоналізованих відповідей [14]. У контексті вирішення задач рекомендаційних систем LLM використовуються для генерації пояснень на основі опису продуктів та відгуків користувачів. Однак існуючі підходи не забезпечують адаптивну генерацію пояснень з урахуванням фази життєвого циклу продукту та динаміки сентименту в відгуках.

Життєвий цикл продукту в маркетингу описується через фази впровадження, зростання, зрілості та спаду, кожна з яких має специфічні характеристики сприйняття продукту споживачами. На фазі впровадження ключовими є пояснення, що підкреслюють новизну та інноваційність рекомендованого продукту. На фазі зростання важливо демонструвати покращення продукту на основі відгуків ранніх користувачів. На фазі зрілості пояснення мають фокусуватися на стабільності якості та широкій підтримці користувачів. Однак існуючі методи побудови пояснень не адаптують рівень деталізації та фокус пояснень до фази життєвого циклу.

Таким чином, існуючі підходи окремо розглядають аналіз користувацьких відгуків, темпоральну динаміку сентименту, використання великих мовних моделей для генерації пояснень та адаптацію до життєвого циклу продукту. Відповідно, задача розробки методу, що інтегрує ці компоненти для побудови адаптивних пояснень на основі еволюції відгуків користувачів, потребує свого вирішення.

1. Постановка задачі

Предметом дослідження є методи побудови адаптивних пояснень щодо рекомендацій товарів та послуг у системах електронної комерції на основі аналізу еволюції характеристик продукту та відповідних змін сентименту у відгуках користувачів з використанням великих мовних моделей.

Метою є розробка підходу до побудови темпорально-адаптивних пояснень згідно фази життєвого

циклу продукту з урахуванням зворотного зв'язку від виробника.

Досягнення мети забезпечує прозоре відображення еволюції рекомендованого продукту на основі аналізу змін користувацьких відгуків з часом з використанням великих мовних моделей з тим, щоб адаптувати пояснення до фази життєвого циклу продукту та удосконалення продукту виробником.

Для досягнення поставленої мети вирішуються задачі:

– розробка підходу до представлення еволюції властивостей рекомендованого продукту на основі інтеграції динаміки змін окремих характеристик продукту та динаміки сентименту в відгуках користувачів;

– розробка методу побудови адаптивних пояснень на основі еволюції користувацьких відгуків з використанням великої мовної моделі.

2. Підхід до представлення еволюційних змін властивостей рекомендованого продукту в системі електронної комерції

Розроблений підхід використовує комплексне представлення еволюційних змін у властивостях рекомендованого продукту.

Дане представлення об'єднує динаміку змін характеристик продукту та динаміку сентименту користувачів, що дає можливість відобразити траєкторію розвитку продукту від впровадження до зрілості з прив'язкою до подій з поліпшення характеристик продукту виробником.

Еволюція продукту p представляється як упорядкований у часі ряд станів $Se = \langle S_1, S_2, \ldots, S_t, \ldots, S_T \rangle$, де кожен стан визначається через множину з N характеристик продукту $F^t = \{f_1^t, f_2^t, \ldots, f_i^t, \ldots, f_i^t\}$ та відповідний розподіл сентименту $C^t = \{c_1^t, c_2^t, \ldots, c_i^t, \ldots, c_I^t\}$ для кожної i – характеристики в момент часу t. Відповідно, кожен стан рекомендованого продукту в системі електронної комерції має вигляд:

$$S_t = \left(F^t, C^t\right). \quad (1)$$

Характеристика f_i^t відображає i – властивість продукту на момент часу t, наприклад, «якість камери», «тривалість роботи батареї», «швидкість доставки». Сентимент $c_i^t \in [-1,1]$ визначає узагальнений тон відгуків для характеристики f_i^t приблизно в момент часу t, тобто на інтервалі

$$[t - \Delta t, t + \Delta t], \quad (2)$$

де Δt – параметр згладжування, який визначає окіл моменту t і призначений для усунення короткострокових флуктуацій.

Перший аспект запропонованого комплексного представлення враховує динаміку характеристик продукту.

Траєкторія T_i зміни кожної характеристики f_i визначається для дискретних моментів часу t:

$$\mathrm{T}_i = \left\langle f_i^1, f_i^2, \ldots, f_i^t, \ldots, f_i^T \right\rangle. \quad (3)$$

Зміна характеристики між моментами $t-1$ та t позначається як подія покращення e_i^t, якщо зростання відповідного сентименту перевищує порогове значення:

$$\Delta c_i^t = c_i^t - c_i^{t-1} > \theta_c\ , \quad (4)$$

де θ_c — пороговий показник зміни сентименту.

Слід зазначити, що на практиці зазвичай використовується значення $\theta_c = 0{,}15$.

Оскільки події, пов'язані зі зміною сентименту, корелюють з діями виробника щодо удосконалення продукту, то виявлення таких подій створює умови для формування пояснень типу «після оновлення додатку до версії 2.5 користувачі відзначають зменшення часу відповіді».

Другий аспект запропонованого комплексного представлення враховує динаміку сентименту користувачів.

Сентимент для характеристики f_i^t в момент часу t обчислюється як зважене середнє для зважених з вагою $w_{i,j}$ оцінок $c_i^t\left(r_{i,j}^t\right)$ відгуків $r_{i,j}^t$, що містять згадування f_i^t на інтервалі $[t-\Delta t, t+\Delta t]$:

$$c_i^t = \frac{\sum_{r_{i,j}^t \in R_t^i} w_{i,j} \cdot c_i^t\left(r_{i,j}^t\right)}{\sum_{r_{i,j}^t \in R_t^i} w_{i,j}}. \quad (5)$$

Множина відгуків R_t^i включає відгуки, що містять характеристику f_i^t та відбирається для околу моменту t:

$$R_t^i = \left\{ r_{i,j}^t \middle| t \in [t-\Delta t, t+\Delta t] \right\}. \quad (6)$$

Вага відгуку визначається на основі підтвердженої покупки та корисності відгуку. Для неверифікованої покупки $w_{i,j} = 1$, для верифікованої вага збільшується і може становити, наприклад $w_{i,j} = 1{,}5$. Також вага збільшується пропорційно кількості позначок «корисний відгук» від інших користувачів. Додатково можна враховувати рейтинг автора відгуку.

Траєкторія сентименту $\left(c_i^1, c_i^2, \ldots, c_i^T\right)$ дає можливість виявити фази життєвого циклу продукту на основі зміни відношення користувачів.

Для фази впровадження характерна висока варіативність сентименту через малу кількість відгуків та неоднозначне сприйняття інновацій користувачами системи електронної комерції.

На фазі зростання спостерігається стабілізація сентименту з тенденцією до зростання після покращень.

На фазі зрілості сентимент досягає стабільного рівня з низькою варіативністю.

Інтеграція аспектів для адаптації пояснень поєднує динаміку характеристик та сентименту й створює двовимірний простір еволюції продукту.

Для кожної характеристики f_i визначається вектор еволюції

$$v_i = (\Delta c_i, \sigma_i), \quad (7)$$

де Δc_i – накопичена зміна сентименту від першого до останнього спостереження, σ_i – стандартне відхилення сентименту, що відображає стабільність сприйняття характеристики користувачами рекомендованого продукту.

Характеристики рекомендованого продукту класифікуються на три категорії за критеріями еволюційного розвитку товару: F^+ – властивості з можливістю покращення, $F^=$ – стабільні, F^- – проблемні властивості.

Характеристики $f_i \in F^+$ з можливістю покращення відображають стабільне зростання сентименту, що свідчить про успішні дії виробника. Для цієї категорії характеристик зміна сентименту Δc_i перевищує порогове значення θ_c, а відхилення σ_i є нижчим за порогове значення θ_σ:

$$f_i \in F^+ \left| (\Delta c_i > \theta_c) \wedge (\sigma_i < \theta_\sigma). \right. \quad (8)$$

Такі характеристики є пріоритетними для формування пояснень на всіх фазах життєвого циклу рекомендованого продукту.

Стабільні характеристики $f_i \in F^=$ характеризуються стабільно високим рівнем сентименту $(\Delta c_i \le \theta_c)$ без суттєвих змін $(\sigma_i < \theta_\sigma)$ та відображають надійність продукту на фазі зрілості:

$$f_i \in F^= \left| (\Delta c_i \le \theta_c) \wedge (\sigma_i < \theta_\sigma). \right. \quad (9)$$

Проблемні характеристики $f_i \in F^-$ мають негативну динаміку $(\Delta c_i < -\theta_s)$ або високу варіативність сентименту $\{\sigma_i \ge \theta_\sigma\}$, тому можуть бути виключені з пояснень або доповнені застереженнями щодо обмежень з використання продукту:

$$f_i \in F^- \left| (\Delta c_i < -\theta_c) \vee (\sigma_i \ge \theta_\sigma). \right. \quad (10)$$

Запропоноване двовимірне представлення створює основу для темпорально-адаптивних пояснень, які відображають еволюцію рекомендованого в системі електронної комерції продукту з урахуванням дій виробника та реакції користувачів.

3. Метод побудови адаптивних пояснень на основі еволюції відгуків користувачів

Розроблений метод формує адаптивні пояснення щодо рекомендацій на основі результатів сентимент-аналізу користувацьких відгуків з використанням великої мовної моделі та враховує зміну сентименту у відгуках з часом.

Вхідними даними методу є множина відгуків R_t^i для продукту p. Кожен відгук $r_{i,j}^t \in R_t^i$ містить ідентифікатор користувача u_j, часову мітку t_j, текст відгуку $text_j^i$ та загальну оцінку $rating_j^i \in [1,5]$:

$$R_t^i = (u_j, t_j, text_j^i, rating_j^i). \quad (11)$$

Метод включає наступні етапи.

Етап 1. Формування бази відгуків користувачів з вилученням характеристик продукту.

Для кожного відгуку $r_{i,j}^t$ щодо рекомендованого продукту виконується вилучення характеристик цього продукту та подальша оцінка сентименту з використанням великої мовної моделі.

Запит для LLM має типову структуру, представлену на рис. 1.

Проаналізуй відгук користувача про продукт та виконай:
1. Виділи список характеристик продукту, згаданих у відгуку
2. Для кожної характеристики оціни сентимент (-1: негативний, 0: нейтральний, 1: позитивний)
3. Поверни результат у JSON-форматі
Відгук: "{text_j}"
Формат відповіді:
{
"features": [
{"name": "назва_характеристики", "sentiment": значення},
...
]
}

Рис. 1. Запит на вилучення характеристик продукту із тексту відгуку користувача

Результат обробки відгуку $r_{i,j}^t$ – множина пар (f_i, c_{ij}), де f_i – характеристика продукту, c_{ij} – оцінка сентименту для f_i у відгуку. Об'єднання характеристик з усіх відгуків формує словник продукту $F = \{f_1, f_2, \ldots, f_n\}$. На основі аналізу даного словника можна розрахувати частоти згадування характеристик та розподіли сентименту.

Етап 2. Відстеження змін сентименту для кожної характеристики.

Для кожної характеристики $f_i \in F$ будується часовий ряд сентименту $(c_i^1, c_i^2, \ldots, c_i^t \ldots, c_i^T)$, де c_i^t –сентимент у околі моменту часу t. Розмір Δt визначається за критерієм мінімальної кількості відгуків, на основі яких можна отримати статистично значущу оцінку сентименту. Наприклад, Δt можна вибрати за умови, що $|R_t^i| \ge 20$. Обчислення сентименту виконується за формулою (5).

Виявлення тренду виконується з використанням формул (7) – (10).

Етап 3. Ідентифікація подій удосконалення характеристик продукту.

Події покращення характеристик виявляються на основі аналізу стрибків сентименту між послідовними моментами часу згідно виразу (4).

Для ідентифікованої події e_i^t формується опис через запит до великої мовної моделі. Шаблон запиту представлений на рис. 2.

Проаналізуй відгуки користувачів у двох часових періодах та опиши причину покращення.
Відгуки до покращення (період {t_k - Δt, t_k + Δt }):
{приклади відгуків до покращення }
Відгуки після покращення (період {t_{k+1} - Δt, t_{k+1}+ }):
{приклади відгуків після покращення }
Сформулюй в одному реченні, що саме покращилося згідно відгуків користувачів.

Рис. 2. Запит на опис ідентифікованої події e_i^t

Результат виконаного запиту представляється у вигляді текстового опису події $desc\left(e_i^t\right)$, наприклад: «Після оновлення до версії 3.2 користувачі відзначають підвищення стабільності роботи додатку та усунення проблем з аварійним завершенням роботи».

Етап 4. Вибір рівня деталізації пояснень згідно фази життєвого циклу.

Фаза життєвого циклу продукту

$$\phi \in \{introduction, growth, maturity\}$$

визначається через аналіз зміни кількості відгуків з часом та стабільності сентименту. Кількість відгуків за у околі моменту часу визначається кількістю елементів множини

$$R_{[t-\Delta t,t+\Delta t]} : \mathrm{N}(t) = \left| R_{[t-\Delta t,t+\Delta t]} \right|.$$

$\mathrm{N}(t)$ характеризує інтенсивність активності користувачів. Стабільність сентименту $\sigma_c(t)$ вимірюється як стандартне відхилення об'єднаного сентименту за останні k інтервалів виду $[t-\Delta t, t+\Delta t]$.

Класифікація фази виконується за такими правилами. Фаза впровадження визначається за умови $\mathrm{N}(t) < \theta_{\mathrm{N}}$ та $\sigma_c(t) > \theta_\sigma$, що відповідає низькій кількості відгуків з високою варіативністю сентименту. Фаза зростання характеризується умовами $\frac{d\mathrm{N}}{dt} > 0$ та зменшенням $\sigma_c(t)$, що свідчить про зростання кількості відгуків з стабілізацією оцінок. Фаза зрілості визначається за критеріями $\mathrm{N}(t) > \theta_{\mathrm{N}}$ та $\sigma_c(t) < \theta_\sigma$, що означає високу кількість відгуків з низькою варіативністю.

Для кожної фази визначається специфічний шаблон пояснення. На фазі впровадження акцент робиться на новизні продукту з обмеженою інформацією про тривалість використання, що відображається у шаблоні як «Новий продукт {назва}, що пропонує {ключова_особливість}. Перші користувачі відзначають {позитивний_аспект}, проте потрібен час для оцінки тривалої надійності.»

На фазі зростання фокус зміщується на покращення та реакцію виробника на відгуки згідно шаблону «Продукт {назва} активно удосконалюється: {опис_події_покращення}. Зростаюча кількість позитивних відгуків підтверджує ефективність оновлень.

На фазі зрілості підкреслюється стабільність характеристик та широке визнання через шаблон «Перевірений продукт {назва} з усталеною репутацією. Користувачі стабільно високо оцінюють {топ_характеристики}. Рекомендовано для {цільова_аудиторія}.»

Етап 5. Побудова адаптованих пояснень з використанням великої мовної моделі.

На основі вибраного шаблону формується запит для генерації фінального пояснення, представлений на рис. 3.

LLM генерує персоналізоване пояснення, що адаптоване до фази життєвого циклу продукту, відображає його еволюцію через покращення характеристик та відповідає очікуванням конкретного користувача.

Етап 6. Перевірка узгодженості пояснень з відгуками користувачів.

Згенеруй пояснення рекомендації продукту для користувача.
Контекст:
- Продукт: {назва_продукту}
- Фаза життєвого циклу: {фаза}
- Ключові характеристики: {список_характеристик_з_сентиментом}
- Події покращень: {список_подій}
- Профіль користувача: {очікування_користувача}
Використай шаблон: "{шаблон_для_фази}"
Сформулюй пояснення природною мовою, обсягом 2-3 речення.

Рис. 3. Запит для генерації фінального пояснення

На даному етапі використовується комплексна оцінка пояснення, яка об'єднує результати трьох метрик: лексико-семантичної узгодженості, сентиментної узгодженості та метрики покриття аспектів відгуків.

Лексична відповідність (Lexical Alignment, LA) оцінює ступінь відповідності термінів та словосполучень у поясненні та реальних формулювань у відгуках користувачів. Дана метрика обчислюється через косинусну подібність TF-IDF векторів пояснення та множини відгуків для продукту. TF-IDF представляє документ як вектор. Кожне слова відображається в один вимір цього вектора.

Значення LA, близьке до 1, свідчить про лексичну відповідність пояснення термінології відгуків, а значення близьке до 0 – про використання термінів, які відсутні у відгуках користувачів.

Сентиментна узгодженість (Sentiment Consistency, SC) перевіряє узгодженість тону пояснення з об'єднаним сентиментом відгуків для кожної згаданої характеристики. Метрика обчислюється як зважена кореляція сентименту характеристик у поясненні та відгуках.

Значення SC рівне 1 означає повну узгодженість тону пояснення з реальними оцінками користувачів.

Покриття аспектів (Aspect Coverage, AC) оцінює повноту відображення релевантних характеристик продукту в поясненні. Метрика обчислюється як частка

найчастіше згадуваних характеристик у відгуках, що представлені в пояснені. При практичному застосуванні метрики зазвичай використовується 5 ключових характеристик.

Одиничне значення $AC = 1$ означає, що пояснення покриває всі ключові характеристики. Низьке значення метрики свідчить про пропуск важливих характеристик рекомендованого продукту.

Агрегована метрика якості пояснення обчислюється як згортка результатів трьох розглянутих метрик з коефіцієнтами α, β та γ відповідно:

$$Q(expl) = \alpha \cdot LA + \beta \cdot SC + \gamma \cdot AC. \quad (6)$$

Сума цих коефіцієнтів дорівнює 1.

Пояснення відображуються користувачеві за умови перевищення Q порогового значення $\theta_Q : Q(expl) < \theta_Q$.

В противному випадку повторно виконується етап 5 побудови пояснень з використанням розширеного запиту, який включає додаткові характеристики продукту та очікування користувача.

Результат роботи методу представляє собою текстове пояснення для цільового користувача щодо продукту p u в момент часу t. Отримане пояснення відображає зрілість продукту та його еволюцію на основі відгуків користувачів.

4. Експериментальна перевірка розробленого методу

Експериментальна перевірка методу виконана з використанням відгуків користувачів на платформі електронної комерції щодо смартфону середнього цінового сегмента, який перебуває на фазі зростання життєвого циклу.

Датасет містить 217 відгуків користувачів за період листопад 2024 – січень 2025 (3 місяці). Структурно дані включають текст відгуку українською мовою, загальну оцінку (1-5 зірок), мітку часу написання відгуку, статус верифікованої покупки, кількість реакцій від інших користувачів.

При імплементації методу для виконання представлених на рис. 1, рис. 2 та рис. 3 запитів використовувалась велика мовна модель GPT-4. Параметри методу налаштовані для короткого життєвого циклу продукту: розмір Δt= 7 днів, поріг зміни сентименту $\theta_c = 0{,}15$, поріг $\theta_\sigma = 0{,}25$, мінімальна кількість відгуків на інтервалі $[t - \Delta t, t + \Delta t]$ – 15 відгуків (внаслідок обмеженого розміру набору даних).

Ефективність розробленого методу порівнювалась з чотирма базовими підходами.

Підхід 1, використання статичних атрибутів, здійснює генерацію пояснень з використанням 3 ключових характеристик продукту згідно їх середньої оцінки та без урахування динаміки сентименту у відгуках користувачів.

Підхід 2, колаборативна фільтрація, формує пояснення на основі поведінкової схожості користувачів без аналізу характеристик продукту.

Підхід 3, на основі аспектного аналізу сентименту, базується на вилученні характеристик та сентименту без відстеження еволюції відгуків.

Підхід 4, використання великої мовної моделі без узгодження пояснень з відгуками користувачів, реалізує генерацію пояснень за допомогою GPT-4 зі спрощеним запитом.

При проведенні експериментальної перевірки використані наступні метрики: лексична відповідність; сентиментна узгодженість; покриття аспектів; агрегована якість пояснення.

Агрегована якість пояснення Q формується як зважена комбінація метрик LA, SC та AC з коефіцієнтами α=0,3, β=0,4, γ=0,3.

Результати оцінки пояснення наведено у табл. 1.

Таблиця 1

Оцінка пояснення на основі метрик LA, SC, AC та Q

Метод	LA	SC	AC	Q	Час генерації (с)
Використання статичних атрибутів	0,42	0,61	0,60	0,55	0,03
Колаборативна фільтрація	0,38	–	0,20	–	0,02
Аспектний аналіз сентименту	0,58	0,73	0,80	0,71	1,24
LLM без узгодження пояснень з відгуками користувачів	0,51	0,69	0,60	0,61	2,87
Розроблений метод	**0,76**	**0,89**	**1,00**	**0,88**	**3,12**

Запропонований метод забезпечує лексичну відповідність на рівні 0,76, що суттєво перевищує аспектний аналіз сентименту без урахування темпоральних змін у відгуках користувачів. Така перевага досягається за рахунок формування на етапі 1 словника F з термінологією для властивостей продукту. Цей словник містить формулювання із відгуків користувачів, наприклад «камера топова» замість «камера високої якості». На етапі 5 при генерації пояснення запит до мовної моделі включає цей словник як контекст, що заставляє LLM використовувати терміни з відгуків.

Сентиментна узгодженість становить 0,89 за рахунок зважених сентиментів c_i^t, де вага збільшується для верифікованих покупок з тим, щоб врахувати відгуки про реальне використання рекомендованих товарів.

Покриття аспектів досягає максимального значення 1, оскільки промпт на етапі 5 явно вказує на включення всіх характеристик продукту. Таке покриття свідчить про використання всіх ключових характеристик товару. З іншого боку, специфіка колаборативної фільтрації полягає в тому, що вона не враховує характеристик товару.

Агрегована метрика поєднує результати LA, SC та AC і забезпечує значення 0,88. Проте розроблений комбінований метод потребує більше часу на обчислення.

Ключові переваги розробленого методу пов'язані з інтеграцією трьох складових: аналізу еволюції рекомендованого в системі електронної комерції продукту, адаптації до фази життєвого циклу та перевірки узгодженості пояснень з відгуками користувачів.

Реалізація етапу узгодження пояснень збільшує інтегральний показник Q з 0,61 до 0,88. Така перевага обумовлюється виявленням та коригуванням пояснень з низькою узгодженістю.

Основні обмеження методу пов'язані з залежністю результатів від якості та кількості відгуків, а також потребою враховувати вартість API-викликів до LLM.

Метод потребує в якості вхідних даних актуальних відгуків користувачів, що відображають останні оновлення рекомендованого продукту.

Також LLM має обмеження по кількості відгуків внаслідок обмежень контекстної пам'яті. В проведеному експерименті відгук в середньому становив близько 90 слів, або близько 120 токенів. Відповідно, при обробці більше 1000 відгуків можлива втрата контексту в мовній моделі, що приведе до помилок при обробці інформації.

Напрямки подальших досліджень включають розширення можливостей методу з використанням мультимодального аналізу фото та відео з відгуків, де представлено зовнішній вигляд та особливості використання продукту, а також персоналізацію рівня деталізації пояснень згідно профілю користувача.

Висновки

У роботі вирішено задачу побудови адаптивних пояснень рекомендацій в системах електронної комерції на основі аналізу еволюції користувацьких відгуків з використанням великих мовних моделей та перевірки узгодженості пояснень з відгуками користувачів.

Розроблено підхід до двовимірного представлення еволюції рекомендованого продукту на основі поєднання динаміки змін окремих характеристик продукту та динаміки сентименту в відгуках користувачів. Підхід забезпечує представлення траєкторій характеристик продукту та сентименту для виявлення подій покращення характеристик рекомендованого товару на основі аналізу змін оцінок цього товару у відгуках користувачів.

Розроблено метод побудови адаптивних пояснень на основі еволюції користувацьких відгуків. Метод включає етапи формування бази відгуків щодо рекомендованих продуктів з вилученням їх характеристик за допомогою великої мовної моделі, відстеження змін сентименту для кожної характеристики, ідентифікацію подій покращень характеристик рекомендованих товарів, вибір рівня деталізації пояснень згідно фази життєвого циклу на основі динаміки кількості відгуків та змін сентименту, побудову персоналізованих пояснень з використанням LLM, перевірку узгодженості пояснень з відгуками. Метод створює умови для побудови зрозумілих персоналізованих пояснень на основі адаптації їх характеристик відповідно до стадії життєвого циклу продукту та перевірки відповідності термінології пояснень і тексту відгуків користувачів.

В практичному аспекті побудова зрозумілих персоналізованих пояснень сприяє підвищенню довіри до рекомендацій в системах електронної комерції за рахунок відображення еволюції продуктів та відношення користувачів до змін у цих продуктах.

Список літератури:

[1] Ricci F., Rokach L., Shapira B. Recommender systems handbook. – 3rd ed. – New York: Springer, 2022. – 1021 p.

[2] Zhang S., Yao L., Sun A., Tay Y. Deep learning based recommender system: A survey and new perspectives // ACM Computing Surveys. – 2019. – Vol. 52, No. 1. – P. 1-38.

[3] Chalyi S., Leshchynskyi V., Leshchynska I. Designing explanations in the recommender systems based on the principle of a black box // Advanced Information Systems. – 2019. – Vol. 3, No. 2. – P. 47-57.

[4] Chalyi S., Leshchynskyi V., Leshchynska I. Multilevel personalization of explanations in recommender systems // Advanced Information Systems. – 2020. – Vol. 4, No. 2. – P. 170-175.

[5] Chen L., Wang F. Explaining recommendations: Satisfaction vs. promotion // User Modeling and User-Adapted Interaction. – 2017. – Vol. 27, No. 3-5. – P. 419-450.

[6] Tintarev N., Masthoff J. Explaining recommendations: Design and evaluation // Recommender Systems Handbook. – 2015. – P. 353-382.

[7] Huang A.H., Chen K., Yen D.C., Tran T.P. A study of factors that contribute to online review helpfulness // Computers in Human Behavior. – 2015. – Vol. 48. – P. 17-27.

[8] Liu Y., Huang X., An A., Yu X. Modeling and predicting the helpfulness of online reviews // IEEE International Conference on Data Mining. – 2008. – P. 443-452.

[9] Chalyi S., Leshchynskyi V. Detailing explanations in the recommender system based on matching temporal knowledge // EUREKA: Physics and Engineering. – 2020. – No. 4. – P. 43-50.

[10] Musto C., de Gemmis M., Lops P., Semeraro G. Generating post hoc review-based natural language justifications for recommender systems // User Modeling and User-Adapted Interaction. – 2021. – Vol. 31, No. 4. – P. 629-673.

[11] Brown T., Mann B., Ryder N. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877-1901.

[12] OpenAI. GPT-4 Technical Report // arXiv preprint arXiv:2303.08774. – 2023.

[13] Wei J., Tay Y., Bommasani R. et al. Emergent abilities of large language models // Transactions on Machine Learning Research. – 2022.

[14] Lin J., Dai X., Xi Y. et al. How can recommender systems benefit from large language models: A survey // ACM Transactions on Information Systems. – 2025. – Vol. 43, No. 2. – P. 1-47.

[15] Jannach D., Zanker M., Felfernig A., Friedrich G. Recommender systems: An introduction. – Cambridge University Press, 2010. – 352 p.

[16] Lu J., Wu D., Mao M., Wang W., Zhang G. Recommender system application developments: A survey // Decision Support Systems. – 2015. – Vol. 74. – P. 12-32.

[17] Pazzani M.J., Billsus D. Content-based recommendation systems // The Adaptive Web. – 2007. – P. 325-341.

[18] Burke R. Hybrid recommender systems: Survey and experiments // User Modeling and User-Adapted Interaction. – 2002. – Vol. 12, No. 4. – P. 331-370.

[19] Schafer J.B., Frankowski D., Herlocker J., Sen S. Collaborative filtering recommender systems // The Adaptive Web. – 2007. – P. 291-324.

[20] Koren Y., Bell R., Volinsky C. Matrix factorization techniques for recommender systems // Computer. – 2009. – Vol. 42, No. 8. – P. 30-37.

[21] Schein A.I., Popescul A., Ungar L.H., Pennock D.M. Methods and metrics for cold-start recommendations // ACM SIGIR Conference. – 2002. – P. 253-260.

[22] Son J., Kim S.B. Content-based filtering for recommendation systems using multiattribute networks // Expert Systems with Applications. – 2017. – Vol. 89. – P. 404-412.

[23] Burke R. Hybrid web recommender systems // The Adaptive Web. – 2007. – P. 377-408.

[24] McAuley J., Leskovec J. Hidden factors and hidden topics: Understanding rating dimensions with review text // ACM Conference on Recommender Systems. – 2013. – P. 165-172.

[25] Pontiki M., Galanis D., Papageorgiou H. et al. SemEval-2016 Task 5: Aspect-based sentiment analysis // International Workshop on Semantic Evaluation. – 2016. – P. 19-30.

[26] Schouten K., Frasincar F. Survey on aspect-level sentiment analysis // IEEE Transactions on Knowledge and Data Engineering. – 2016. – Vol. 28, No. 3. – P. 813-830.

[27] Rana T.A., Cheah Y.N. Aspect extraction in sentiment analysis: comparative analysis and survey // Artificial Intelligence Review. – 2016. – Vol. 46, No. 4. – P. 459-483.

Надійшла до редколегії 26.09.2025

УДК 004.9 DOi 10.30837/ bi.2025.2(103).13

Г. А. Плєхова

ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520

ЗАСТОСУВАННЯ АЛГЕБРО-ЛОГІЧНОГО МОДЕЛЮВАННЯ В УМОВАХ ІНТЕЛЕКТУАЛІЗАЦІЇ ПРИЙНЯТТЯ РІШЕНЬ НЕПОВНОГО ВИЗНАЧЕННЯ ІНФОРМАЦІЇ

У статті розглянуто застосування алгебро-логічного моделювання в умовах інтелектуалізації прийняття рішень неповного визначення інформації. В роботі розглянута можливість зменшення обсягу обчислень в системному управлінні складною системою: обговорена процедура відбору ознак, де кількість ознак може бути зменшена. Наведено приклад використання математичного апарату теорії інтелекту, методу компараторної ідентифікації та інструментарію алгебри скінченних предикатів, який полягає у побудові моделі ідентифікації медично-діагностичних параметрів у вигляді системи предикатних рівнянь, при розв'язанні яких маємо інтерпретацію медичних знань у певній області.

ПРИЙНЯТТЯ РІШЕНЬ, МОДЕЛЮВАННЯ, ОЗНАКА, ПРЕДИКАТ, КВАНТОР, СКРИНІНГ

G. A. Pliekhova. Application of algebraic-logical modeling in conditions of intellectualization of decision-making with incomplete information definition. The article considers the application of algebraic-logical modeling in the conditions of intellectualization of decision-making of incomplete information definition. The paper considers the possibility of reducing the volume of calculations in the system management of a complex system: the procedure for feature selection is discussed, where the number of features can be reduced. An example of using the mathematical apparatus of the theory of intelligence, the method of comparative identification, and the tools of finite predicate algebra is given, which consists in building a model for identifying medical diagnostic parameters in the form of a system of predicate equations, when solving which we have an interpretation of medical knowledge in a certain area.

DECISION MAKING, MODELING, SIGN, PREDICATE, QUANTIFIER, SCREENING

Вступ

Реалії сьогодення свідчать про необхідність створення складних систем стійких до впливу негативних (агресивних) факторів зовнішнього середовища. Особливу важливу роль відіграють інфокомунікаційні мережеві системи (ІМС), які забезпечують передачу, розподілену обробку інформації для прийняття відповідальних рішень по управлінню економікою країни та плануванню оборонних дій.

Метою дослідження є створення методологічних основ розробки іфокомунікаційної мережевої системи стійкої до впливу деструктивних факторів зовнішнього середовища.

В усіх галузях, які працюють з базами знань, використовуються інформаційні технології та виробляють або споживають дані. Якість цих даних має вирішальне значення для ефективності процесів підтримки прийняття рішень. Актуальними напрямками розвитку ІТ є: інтерпретація знань, їх вилучення з різних джерел, інтелектуальна обробка даних, формування якісної для прийняття відповідних управляючих рішень. Інформаційний скринінг представляє з себе процес розподіленої семантичної обробки слабоструктурованих, неповних, неоднозначних даних з метою проведення аналізу ознак і виявлення закономірностей та невідповідностей [1].

1. Зменшення обсягу обчислень в системному управлінні складною системою

Розглянемо процедуру відбору ознак, де кількість ознак може бути зменшена. Тут ми можемо зіткнутися з наступними проблемами.

1. Нам може знадобитися знайти деякі набори значень ознак, які нас цікавлять, де є принаймні одне значення несуттєвих ознак, для яких існує принаймні один набір значень суттєвих ознак. У цьому випадку ми застосовуємо квантор існування до множини несуттєвих значень.

2. Нам може знадобитися знайти деякі набори значень ознак, де для будь-якого набору несуттєвих ознак існує принаймні один розв'язок рівняння. У цьому випадку ми застосовуємо універсальний квантор до несуттєвих змінних.

3. Нам може знадобитися знайти деякі набори значень ознак, які задовольняють рівнянню за умови, що несуттєві ознаки набувають певних конкретних значень.

Нехай предикат P залежить від змінних $x, y, \ldots, z$. Визначимо оператор підстановки $a(P)$ (a а належить області визначення змінної x), який діє на предикат P наступним чином:

$$a(P(x, y, \ldots, z) = P(a, y, \ldots, z).$$

Будемо називати оператор підстановки обмежувальним, якщо виконується наступна умова

$$(a, y, \ldots, z) \to P(x, y, \ldots, z)$$

для всіх $x, y, \ldots, z$.

Визначимо дистрибутив оператора заміщення, якщо умова виконується

$$P(a, y, \ldots, z) \leftarrow P(x, y, \ldots, z)$$

для всіх $x, y, \ldots, z$.

Інтерпретуючи знання, представлені цією імплікацією, можна сказати, що оператори заміщення посилюють логічний зв'язок між дискретними ознаками, а оператори розподільчої підстановки послаблюють

цей зв'язок, зсуваючи відношення між ознаками довільним чином.

Розглянемо предикат P наступним чином:

$$P(x,y,\ldots,z) = x^{a_1}P_1(y,\ldots,z) \vee x^{a_2}P_2(y,\ldots,z) \vee \ldots$$
$$\vee x^{a_n}P_n(y,\ldots,z).$$

Тоді:

$$a_1(P) = P_1(y,\ldots,z) = x^{a_1}P_1(y,\ldots,z) \vee x^{a_2}P_1(y,\ldots,z) \vee \ldots$$
$$\vee x^{a_n}P_1(y,\ldots,z).$$

Очевидно, що предикат $a_1(P)$ буде скорочуватись, якщо $P_1 \to P_i \forall i = 1,2,\ldots,n$.

Оператор $a_1(P)$ буде розподіляти, якщо

$$P_1 \leftarrow P_i \forall i = 1,2,\ldots,n.$$

Розглянемо приклади застосування оператора a_1 до предиката $P(x,y)$, де змінні x,y та z мають області визначення $\{a_1,a_2\}$, $\{b_1,b_2\}$ і $\{c_1,c_2\}$ відповідно.

Нехай

$$P = x^{a_1}y^{b_1}z^{c_1} \vee x^{a_2}y^{b_1}z^{c_2} \vee x^{a_2}y^{b_1}z^{c_1}.$$

Тоді:

$$a_1(P) = y^{b_1}z^{c_1} = \left(x^{a_1} \vee x^{a_2}\right) \& \, y^{b_1}z^{c_1} =$$
$$= x^{a_1}y^{b_1}z^{c_1} \vee x^{a_2}y^{b_1}z^{c_1}.$$

За винятком диз'юнктів, які містить предикат $a_1(P)$, він містить ще один диз'юнкт $x^{a_2}y^{b_1}c^{c_1}$, тобто оператор a_1 є обмежуючим для предиката P. Згідно з введеними означеннями, у наведеному прикладі $P_1 = y^{b_1}z^{c_1}$, $P_2 = y^{b_1}z^{c_2} \vee y^{b_1}z_1^{c_1}$. Звідси очевидно, що $P_1 \to P_2$. Розглянемо тепер предикат:

$$P = x^{a_1}y^{b_1}z^{c_1} \vee x^{a_1}y^{b_1}z^{c_2} \vee x^{a_2}y^{b_1}z^{c_1},$$
$$a_1(P) = y^{b_1}z^{c_1} \vee y^{b_1}z^{c_2} = \left(x^{a_1} \vee x^{a_2}\right) \&$$
$$\& \left(y^{b_1}z^{c_1} \vee y^{b_1}z^{c_2}\right) = x^{a_1}y_1b_1c_1 \vee x^{a_2}y^{b_1}z^{c_2} \vee$$
$$\vee x^{a_2}y^{b_1}z^{c_1} \vee xx^{a_2}b_1b^{c_2}.$$

Оператор a_1 для цього предиката, очевидно, є розподільчим. У цьому прикладі:

$P_1 = y^{b_1}z^{c_1} \vee y^{b_1}z^{c_2}$, $P_2 = y^{b_1}z^{c_2}$, тобто $P_1 \leftarrow P_2$.

Для того, щоб відповісти на друге питання, необхідно виключити з вихідного рівняння всі змінні, крім розглянутих, і дослідити отримане рівняння з меншою кількістю змінних, що описує всі допустимі набори значень ознак. У роботі [2] розглянуто досить широкий клас предикатів, для яких можна вказати ефективний алгоритм виключення змінних без збільшення розміру вихідної формули. Тут ми розширюємо цей клас, додаючи деякі додаткові властивості. Розглянемо наступні властивості квантора існування:

1. $\exists x x^a = 1$.
2. $\exists x \neg x^a = 1$.
3. $\exists x\left(\neg\left(P(x)Q(x)\right)\right) = \exists x \neg P(x) \vee \exists x \neg Q(x)$.
4. $\exists x\left(P(x) \vee Q(x)\right) = \exists x P(x) \vee \exists x Q(x)$.
5. $\exists x(P(x) \& Q(y) = \exists x P(x) \& Q(y)$.
6. $\exists y\left(P(x) \to Q(y)\right) = P(x) \to \exists y Q(y)$.
7. $\exists y\left(P(x) \to Q(y)\right) = P(x) \to \exists y Q(y)$.

8. Припустимо. $P_i(x) \& P_j(x) = 0, i \neq j$, $i,j = 1,2,\ldots,k$. Припустимо:

$$\exists y\left(\left(P_1(x) \to Q_1(y)\right) \& \left(P_2(x) \to Q_2(y)\right) \& \ldots\right.$$
$$\left. \& \left(P_k(x) \to Q_k(y)\right)\right) = \left(P_1(x) \to \exists y Q_1(y)\right) \&$$
$$\& \left(P_2(x) \to \exists y Q_2(y)\right) \& \ldots \& \left(P_k(x) \to \exists y Q_k(y)\right).$$

9. Якщо тотожність $P_i(x) \equiv 0$ не є істинною для будь-якої $i = 1,2,\ldots,k$ і $P_i(x) \& P_j(x) = 0, i \neq j$, $i,j = 1,2,\ldots,k$, то

$$\exists x\left(\left(P_1(x) \to Q_1(y)\right) \& \left(P_2(x) \to Q_2(y)\right) \& \ldots\right.$$
$$\left. \& \left(P_k(x) \to Q_k(y)\right)\right) = Q_1(y) \vee Q_2(y) \vee \ldots \vee Q_k(y).$$

Перелічені вище властивості дозволяють описати широкий клас скінченних предикатів (відповідно до рівнянь), визначених на множині змінних $\{x,y,\ldots,z\}$, для яких легко знайти зв'язки між вибраними змінними без збільшення розміру вихідних формул. Означимо такий клас рекурсивно.

1. Всі «розпізнавання» $x^a, x^b, \ldots, x^c$ ($a,b,\ldots,c$ – символи, що належать домену для змінної x) належать до Δ_x.

2. Всі заперечення $\neg x^a, \neg x^b, \ldots, \neg x^c$ належать Δ_x.

3. Якщо предикати $\neg P(x), \neg Q(x)$ належать до Δ_x, то предикат $\neg\left(P(x)Q(x)\right)$ належить до Δ_x.

4. Будь-який предикат, що не залежить від змінної x, належить множині Δ_x.

5. Якщо предикати P_1 та P_2 належать до Δ_x, то предикат $P = P_1 \vee P_2$ належить до Δ_x.

6. Якщо предикат P_1 належить до Δ_x і предикат P_2 не залежить від x, то предикат $P = P_1 \& P_2$ належить до Δ_x.

7. Якщо предикат P_1 не залежить від x і предикат P_2 належить до Δ_x, то предикат $P = P_1 \to P_2$ належить до Δ_x.

8. Нехай предикати $P_1, P_2, \ldots, P_k$ не залежать від x; $P_i \& P_j = 0$ для $i \neq j$, $i,j = 1,2,\ldots,k$, предикати $Q_1, Q_2, \ldots, Q_k$ належать до Δ_x; то:

$$P = \left(P_1 \to Q_1\right) \& \left(P_2 \to Q_2\right) \& \ldots \& \left(P_k \to Q_k\right)$$

належить Δ_x.

9. Якщо предикати $P_1, P_2, \ldots, P_k$ залежать тільки від x, $P_i \& P_j = 0$ для $i \neq j$, $i,j = 1,2,\ldots,k$; для будь-якого $i = 1,2,\ldots,k$ тотожність $P_i \equiv 0$ не є істинною; предикати $Q_1, Q_2, \ldots, Q_k$ не залежать від x; тоді предикат

$$P = \left(P_1 \to Q_1\right) \& \left(P_2 \to Q_2\right) \& \ldots \& \left(P_k \to Q_k\right)$$

належить Δ_x.

Також може виникнути потреба виключити зайві змінні за допомогою універсального квантора. У цьому випадку ми можемо скористатися наступними властивостями цього квантора:

1. $\forall x x^a = 0$.
2. $\forall x \neg x^a = 0$.
3. $\forall x \neg\left(P(x) \vee Q(x)\right) = \forall x \neg P(x) \& \forall x \neg Q(x)$.

4. $\forall x\big(P(x)\,\&\,Q(x)\big)=\forall xP(x)\,\&\,\forall xQ(x)$.

5. $\forall x(P(x)\vee Q(y)=\forall xP(x)\vee Q(y)$.

6. $\forall y\big(P(x)\,\&\,Q(y)\big)=P(x)\,\&\,\forall yQ(y)$.

7. Припустимо, що:

$$P_i(x)\,\&\,P_j(x)=0,\ i\neq j, i,j=1,2,\ldots,k,$$

тоді:

$$\forall y\Big(\big(P_1(x)\,\&\,Q_1(y)\big)\vee\big(P_2(x)\,\&\,Q_2(y)\big)\vee\ldots$$
$$\vee\big(P_k(x)\,\&\,Q_k(y)\big)\Big)=\big(P_1(x)\,\&\,\forall yQ_1(y)\big)\vee$$
$$\vee\big(P_2(x)\,\&\,\forall yQ_2(y)\big)\vee\ldots\vee\big(P_k(x)\,\&\,\forall yQ_k(y)\big).$$

8. Якщо ідентичність $P_i(x)\equiv 0$ не є істинною для будь-якої $i=1,2,\ldots,k$ і $P_i(x)\,\&\,P_j(x)=0$ для $i\neq j$, $i,j=1,2,\ldots,k$, то

$$\forall x\Big(\big(P_1(x)\,\&\,Q_1(y)\big)\vee\big(P_2(x)\,\&\,Q_2(y)\big)\vee\ldots$$
$$\vee\big(P_k(x)\,\&\,Q_k(y)\big)\Big)=Q_1(y)\,\&\,Q_2(y)\,\&\ldots\&\,Q_k(y).$$

Ми можемо рекурсивно визначити клас предикатів Σ_x, з якого можна виключити змінну x без збільшення розміру формули:

1. Всі «впізнавання» $x^a,x^b,\ldots,x^c$ належать Σ_x.

2. Всі заперечення $\neg x^a,\neg x^b,\ldots,\neg x^c$, які не залежать від x, належать Σ_x.

3. Якщо $\neg P_1$ і $\neg P_2$ належать до Σ_x, то $\neg(P_1\vee P_2)$ належить до Σ_x.

4. Якщо предикати P_1 та P_2 належать до Σ_x, то предикат $P=P_1\,\&\,P_2$ належить до Σ_x.

5. Якщо предикат P_1 належить до Σ_x і предикат P_2 не залежить від x, то предикат $P=P_1\vee P_2$ належить до Σ_x.

6. Якщо предикат P_1 не залежить від x і предикат P_2 належить до Σ_x, то предикат $P=P_1\,\&\,P_2$ належить до Σ_x.

7. Припустимо, що предикати $P_1,P_2,\ldots,P_k$ не залежать від x, $P_i\,\&\,P_j=0$ для $i\neq j$, $i,j=1,2,\ldots,k$; предикати $Q_1,Q_2,\ldots,Q_k$ належать до Σ_x, тоді

$$P=(P_1\,\&\,Q_1)\vee(P_2\,\&\,Q_2)\vee\ldots\vee(P_k\,\&\,Q_k)$$

належить Σ_x.

8. Якщо предикати $P_1,P_2,\ldots,P_k$ залежать тільки від x, $P_i\,\&\,P_j=0$ для $i\neq j$, $i,j=1,2,\ldots,k$; для будь-якого $i,j=1,2,\ldots,k$ тотожність $P_i\equiv 0$ не є істинною, предикати $Q_1,Q_2,\ldots,Q_k$ не залежать від x, тоді предикат $P=(P_1\,\&\,Q_1)\vee(P_2\,\&\,Q_2)\vee\ldots\vee(P_k\,\&\,Q_k)$ належить множині Σ_x.

Приклад використання цього методу показаний у наступному розділі на матеріалі інформаційного скринінгу медичної документації.

2. Приклад застосування алгебро-логічного моделювання для в умовах інтелектуалізації прийняття рішень неповного визначення інформації

Розглянемо задачу розподіленого аналізу певної медичної документації, за допомогою інфокомунікаційної мережі, а саме даних, які зберігаються в медичних картках пацієнтів. Такі дані, в цілому, характеризуються такими особливостями: неповнота, застарілість, суперечливість та ін. Така ситуація призводить до наявності більш одного тлумачення, тобто до неоднозначності інформації. Використання результатів розподіленої обробки неоднозначної інформації з використанням інформаційної мережі показав, що, інформація, яка міститься в медичній документації, при низькій якості даних є причиною невірних управлінських і медичних рішень. У зв'язку з цим актуальним є впровадження моделей і методів обробки медичних даних, що дозволяють підвищити якість інформації для прийняття рішень, а саме: підвищити повноту даних шляхом видобування додаткової інформації, підвищити точність даних за рахунок використання лише актуальної інформації, вилучити суперечливу інформацію, сформувати релевантну сукупність даних.

Методи досліджень для такого типу задач ґрунтуються на використанні дистанційній, розподіленій обробки даних з використанням методів системного аналізу, теорії прийняття рішень, методів інтелектуального аналізу даних, методів теорії інтелекту. А саме: алгебро-логічний метод моделювання, метод компараторної ідентифікації був використаний для розподіленої інтелектуальної обробці даних з медичної карти пацієнта для інформаційного скринінгу медичної документації; математичний апарат методів оптимізації використовується при вирішенні задачі планування лікувально-профілактичних заходів на основі технології інформаційного скринінгу медичної документації; сервісно-орієнтований підхід використовувався для розробки інструментальних засобів вирішення задач дослідження.

У роботах [1, 3] запропоновано і розроблено метод інформаційного скринінгу медичної документації. Запропоновано також використання методу компараторної ідентифікації для вирішення задачі інформаційного скринінгу медичної документації. Розроблено модель ідентифікації медично-діагностичних параметрів.

Так, наприклад, для підвищення якості інформації для прийняття медичних рішень сформульовано й обґрунтовано метод інформаційного скринінгу, який складається з використання сукупності моделей передачі та розподіленої обробки даних з використанням інформаційної мережі (рис. 1).

Для вирішення задачі розподіленого інформаційного скринінгу певної медичної документації (наприклад, медичних карток, які заповнюються сімейним лікарем), в якості методу моделювання предметних знань запропоновано використання алгебро-логічного методу моделювання системи ознак, який базується на алгебрі скінченних предикатів. Загальна схема використання запропонованого методу полягає в на-

ступному: на основі інформації з медичних карток та інших джерел медичної інформації, яка передається в інфокомунікаційних мережах, формується множина ознак, яка моделюється предикатними рівняннями. У залежності від того, які результати потрібно отримати, реакції предикатів піддаються семантичній обробці та групуються в набір агрегованих ознак, які в подальшому є основою для прийняття медичних рішень (рис. 1).

Рис. 1. Метод розподіленого інформаційного скринінгу медичної документації з використанням інформаційної мережі

Однією з головних умов застосування методу компараторної ідентифікації – це дискретність, скінченність і детермінованість об'єктів предметного простору. В даному випадку предметний простір U є декартовим добутком множини об'єктів з предметної області $U_1 \times U_2 \times ... \times U_n$: медичних даних, карток пацієнтів, даних клінічного моніторингу тощо. За допомогою предиката $P(x_1, x_2, ..., x_m)$ описується будь-яке відношення, яке задане на предметному просторі $U = U_1 \times U_2 \times ... \times U_n$. На мові алгебри скінченних предикатів, цей предикат формалізується за допомогою предикату впізнавання, а також базисних операцій кон'юнкції і диз'юнкції. Предикат впізнавання моделює здатність людини однозначно віднести об'єкт, що розглядається, до одного з двох класів. Предикат дорівнює 1 при впізнанні об'єкту та 0 – в іншому випадку. Введені предикатні змінні зв'язуються логічними рівняннями, спільний розв'язок яких дає віднесення елементів множини розглянутих об'єктів до певного класу, тобто визначення групи ризику пацієнта щодо наявності певних захворювань.

План експерименту пропонується такий. Ми використовуємо реальні медичні дані і кодуємо їх за допомогою предикатних рівнянь. Зауважимо, що хоча деякі змінні можуть набувати значень «невідомо», це все ж таки випадок закритого світу, оскільки «невідомо» означає лише значення з алфавіту, на якому визначена змінна. Кожна область для будь-якої змінної є замкненою. Після того, як ми написали систему рівнянь за допомогою експертів, ми починаємо видаляти змінні, які вважаємо несуттєвими в даний момент. Це не означає, що в інших випадках інші змінні будуть вважатися несуттєвими. Важливі змінні – це ті, для яких ми хочемо визначити логічні зв'язки На виході ми отримуємо рівняння, з якого видалено несуттєві змінні. Отримане рівняння є простішим, ніж вихідна система, і можна простіше проаналізувати зв'язки між суттєвими змінними.

Якщо розглядати розподілений інформаційний скринінг медичних даних для оцінки розвитку та профілактики захворювань серця і судин [4], то можна виділити набір ознак для формалізації скринінгових процедур. Розглянемо такі ознаки та їх значення:

Стать: $X_1 = \{x_1^1, x_1^2\}$, де x_1^1 означає жінка, x_1^2 означає чоловік.

Вік: $X_2 = \{x_2^1, x_2^2, x_2^3\}$, де x_2^1 менше 40 років, x_2^2 від 40 до 50 років, x_2^3 більше 50 років.

Цукровий діабет: $X_3 = \{x_3^1, x_3^2, x_3^3, x_3^4\}$, де x_3^1 – так, x_3^2 – ні (фактичний діагноз), x_3^3 – ні (діагноз не встановлений), x_3^4 – невідомо.

Артеріальна гіпертензія: $X_4 = \{x_4^1, x_4^2, x_4^3, x_4^4\}$, де x_4^1 – так, x_4^2 – ні (фактичний діагноз), x_4^3 – ні (діагноз не встановлений), x_4^4 – невідомо.

Проблеми з нирками: $X_5 = \{x_5^1, x_5^2, x_5^3\}$, де x_5^1 – так, x_5^2 – ні, x_5^3 – невідомо.

Тахікардія: $X_6 = \{x_6^1, x_6^2, x_6^3, x_6^4, x_6^5\}$, де x_6^1 – так (фактичний діагноз), x_6^2 – так (діагноз не встановлений), x_6^3 – ні (справжній діагноз), x_6^4 – ні (діагноз не встановлений), x_6^5 – невідомо.

Спадковість щодо захворювань серця та судин: $X_7 = \{x_7^1, x_7^2, x_7^3\}$, де x_7^1 – так, x_7^2 – ні, x_7^3 – невідомо.

Куріння: $X_8 = \{x_8^1, x_8^2, x_8^3\}$, де x_8^1 – так, x_8^2 – ні, x_8^3 – невідомо.

Проблеми з алкоголем: $X_9 = \{x_9^1, x_9^2, x_9^3\}$, де x_9^1 – так, x_9^2 – ні, x_9^3 – невідомо.

Гіподинамія: $X_{10} = \{x_{10}^1, x_{10}^2, x_{10}^3\}$, де x_{10}^1 – так, x_{10}^2 – ні, x_{10}^3 – невідомо.

Ці особливості дозволяють розробити модель ідентифікації діагностичних параметрів, за допомогою якої можна визначити групу здоров'я пацієнта $R = \{r_1, r_2, r_3, r_4\}$, де r_1 – низький ризик захворювань серця та судин, r_2 – помірний ризик, r_3 – високий ризик, r_4 – дуже високий ризик.

Для визначення групи здоров'я використовується набір агрегованих ознак $Q_1 - Q_3$, де Q_1 виражається через X_1 та X_2, Q_2 виражається через $X_7 - X_{10}$, Q_3 виражається через $X_3 - X_6$.

Значення кожної групи здоров'я та кожної агрегованої ознаки розподілено на чотири класи згідно з відповідною медико-технологічною документацією (уніфікованим клінічним протоколом та локальними протоколами, що стосуються профілактики хвороб серця і судин).

Система рівнянь для формування, наприклад, ознаки Q_2 має вигляд:

$$\begin{cases} q_2^1 = x_7^2x_8^2(x_9^2 \vee x_9^3(x_{10}^2 \vee x_{10}^3)) \vee x_7^2x_8^3x_9^2x_{10}^2 \vee \\ \quad \vee x_7^3x_8^2x_{10}^2(x_9^2 \vee x_9^3), \\ q_2^2 = x_7^2(x_8^1(x_9^1x_{10}^2 \vee x_9^2) \vee x_9^3(x_8^1x_{10}^2 \vee x_8^2x_{10}^1)) \vee \\ \quad \vee (x_7^2(x_8^2x_9^1 \vee x_8^3x_9^2) \vee (x_7^2x_9^3 \vee x_7^3x_9^1)x_8^3x_{10}^2 \vee \\ \quad \vee (x_7^2x_8^3 \vee x_7^3x_8^2)x_9^1(x_{10}^2 \vee x_{10}^3) \vee \\ \quad \vee x_7^3x_8^2(x_9^2 \vee x_9^3))(x_{10}^1 \vee x_{10}^3) \vee \\ \quad \vee x_7^3x_8^1(x_9^2x_{10}^2 \vee x_9^3) \vee x_7^3x_8^3x_9^2(x_{10}^1 \vee x_{10}^2), \\ q_2^3 = x_7^1x_{10}^2(x_8^1x_9^2 \vee x_8^2(x_9^1 \vee x_9^2)) \vee \\ \quad \vee (x_7^1x_9^3(x_8^1 \vee x_8^2) \vee (x_7^1x_8^3 \vee x_7^3x_8^1)x_9^1)(x_{10}^2 \vee x_{10}^3) \vee \\ \quad \vee x_7^1x_8^3(x_9^2 \vee x_9^3) \vee (x_7^2(x_8^1(x_9^1 \vee x_9^3) \vee x_8^3x_9^3) \vee \\ \quad \vee (x_7^2x_8^3 \vee x_7^3x_8^2)x_9^1x_{10}^1 \vee \\ \quad \vee x_7^3(x_8^1x_9^2 \vee x_8^3x_9^1))(x_{10}^1 \vee x_{10}^3) \vee x_7^3x_8^3(x_9^2x_{10}^3 \vee x_9^3), \\ q_2^4 = x_7^1x_9^3x_{10}^1(x_8^1 \vee x_8^2) \vee (x_7^1x_8^3 \vee x_7^3x_8^1)x_9^1x_{10}^1 \vee \\ \quad \vee (x_7^1x_8^2x_9^1 \vee x_7^1x_9^2(x_8^1 \vee x_8^2))(x_{10}^1 \vee x_{10}^3) \vee x_7^1x_8^1x_9^1. \end{cases}$$

Модель ідентифікації медично-діагностичних параметрів на прикладі визначення класу ризику представлено у вигляді системи предикатних рівнянь:

$$\begin{cases} r_1 = q_1^1q_2^1(q_3^1 \vee q_3^2) \vee (q_1^1q_2^2 \vee (q_1^2 \vee q_1^3)q_2^1)q_3^1, \\ r_2 = q_1^1(q_2^1q_3^3 \vee q_2^2q_3^2) \vee (q_1^1(q_2^3 \vee q_2^4) \vee q_1^2(q_2^2 \vee q_2^3) \vee \\ \quad \vee q_1^3q_2^2 \vee q_1^4(q_2^1 \vee q_2^2))(q_3^1 \vee q_3^2) \vee \\ \quad \vee (q_1^2 \vee q_1^3)q_2^1(q_3^2 \vee q_3^3) \vee (q_1^2q_2^4 \vee (q_1^3 \vee q_1^4)q_2^3)q_3^1, \\ r_3 = q_2^1q_3^4 \vee (q_1^1 \vee q_1^2 \vee q_1^3)(q_2^2 \vee q_2^3)(q_3^3 \vee q_3^4) \vee \\ \quad \vee q_1^3q_2^2(q_2^3 \vee q_2^4) \vee (q_1^3 \vee q_1^4)q_2^4q_3^1 \vee \\ \quad \vee (q_1^1q_2^4 \vee q_1^4(q_2^1 \vee q_2^2))q_3^3 \vee (q_1^2q_2^4 \vee q_1^4q_2^3)(q_3^2 \vee q_3^3), \\ r_4 = (q_1^1 \vee q_1^2)q_2^4q_3^4 \vee q_1^3q_2^4(q_3^3 \vee q_3^4) \vee q_1^4q_3^4(q_2^2 \vee q_2^3) \vee \\ \quad \vee q_1^4q_2^4(q_3^2 \vee q_3^3 \vee q_3^4). \end{cases}$$

Остаточна класифікація може бути виражена наступною системою:

$$\begin{cases} r_1 = q_1^1q_2^1\left(q_3^1 \vee q_3^2\right) \vee \left(q_1^1q_2^2 \vee \left(q_1^2 \vee q_1^3\right)q_2^1\right)q_3^1, \\ r_2 = q_1^1\left(q_2^1q_3^3 \vee q_2^2q_3^2\right) \vee \left(q_1^1\left(q_2^3 \vee q_2^4\right) \vee q_1^2\left(q_2^2 \vee q_2^3\right) \vee q_1^3q_2^2 \vee q_1^4\left(q_2^1 \vee q_2^2\right)\right)\left(q_3^1 \vee q_3^2\right) \vee \\ \qquad \vee \left(q_1^2 \vee q_1^3\right)q_2^1\left(q_3^2 \vee q_3^3\right) \vee \left(q_1^2q_2^4 \vee \left(q_1^3 \vee q_1^4\right)q_2^3\right)q_3^1, \\ r_3 = q_2^1q_3^4 \vee \left(q_1^1 \vee q_1^2 \vee q_1^3\right)\left(q_2^2 \vee q_2^3\right)\left(q_3^3 \vee q_3^4\right) \vee q_1^3q_2^2\left(q_2^3 \vee q_2^4\right) \vee \left(q_1^3 \vee q_1^4\right)q_2^4q_3^1 \vee \\ \qquad \vee \left(q_1^1q_2^4 \vee q_1^4\left(q_2^1 \vee q_2^2\right)\right)q_3^3 \vee \left(q_1^2q_2^4 \vee q_1^4q_2^3\right)\left(q_3^2 \vee q_3^3\right), \\ r_4 = \left(q_1^1 \vee q_1^2\right)q_2^4q_3^4 \vee q_1^3q_2^4\left(q_3^3 \vee q_3^4\right) \vee q_1^4q_3^4\left(q_2^2 \vee q_2^3\right) \vee q_1^4q_2^4\left(q_3^2 \vee q_3^3 \vee q_3^4\right). \end{cases}$$

Дослідимо логічні зв'язки між дискретними ознаками $x_1 - x_{10}$. Перш за все, перепишемо систему предикатних рівнянь у наступному вигляді:

$$\begin{aligned} &P(q_2,x_1,\ldots,x_{10}) = q_2^1\left(x_7^2x_8^2\left(x_9^2 \vee x_9^3\left(x_{10}^2 \vee x_{10}^3\right)\right) \vee x_7^2x_8^3x_9^2x_{10}^2 \vee x_7^3x_8^2x_{10}^2\left(x_9^2 \vee x_9^3\right)\right) \vee \\ &\vee q_2^2(x_7^2\left(x_8^1\left(x_9^1x_{10}^2 \vee x_9^2\right) \vee x_9^3\left(x_8^1x_{10}^2 \vee x_8^2x_{10}^1\right)\right) \vee \left(x_7^2\left(x_8^2x_9^1 \vee x_8^3x_9^2\right) \vee \left(x_7^2x_9^3 \vee x_7^3x_9^1\right)x_8^3x_{10}^2 \vee \\ &\vee\left(x_7^2x_8^3 \vee x_7^3x_8^2\right)x_9^1\left(x_{10}^2 \vee x_{10}^3\right) \vee x_7^3x_8^2\left(x_9^2 \vee x_9^3\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee x_7^3x_8^1\left(x_9^2x_{10}^2 \vee x_9^3\right) \vee \\ &\vee x_7^3x_8^3x_9^2\left(x_{10}^1 \vee x_{10}^2\right)) \vee q_2^3(x_7^1x_{10}^2\left(x_8^1x_9^2 \vee x_8^2\left(x_9^1 \vee x_9^2\right)\right) \vee \left(x_7^1x_9^3\left(x_8^1 \vee x_8^2\right) \vee \left(x_7^1x_8^3 \vee x_7^3x_8^1\right)x_9^1\right)\left(x_{10}^2 \vee x_{10}^3\right) \vee \\ &\vee x_7^1x_8^3\left(x_9^2 \vee x_9^3\right) \vee \left(x_7^2\left(x_8^1\left(x_9^1 \vee x_9^3\right) \vee x_8^3x_9^3\right) \vee \left(x_7^2x_8^3 \vee x_7^3x_8^2\right)x_9^1x_{10}^1 \vee \\ &\vee x_7^3\left(x_8^1x_9^2 \vee x_8^3x_9^1\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee x_7^3x_8^3\left(x_9^2x_{10}^3 \vee x_9^3\right)) \vee \\ &\vee q_2^4\left(x_7^1x_9^3x_{10}^1\left(x_8^1 \vee x_8^2\right) \vee \left(x_7^1x_8^3 \vee x_7^3x_8^1\right)x_9^1x_{10}^1 \vee \left(x_7^1x_8^2x_9^1 \vee x_7^1x_9^2\left(x_8^1 \vee x_8^2\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee x_7^1x_8^1x_9^1\right) = 1. \end{aligned}$$

Видно, що цей предикат належить до класу Δ_{x7}. Дослідимо зв'язок між усіма змінними, крім x_7. Таке виключення дасть нам зв'язок між змінними $q_2,x_1,\ldots,x_6,x_8,x_9,x_{10}$:

$$F = \exists x_7 P(q_2, x_1, \ldots, x_{10}) = q_2^1\left(x_8^2\left(x_9^2 \vee x_9^3\left(x_{10}^2 \vee x_{10}^3\right)\right) \vee x_8^3 x_9^2 x_{10}^2 \vee x_8^2 x_{10}^2\left(x_9^2 \vee x_9^3\right)\right) \vee$$
$$\vee q_2^2\left(x_8^1\left(x_9^1 x_{10}^2 \vee x_9^2\right) \vee x_9^3\left(x_8^1 x_{10}^2 \vee x_8^2 x_{10}^1\right)\right) \vee \left(\left(x_8^2 x_9^1 \vee x_8^3 x_9^2\right) \vee \left(x_9^3 \vee x_9^1\right) x_8^3 x_{10}^2 \vee\right.$$
$$\left. \vee\left(x_8^3 \vee x_8^2\right) x_9^1\left(x_{10}^2 \vee x_{10}^3\right) \vee x_8^2\left(x_9^2 \vee x_9^3\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee x_8^1\left(x_9^2 x_{10}^2 \vee x_9^3\right) \vee x_8^3 x_9^2\left(x_{10}^1 \vee x_{10}^2\right)) \vee$$
$$\vee q_2^3(x_{10}^2\left(x_8^1 x_9^2 \vee x_8^2\left(x_9^1 \vee x_9^2\right)\right) \vee \left(x_9^3\left(x_8^1 \vee x_8^2\right) \vee \left(x_8^3 \vee x_7^3 x_8^1\right) x_9^1\right)\left(x_{10}^2 \vee x_{10}^3\right) \vee$$
$$\vee x_8^3\left(x_9^2 \vee x_9^3\right) \vee \left(\left(x_8^1\left(x_9^1 \vee x_9^3\right) \vee x_8^3 x_9^3\right) \vee \left(x_8^3 \vee x_8^2\right) x_9^1 x_{10}^1 \vee \left(x_8^1 x_9^2 \vee x_8^3 x_9^1\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee x_8^3\left(x_9^2 x_{10}^3 \vee x_9^3\right)) \vee$$
$$\vee q_2^4(x_9^3 x_{10}^1(x_8^1 \vee x_8^2) \vee (x_8^3 \vee x_8^1) x_9^1 x_{10}^1 \vee (x_8^2 x_9^1 \vee x_9^2(x_8^1 \vee x_8^2))(x_{10}^1 \vee x_{10}^3) \vee x_8^1 x_9^1) = 1.$$

Слід зазначити, що розмір вихідної формули не збільшився, що пояснюється тим, що предикат $P(q_2, x_1, \ldots, x_{10})$ належить до Δ_{x7}.

Припустимо, нас цікавить зв'язок між q_2, x_9, x_{10}. Вилучимо з предиката інші ознаки $F(x_1, \ldots, x_{10})$:

$$G(q_2, x_9, x_{10}) = \exists x_1 \exists x_2 \exists x_3 \exists x_4 \exists x_5 \exists x_6 \exists x_7 \exists x_8 P(q_2, x_1, \ldots, x_{10}) = q_2^1\left(\left(x_9^2 \vee x_9^3\left(x_{10}^2 \vee x_{10}^3\right)\right) \vee x_9^2 x_{10}^2 \vee x_{10}^2\left(x_9^2 \vee x_9^3\right)\right) \vee$$
$$\vee q_2^2\left(\left(x_9^1 x_{10}^2 \vee x_9^2\right) \vee x_9^3\left(x_{10}^2 \vee x_{10}^1\right)\right) \vee \left(\left(x_9^1 \vee x_9^2\right) \vee \left(x_9^3 \vee x_9^1\right) x_{10}^2 \vee x_9^1\left(x_{10}^2 \vee x_{10}^3\right) \vee \left(x_9^2 \vee x_9^3\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee \left(x_9^2 x_{10}^2 \vee x_9^3\right) \vee$$
$$\vee x_9^2\left(x_{10}^1 \vee x_{10}^2\right)) \vee q_2^3(x_{10}^2\left(x_9^1 \vee x_9^2\right) \vee x_9^3\left(x_{10}^2 \vee x_{10}^3\right) \vee \left(x_9^2 \vee x_9^3\right) \vee \left(\left(x_9^1 \vee x_9^3\right) \vee x_9^1 x_{10}^1 \vee\right.$$
$$\left. \vee\left(x_9^2 \vee x_9^1\right)\right)\left(x_{10}^1 \vee x_{10}^3\right) \vee \left(x_9^2 x_{10}^3 \vee x_9^3\right)) \vee q_2^4(x_9^3 x_{10}^1 \vee x_9^1 x_{10}^1 \vee (x_9^1 \vee x_9^2)(x_{10}^1 \vee x_{10}^3) \vee x_9^1) = 1.$$

Ми скоротили вихідну формулу і отримали простішу залежність між обраними ознаками. Після того, як отримано необхідну залежність, ми можемо розв'язати отримане рівняння з однією або кількома цільовими змінними. В залежності від того розв'язку системи предикатних рівнянь, до якого класу віднесено медичну картку, лікарем буде розроблятися комплекс лікувально-профілактичних процедур та набір рекомендацій для підтримання здоров'я пацієнта в належному стані.

Висновки

Таким чином у проведеному дослідженні наведений приклад демонструє використання математичного апарату теорії інтелекту, методу компараторної ідентифікації та інструментарію алгебри скінченних предикатів на моделі ідентифікації медично-діагностичних параметрів у вигляді системи предикатних рівнянь, при розв'язанні яких маємо інтерпретацію медичних знань у певній області.

Список літератури

[1] Melnik K. Towards medical screening information technology: the healthgrid-based approach / K. Melnik, O. Cherednichenko, V. Glushko // Information Systems: Methods, Models, and Applications. – 2013. – Vol. 117. – pp. 202-204.

[2] Smelyakov K. The Neural Network Models Effectiveness for Face Detection and Face Recognition / K. Smelyakov, A. Chupryna, O. Bohomolov, N. Hunko // Proceedings of the 2021 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream '2021) (22 April 2021, Vilnius, Lithuania). – 2021 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2021. – pp. 1-7.

[3] Ситніков Д.Є. Виявлення суттєвих особливостей даних шляхом складання та маніпулювання логічними рівняннями / Д.Є. Ситніков, Б. Д'Круз, П.Є. Ситнікова // Data Mining II. – WIT Press, 2000. – С. 241-248.

[4] Ameri F. Product lifecycle management: closing the knowledge loops / F. Ameri, D. Dutta // Computer-Aided Design and Applications. – 2005. – No. 2(5). – pp. 577–590.

Надійшла до редколегії 22.10.2025

УДК 004.8:004.9 DOi 10.30837/ bi.2025.2(103).14

С. Ф. Чалий[1], І. О. Лещинська[1]

[1]ХНУРЕ, м. Харків, Україна, serhii.chalyi@nure.ua, ORCID iD: 0000-0002-9982-9091

[1]ХНУРЕ, м. Харків, Україна, iryna.leshchynska@nure.ua, ORCID iD: 0000-0002-8737-4595

МЕТОД ПОБУДОВИ НЕЙРОСИМВОЛЬНОГО ПРЕДСТАВЛЕННЯ МЕНТАЛЬНОЇ МОДЕЛІ РІШЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

Розглянуто методи побудови ментальних моделей рішень інтелектуальних систем на основі інтеграції нейромережевих і символьних компонентів. Розроблено метод побудови нейросимвольного представлення ментальної моделі, який базується на двошаровій нейросимвольній архітектурі з можливістю виявлення прихованих ознак, відбору значущих ознак та механізмом нейросимвольного перетворення для відображення скритих представлень у символьні концепції. Метод містить етапи вилучення прихованих ознак, відбору ознак на основі уваги, нейросимвольного перетворення, побудови орієнтованого ациклічного графа для каузальної структури та перевірки каузальності з використанням лінійної темпоральної логіки. Метод створює умови для автоматизованого виявлення індивідуальних ментальних моделей користувачів із можливостями їх інтерпретації згідно з особливостями предметної області, а також побудови персоналізованих пояснень у системах пояснювального штучного інтелекту.

МЕНТАЛЬНІ МОДЕЛІ, НЕЙРОСИМВОЛЬНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ПОЯСНЮВАНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ОРІЄНТОВАНІ АЦИКЛІЧНІ ГРАФИ, ЛІНІЙНА ТЕМПОРАЛЬНА ЛОГІКА, ПЕРСОНАЛІЗОВАНІ ПОЯСНЕННЯ

S. F. Chalyi, I. O. Leshchynska. Method for constructing neurosymbolic representation of mental model of intelligent system decision. Methods for constructing mental models of intelligent system decisions based on integration of neural network and symbolic components are considered. A method for constructing neurosymbolic representation of mental model has been developed, based on dual-layer neurosymbolic architecture with capability for latent feature identification, significant feature selection, and neural-symbolic transformation mechanism for mapping hidden representations to symbolic concepts. The method includes stages of latent feature extraction, attention-based feature selection, neural-symbolic transformation, directed acyclic graph construction for causal structure representation, and causality verification using linear temporal logic. The method creates conditions for automated identification of individual user mental models with capabilities for their interpretation according to domain specifics, as well as construction of personalized explanations in explainable artificial intelligence systems.

MENTAL MODELS, NEUROSYMBOLIC ARTIFICIAL INTELLIGENCE, EXPLAINABLE ARTIFICIAL INTELLIGENCE, DIRECTED ACYCLIC GRAPHS, LINEAR TEMPORAL LOGIC, PERSONALIZED EXPLANATIONS

Вступ

Інтелектуальні інформаційні системи поєднують переваги традиційних інформаційних систем та систем штучного інтелекту при вирішенні комплексних задач у фінансовій сфері, промисловості, транспортній галузі [1]. Такі системи використовують моделі машинного навчання, що утруднює розуміння логіки їх роботи для користувачів [2]. Для забезпечення прозорості інтелектуальних інформаційних систем на сьогодні використовуються методи пояснювального штучного інтелекту (XAI) [3]. Для побудови зрозумілих пояснень в рамках XAI можуть бути використані ментальні моделі, які є внутрішніми представленнями користувачів про те, як працює інтелектуальна система, які каузальні залежності вона використовує для прийняття рішень [4, 5]. Ментальна модель відображає розуміння користувачем логіки роботи системи та обумовлює інтерпретацію отриманих рішень [6]. Використання ментальних моделей дає можливість адаптувати пояснення рішень інтелектуальної системи відповідно до рівня знань користувача та до його очікувань щодо можливостей застосування цього рішення [7].

Існуючі підходи до побудови ментальних моделей базуються переважно на використанні нейромережних та символьних методів [8, 9]. Перші дають можливість персоналізувати модель для користувача, проте представлені у вигляді чорної скриньки, що утруднює пояснення представлених в моделі залежностей [10]. Символьні методи використовують каузальне міркування і тому дають можливість сформувати явні залежності, які можуть бути безпосередньо інтерпретовані користувачем [11]. Однак ці методи потребують додаткових експертних знань для адаптації правил до індивідуальних потреб користувачів.

Таким чином, поєднання переваг обох підходів в рамках нейросимвольних архітектур створює умови для побудови пояснювальних персоналізованих ментальних моделей, що і свідчить про актуальність теми даного дослідження.

Нейросимвольні підходи до пояснювального штучного інтелекту реалізують інтеграцію нейромережних та символьних компонентів з використанням трьох основних парадигм [8, 9]. Перша парадигма, нейросимвольне перетворення, полягає у вилученні символьних правил з навчених нейронних мереж. Друга парадигма,

символьно-нейронне вбудовування, полягає у введенні символьних обмежень у процес навчання нейронних мереж [12]. Третя парадигма, використання гібридних архітектур, передбачає паралельну роботу нейромережевого та символьного компонентів [9].

Моделі концептуального вузького місця (Concept Bottleneck Models) використовують інтерпретовані концепти як проміжний шар між входом та виходом мережі [12]. Обмеженням даного підходу є статичність символьних компонентів та відсутність персоналізації під індивідуальні ментальні моделі користувачів [6].

Методи побудови ментальних моделей у когнітивній психології визначають ментальні моделі як внутрішні представлення зовнішнього світу, які користувачі використовують для міркування та прогнозування. Витягування ментальних моделей (mental model elicitation) виконується за допомогою структурованих інтерв'ю та побудови концептуальних схем моделей [4, 13]. Підходи на основі аналізу поведінки [14] аналізують патерни взаємодії користувачів з інтелектуальною системою для імпліцитного виявлення ментальних моделей. Обмеженням цього підходу виступає відсутність масштабованості ручного витягування та неможливість забезпечити пояснюваність каузальних зв'язків у виявлених моделях [6].

Двошарові архітектури в штучному інтелекті розділяють систему на нейромережний шар для адаптивного навчання та символьний шар для верифікованого міркування з двонаправленим потоком інформації [9, 10].

Фреймворк каузального міркування (causal reasoning framework) використовує каузальні байєсівські мережі у символьному шарі для верифікації каузальних залежностей. Обмеженням цієї архітектури є відсутність специфікації для побудови ментальних моделей у системах пояснювального штучного інтелекту [7].

Для визначення каузальних залежностей у ментальних моделях використовують темпоральні знання, які фіксують часові послідовності змін станів керованого об'єкта [15]. Автоматизоване керування базами знань забезпечують темпоральні правила у логіко-ймовірнісній формі, що використовують оператори лінійної темпоральної логіки (LTL): NeXt, Future, Until [16]. Модель представлення темпоральних знань містить множину фактів виникнення станів, темпоральні відношення між фактами, а також операції над фактами, що дає можливість відобразити багатоваріантність рішень із заданим ступенем деталізації для відповідного ієрархічного рівня організації. Темпоральні залежності дають можливість формалізувати послідовність керуючих дій у часі та верифікувати каузальні зв'язки через темпоральні обмеження [15, 16].

Таким чином, існуючі підходи окремо виявляють латентні ознаки, реалізують каузальне міркування, а також виконують темпоральну перевірку з використанням лінійної темпоральної логіки. Відповідно, задача розробки підходу, що поєднує виявлення ознак, каузальне міркування та темпоральну верифікацію для побудови нейросимвольного представлення ментальної моделі потребує свого вирішення.

1. Постановка задачі

Метою є розробка підходу до побудови двошарової нейросимвольної архітектури, яка забезпечує інтерпретованість залежностей в ментальній моделі, автоматизацію виявлення індивідуальних ментальних моделей та можливість пояснюваності моделі на основі інтеграції компонентів глибокого навчання й каузального графа з подальшою перевіркою узгодженості у часі з використанням темпоральної логіки.

Для досягнення поставленої мети вирішуються такі задачі:

– розробка підходу до побудови нейромережного та символьного шарів ментальної моделі рішення інтелектуальної системи з урахуванням можливості побудови персоналізованого пояснення;

– розробку методу побудови ментальної моделі на основі двошарової архітектури з варіаційним автокодувачем для вилучення прихованих ознак, механізмом уваги для відбору значущих ознак, нейросимвольним перетворенням для відображення прихованих ознак у символьне представлення, орієнтованим ациклічним графом для відображення каузальної структури моделі з використанням лінійної темпоральної логіки для перевірки каузальних залежностей.

2. Підхід до побудови нейромережного та символьного шарів ментальної моделі рішення інтелектуальної системи

Розроблений підхід орієнтований на формування двошарової архітектури, яка інтегрує нейромережний шар, призначений для адаптивного виявлення прихованих ознак з поведінкових даних користувачів, та символьний шар, призначений для реалізації причинно-наслідкового міркування з можливістю подальшої перевірки засобами темпоральної логіки.

Запропонована двошарова нейросимвольна архітектура включає такі основні компоненти: варіаційний автокодувач; багатоголовий механізм уваги; нейросимвольний перетворювач; орієнтований ациклічний граф. Варіаційний автокодувач використовується для ймовірнісного виявлення латентних ознак. Багатоголовий механізм уваги реалізує відбір значущих ознак на основі фільтрації поведінкового шуму. Нейросимвольний перетворювач відображує латентні представлення у символьні концепції. Орієнтований ациклічний граф призначений для побудови каузальної структури. Перевірка каузальних залежностей виконується з використання лінійної темпоральної логіки шляхом фільтрації хибних залежностей.

Розглянемо ключові особливості запропонованого підходу.

Нейромережевий шар приймає на вхід послідовність $\mathbf{X}$ векторів спостережень $\mathbf{x}_i$ у моменти часу i: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_I\}$. Наприклад, при вирішенні задачі підтримки прийняття рішень з медичної діагностики вектор $\mathbf{x}_i$ може містити тип дії (перегляд симптому, замовлення тесту, вибір діагнозу), темпоральну мітку, параметри взаємодії з інтерфейсом системи. Варіаційний автокодувач кодує поведінкову послідовність у латентний простір. Ймовірнісний характер латентного простору дає можливість розрізняти користувачів з різним рівнем впевненості: новачки мають високу невизначеність (різноманітна поведінка), експерти – низьку (представлену стабільними патернами).

Регуляризація виконується з використанням дивергенції Кульбака-Лейблера (KL-divergence) між апостеріорним розподілом $q_\phi(z \mid X)$ та апріорним розподілом $p(z)$, шо дає можливість виконати інтерполяцію між ментальними моделями різних користувачів. Функція втрат $\mathcal{L}_{VAE}$ варіаційного автокодувача має вигляд:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|X)}\left[\log p_\theta(X \mid z)\right] - \\ -\beta \cdot D_{KL}(q_\phi(z \mid X) \,\|\, p(z)). \tag{1}$$

Формула (1) складається з двох доданків. Перший доданок $\mathbb{E}_{q_\phi(z|X)}[logp_\theta(X \mid z)]$ – це математичне сподівання $\mathbb{E}_{q_\phi(z|X)}$ логарифма ймовірності відновлення вхідної послідовності $X \mid z$ з латентного коду z декодером з параметрами θ, що забезпечує точність кодування поведінкових патернів.

Високе значення першого доданку свідчить про точне відновлення поведінкових патернів. Наприклад, якщо користувач спочатку переглядає симптоми, потім замовляє тести, автокодувач має коректно відновити цю послідовність з латентного представлення.

Другий доданок $D_{KL}(q_\phi(z \mid X) \,\|\, p(z))$ – це дивергенція Кульбака-Лейблера між апостеріорним розподілом (навчений розподіл латентних кодів для даного користувача) та апріорним розподілом (стандартний нормальний розподіл). Дивергенція Кульбака-Лейблера визначає, наскільки розподіл $q_\phi(z \mid X)$ відрізняється від $p(z)$. Цей доданок виконує регуляризацію для отримання латентного простору без ізольованих кластерів. Параметр β контролює баланс між відновленням і регуляризацією. Типове значення $\beta = 1$. При $\beta > 1$ підвищується можливість інтерпретувати окремі розмірності латентного простору.

Багатоголовий механізм уваги обробляє латентні представлення $\mathbf{z}$ через H паралельних голів уваги для відбору значущих ознак. Для кожної голови уваги та латентного представлення $\mathbf{z}$ традиційно обчислюються матриці запитів $\mathbf{Q}$, ключів $\mathbf{K}$ та значень $\mathbf{V}$ розмірністю d_k.

Увага Attention($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$) для кожної голови обчислюється таким чином:

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2}$$

Вираз (2) відображає механізм вибіркової уваги. Функція softmax перетворює схожості у ваги уваги (сума ваг дорівнює 1). Скалярний добуток $\mathbf{Q}\mathbf{K}^\top$ відображає схожість між запитами та ключами. Цей скалярний добуток може приймати великі значення при збільшенні розмірності матриць. Результатом може бути насичення функції softmax та виникнення проблеми зникаючих градієнтів. Масштабування на $\sqrt{d_k}$ нормалізує ці значення, забезпечуючи стабільність градієнтів при великих розмірностях матриць. Множення на $\mathbf{V}$ дає можливість відібрати значущі ознаки. Так, одна голова уваги може фокусуватися на темпоральних патернах, визначаючи порядок дій, а інша – на типах дій, наприклад діагностичних тестах або терапевтичних рішеннях.

Після того, як кожна з голів уваги обробила латентне представлення згідно (2), виконується об'єднання (конкатенація) отриманих векторів h_j розміру d_k у один довгий вектор $Concat\left(h_1,\ldots,h_H\right)$ розмірності $H \cdot d_k$. Об'єднаний вектор перемножується з проекційною матрицею $\mathbf{W}^O$, в результаті чого виконується агрегація інформації від усіх голів уваги, а також відновлюється сумісність розмірностей з латентним представленням $\mathbf{z}$. Результуюче представлення $\mathbf{z}_{\text{salient}}$ має вигляд:

$$\mathbf{z}_{\text{salient}} = \text{Concat}\left(h_1,\ldots,h_H\right)\mathbf{W}^O. \tag{3}$$

Відфільтроване представлення $\mathbf{z}_{\text{salient}}$ містить лише значущі ознаки для символьного шару, що створює умови для інтерпретації рішення.

Нейросимвольне перетворення відображає латентні ознаки $\mathbf{z}_{\text{salient}}$ на множину символьних концептів $C = \{c_1, c_2, \ldots, c_L\}$. Попередньо визначається концептуальний словник $V = \{v_1, v_2, \ldots, v_R\}$ предметної області.

Наприклад, для медичної системи підтримки прийняття рішень словник може містити концепти: «аналіз симптомів», «диференційна діагностика», «призначення тестів», «оцінка ризиків», «вибір терапії».

Нейромережний класифікатор обчислює розподіл ймовірностей над концептуальним словником:

$$p(c \mid \mathbf{z}_{\text{salient}}) = \text{softmax}\left(f\left(\mathbf{z}_{\text{salient}}\right)\right). \tag{4}$$

Зокрема, ймовірність $p(c_i \mid \mathbf{z}) = 0{,}85$ означає, що латентні ознаки з впевненістю 0,85 відповідають концепту c_l.

При побудові символьного шару відбираються концепти з найвищими ймовірностями. Кожному концепту присвоюється семантичне вбудовування e_{c_l} з попередньо навченого ембеддінгу для структурованого представлення у побудові графа.

Можливість двонаправленої інтеграції створює умови для узгодженості між нейромережним та сим-

вольним шарами. Така інтеграція полягає у зворотній перевірці, коли символьні обмеження з графа можуть бути використані для регуляризації навчання нейронної мережі.

Символьний шар призначений, щоб сформувати орієнтований ациклічний граф $G=(C,E)$ для множини символьних концептів C з тим, щоб явно представити ментальну модель за допомогою каузальних залежностей.

Для кожної пари концептів (c_l,c_m) обчислюється показник $s(c_l \to c_m)$ на основі перевірки можливої каузальної залежності між ними. В даному випадку оцінюється каузальна залежність між концептами на основі поведінкових даних, коли за концептом c_l безпосередньо (тобто з використанням оператора $Next$) слідує концепт c_m:

$$s(c_l \to c_m)=\frac{1}{N}\sum_{n=1}^{N}\left(1\ iff\ \left((c_l,n)\,Next(c_m,n)\right)\right). \qquad (5)$$

Згідно (5) каузальний зв'язок між концептами у метальній моделі оцінюється на основі послідовності виявлення концептів (c_l,c_m) у вхідних даних. Тобто, якщо у поведінковій послідовності концепт c_l (наприклад, «аналіз симптомів») передує в часі концепту c_m (наприклад, «призначення тестів»), то значення функції під знаком суми дорівнюватиме 1. В іншому випадку значення дорівнюватиме 0. Усереднення за всіма отриманими послідовностями визначає частоту спільної появи пар (c_l,c_m) за умови, що c_l передує c_m. Орієнтоване ребро $c_l \to c_m$ додається до графа G, якщо $s(c_l \to c_m)$ перевищує порогове значення, а також відсутня ациклічність, тобто не існує шляху по графу у зворотному напрямку. Такий підхід забезпечує причинно наслідкове упорядкування пар (c_l,c_m) у графі. Тобто за умови, що c_l передує c_m, c_l можна розглядати як причину для c_m.

Перевірка каузальності через лінійну темпоральну логіку дає можливість підтвердити коректність каузальних дуг графа G. Для кожної каузальної дуги $c_l \to c_m$ формується обмеження лінійної темпоральної логіки:

$$G(c_l\ F\ c_m), \qquad (6)$$

де темпоральний оператор G означає, що послідовність $c_l \to c_m$ має виконуватись для всіх траєкторій, а темпоральний оператор F задає послідовність «c_l передує c_m у деякій наступній точці у майбутньому».

Відповідно, формула (6) має значення «якщо відбувається c_l, то пізніше відбудеться c_m».

Також у процесі перевірки обчислюється відношення послідовностей, для яких виконується умова (6), до загальної кількості послідовностей. Ребра зі значенням відношення меншим за порогове видаляються з графа як хибні залежності, які були сформовані внаслідок випадкових збігів у даних.

3. Метод побудови нейросимвольного представлення ментальної моделі рішення інтелектуальної системи

Розроблений метод формує ментальну модель на основі представленої двошарової архітектури. Метод включає наступні етапи.

Етап 1. Вилучення прихованих ознак варіаційним автокодувачем.

Вхідна послідовність, що відображає дії користувача, кодується у латентний простір. Декодер реконструює вхідну послідовність з латентного представлення для навчання енкодера. Регуляризація з використанням дивергенції Кульбака-Лейблера (вираз 1) забезпечує відсутність ізольованих кластерів у латентному просторі і, відповідно, створює умови для плавної інтерполяції.

Етап 2. Відбір ознак на основі уваги для фільтрації поведінкового шуму.

Латентне представлення обробляється багатоголовим механізмом уваги з метою виявлення значущих ознак. Для кожної голови уваги обчислюються запити, ключі та значення. Ваги уваги для кожної голови обчислюються за формулою (2). Виходи голів уваги агрегуються через конкатенацію (3). Фільтрація поведінкового шуму відбувається на основі відкидання ознак з низькими вагами уваги. В результаті знижується розмірність входу для символьного шару.

Етап 3. Нейросимвольне перетворення через відображення прихованих ознак у символьне представлення.

Відфільтровані латентні ознаки відображаються на множину символьних концептів через класифікатор. Класифікатор обчислює розподіл ймовірностей над попередньо визначеним концептуальним словником предметної області. Відбираються концепти з найвищими ймовірностями. Кожному концепту присвоюється семантичне вбудовування з попередньо навченого ембеддінгу для структурованого представлення у побудові графа.

Етап 4. Побудова орієнтованого ациклічного графа для відображення каузальної структури ментальної моделі.

На базі множини символьних концептів будується орієнтований ациклічний граф, який представляє каузальну структуру ментальної моделі. Для кожної пари концептів обчислюється оцінка каузальності (5).

Орієнтоване ребро додається до орієнтованого графа, якщо оцінка каузальності перевищує порогове значення (наприклад, 0,5), а також шлях у зворотному напрямку відсутній.

Етап 5. Перевірка орієнтованого ациклічного графа з використанням лінійної темпоральної логіки.

Каузальні ребра графа перевіряються на відповідність обмеженням, представленим формулами лінійної темпоральної логіки. Для кожного каузального ребра формується обмеження у вигляді правила типу Future

лінійної темпоральної логіки: $G(c_l \ F \ c_m)$. Порушення темпоральних обмежень свідчать про некоректні каузальні залежності. При перевірці обмежень обчислюється відношення послідовностей, що задовольняють обмеженню, до загальної кількості послідовностей. Ребра графа, що мають значення відношення нижче, ніж порогове значення (наприклад, 0,8), вилучаються з графа, оскільки вони моделюють нестійкі каузальні залежності.

Отриманий в результаті імплементації методу направлений граф дає можливість пояснити зв'язки між елементами рішення у ментальній моделі.

4. Експериментальна перевірка розробленого методу

Експериментальна перевірка розробленого методу виконана з використанням синтетичних медичних даних, отриманих з використанням рушія Synthea (Synthetic Health Data Engine). Ці дані моделюють клінічні траєкторії пацієнтів на основі реальних епідеміологічних даних та клінічних протоколів США. Набір даних представлено в офіційному репозиторії Synthea за посиланням https://mitre.box.com/shared/static/aw9po06ypfb9hrau4jamtvtz0e5ziucz.zip. З повного набору даних для експерименту відібрано таблицю conditions.csv, яка містить часові послідовності станів пацієнтів (діагнози, соціальні та поведінкові фактори здоров'я) з повними часовими мітками початку (START) та завершення (STOP) кожного стану, ідентифікаторами пацієнта (PATIENT) та епізоду надання допомоги (ENCOUNTER), а також кодами та описами станів

Набір даних містить 38 094 записів станів для 1 147 пацієнтів з 26 904 унікальними епізодами надання допомоги та 202 унікальними кодами станів. Середня кількість станів на одного пацієнта дорівнює 33,2. Дані охоплюють період з 1944 до 2024 року, що забезпечує можливість перевірки темпоральних обмежень. Для кожного пацієнта задано впорядковану послідовність станів згідно з міткою START, що відповідає поведінковим послідовностям користувачів у постановці задачі методу.

При проведенні експерименту використано спрощену реалізацію запропонованої двошарової архітектури, яка забезпечує повну реалізацію методу, але дає можливість знизити обчислювальні витрати.

Для формування вхідних ознак нейромережевого шару кожний стан класифіковано за тривалістю у три категорії: гострі короткострокові стани (acute_short, тривалість менше 30 днів), середньострокові стани (medium_term, тривалість від 30 до 365 днів) та хронічні/тривалі стани (chronic_or_open, тривалість більше 365 днів або відсутність дати завершення STOP). Ця класифікація відповідає доменній інтерпретації у медичній практиці: гострі епізодічні захворювання (вірусні інфекції, травми), середньострокові курси лікування (реабілітація, терапія) та хронічні стани (гіпертензія, діабет, серцева недостатність). Для кожного пацієнта обчислено вектор з семи ознак: частки трьох типів станів у його послідовності (acute_prop, medium_prop, chronic_prop), середня та максимальна тривалість станів (avg_dur, max_dur), кількість унікальних епізодів надання допомоги (n_enc) та часовий розмах між першим та останнім станом (span_days).

У процесі експериментальної перевірки оцінювались можливість пояснити рішення з використанням нейросимвольного представлення ментальної моделі, а також точність персоналізації, темпоральна узгодженість та час виконання. При оцінці можливості пояснити рішення використана шкала від 1 до 5, де 5 відповідає повній прозорості каузальних залежностей без потреби додаткових пояснень. Точність персоналізації оцінювалась як якість кластеризації латентних представлень. Для оцінки використано коефіцієнт силуету у відсотковій шкалі, який визначає, наскільки кожен елемент даних підходить до свого кластера у порівнянні з іншими кластерами. Темпоральна узгодженість розраховується як відношення кількості шляхів лікування, для яких виконана формула лінійної темпоральної логіки, до загальної кількості шляхів лікування. Оскільки кожний шлях лікування у наборі даних пов'язаний із окремим пацієнтом, то розрахунок виконано по пацієнтам. Обчислювальні витрати при проведенні експерименту оцінювались через час виконання у секундах в розрахунку на одного користувача.

Порівняння розробленого методу виконано для трьох базових підходів. В рамках першого підходу використана нейромережна архітектура без символьного шару та без темпоральної перевірки. Другий базовий підхід базується на використання лише правил виду: «якщо chronic_prop > 0,6, то кластер «хронічний»; «якщо acute_prop > 0,4, то кластер «гострий»; інакше кластер «змішаний». Тобто даний підхід є типовим для класичних експертних систем із апріорно заданими правилами. В рамках третього базового підходу використана гібридна архітектура із типовою кластеризацією, без механізму уваги, без проекції у латентний простір та без фільтрації значущих ознак. Тобто дана архітектура не містить механізму інтеграції шарів.

Результати експериментальної перевірки наведено у табл. 1.

Оцінка пояснення для розробленого методу становить 4,3 з 5,0, що суттєво перевищує можливості чистої нейромережної архітектури (3,2 з 5,0) завдяки явним символьним концептам та каузальному графу переходів між типами станів. Чиста символьна архітектура має найвищий рівень зрозумілості пояснень (4,5 з 5,0) внаслідок використання детермінованих правил, але поступається розробленому методу за точністю персоналізації. Нейромережна архітектура показує таку саму точність персоналізації (77,6%), але не надає можливості побудови інтерпретованих ментальних

моделей внаслідок відсутності символьного шару. Гібридна архітектура без механізму уваги демонструє нижчий рівень персоналізації внаслідок шуму у вхідних ознаках, оскільки цей шум не було відфільтровано з використанням механізму уваги.

Таблиця 1

Порівняння методів побудови ментальних моделей на даних Synthea

Метод	Оцінка пояснюваності (1–5)	Точність персоналізації (%)	Темпоральна узгодженість (%)	Час виконання (сек/користувач)
Нейромережна архітектура	3,2	77,6	–	3,1
Символьна архітектура	4,5	78,2	80,4	12,3
Гібридна архітектура без уваги	3,6	75,2	81,2	8,9
Розроблена двошарова архітектура	**4,3**	**77,6**	**82,8**	**6,8**

Темпоральна узгодженість розробленого методу досягає 82,8%, що відповідає шляхам лікування з періодичними гострими епізодами на фоні хронічних станів. Зокрема, кластер «глибоко хронічні» має найвище значення темпоральної узгодженості на рівні 89,0%, кластер «стабільні хронічні» має значення 80,8%, а кластер «епізодичні» – 64,7%. Середньозважена темпоральна узгодженість 82,8% забезпечується внаслідок фільтрації хибних залежностей, лише пацієнти зі стійкими патернами переходів між станами задовольняють темпоральному правилу (6). Символьна архітектура має нижчу темпоральну узгодженість внаслідок того, що правила не є персоналізованими. Гібридна архітектура без механізму уваги має проміжне значення узгодженості у порівнянні з розробленим методом та підходом на основі правил. Темпоральна узгодженість для нейромережної архітектури не розраховувалась внаслідок відсутності символьних концептів й відповідних формальних залежностей, представлених засобами темпоральної логіки.

Обчислювальна ефективність розробленого методу має середнє значення, становить 6,8 секунд на одного користувача. Такі витрати часу є допустимими для пакетної обробки даних. Нейромережна архітектура має найменші витрати часу на користувача, через те, що не формувався направлений ациклічний граф й не виконувалась темпоральна перевірка. Використання правил пов'язано з найбільшими витратами часу внаслідок перевірки всіх правил на всіх послідовностях без фільтрації цих правил. Гібридна архітектура без уваги показує проміжний час виконання (8,9 секунди на користувача).

Порівняльний аналіз по кластерам показує, що кластер «глибоко хронічні пацієнти», який включає 484 особи, має найвищу темпоральну узгодженість 89%. Причина такого рівня узгодженості полягає в тому, що з 484 пацієнтів з хронічними станами 431 особа не має трьох і більше послідовних гострих епізодів після появи хронічного стану. Кластер 0 «стабільні хронічні пацієнти» включає 417 осіб та має рівень темпоральної узгодженість 80,8%. Тобто гострі епізоди виникають, але не формують довгі кластери, ймовірно завдяки менеджменту. Кластер «епізодичні пацієнти» включає 246 осіб та має найнижчу темпоральну узгодженість 64,7%, що свідчить про відсутність постійного контролю хронічних станів.

Таким чином, розроблений метод забезпечує можливість формування пояснень за рахунок символьного шару з трьома інтерпретованими ментальними моделями: «глибоко хронічні пацієнти», «стабільні хронічні пацієнти», «епізодичні пацієнти». Перехід між типами станів представляється у вигляді каузального графа. Програмно згенерований при проведенні експерименту граф ментальної моделі наведено на рис. 1. Використано мову програмування Python.

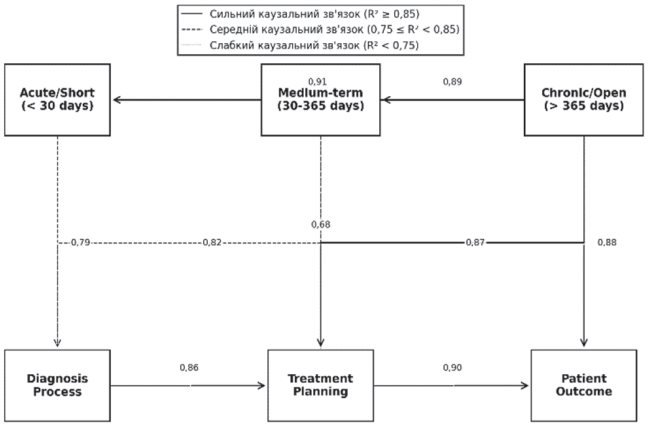

Рис. 1. Граф ментальної моделі

Представлений граф на верхньому рівні містить каузальні залежності між шістьма концептами ментальної моделі, які представляють типи захворювань – Acute/Short, Medium-term, Chronic/Open. Нижній рівень граф відображає клінічні процеси – Diagnosis Process, Treatment Planning, а також та результат лікування – Patient Outcome.

Граф включає 8 залежностей, виділених за силою зв'язку. Суцільна лінія позначає сильні каузальні зв'язки. Ми виявили п'ять сильних каузальних залежностей.

Залежність Chronic/Open – Acute/Short свідчить, що загальний стан приводить в майбутньому до виникнення загострень у захворюванні. Схожа каузальна залежність Chronic/Open – Medium-term демонструє перехід від хронічного стану до середньострокових епізодів. Найсильніша каузальна залежність Chronic/Open – Treatment підтверджує, що хронічні стани

приводять до ініціювання лікування, що відповідає клінічним протоколам.

Процес із залежністю Chronic/Open – Medium-term представляє собою традиційний клінічний потік робіт, в якому діагностика є передумовою для планування лікування.

Зв'язок Treatment Planning – Patient Outcome характеризує вплив лікування на результат. Велика вага правила підтверджує ефективність терапевтичного втручання.

Штрихова лінія маркує середні каузальні зв'язки. Зв'язок Acute/Short – Diagnosis (кореляція = 0,78) між гострими станами пацієнтів та діагностуванням описує типову ситуацію, коли пацієнти звертаються за невідкладною допомогою. Середня сила каузального зв'язку пояснюється випадками самолікування або відкладеного звернення.

Зв'язок Medium-term – Diagnosis (кореляція = 0,82) для середньострокових захворювань вказує на системний підхід до обстеження пацієнтів при тривалих симптомах хвороби.

Слабкий каузальний зв'язок Acute/Short – Patient Outcome вказує на високу варіативність результатів лікування при гострих станах. Такий зв'язок узгоджується з клінічною практикою, де прогноз залежить від таких чинників, як своєчасність звернення та стан здоров'я.

Таким чином, на основі аналізу графа ментальної моделі експерт може безпосередньо оцінити коректність віднесення пацієнта до кластера «глибоко хронічні» на основі частки його хронічних станів та протяжності клінічного шляху в часі. Персоналізація результатів реалізується через адаптивне виявлення прихованих ознак у латентному просторі, що дозволяє виявляти поведінкові патерни кожного пацієнта без додаткового налаштування правил. Кластеризація на латентних координатах виявила три стани пацієнта без попереднього знання про їх існування. Темпоральна узгодженість реалізується шляхом перевірки на відповідність формулам лінійної темпоральної логіки. Незалежність від предметної області забезпечується шляхом заміни концептуального словника. Наприклад, замість медичних станів можна використовувати відомі типи дій з підтримки прийняття рішень.

При проведенні експерименту були використані синтетичні медичні дані, які моделюють життєві траєкторії пацієнтів з високою варіативністю, в тому числі враховують випадкові травми, інфекції, а також соціальні фактори. Ця особливість темпоральних даних знижує темпоральну узгодженості порівняно з поведінковими логами медичної системи підтримки рішень, оскільки реальні логи медичних систем з фіксованими протоколами прийняття рішень мають більш стабільні темпоральні патерни без випадкових зовнішніх подій. Тобто метод демонструє темпоральну узгодженість 82,8 з використанням шляхів лікування з високою варіативністю. Проте на структурованих поведінкових логах медичних систем така узгодженість може бути суттєво вищою.

Метод потребує вхідних даних з темпоральними мітками при вирішенні задачі побудови персоналізованих пояснень.

Перспективи подальших досліджень включають адаптацію розробленого методу для інкрементного навчання в режимі онлайн при побудові ментальних моделей, а також інтеграцію з методами активного навчання для інтерактивного уточнення ментальних моделей в процесі діалога з користувачем.

5. Висновки

Запропоновано підхід до інтеграції нейромережного та символьного шарів для побудови ментальних моделей рішень з урахуванням вимог формування пояснень та адаптивності. Пояснення з використанням ментальної моделі базується на каузальних ациклічних графах з перевіркою отриманих казуальних залежностей за допомогою лінійної темпоральної логіки за умови збереження нейромережної адаптації до індивідуальних шаблонів поведінки користувача.

Розроблено метод побудови ментальної моделі на основі двошарової архітектури. Метод містить п’ять етапів: вилучення прихованих ознак варіаційним автокодувачем; відбір ознак на основі уваги для фільтрації поведінкового шуму; нейросимвольне перетворення через відображення прихованих ознак у символьне представлення; побудова орієнтованого ациклічного графа для відображення каузальної структури моделі; валідація з використанням лінійної темпоральної логіки. Метод створює умови для автоматизованого виявлення індивідуальних ментальних моделей користувачів із можливістю сформувати пояснення згідно особливостей предметної області.

Експериментальна перевірка на синтетичних медичних даних Synthea показала, що розроблений метод забезпечує можливість побудови персоналізованих пояснень на основі інтерпретації ментальної моделі користувача інтелектуальної системи.

Список літератури:

[1] Kautz H. The third AI summer: AAAI Robert S. Engelmore memorial lecture / H. Kautz // AI Magazine. – 2022. – Vol. 43, No. 1. – P. 93–104. DOI: https://doi.org/10.1002/aaai.12036.

[2] Gunning D. DARPA’s explainable artificial intelligence (XAI) program / D. Gunning, D. Aha // AI Magazine. – 2019. – Vol. 40, No. 2. – P. 44–58. DOI: https://doi.org/10.1609/aimag.v40i2.2850.

[3] Chalyi S. Externalization of tacit knowledge in the mental model of a user of an artificial intelligence system / S. Chalyi, I. Leshchynska // Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies. – 2024. – Vol. 1. – P. 91–96. https://doi.org/10.20998/2079-0023.2024.01.15.

[4] Johnson-Laird P. N. Mental models and human reasoning / P. N. Johnson-Laird // Proceedings of the National Academy of Sciences. – 2010. – Vol. 107, No. 43. – P. 18243–18250. DOI: https://doi.org/10.1073/pnas.1012933107.

[5] Hoefer M. Designing AI systems for mental model development / M. Hoefer, A. Felfernig // CEUR Workshop Proceedings. – 2025. – Vol. 3957. – P. 9–14.

[6] Bansal G. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff / G. Bansal, T. Wu, J. Zhou, et al. // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, No. 1. – P. 2429–2437. DOI: doi: 10.1609/aaai.v33i01.33012429.

[7] Чалий С.Ф., Лещинська І.О. Уточнення ментальної моделі рішення на основі доповнення вхідних даних в задачі формування пояснень в інтелектуальній системі. АСУ та прилади автоматики. – 2024. – Вип. 182. – С. 66-72. https://doi.org/ 10.30837/0135-1710.2024.182.066

[8] Sarker M. K. Neuro-symbolic artificial intelligence: Current trends / M. K. Sarker, L. Zhou, A. Eberhart, et al. // 10.48550/arXiv.2105.05330.

[9] Hitzler P. Neuro-symbolic integration for AI / P. Hitzler, A. Eberhart, M. Ebrahimi, et al. // Neuro-Symbolic Artificial Intelligence: The State of the Art. – National Science Review. 9. 10.1093/nsr/nwac035.

[10] Mao J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision / J. Mao, C. Gan, P. Kohli, et al. // 10.48550/arXiv.1904.12584. – 2019.

[11] Verma S. Counterfactual Explanations for Machine Learning: A Review/ Verma, Sahil & Dickerson, John & Hines, Keegan. // 10.48550/arXiv.2010.10596. – 2020.

[12] Koh P. W. Concept bottleneck models / P. W. Koh, T. Nguyen, Y. S. Tang, et al. // Proceedings of the 37th International Conference on Machine Learning (ICML). – 2020. – P. 5338–5348.

[13] Чалий С. Ф. Концептуальна ментальна модель пояснення в системі штучного інтелекту / С. Ф. Чалий, І. О. Лещинська // Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології. – 2023. – № 1(9). – С. 70–75. DOI: https://doi.org/10.20998/2079-0023.2023.01.11

[14] Chalyi S. A dynamic explanation model for human-computer interaction in an artificial intelligence system / S. Chalyi, I. Leshchynska // Advanced Information Systems. – 2020. – Vol. 4(4) 1. – P. 114–119. https://doi.org/10.20998/2522-9052.2020.4.16

[15] Чала О. В. Модель узагальненого представлення темпоральних знань для інтелектуальних систем підтримки прийняття рішень / О. В. Чала // Вісник Національного технічного університету «ХПІ». Системний аналіз, управління та інформаційні технології. – 2020. – Вип. 1(3). – С. 14–18. DOI: 10.20998/2079-0023.2020.01.03

[16] Levykin V. Development of a method of probabilistic inference of sequences of business process activities to support business process management / V. Levykin, O. Chala // Eastern-European Journal of Enterprise Technologies. – 2018. – Vol. 5, No. 3(95). – P. 16–24. DOI: https://doi.org/10.15587/1729-4061.2018.142664.

Надійшла до редколегії 30.10.2025

УДК 004.8:004.42:681.518 DOi 10.30837/ bi.2025.2(103).15

В. В. Мяляренко[1], О. Ю. Чередніченко[2]

[1]НТУ ХПІ, м. Харків, Україна, vladyslav.maliarenko@cs.khpi.edu.ua,
ORCID iD: 0009-0009-6064-061X

[2]НТУ ХПІ, м. Харків, Україна, olga.cherednichenko@khpi.edu.ua,
ORCID iD: 0000-0002-9391-5220

МОДЕЛІ ОБРОБКИ ТЕКСТОВИХ БІЗНЕС-ПРАВИЛ У СИСТЕМАХ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

У статті розглянуто проблему формалізації та автоматизованої обробки текстових бізнес-правил у системах підтримки прийняття рішень з урахуванням наявної структури даних та контексту предметної області. Запропоновано математичні моделі, що охоплюють три ключові етапи обробки правил: визначення структурних компонентів текстового бізнес-правила, побудову формальної таблиці рішень у нотації DMN та валідацію синтаксису й семантики згенерованої моделі за правилами логічної узгодженості. Модель структурного аналізу забезпечує формальне виділення умов, дій та залежностей із урахуванням доступних даних; модель генерації DMN визначає відповідність текстових конструкцій елементам таблиці рішень і враховує контекст системи підтримки рішень; модель валідації дозволяє виявляти логічні суперечності, неповноту та помилки узгодженості у формальній моделі. Представлено прототип програмної системи, що реалізує запропоновані моделі та дозволяє проводити експериментальне тестування їхньої ефективності. Результати експериментів демонструють коректність формалізації, повноту відображення бізнес-правил у DMN-таблиці та дієвість автоматичного виявлення структурних, синтаксичних і семантичних помилок.

БІЗНЕС-ПРАВИЛА, DMN, СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, ФОРМАЛЬНІ МОДЕЛІ, СТРУКТУРА ДАНИХ, СИНТАКСИЧНА ВАЛІДАЦІЯ, СЕМАНТИЧНА ВАЛІДАЦІЯ, АВТОМАТИЗОВАНЕ МОДЕЛЮВАННЯ

V. V. Maliarenko, O. Y. Cherednichenko. Models for Processing Textual Business Rules in Decision Support Systems. The paper addresses the problem of formalization and automated processing of textual business rules in decision support systems, taking into account the existing data structure and the contextual constraints of the decision-making domain. The study proposes mathematical models covering three key stages: identifying the structural components of a textual business rule, generating a formal DMN decision table, and validating the syntax and semantics of the resulting model in accordance with logical consistency rules. The structural analysis model extracts conditions, actions, and dependencies while incorporating the available data schema; the DMN generation model maps textual constructs to decision table elements with respect to the context of the decision support system; the validation model detects logical inconsistencies, incompleteness, and coherence errors in the formalized model. A prototype implementation is presented, enabling experimental evaluation of the proposed models. The results demonstrate correct formalization, completeness of business rule transformation into the DMN table, and effective detection of structural, syntactic, and semantic errors in the generated decision model.

BUSINESS RULES, DMN, DECISION SUPPORT SYSTEMS, FORMAL MODELS, DATA STRUCTURE, SYNTACTIC VALIDATION, SEMANTIC VALIDATION, AUTOMATED MODELING

Вступ

Сучасні системи підтримки прийняття рішень дедалі частіше спираються на явні бізнес-правила, які визначають логіку вибору альтернатив у складних організаційних і технічних системах. Стандарт Decision Model and Notation (DMN) надає формальну основу для подання такої логіки у вигляді виконуваних таблиць рішень, що сприяє прозорості, керованості та повторному використанню моделей рішень. Разом з тим на практиці бізнес-правила зазвичай формулюються у природній мові, а їх перетворення у формальні DMN-моделі потребує значних ручних зусиль і залишається джерелом помилок.

Останні дослідження демонструють потенціал використання штучного інтелекту для автоматизації подання логіки рішень і генерації DMN-артефактів з текстових описів [1]. Зокрема, показано, що поєднання мовних моделей зі структурними обмеженнями дозволяє значно скоротити час побудови таблиць рішень і підвищити їхню узгодженість зі схемою даних. Водночас проблема семантичної коректності згенерованих моделей, їхньої повноти та несуперечності залишається відкритою і потребує систематичного підходу до валідації.

У цій роботі пропонується комплексний підхід, який розглядає ідентифікацію семантики, генерацію DMN-таблиць і їх тестоорієнтовану валідацію як єдиний узгоджений процес.

1. Аналіз стану вирішеності задачі

Автоматизація формалізації бізнес-правил і побудови моделей рішень є активним напрямом досліджень у галузях систем підтримки прийняття рішень, бізнес-процесного моделювання та програмної інженерії. Застосування штучного інтелекту для генерації формальних представлень логіки рішень, зокрема у нотації DMN, продемонструвало свою перспективність

у зменшенні ручної роботи та підвищенні узгодженості моделей [1]. Водночас стандарт DMN визначає лише формат і семантику моделей рішень, не регламентуючи методів їх отримання з неформальних текстових описів [2].

У роботах, присвячених інтеграції DMN у бізнес-процеси, показано, що винесення логіки рішень у окремі DMN-артефакти підвищує прозорість і керованість процесів, зменшує складність BPMN-моделей і полегшує супровід рішень [3]. Водночас побудова DMN-таблиць у таких підходах, як правило, здійснюється вручну або напівавтоматично, що обмежує їх застосовність для великих і динамічних наборів правил.

Значна частина досліджень спрямована на вилучення логіки рішень із природномовних текстів за допомогою методів обробки природної мови. Запропоновані підходи використовують як правило-орієнтовані NLP-техніки, так і нейронні моделі для ідентифікації умов, дій і залежностей між ними [4], [5]. Такі методи демонструють здатність відновлювати елементи моделей рішень або діаграми вимог до рішень, однак зазвичай не забезпечують автоматичної генерації повністю виконуваних DMN-таблиць і не гарантують семантичної узгодженості з наявною схемою даних.

Поява великих мовних моделей (LLM) суттєво розширила можливості автоматичної генерації структурованих артефактів із тексту. Дослідження показують, що сучасні LLM здатні генерувати табличні структури та формалізовані правила, у тому числі у форматі, наближеному до DMN [6], [7]. Водночас такі моделі схильні до галюцинацій, логічних суперечностей і порушення структурних обмежень, що робить неможливим їх пряме використання без додаткових механізмів контролю.

Для підвищення структурної коректності результатів генерації запропоновано schema-aware підходи, які передбачають ін'єкцію схем даних, використання формальних форматів виводу та обмежень на синтаксис згенерованих структур [8]. Хоча ці методи покращують відповідність результатів формальним вимогам, вони не вирішують проблему семантичної коректності логіки рішень і не гарантують повноти покриття можливих комбінацій вхідних умов.

Окремим напрямом досліджень є Retrieval-Augmented Generation (RAG), що поєднує мовні моделі з пошуком релевантного зовнішнього контексту. Залучення доменних документів і правил дозволяє зменшити кількість пропусків і підвищити повноту згенерованих моделей [9], [10]. Разом з тим у літературі відзначається, що без додаткових механізмів семантичної фільтрації RAG може вводити суперечливі або нерелевантні знання, що негативно впливає на якість рішень [11].

Валідація DMN-таблиць є добре дослідженою задачею. Існуючі підходи охоплюють виявлення неповноти, перекриття правил, суперечностей і надлишковості за допомогою формальних методів та виконуваних рушіїв DMN [12], [13]. Проте більшість таких робіт розглядають валідацію як ізольований етап, який застосовується після побудови моделі, і не інтегрують її безпосередньо в процес автоматизованої генерації.

Отже, аналіз сучасних досліджень свідчить про наявність розриву між методами інтерпретації текстових бізнес-правил, генерацією формальних DMN-структур і їх систематичною семантичною валідацією. Це обґрунтовує необхідність комплексного підходу, у якому ідентифікація семантики, генерація моделей рішень і їх тестоорієнтована валідація розглядаються як єдиний узгоджений процес.

Текстові бізнес-правила, що використовуються у системах підтримки прийняття рішень, за своєю природою є неформальними специфікаціями логіки рішень. Вони формулюються природною мовою, орієнтовані на експертів предметної області та призначені для комунікації намірів і політик. Водночас практичне використання таких правил у програмних системах потребує їхнього перетворення у формальні, машинно-інтерпретовані моделі, зокрема у вигляді таблиць рішень DMN. Таким чином, ключовою проблемою є розрив між неформальним текстовим описом правила та його формальною реалізацією у вигляді виконуваної моделі прийняття рішень.

З точки зору теорії програмної семантики, текстове бізнес-правило можна розглядати як неформальну специфікацію програми, що задає відображення з простору вхідних станів у простір результатів рішення. Важливою особливістю цієї задачі є те, що відновлення семантики не може виконуватися у відриві від контексту використання правила. Інтерпретація тексту суттєво обмежується наявною схемою даних, яка визначає допустимі вхідні змінні, їх типи та домени, а також можливі результати рішення. Таким чином, постановку задачі ідентифікації бізнес-правил доцільно розглядати як задачу відновлення формальної семантики з неформального текстового опису з урахуванням типової та доменної семантики системи.

Метою даного дослідження є розроблення та експериментальна оцінка формалізованого підходу до автоматизованої обробки текстових бізнес-правил, який забезпечує відновлення їх формальної семантики та генерацію виконуваних DMN-таблиць рішень з контролем коректності, повноти та несуперечності результатів.

2. Методологія дослідження

Дослідження побудовано за принципом послідовного уточнення семантики. На першому етапі виконується ідентифікація бізнес-правил, у межах якої текстові описи інтерпретуються як джерело абстрактної семантики рішення. Цей етап формалізується як задача відновлення формального правила з тексту за

наявності обмежень схеми даних і доменного контексту. Для зменшення неоднозначності інтерпретації використовується підхід RAG, який дозволяє інтегрувати зовнішній доменний контекст у процес семантичного аналізу тексту. У результаті формується множина формальних правил, що визначають умови застосування та очікувані результати рішення.

Другий етап полягає у генерації таблиць рішень DMN як семантичного уточнення ідентифікованих формальних правил. Генерація розглядається не як механічне відображення логічних формул у табличну форму, а як процес побудови операційної семантики, що є семантично еквівалентною абстрактній специфікації. На цьому етапі виконується декомпозиція умов правил, визначення вхідних і вихідних атрибутів, а також фіксація політики виконання рішень відповідно до стандарту DMN. Згенеровані таблиці проходять синтаксичну перевірку, що гарантує їхню виконуваність.

Третій етап присвячено валідації згенерованих DMN-таблиць. На відміну від традиційних підходів, орієнтованих на ручний аналіз або формальну верифікацію, у роботі запропоновано тестоорієнтовану методологію валідації, яка відтворює практики, що застосовуються експертами предметної області. Ключовою особливістю цього етапу є незалежна генерація тестового покриття на основі тих самих джерел знань, що й генерація таблиць, але без використання їхньої структури. Тестові випадки формуються як конкретизації очікуваної семантики рішення для репрезентативних вхідних станів з урахуванням схеми даних і доменних обмежень.

Процес валідації організовано ітеративно. Після успішної синтаксичної перевірки DMN-таблиця тестується на згенерованому наборі тестів, а результати порівнюються з очікуваними значеннями, отриманими з референтної семантичної моделі. Якщо таблиця не задовольняє задані порогові критерії точності та покриття, ініціюється цикл уточнення, який може включати повторну генерацію тестів або повторну генерацію DMN-таблиці. Такий підхід дозволяє поступово зменшувати семантичні розбіжності між абстрактною специфікацією та операційною моделлю за обмеженого бюджету обчислювальних спроб.

Запропоновано інтегрований конвеєр, який поєднує стохастичні компоненти, засновані на великих мовних моделях, із детермінованими процедурами перевірки та виконання рішень. Це дозволяє зберегти інтерпретованість і керованість процесу, водночас зменшуючи обсяг ручної роботи та ризик помилок. Такий підхід забезпечує відтворювану основу для експериментального оцінювання та розроблення промислових систем керування бізнес-правилами та підтримки прийняття рішень.

Формалізація задачі спирається на інтерпретацію текстового правила та розглядає процес ідентифікації як відновлення формального семантичного представлення, узгодженого з наявною схемою даних і контекстом предметної області. Для цього необхідно чітко визначити вхідні дані, простір допустимих формальних правил та критерії коректності інтерпретації.

Нехай $t \in \Sigma^*$ позначає текстове бізнес-правило, сформульоване природною мовою. Вважається, що текст сам по собі не визначає однозначної формальної семантики, а задає множину потенційних інтерпретацій, допустимих з точки зору мовного формулювання. Інтерпретація цього тексту здійснюється не ізольовано, а з урахуванням формальної схеми даних системи підтримки прийняття рішень та контексту предметної області. Схема даних задається як трійка $\mathcal{S} = (\mathcal{A}, \mathcal{D}, \tau)$, де $\mathcal{A}$ є множиною атрибутів, $\mathcal{D}$ – множиною доменів значень, а $\tau : \mathcal{A} \to \mathcal{D}$ є відображенням типізації атрибутів. Контекст предметної області $\mathcal{C}$ розглядається як сукупність доменних тверджень, обмежень і визначень, які уточнюють семантику термінів, що використовуються у тексті правила, та накладають додаткові обмеження на допустимі інтерпретації.

Метою ідентифікації є побудова формального бізнес-правила. Простір формальних бізнес-правил визначається як множина об'єктів вигляду $r = \Phi, \Psi, \kappa$, де Φ є логічною формулою умов застосування правила, визначеною над атрибутами зі схеми $\mathcal{S}$, Ψ є множиною дій або результатів рішення, а κ містить метадані, зокрема інформацію про джерело правила, його пріоритет або ступінь впевненості. Формула умов Φ інтерпретується як предикат над простором вхідних станів, тоді як дії Ψ визначають відображення з простору вхідних значень у простір результатів рішення.

З формальної точки зору задача ідентифікації полягає у знаходженні такого правила r^*, яке є допустимою інтерпретацією тексту t за наявних обмежень схеми та контексту. Це відображення можна подати у вигляді функції

$$f : (t, \mathcal{S}, \mathcal{C}) \to r^*,$$

де r^* належить простору формальних правил і задовольняє умови семантичної та типової коректності. Допустимість формального правила визначається, по-перше, узгодженістю зі схемою даних, що означає використання виключно атрибутів, визначених у $\mathcal{S}$, та коректність усіх логічних і обчислювальних операцій відносно їхніх типів. По-друге, правило повинно бути узгодженим з контекстом предметної області, тобто не порушувати політики, зафіксовані в $\mathcal{C}$.

Таким чином, формальна постановка задачі ідентифікації бізнес-правил створює основу для побудови моделей генерації та валідації рішень, а також для аналізу якості й повноти бізнес-логіки.

3. Формальні моделі обробки текстових бізнес-правил

Для побудови математичної моделі необхідно формально визначити простори вхідних даних, простір допустимих семантичних інтерпретацій та критерії вибору оптимального правила.

Нехай $\mathcal{A}$ є скінченною множиною атрибутів, визначених у схемі даних системи підтримки прийняття рішень, а $\mathcal{D}$ – множиною доменів значень, асоційованих з цими атрибутами. Типізація атрибутів задається відображенням $\tau : \mathcal{A} \to \mathcal{D}$. Простір вхідних станів рішення визначається як декартів добуток доменів вхідних атрибутів і позначається $\mathcal{X} = \prod_{a \in \mathcal{A}_\mathcal{I}} \tau(a)$, де $\mathcal{A}_I \subseteq \mathcal{A}$ – множина атрибутів, що використовуються як умови. Аналогічно, простір результатів рішення визначається як $\mathcal{Y} = \prod a \in \mathcal{A}_\mathcal{O} \tau(a)$, де $\mathcal{A}_\mathcal{O} \subseteq \mathcal{A}$ – множина атрибутів, що використовуються як результати.

Формальне бізнес-правило задається у вигляді пари $r = (\Phi, \Psi)$, де Φ є логічною формулою над змінними з $\mathcal{A}_\mathcal{I}$, а Ψ – відображенням, що кожному допустимому вхідному стану ставить у відповідність результат рішення. Формула умов Φ інтерпретується як предикат

$$\Phi : \mathcal{X} \to \text{true}, \text{false},$$

тоді як дія Ψ визначається як часткова функція

$$\Psi : \mathcal{X} \rightharpoonup \mathcal{Y}.$$

Відповідно, семантика правила r задається як часткова функція

$$[\![r]\!] : \mathcal{X} \rightharpoonup \mathcal{Y},$$

яка визначена для тих і лише тих вхідних станів, для яких виконується умова Φ.

Текстове бізнес-правило t розглядається як джерело семантичної інформації, яке визначає функцію подібності між текстом та формальним правилом. Нехай $sim(t,r)$ є числовою мірою семантичної відповідності формального правила r та тексту t, що відображає ступінь узгодженості логічної структури та термінів правила з природномовним описом з урахуванням контексту $\mathcal{C}$. Тоді задача ідентифікації формулюється як задача оптимального вибору:

$$r^* = \arg\max_{r \in \mathcal{R}_{\text{valid}}} \text{sim}(t,r),$$

де оптимальне правило r^* є найкращою семантичною апроксимацією текстового опису серед усіх допустимих формальних інтерпретацій.

У практичній постановці ідентифікації бізнес-правил функція семантичної відповідності $sim(t,r)$ повинна одночасно відображати близькість формального правила r до змісту тексту t, узгодженість із схемою даних $\mathcal{S}$ та несуперечність доменному контексту $\mathcal{C}$. У підході на основі RAG ці три компоненти інтегруються в єдину цільову функцію через явне моделювання релевантного підконтексту $K \subseteq \mathcal{C}$, який вибирається механізмом пошуку й подається на вхід мовній моделі разом із текстом правила та інформацією про схему. Це дозволяє розглядати ідентифікацію як задачу максимізації апостеріорної правдоподібності формальної інтерпретації за умов обмеженого контексту.

Нехай оператор пошуку для заданого правила t повертає множину фрагментів контексту $K = \{k_j\}_{j=1}^m$, відібраних з $\mathcal{C}$. K визначається як результат максимізації семантичної релевантності в ембединг-просторі, тобто

$$K = \arg\max_{|K|=m} \sum_{k \in K} sim_{emb}\left(e(t), e(k)\right),$$

де $e(\cdot)$ – відображення у векторний простір, а sim_{emb} – міра близькості (наприклад, косинусна). Важливо, що ця подібність використовується лише для формування K, тоді як цільова функція ідентифікації оперує вже семантикою кандидата r.

Далі вводиться ймовірнісна інтерпретація генератора. Нехай $P_\theta(rt, \mathcal{S}, K)$ – умовний розподіл, що реалізується мовною моделлю з параметрами θ, яка генерує кандидатні правила r за умови тексту t, схеми $\mathcal{S}$ та контексту K. Тоді природною конкретизацією є визначення

$$sim(t,r) \equiv \log P_\theta(rt, \mathcal{S}, K) - \Omega(r; \mathcal{S}, K),$$

де Ω – штрафний функціонал.

Щоб зробити інтерпретованим і контрольованим, його доцільно розкласти на компоненти, що відповідають ключовим вимогам до правила: схемній узгодженості, типовій коректності, контекстній узгодженості та структурній придатності до подальшої компіляції в DMN. У цьому випадку можна записати

$$\Omega(r; \mathcal{S}, K) = \lambda_1, \pi_{\text{link}}(r; \mathcal{S}) + \lambda_2, \pi_{\text{type}}(r; \mathcal{S}) + \\ + \lambda_3, \pi_{\text{ctx}}(r; K) + \lambda_4, \pi_{\text{struct}}(r),$$

де кожний $\pi(\cdot) \geq 0$ є мірою порушення відповідної властивості, а λ_i задають ваги.

Компонент π_{link} формалізує вимогу, що всі змінні, використані у формулі умов і діях правила, мають бути зіставлені з атрибутами зі схеми. Компонент π_{type} відповідає за типову коректність: усі предикати в Φ та всі присвоєння у Ψ мають бути визначені для типів, що задані τ. У контексті DMN це також інтерпретується як умова коректності FEEL-виразів на рівні типів. Компонент π_{ctx} моделює узгодженість правила з контекстом. Його можна визначати як штраф за порушення доменних інваріантів або як штраф за відсутність логічної підтримки ключових тверджень у K. Нарешті, π_{struct} відповідає за структурну придатність правила до подальшої компіляції в DMN. Ідея полягає в тому, що навіть семантично коректне правило може бути сформульоване у вигляді, що ускладнює нормалізацію (наприклад, надлишкова вкладеність, неявні залежності, змішування кількох рішень в одному правилі). Типовим прикладом формалізації є штраф за складність формули умов, зокрема за кількість кон'юнктів/диз'юнктів та глибину дерева.

Після введення цієї конкретизації явно визначається, що LLM у зв'язці з RAG виконує роль стохастичного семантичного "декодера" формального правила, тоді як штрафний функціонал включає формальні вимоги схеми даних і предметної області, перетворюючи задачу на керований процес відновлення семантики, придатний до подальшої генерації та валідації DMN-моделей.

Після етапу ідентифікації бізнес-правила результатом обробки текстового опису є формальне правило, яке задає абстрактну семантику рішення у вигляді логічної умови застосування та відповідної дії або множини дій. Таке правило є машинно-інтерпретованим, однак воно ще не є безпосередньо виконуваною моделлю у складі системи підтримки прийняття рішень. Наступним кроком є перетворення цієї абстрактної семантики у конкретну операційну форму, сумісну зі стандартами виконання рішень, зокрема у вигляді таблиці рішень Decision Model and Notation. З концептуальної точки зору ця задача полягає у семантичному уточненні формального правила, тобто у переході від декларативного опису логіки до структурованої, детермінованої та виконуваної моделі.

У термінах програмної семантики формальне бізнес-правило, отримане на попередньому етапі, можна розглядати як абстрактну специфікацію програми прийняття рішення. Воно визначає, за яких умов повинні виконуватися певні дії, але не фіксує конкретний спосіб організації цієї логіки у вигляді окремих альтернативних випадків, порядку перевірки умов чи механізму розв'язання конфліктів. Натомість DMN-таблиця рішень задає операційну семантику цієї специфікації, явно визначаючи множину рядків правил, структуру вхідних та вихідних змінних, а також політику обробки ситуацій, коли кілька правил можуть бути застосовними одночасно. Таким чином, генерація DMN-таблиці є процесом уточнення семантики, у якому усувається недовизначеність, притаманна абстрактному правилу, і вводяться всі необхідні деталі для виконання.

Якісно задача генерації таблиці рішень полягає у побудові такої DMN-структури, яка є семантично еквівалентною ідентифікованому формальному правилу в межах допустимого простору вхідних станів, визначеного схемою даних і контекстом предметної області. Це означає, що для кожного допустимого набору вхідних значень таблиця рішень повинна або породжувати той самий результат, що й абстрактне правило, або, у випадках неповної специфікації, явно фіксувати відсутність застосовного рішення. Генерація таблиці не повинна вводити нових рішень, не передбачених семантикою правила, і не повинна втрачати жодних рішень, які правило дозволяє.

Водночас семантичне уточнення не є тривіальним механічним перетворенням. Формальне правило може містити складні логічні формули з кон'юнкціями, диз'юнкціями та неявними залежностями між умовами, тоді як DMN-таблиця вимагає явної декомпозиції логіки на набір дискретних рядків, кожен з яких відповідає певній комбінації вхідних умов. Тому генерація таблиці рішень передбачає нормалізацію умов, виділення релевантних вхідних атрибутів, розгортання складних логічних виразів у множину альтернативних випадків і, за потреби, доповнення таблиці службовими рядками для забезпечення повноти або коректної обробки винятків.

Суттєвим аспектом цієї задачі є вибір та фіксація політики виконання рішень, зокрема політики обробки збігів правил. У формальному правилі порядок або спосіб застосування умов може бути неявним або взагалі не визначеним, тоді як у DMN він має бути зафіксований через відповідну hit policy. Таким чином, генерація таблиці рішень включає не лише структурну декомпозицію логіки, а й прийняття семантичних рішень щодо того, як інтерпретувати потенційні конфлікти або перекриття умов, спираючись на метадані правила та доменний контекст.

Отже, задачу генерації DMN-таблиці доцільно розглядати як задачу побудови операційної семантики, яка є коректним і повним семантичним уточненням абстрактної семантики ідентифікованого бізнес-правила. Така постановка створює концептуальне підґрунтя для подальшої математичної формалізації процесу генерації таблиць рішень, у якій буде чітко визначено умови семантичної еквівалентності, критерії повноти та обмеження, накладені стандартом DMN і структурою даних системи підтримки прийняття рішень.

Математична модель генерації таблиці рішень ґрунтується на трактуванні формального бізнес-правила як абстрактної семантики рішення та DMN-таблиці як її семантичного уточнення. Метою генерації є побудова такої таблиці рішень, семантика якої є еквівалентною семантиці ідентифікованого правила в межах допустимого простору вхідних станів, визначеного схемою даних і доменними обмеженнями.

Нехай формальне правило задане у вигляді $r = (\Phi, \Psi)$, де Φ є логічною формулою над атрибутами зі схеми даних, а Ψ – відображенням, що визначає результат рішення для станів, у яких умова Φ виконується. Семантика цього правила визначається як часткова функція

$$[\![r]\!] : \mathcal{X} \rightharpoonup \mathcal{Y},$$

яка для кожного вхідного стану $x \in \mathcal{X}$ визначена тоді й лише тоді, коли $x \vDash \Phi$.

DMN-таблиця рішень формалізується як скінченна структура

$$\mathcal{D} = \langle I, O, R, H, \sigma \rangle,$$

де $I \subseteq \mathcal{A}$ є множиною вхідних атрибутів, $O \subseteq \mathcal{A}$ – множиною вихідних атрибутів, $R = 1,\ldots,m$ – множиною рядків таблиці, H – політика обробки збігів правил

(hit policy), а σ задає для кожного рядка $j \in R$ пару (Φ_j, Ψ_j), що відповідає умовам і результатам цього рядка. Кожна формула Φ_j є кон'юнкцією локальних обмежень над вхідними атрибутами, а Ψ_j – конкретизацією значень вихідних атрибутів.

Семантика DMN-таблиці визначається як відображення

$$[\![\mathcal{D}]\!] : \mathcal{X} \rightharpoonup \mathcal{Y}.$$

Задача генерації DMN-таблиці формулюється як задача побудови такої структури $\mathcal{D}$, що семантика таблиці є еквівалентною семантиці формального правила на допустимій області визначення. Формально це означає виконання умови семантичної еквівалентності

$$\forall x \in \mathcal{X}_C : \ [\![r]\!](x) = [\![\mathcal{D}]\!](x),$$

де $\mathcal{X}_C \subseteq \mathcal{X}$ – множина вхідних станів, допустимих з точки зору контексту предметної області.

Таким чином, математична модель генерації DMN-таблиці формалізує цей етап як задачу побудови операційної семантики, що є коректним семантичним уточненням формального бізнес-правила.

Після генерації DMN-таблиці бізнес-правило отримує форму, придатну до виконання у складі системи підтримки прийняття рішень. Проте сам факт успішної генерації таблиці не гарантує її коректності з точки зору семантики рішення, повноти охоплення допустимих ситуацій або узгодженості з доменними обмеженнями. Тому необхідним завершальним етапом конвеєра є валідація таблиці рішень, яка має на меті перевірити, що отримана операційна модель дійсно реалізує очікувану логіку рішення та не містить прихованих дефектів.

Якісно задачу валідації доцільно розглядати як перевірку семантичної коректності DMN-таблиці відносно трьох взаємопов'язаних джерел вимог. По-перше, таблиця повинна бути узгодженою з формальним бізнес-правилом, яке слугувало основою для її генерації. Це означає, що таблиця не повинна породжувати результатів, не передбачених абстрактною семантикою правила, і не повинна втрачати допустимі результати для жодного вхідного стану, на якому правило визначене. По-друге, таблиця повинна відповідати схемі даних системи підтримки прийняття рішень, зокрема використовувати лише визначені атрибути, коректно оперувати їхніми типами та доменами і не створювати внутрішніх типових суперечностей. По-третє, таблиця повинна бути узгодженою з контекстом предметної області, що включає доменні інваріанти, політики та обмеження, які не завжди явно представлені у формальній структурі правила.

З точки зору семантики програм, валідацію DMN-таблиці можна інтерпретувати як перевірку того, що операційна семантика, реалізована таблицею, є коректним семантичним уточненням абстрактної семантики бізнес-правила. Це передбачає аналіз поведінки таблиці на всьому допустимому просторі вхідних станів і виявлення ситуацій, у яких операційна модель поводиться неочікувано або неоднозначно. Зокрема, навіть за наявності семантичної еквівалентності на рівні окремих правил, таблична форма може містити конфлікти між рядками, перекриття умов або неохоплені комбінації вхідних значень, що призводять до непередбачуваної або некоректної поведінки під час виконання.

Суттєвим аспектом валідації є перевірка повноти таблиці рішень. У контексті систем підтримки прийняття рішень повнота означає, що для кожного допустимого вхідного стану, визначеного схемою даних і доменними обмеженнями, таблиця або породжує однозначний результат, або явно сигналізує про відсутність застосовного рішення. Неповні таблиці створюють ризик неявних помилок під час експлуатації системи, оскільки поведінка в непокритих випадках часто залежить від реалізації виконавчого рушія або зовнішньої логіки обробки помилок.

Іншим критично важливим аспектом є виявлення внутрішніх суперечностей і неоднозначностей. Таблиця рішень може містити кілька рядків, умови яких перекриваються, але результати є несумісними або обробляються некоректно з точки зору обраної політики виконання. Навіть якщо така таблиця формально виконується рушієм DMN, її семантика може бути неочевидною або суперечливою з точки зору бізнес-логіки. Тому якісна валідація повинна виявляти не лише явні синтаксичні помилки, а й глибші семантичні дефекти, пов'язані з конфліктами правил.

Отже, задачу валідації таблиць рішень DMN доцільно розглядати як завершальний етап семантичного аналізу, який перевіряє відповідність операційної моделі рішень абстрактній семантиці бізнес-правил, структурі даних системи та доменним обмеженням. Така постановка створює основу для подальшої математичної формалізації процедур валідації, зокрема для формального означення повноти, несуперечності та семантичної коректності DMN-таблиць рішень.

Нехай $\mathcal{S} = (\mathcal{A}, \mathcal{D}, \tau)$ – схема даних, $\mathcal{C}$ – доменний контекст, $T = \{t_i\}_{i=1}^{N}$ – множина текстових правил. За результатом етапу ідентифікації отримуємо формальну семантичну специфікацію рішення у вигляді множини правил $\mathcal{R} = \{r_i\}_{i=1}^{N}$, яка створює «еталонну» семантику $F_{\mathcal{R}} : \mathcal{X} \rightharpoonup \mathcal{Y}$ на допустимій області $\mathcal{X}_C \subseteq \mathcal{X}$. Далі генератор будує DMN-таблицю $\mathcal{D}$, яка задає операційну семантику $F_{\mathcal{D}} : \mathcal{X} \rightharpoonup \mathcal{Y}$.

Ключовий елемент підходу – незалежний генератор тестів, який використовує лише $(T, \mathcal{S}, \mathcal{C})$ або еквівалентно $(\mathcal{R}, \mathcal{S}, \mathcal{C})$, але не залежить від конкретної таблиці $\mathcal{D}$. Формально тест визначимо як пару $\xi = (x, \mathcal{O})$, де $x \in \mathcal{X}_C$ – вхідний стан, а $\mathcal{O} \subseteq \mathcal{Y}$ – множина припустимих очікуваних результатів. Використання

множини $\mathcal{O}$, а не одного значення, дозволяє коректно працювати з випадками недовизначеності або альтернатив у текстових правилах, а також із ситуаціями, коли бізнес-логіка визначає лише частину виходів. Тоді тестовий набір $\Xi = \{\xi_k\}_{k=1}^{m}$ є скінченною емпіричною специфікацією очікуваної поведінки рішення.

Генерацію тестів можна подати як стохастичний оператор

$$TestGen_{\eta} : (T, \mathcal{S}, \mathcal{C}) \Rightarrow \Xi,$$

параметризований η (наприклад, розмір набору, стратегія покриття, баланс типових і граничних випадків). При цьому *TestGen* можна інтерпретувати як вибір розподілу $Q\eta$ над $\mathcal{X}_{\mathcal{C}}$ та побудову тестів шляхом семплінгу $x \sim Q$ і виведення очікуваних результатів $\mathcal{O}(x)$ з первинних артефактів правил і контексту. Узагальнено,

$$\xi = (x, \mathcal{O}(x)),\ x \sim Q_{\eta},\ \mathcal{O}(x) = Oracle(T, \mathcal{S}, \mathcal{C}; x),$$

де *Oracle* — процедура, що синтезує очікування (через формальну семантику $\mathcal{R}$, через LLM+RAG-інтерпретацію, або їх комбінацію з консервативними правилами).

На основі запропонованих моделей ідентифікації бізнес-правил, генерації таблиць рішень та тестової валідації можна сформулювати інтегрований алгоритм, який реалізує повний конвеєр автоматизованої обробки текстових бізнес-правил у виконувані та перевірені DMN-таблиці (рис. 1). Вхідними даними алгоритму є множина текстових бізнес-правил T, схема даних системи підтримки прийняття рішень $\mathcal{S}$, доменний контекст $\mathcal{C}$, а також параметри керування процесом, зокрема пороги прийнятності тестової валідації та бюджети ітерацій. Вихідним результатом є або прийнята DMN-таблиця рішень, семантично узгоджена з бізнес-правилами, або множина контрприкладів, які вказують на неможливість досягнення прийнятної моделі за заданих обмежень.

Algorithm 1: End-to-end pipeline for DMN generation and validation

Input: Textual business rules T, data schema $\mathcal{S}$, domain context $\mathcal{C}$, thresholds (α, γ), budgets (B_T, B_D)
Output: Accepted DMN decision table $\mathcal{D}$ or counterexamples $\mathcal{F}$

```
R ← IdentifyRules(T, S, C);
g ← 0;
while g < B_D do
    D ← GenerateDMN(R, S, C);
    if not SynValid(D, S) then
        g ← g + 1;
        continue;
    Ξ ← TestGen(R, S, C);
    k ← 0;
    while k < B_T do
        F ← ExecuteTests(D, Ξ);
        if Accept(D, F, α, γ) then
            return D;
        if UpdateTests(F) then
            Ξ ← RefineTests(Ξ, F);
            k ← k + 1;
        else
            break;
    g ← g + 1;
return F;
```

Рис. 1. Алгоритм обробки бізнес-правил

Алгоритм починається з етапу ідентифікації бізнес-правил, на якому текстові правила інтерпретуються як неформальні специфікації семантики рішення. За допомогою конвеєра LLM+RAG виконується побудова множини формальних правил $\mathcal{R}$, узгоджених зі схемою даних і доменним контекстом. На цьому етапі формується абстрактна семантика рішення, яка надалі використовується як основне джерело очікуваної поведінки системи. На наступному етапі з абстрактної семантики $\mathcal{R}$ генерується кандидатна DMN-таблиця рішень $\mathcal{D}$. Генерація здійснюється як семантичне уточнення формальних правил і включає вибір вхідних і вихідних атрибутів, декомпозицію умов у множину рядків та фіксацію політики виконання. Згенерована таблиця підлягає синтаксичній перевірці відповідно до стандарту DMN. Таблиці, що не проходять цей етап, відхиляються без подальшого тестування, і процес повертається до повторної генерації.

У практичних сценаріях використання DMN-таблиць рішень їхня коректність рідко перевіряється шляхом повного формального аналізу семантики. Натомість, поширеним підходом є валідація через тестування, коли експерт предметної області або аналітик формує набір тестових випадків, що відображають типові, граничні та критичні ситуації, і перевіряє поведінку таблиці на цих прикладах. Такий підхід інтуїтивно зрозумілий, добре масштабується на складні бізнес-правила та дозволяє виявляти практично значущі дефекти, навіть якщо формальна модель є складною або неповною. Водночас ручне формування тестів є трудомістким і залежить від досвіду експерта, що обмежує відтворюваність і автоматизацію процесу.

У цій роботі валідація таблиць рішень пропонується як автоматизований процес тестування, у якому тестове покриття генерується безпосередньо з тих самих джерел знань, що й сама таблиця рішень, а саме з текстових бізнес-правил, схеми даних та контексту предметної області. Ключовою ідеєю є принципова незалежність тестового покриття від конкретної згенерованої DMN-таблиці. Тести не виводяться з табличної структури і не повторюють її логіку, а натомість репрезентують очікувану семантику рішення, відновлену з первинних артефактів бізнес-логіки. Завдяки цьому тестування стає інструментом виявлення семантичних дефектів генерації, а не формальним підтвердженням внутрішньої узгодженості таблиці.

Якісно кожен тестовий випадок можна інтерпретувати як конкретизацію очікуваної поведінки бізнес-правила для певного вхідного стану. Вхідні значення тесту формуються з урахуванням схеми даних, типів атрибутів і доменних обмежень, тоді як очікуваний результат або множина допустимих результатів визначається семантикою правила та контекстом предметної області. Таким чином, тестовий набір фактично реалізує емпіричну специфікацію семантики рішення,

проти якої перевіряється операційна модель у вигляді DMN-таблиці.

Процес валідації передбачає, що згенерована DMN-таблиця спочатку проходить синтаксичну перевірку та перевірку типів, які гарантують її формальну виконуваність у рушії DMN. Після цього таблиця піддається тестуванню на згенерованому наборі тестових випадків. Для кожного тесту обчислюється результат виконання таблиці, який порівнюється з очікуваним результатом, визначеним тестом. Таблиця вважається семантично коректною, якщо вона проходить тестування з достатнім рівнем покриття та не демонструє критичних розбіжностей з очікуваною поведінкою.

Оскільки тестове покриття є скінченним наближенням до повної семантичної перевірки, процес валідації організовується ітеративно та з урахуванням порогових критеріїв. Для тестування встановлюється обмеження на кількість тестових спроб або на рівень досягнутого покриття простору допустимих вхідних станів. Якщо згенерована таблиця не проходить тестування в межах заданого порогу, система може ініціювати повторну генерацію тестів з метою уточнення перевірки або повторну генерацію DMN-таблиці з урахуванням виявлених невідповідностей. Таким чином формується замкнений цикл уточнення, у якому тестування виступає інструментом зворотного зв'язку між очікуваною семантикою правила та її операційною реалізацією.

Такий підхід дозволяє розглядати валідацію DMN-таблиць як процес стохастичної семантичної перевірки, що поєднує формальні обмеження схеми даних і контексту з практикою тестування, прийнятою у прикладних системах. Він зберігає інтерпретованість ідеї валідації для експертів предметної області, водночас створюючи основу для автоматизації та формального аналізу якості згенерованих моделей рішень. Це, у свою чергу, відкриває можливість подальшої математичної формалізації процесу тестового покриття, критеріїв прийнятності та ітеративного вдосконалення DMN-таблиць.

Центральною передумовою запропонованого підходу до валідації є наявність незалежного джерела очікуваної семантики рішення, яке використовується для побудови тестових випадків і перевірки згенерованих DMN-таблиць. Це джерело формалізується у вигляді референтної семантичної моделі, яка для заданого вхідного стану визначає допустимі результати рішення на основі текстових бізнес-правил, схеми даних і доменного контексту. Принципово важливо, що ця референтна модель не використовує структуру DMN-таблиці та не аналізує її рядки, що забезпечує незалежність тестового покриття від результату генерації.

Референтна семантична модель розглядається як процедура, що інтерпретує ідентифіковану абстрактну семантику бізнес-правил у конкретних точках простору вхідних станів. Для кожного допустимого вхідного набору значень вона повертає множину допустимих результатів, що відображає можливу альтернативність у початкових текстових правилах. Така постановка дозволяє коректно працювати з правилами, які частково визначають поведінку системи, наприклад задають лише обмеження або політики, але не фіксують точного значення виходу.

Генератор тестових випадків використовує референтну семантичну модель як допоміжний компонент, який виконує окрему задачу побудови репрезентативного тестового покриття. З точки зору семантики, кожен тестовий випадок можна інтерпретувати як точкову специфікацію поведінки правила, тобто як перевірку того, що для конкретного вхідного стану операційна модель рішення не суперечить абстрактній семантиці, відновленій з тексту. Сукупність таких тестів утворює скінченне емпіричне наближення до повної семантичної перевірки.

Отже, після успішної синтаксичної перевірки запускається етап валідації. Незалежно від структури таблиці генерується тестове покриття $\mathcal{T}$ на основі $(\mathcal{R},\mathcal{S},\mathcal{C})$ із використанням семантичної моделі очікувань. Кожен тестовий випадок виконується на таблиці $\mathcal{D}$, а отримані результати порівнюються з допустимими очікуваннями. За результатами тестування обчислюються метрики точності та покриття, на основі яких приймається рішення про прийнятність таблиці. Якщо таблиця задовольняє порогові критерії, алгоритм завершується успішно, повертаючи $\mathcal{D}$ як валідну модель рішення. Якщо ж ні, то алгоритм переходить до ітеративного режиму уточнення. Залежно від характеру виявлених помилок ініціюється або регенерація тестового покриття, або повторна генерація DMN-таблиці. Цей процес повторюється до досягнення прийнятного результату або до вичерпання заданого бюджету ітерацій. Узагальнений псевдокод алгоритму наведено на рис.1.

Отже, запропонований алгоритм узагальнює всі розглянуті вище моделі в єдиний формальний процес. Він поєднує інтерпретацію неформальних бізнес-правил, семантично вмотивовану генерацію DMN-таблиць та практично орієнтовану валідацію. Така структура створює основу для подальшого експериментального аналізу ефективності та якості запропонованого методу.

Таким чином, розроблена формалізація визначає абстрактну модель процесу семантичного уточнення текстових бізнес-правил і множину припущень щодо взаємодії її компонентів. Для перевірки практичної застосовності цих припущень і тестування ефективності підходу в наступному розділі подано експериментальну оцінку на репрезентативних бізнес-сценаріях.

4. Експериментальне тестування

Метою експериментів є оцінка запропонованого підходу, а також емпіричне порівняння двох стратегій тестоорієнтованої валідації: з використання фіксованого набору тест-кейсів та з незалежною генерацією тест-кейсів і DMN-таблиць на кожній ітерації валідаційного циклу. Експерименти виконувалися з використанням прототипу програмної системи. Архітектура системи (рис. 2) відображає поділ процесу на етапи ідентифікації семантики, генерації DMN-таблиць і валідації, у яких текстові бізнес-правила, схема даних і доменний контекст використовуються як спільне семантичне джерело.

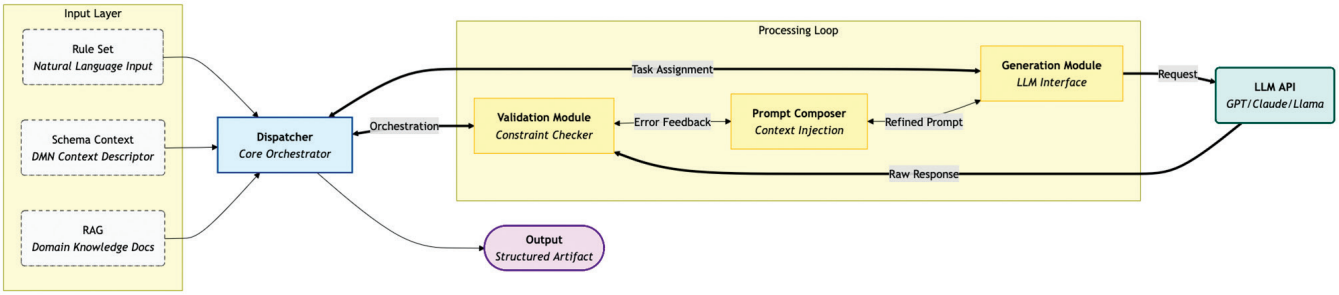

Рис. 2. Архітектура прототипу інтелектуального конвеєра обробки бізнес-правил

Прототип реалізовано з використанням LangChain для оркестрації мовних моделей і механізмів RAG та мовної моделі Claude Sonnet 4.5. За допомогою Prompt Composer формуються структуровані запити до мовної моделі. Перевірка виконується Validation Module, який послідовно здійснює синтаксичну, семантичну та логічну валідацію, включно з виконанням DMN-таблиць на згенерованих тест-кейсах.

Як репрезентативний сценарій використано набір бізнес-правил для автоматичного призначення продуктів іпотеки із політикою прийняття рішень FIRST. Правила визначають умови вибору між трьома можливими результатами: Fixed30, ARM7/ та ManualReview, на основі атрибутів заявника, зокрема кредитного рейтингу, співвідношення боргу до доходу (DTI), Loan-to-Value (LTV), стабільності доходу та поведінкових уподобань тощо. Для кожного експериментального прогону система генерувала до 200 тест-кейсів. Кожен тест-кейс містив вхідний об'єкт і множину очікуваних результатів, що дозволяло враховувати можливу недовизначеність семантики бізнес-правил.

Рисунок 3 ілюструє операційну інтерпретацію формальної моделі валідації DMN-таблиць, зокрема незалежну генерацію тестового покриття та моделей рішень. Було проведено два типи експериментів.

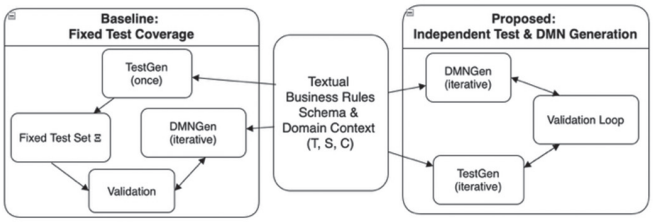

Рис. 3. Порівняння стратегій тестової валідації DMN-таблиць

У першому експерименті тест-кейси генерувалися один раз перед початком генерації DMN-таблиці та використовувалися як незмінний еталон для всіх подальших ітерацій валідації. Передбачалося, що такого набору буде достатньо для перевірки коректності багаторазово згенерованих DMN-таблиць. У 9 випадках система не змогла успішно пройти валідацію. Основні типи помилок включали некоректну множину можливих результатів (порожній список результатів) та семантичні невідповідності між очікуваним і фактичним значенням вихідної змінно product. Більшість успішних випадків завершувалися з першої спроби генерації DMN, однак наявність стабільних помилок вказала на обмеженість підходу з фіксованим тестовим покриттям .

У другому експерименті було застосовано альтернативну стратегію, за якої тест-кейси і DMN-таблиця генерувалися незалежно на кожній ітерації валідаційного циклу. Обидва артефакти формувалися з одного й того самого набору вхідних бізнес-правил, схемного та доменного контексту, але без взаємного використання структури. Результати показали, що за цієї стратегії всі 200 тестових прогонів завершилися успішно без жодного проваленого кейсу. Більшість DMN-таблиць проходили валідацію з першої або другої спроби, а максимальна кількість ітерацій не перевищувала трьох. Середній час обробки одного тестового кейсу склав 67 секунд, що є прийнятним для інтерактивного або напівавтоматизованого сценарію використання.

Порівняльний аналіз двох експериментів демонструє, що фіксований набір тест-кейсів не забезпечує достатнього покриття семантичного простору бізнес-правил. Натомість незалежна генерація тестів і таблиць створює ефект взаємної семантичної перевірки, за якого обидва артефакти повинні узгодитися з однією й тією ж абстрактною специфікацією. Отримані результати підтверджують, що тестове покриття повинно розглядатися як незалежне наближення

до референтної семантичної моделі рішення. Отже, доведена доцільність використання тестоорієнтованої валідації як ключового компонента автоматизованих систем формалізації бізнес-правил.

Висновки та обговорення

У цій роботі розглянуто задачу автоматизованої обробки текстових бізнес-правил як задачу відновлення формальної семантики рішень з неформального опису та її подальшого уточнення до виконуваних моделей у нотації DMN. На відміну від підходів, орієнтованих на безпосередню генерацію таблиць рішень, запропонований підхід розглядає ідентифікацію семантики, генерацію операційної моделі та її валідацію як єдиний узгоджений процес, формалізований у вигляді послідовних моделей і керованого ітеративного конвеєра.

Отримані результати підтверджують, що використання великих мовних моделей у поєднанні з RAG є ефективним засобом для інтерпретації текстових бізнес-правил і побудови початкових формальних представлень логіки рішень. Водночас експериментальна оцінка показала, що пряма генерація DMN-таблиць, навіть за умови врахування схемних обмежень, не гарантує семантичної коректності, повноти та несуперечності моделей рішень.

Ключовим результатом роботи є обґрунтування та експериментальне підтвердження ефективності тестоорієнтованої валідації, у межах якої тестове покриття генерується незалежно від DMN-таблиці, але з того самого семантичного джерела. Такий підхід дозволяє інтерпретувати тестовий набір як окрему проєкцію референтної семантичної моделі рішення і використовувати його для виявлення прихованих семантичних помилок, які не проявляються на рівні синтаксичної або структурної перевірки. Порівняння двох експериментальних стратегій показало, що незалежна генерація тестів і DMN-таблиць забезпечує стабільнішу конвергенцію та зменшує кількість некоректних моделей у фінальному результаті. Запропонований підхід розвиває і доповнює попередні результати з автоматизації подання логіки рішень, зокрема роботу [1], у якій було продемонстровано можливість використання штучного інтелекту для ефективної генерації DMN-артефактів.

Разом з тим робота має певні обмеження. По-перше, запропонований підхід покладається на якість текстових бізнес-правил і доменного контексту, що подаються на вхід системи. Неоднозначні або суперечливі формулювання можуть призводити до розширення множини допустимих результатів у референтній семантичній моделі. По-друге, використання великих мовних моделей зумовлює стохастичність процесу генерації та залежність від обчислювальних ресурсів, що може обмежувати застосування підходу в сценаріях із жорсткими вимогами до часу виконання.

Подальші дослідження доцільно спрямувати на формалізацію критеріїв зупинки ітеративного процесу валідації, автоматичну адаптацію бюджетів генерації та розширення підходу на складніші типи DMN-моделей, зокрема з багаторівневими діаграмами вимог до рішень і складними політиками агрегації результатів. Окремим перспективним напрямом є інтеграція формальних методів верифікації з тестоорієнтованою валідацією для подальшого підвищення гарантій семантичної коректності.

Список літератури:

[1] Cherednichenko O. AI-Based Efficient Automation of Decision Logic Representation / O. Cherednichenko, V. Maliarenko // Proc. of International Conference on Business Information Systems. – 2026. – P. 303–315. – doi: 10.1007/978-3-032-08614-3_19.

[2] Object Management Group. Decision Model and Notation (DMN), Version 1.6 // OMG Specification. – 2023.

[3] Batoulis K. Decision modeling and business process management / K. Batoulis, M. Weske // Business Process Management Journal. – 2018. – Vol. 24(2). – P. 451–470.

[4] Etikala S. Text2Dec: Extracting decision logic from natural language / S. Etikala, S. Goossens, J. De Smedt // Proc. of International Conference on Business Process Management (BPM). – 2020. – P. 303–319.

[5] Goossens S. Automatic extraction of decision models from textual descriptions / S. Goossens, J. De Smedt, J. Vanthienen // Decision Support Systems. – 2019. – Vol. 123. – P. 113–129.

[6] Goossens S. Evaluating large language models for DMN decision table generation / S. Goossens, J. De Smedt // Proc. of IEEE International Conference on Business Informatics. – 2023. – P. 1–8.

[7] Berkovitch S. Benchmarking large language models for structured table generation / S. Berkovitch, M. van der Aalst // Proc. of CAiSE. – 2023. – P. 145–160.

[8] Shorten R. Structured generation with schema constraints / R. Shorten, J. Kelleher // Proc. of EMNLP. – 2022. – P. 5621–5633.

[9] Li J. Retrieval-augmented generation: A survey / J. Li, Y. Liu, Z. Li // ACM Computing Surveys. – 2023. – Vol. 56(4). – P. 1–38.

[10] Jeong S. Schema-aware retrieval for enterprise decision systems / S. Jeong, H. Kim // Information Systems. – 2022. – Vol. 107. – P. 101–119.

[11] Liu P. On the challenges of contextual relevance in retrieval-augmented generation / P. Liu, X. Zhang // Proc. of ACL. – 2023. – P. 412–423.

[12] Calvanese D. Detecting incompleteness and inconsistency in DMN decision tables / D. Calvanese, M. Montali // Proc. of KR. – 2018. – P. 318–327.

[13] Corea C. Decision logic verification with Camunda DMN / C. Corea, M. Dumas // Software and Systems Modeling. – 2021. – Vol. 20(3) . – P. 965–985.

Надійшла до редколегії 12.11.2025

УДК 004.93:: 355.58 DOi 10.30837/ bi.2025.2(103).16

І. П. Гамаюн[1], Г. А. Плєхова[2], М. В. Костікова[3], Д. О. Плєхов[4], Р. Б. Багмут[5]

[1]НТУ «Харківський політехнічний інститут», м. Харків, Україна, ihor.hamaiun@khpi.edu.ua, ORCID iD: 0000-0003-2099-4658

[2]ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520

[3]ХНАДУ, м. Харків, Україна, kmv_topaz@ukr.net, ORCID iD: 0000-0001-5197-7389

[4]ХНАДУ, м. Харків, Україна, plehov@gmail.com, ORCID iD: 0009-0004-7873-1716

[5]ХНАДУ, м. Харків, Україна, bagmutroman58@gmail.com, ORCID iD: 0009-0003-1255-5097

СИНТЕЗ КОМП'ЮТЕРНОЇ ІНФОРМАЦІЙНО-АНАЛІТИЧНОЇ СИСТЕМИ ПО НАДЗВИЧАЙНИМ СИТУАЦІЯМ. ЧАСТИНА 1

У роботі розглянуто концепцію синтезу комп'ютерної інформаційно-аналітичної системи для моніторингу, аналізу та підтримки прийняття рішень під час надзвичайних ситуацій. Робота містить огляд класифікацій надзвичайних ситуацій, аналіз існуючих інформаційних систем, а також обгрунтування вибору архітектури майбутньої системи. Метою роботи є синтез комп'ютерної інформаційно-аналітичної системи, яка дозволяє в реальному часі аналізувати поточну ситуацію, прогнозувати розвиток подій та надавати рекомендації для реагування на надзвичайні ситуації.

НАДЗВИЧАЙНІ СИТУАЦІЇ, ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА, МАШИННЕ НАВЧАННЯ, СИНТЕЗ СИСТЕМИ, МОДЕЛЮВАННЯ, РЕАГУВАННЯ, ПІДТРИМКА ПРИЙНЯТТЯ РІШЕНЬ

I. P. Gamayun, G. A. Pliekhova, M. V. Kostikova, D. O. Pliekhov, R. B. Bagmut. Synthesis of computer information and analytical system for emergencies. Part 1. The paper explores the concept of synthesising a computer information and analytical system for monitoring, analysing, and supporting decision-making during emergencies. The paper includes a review of emergency classifications, an analysis of existing information systems, and a justification for choosing the architecture of the future system. The goal of this work is to develop a computer information and analytical system that enables real-time analysis of the current situation, forecasts the development of events, and provides recommendations for responding to emergencies.

EMERGENCIES, INFORMATION AND ANALYTICAL SYSTEM, MACHINE LEARNING, SYSTEM SYNTHESIS, MODELING, RESPONSE, DECISION-MAKING SUPPORT

Вступ

Актуальність теми обумовлена зростаючою частотою та інтенсивністю надзвичайних ситуацій (НС), пов'язаних із природними катаклізмами, техногенними аваріями, військовими діями. У таких умовах особливо важливою є здатність швидко отримувати достовірну інформацію, аналізувати її та ухвалювати обгрунтовані рішення. Розробка комп'ютерної інформаційно-аналітичної системи здатної автоматично здійснювати збір, обробку та інтерпретацію даних під час НС є актуальною науковою і практичною задачею [1-3].

Першочергові завдання роботи:

– дослідити класифікацію НС та інформаційні потреби в умовах їх виникнення;

– проаналізувати сучасні ІТ-рішення у сфері НС;

– розробити архітектуру комп'ютерної інформаційно-аналітичної системи.

1. 1. Аналіз предметної області

1.1. Класифікація надзвичайних ситуацій

Надзвичайні ситуації становлять серйозну загрозу життю, здоров'ю населення, об'єктам інфраструктури та навколишньому середовищу. Вони виникають унаслідок природних явищ, людської діяльності або соціальних конфліктів. Для розробки ефективних систем реагування і підтримки рішень важливо здійснити системну класифікацію НС, яка дозволяє формалізувати характеристики загроз, їхні джерела та сценарії розвитку.

Загальноприйнята класифікація поділяє НС на чотири основні категорії: природні, техногенні, соціальні та воєнні. Природні НС включають геофізичні та метеорологічні події, які відбуваються незалежно від людської діяльності. Це можуть бути повені, землетруси, зсуви, снігові лавини, урагани та інші екстремальні погодні умови. Їхня особливість полягає в складності прогнозування та масштабності наслідків. Техногенні НС, своєю чергою, є наслідком людської діяльності – зокрема аварій на виробництві, пожеж, транспортних катастроф або викидів небезпечних речовин. Саме вони часто мають локальний характер, але можуть переходити у міжрегіональні кризові стани.

Соціальні та воєнні НС об'єднує те, що їх причиною є напруження в суспільстві або прямий конфлікт. До соціальних зараховують масові заворушення, терористичні акти, панічні настрої внаслідок фейкових повідомлень, блокування критичної інфраструктури. Воєнні – це бойові дії, окупація територій, диверсії на об'єктах енергетики або транспорту, а також використання зброї масового ураження. Усі ці категорії можуть мати міждисциплінарний характер, наприклад, поєднання природної катастрофи з техногенною (землетрус – руйнування АЕС) або воєнної з техногенною (пожежа на складі боєприпасів).

Окрім класифікації за джерелом походження, в управлінні НС важливо враховувати масштаб події (місцевий, регіональний, національний, транскордонний), її динаміку, потенційний вплив на критичну інфраструктуру та можливість передбачення. Усі ці характеристики мають бути відображені в інформаційно-аналітичній системі, яка автоматизує процес оцінки ситуації.

Порівняння основних типів надзвичайних ситуацій за ключовими критеріями надано у табл. 1.

Таблиця 1

Порівняльна таблиця типів надзвичайних ситуацій

Критерій	Природні НС	Техногенні НС	Соціальні НС	Воєнні НС
Джерело виникнення	Природні явища	Людська діяльність / техніка	Соціальні процеси	Збройні сили / воєнні формування
Передбачуваність	Низька / Середня	Середня	Висока (деякі раптові)	Середня
Приклад	Землетрус, повінь, ураган	Вибух на заводі, аварія на АЕС	Терористичний акт, масові заворушення	Ракетний обстріл, диверсія, наступальна операція
Масштаб впливу	Регіональний / глобальний	Місцевий / регіональний	Місцевий / державний	Державний / глобальний
Тривалість дії	Від годин до тижнів	Від хвилин до днів	Від годин до місяців	Від днів до років
Об'єкти ураження	Населення, природа, інфраструктура	Виробництво, транспорт, енергетика	Людські маси, органи влади, інформаційні системи	Військові та цивільні об'єкти, інфраструктура
Наявність комбінованих ефектів	Часто	Часто	Іноді	Часто
Приклад ІАС-рішень	Прогнозування катаклізмів, евакуація	Моніторинг виробництв, автоматизоване блокування	Аналіз соцмереж, виявлення паніки	Системи ППО, військове картографування

1.2. Інформаційні потреби при надзвичайних ситуаціях

Ефективне реагування на надзвичайні ситуації безпосередньо залежить від наявності точної, актуальної та релевантної інформації. У сучасному динамічному середовищі, де масштаби загроз можуть змінюватися в режимі реального часу, інформація набуває стратегічного значення. Вона є основою для формування ситуаційної обізнаності, яка включає розуміння таких критично важливих аспектів: що саме відбувається, де це відбувається, які можливі наслідки, який масштаб ураження та які ресурси можуть бути залучені для реагування.

Одним з ключових завдань під час НС є забезпечення ситуаційної обізнаності – багатовимірного розуміння оперативної обстановки. Це охоплює такі компоненти:

– локалізація джерела загрози (геопросторові координати);

– оцінка динаміки події (темп поширення, інтенсивність);

– визначення зон потенційного ураження;

– аналіз інфраструктурної доступності (дороги, мости, комунікації);

– характеристика постраждалого населення (вік, стан здоров'я, мобільність).

Для досягнення цього використовуються широкі інформаційні джерела:

– супутникові знімки високої роздільності для виявлення масштабу змін у ландшафті;

– дані сенсорних мереж (IoT), зокрема детектори диму, сейсмодатчики, метеорологічні станції;

– системи раннього попередження (автоматичні й ручні);

– контент зі ЗМІ та соціальних мереж (моніторинг подій, фото- та відеофіксація);

– звіти очевидців – інструмент для верифікації й деталізації подій.

Сучасні системи підтримки прийняття рішень повинні бути здатні об'єднувати ці розрізнені джерела в уніфіковану аналітичну модель ситуації. Наприклад, дані про геолокацію можуть бути використані для автоматизованого визначення потенційних зон евакуації, обхідних шляхів, місць для розгортання польових госпіталів чи укриттів. Інформація про погодні умови та стан інфраструктури дозволяє оцінити ризики для рятувальних підрозділів.

Іншою важливою складовою є аналіз вразливості населення. Система має враховувати:

– наявність осіб з інвалідністю;

– літніх людей та дітей;

– мовні бар'єри (у випадку наявності іноземців або нацменшин);

– психоемоційний стан та поведінкові особливості громадян.

У цьому контексті критично важливо забезпечити ефективну комунікацію – через інтегровані модулі зв'язку система поширює сповіщення, рекомендації щодо дій, надає навігацію до безпечних зон, а також забезпечує зворотній зв'язок з користувачами (наприклад, через мобільні додатки або SMS-інформування).

Система інформаційної підтримки повинна виконувати чотири ключові функції:

1. Моніторинг – безперервне спостереження за подіями у просторі та часі.

2. Аналітика – обробка інформації для виявлення закономірностей та відхилень.

3. Прогнозування – побудова ймовірнісних сценаріїв розвитку подій.

4. Інтеграція – синхронізація з державними, регіональними й міжнародними базами даних, GIS-системами, сервісами телеметрії та базами медичних установ.

Таким чином, інформаційна система при надзвичайних ситуаціях є не просто технічним інструментом, а центральною платформою прийняття рішень, що базується на багатовимірному аналізі ситуації, людському факторі та реальних можливостях інфраструктури реагування.

1.3. Аналіз існуючих систем

У сфері реагування на надзвичайні ситуації функціонує низка критично важливих глобальних, регіональних та національних інформаційних систем, покликаних підтримувати прийняття рішень та візуалізувати ситуаційну обстановку. Серед них варто відзначити:

Європейську службу Copernicus EMS (Emergency Management Service): Надає безплатні швидкі та детальні карти супутникового моніторингу для подій природного або техногенного характеру. Забезпечує оперативний аналіз масштабів руйнувань та постраждалих територій.

Платформу ESRI Disaster Response: Побудована на основі ArcGIS, широко використовується глобально. Забезпечує створення інтерактивних карт, аналітичних панелей (dashboards) та статистичних оглядів. Сильна сторона – інтеграція з різними джерелами даних, проте часто вимагає значних ліцензійних витрат та глибокої експертизи ArcGIS.

Систему GDACS (Global Disaster Alert and Coordination System): Надає автоматизовані попередження та оцінки наслідків великих катастроф по всьому світу, сприяючи міжнародній координації допомоги.

В Україні функціонує низка державних платформ, серед яких системи ДСНС (Державна служба надзвичайних ситуацій) та інтегровані платформи територіальних громад. Вони забезпечують базовий облік сил і засобів цивільного захисту, відстеження інцидентів та підготовку звітності. Проте аналіз виявляє суттєві обмеження:

1. Низький рівень автоматизації: Значна частина процесів (збір даних, звітність) залишається ручною, що уповільнює реагування та збільшує ризик помилок.

2. Обмежена аналітика в реальному часі: Системи не забезпечують глибокого аналізу великих масивів даних для миттєвої оцінки динаміки ситуації чи прогнозування розвитку подій.

3. Відсутність інтеграції: Критичною проблемою є силос даних та відсутність інтеграції з ключовими відкритими джерелами: API супутникових моніторингів (Copernicus, NASA FIRMS), метеорологічних сервісів (наприклад., Open-Meteo), даних соціальних медіа (для виявлення сигналів про НС), а також з датчиками IoT. Це унеможливлює формування єдиної цілісної ситуаційної картини (Common Operational Picture – COP).

4. Неадаптивність: Багато систем є жорстко фіксованими («монолітними») і не підтримують швидку адаптацію чи перепрофілювання під нові, непередбачені сценарії НС.

5. Обмежене використання ШІ/ML: Аналітика ґрунтується переважно на традиційних алгоритмах або ручній обробці. Потенціал ШІ для прогнозування, автоматичної класифікації звітів, аналізу зображень/відео з місця події або обробки неструктурованого тексту з соцмереж використовується мінімально.

6. Залежність від власницьких технологій: Використання закритих платформ ускладнює масштабування, кастомізацію та інтеграцію.

7. Інформаційне перевантаження та якість даних – системи не завжди ефективно фільтрують та валідують вхідні дані, що може призводити до «шуму» та ускладнювати прийняття рішень.

Ці обмеження обумовлюють нагальну потребу в розробці нової покоління комп'ютерної інформаційно-аналітичної системи (КІАС) для НС в Україні. Така система повинна бути:

Адаптивною та гнучкою: Модульна архітектура для швидкого налаштування під будь-які сценарії НС та конкретні регіональні/галузеві потреби.

Інтегрованою: Побудована на відкритих стандартах та API, здатна легко агрегувати дані з різноманітних джерел: державних реєстрів, супутників, метеоданих, датчиків IoT, соціальних медіа, мобільних додатків населення та оперативних платформ (наприклад, ЦПР/ЦЗР).

Інтелектуальною: Максимально використовувати ШІ/ML для:

а) Автоматизації рутинних завдань (збір, валідація, класифікація даних).

б) Аналізу великих даних (включаючи неструктуровані – текст, фото, відео) в реальному часі.

в) Прогнозування розвитку ситуації, оцінки ризиків та наслідків.

г) Генерації рекомендацій для прийняття рішень.

Візуально ефективною: Надавати інтуїтивні інструменти для побудови комплексних ситуаційних картин (COP), інтерактивних карт, аналітичних панелей з можливістю глибокого «занурення» в дані (drill-down).

Масштабованою та доступною: Працювати на різних рівнях (від громади до національного), мати веб-інтерфейс та мобільні додатки для роботи в полі. Інтерфейси повинні бути інтуїтивними для користувачів з різним рівнем технічної підготовки.

Захищеною: Забезпечувати високий рівень кібербезпеки та відповідати вимогам захисту персональних даних (табл. 2 – 3).

Таблиця 2

Порівняльний аналіз ключових систем для управління НС

Характеристика	Copernicus EMS	ESRI Disaster Response	Типові Українські Системи (ДСНС, громади)	Ідеальна Нова КІАС (Потрібна)
Основна функція	Супутниковий моніторинг, карти руйнувань	ГІС-платформа, інтерактивні карти, дашборди	Облік сил/засобів, відстеження інцидентів, звітність	Уніфікована платформа ситуаційної обізнаності, аналітики та управління
Географія охоплення	Глобальна (фокус на активації по запиту)	Глобальна (ліцензійна)	Національна / Регіональна / Локальна	Національна (з масштабуванням на регіон/громаду)
Тип даних	Супутникові знімки, геопросторові дані	ГІС-дані, оперативні дані, статистика	Структуровані оперативні дані, звіти	Всі типи: геопросторові, сенсорні (IoT), оперативні, соцмедіа, метео, супутникові, неструктуровані (текст, фото)
Інтеграційні можливості (API)	Обмежені публічні API	Потужні API (вимагають експертизи ArcGIS)	Обмежені / Відсутні	Відкриті API (REST, GraphQL) для всіх компонентів, стандарти (OGC, SensorThings)
Автоматизація / ШІ/ML	Автоматизована обробка знімків	Базова аналітика, інструменти для ШІ (вимагають налаштування)	Мінімальна	Висока: Автоматичний збір/валідація, прогнозування, аналіз зображень/тексту, генерація рекомендацій
Адаптивність	Спеціалізована служба	Конфігурована платформа (але потребує часу/експерта)	Жорстка, важко змінюється	Модульна, гнучка: Швидке додавання сценаріїв, джерел даних, аналітичних моделей
Оперативність оновлення	Висока (після активації)	Залежить від налаштувань та даних	Низька / Середня (частіше ручне оновлення)	Висока (реальний час / near real-time)
Доступність / Юзабіліті	Веб-доступ до карт	Потужний, але складний інтерфейс (для ГІС-фахівців)	Спеціалізовані інтерфейси (різний рівень зручності)	Інтуїтивні веб- та мобільні інтерфейси для всіх рівнів користувачів
Вартість (для кінцевого користувача)	Безкоштовна (для активацій)	Висока (ліцензії, ПЗ, експертиза)	Розробка/підтримка за бюджетом	Комбінація: бюджетна розробка/підтримка, можливі хмарні сервіси

Таблиця 3

Ключові вимоги до нової КІАС нового покоління

Вимога	Опис	Приклад / значення
Відкритість та інтеграція	Побудова на відкритих стандартах, API для легкої інтеграції будь-яких джерел даних та систем.	Використання OGC стандартів (WMS, WFS), SensorThings API, REST/GraphQL API.
Модульність та адаптивність	Гнучка архітектура, що дозволяє швидко додавати нові модулі, сценарії НС, джерела даних, аналітичні моделі без перебудови ядра системи.	Можливість додати модуль моніторингу повеней чи лісових пожеж за тиждень.
Розширена аналітика на базі ШІ/ML	Використання машинного навчання та штучного інтелекту для автоматизації, прогнозування, аналізу складних даних, генерації інсайтів.	Прогноз поширення диму пожежі, автоматична класифікація звернень з соцмереж, аналіз супутникових знімків на предмет руйнувань.
Ситуаційна обізнаність (COP)	Формування єдиної, цілісної та актуальної в реальному часі карти події з усіма доступними даними для всіх рівнів прийняття рішень.	Інтерактивна карта з накладанням: зон ураження, сил реагування, метеоданих, потоків населення.
Масштабованість	Здатність обробляти зростаючі обсяги даних та обслуговувати зростаючу кількість користувачів без втрати продуктивності.	Робота під час великомасштабної катастрофи (наприклад, землетрус, велика пожежа).
Користувацький досвід (UX/UI)	Інтуїтивні, зрозумілі інтерфейси (веб, мобільні) для користувачів з різним досвідом (від рятувальника до керівника).	Простий мобільний додаток для фіксації інциденту на місці з фото/координатами.
Кібербезпека та захист даних	Вбудовані механізми захисту від кібератак, забезпечення цілісності та конфіденційності даних, відповідність законодавству.	Шифрування даних, контроль доступу на основі ролей (RBAC), аудит дій.
Підтримка рішень	Надання не тільки даних, але й аналітичних висновків, моделювання наслідків, рекомендацій для оптимізації ресурсів.	Рекомендації щодо оптимального розгортання сил на основі прогнозу розвитку пожежі.

Таким чином, результати аналізу підтверджують критичну потребу в оновленій комп'ютерній інформаційно-аналітичній системі нового покоління. Така система має стати уніфікованим цифровим середовищем, здатним об'єднувати просторові, часові, сенсорні, соціальні та оперативні дані у єдиний інформаційний потік. Це дозволить забезпечити якісно новий рівень ситуаційної обізнаності, швидкості та ефективності реагування на НС в Україні, а також може стати цінним досвідом для інших країн (табл. 4).

Таблиця 4

Комп'ютерна інформаційна аналітична система

Скорочення	Повна назва	Пояснення
К	Комп'ютерна	Базується на сучасних ІТ-технологіях (хмара, ШІ, Big Data)
І	Інформаційна	Збирає, зберігає та управляє даними з різних джерел
А	Аналітична	Забезпечує глибоку обробку даних (аналітика, прогнозування, моделювання)
С	Система	Цілісний програмно-апаратний комплекс з чіткою архітектурою

Ключові функції такої системи:

1. Агрегація даних: Об'єднання інформації з супутників, сенсорів, соцмереж, держреєстрів тощо.

2. Аналітика в реальному часі: Використання ШІ для прогнозування ризиків, оцінки наслідків, класифікації інцидентів.

3. Візуалізація: Створення інтерактивних карт, аналітичних панелей (dashboards), ситуаційних звітів.

4. Підтримка прийняття рішень: Генерація рекомендацій щодо розгортання ресурсів, евакуації, координації служб.

KIAC нового покоління усуває обмеження існуючих систем (низька автоматизація, відсутність інтеграції з API), реалізує модульність, відкриті стандарти та AI-аналітику для НС.

2. Синтез комп'ютерної інформаційно-аналітичної системи

У сучасному світі надзвичайні ситуації, такі як природні катастрофи, техногенні аварії чи соціальні кризи, можуть виникати несподівано та потребують швидкого реагування. Ефективне управління цими подіями відіграє ключову роль у зменшенні збитків та порятунку людських життів. Сьогодні технології стають незамінним інструментом для підвищення можливостей екстрених служб, дозволяючи оперативно обробляти інформацію та координувати дії в умовах криз. Цей розділ присвячена новій комп'ютерній інформаційно-аналітичній системі, яка була розроблена спеціально для потреб управління надзвичайними ситуаціями.

Зростаюча складність і частота надзвичайних ситуацій вимагають сучасних рішень для збору даних, їх аналізу та прийняття обґрунтованих рішень у реальному часі. Традиційні підходи часто виявляються недостатньо ефективними через затримки в обробці інформації та відсутність належної координації. KIAC долає ці обмеження завдяки інтеграції передових технологій, таких як штучний інтелект, аналітика великих даних і системи зв'язку в реальному часі. У цьому розділі ми розглянемо основні принципи функціонування KIAC, її архітектуру та компоненти, а також те, як система сприяє кращій взаємодії між командами екстреного реагування. Окрім цього, ми проаналізуємо, як впровадження KIAC може скоротити час реагування та підвищити загальну ефективність управління кризовими ситуаціями.

2.1. Архітектура системи

KIAC побудована за мікросервісним підходом, що забезпечує високу гнучкість, масштабованість та легкість у розгортанні та оновленні окремих компонентів. Система складається з п'яти основних компонентів:

- сервіс збору та агрегації даних;
- сховище даних (PostgreSQL + PostGIS);
- аналітичний модуль;
- модуль машинного навчання;
- візуалізаційний інтерфейс.

Ці компоненти взаємодіють між собою та із зовнішніми джерелами даних і користувачами, як показано на рис. 1.

Сервіс збору та агрегації даних відповідає за отримання інформації з різних зовнішніх джерел, таких як сенсори, бази даних екстрених служб, соціальні мережі тощо. Він забезпечує первинну обробку та агрегацію даних, після чого зберігає їх у сховищі даних для подальшого використання іншими компонентами системи.

Сховище даних (PostgreSQL + PostGIS). Сховище даних є центральним компонентом системи, де зберігаються всі зібрані дані, аналітичні результати та прогнози. Використання PostgreSQL з розширенням PostGIS дозволяє ефективно працювати з геопросторовими даними, що є критично важливим для управління надзвичайними ситуаціями, пов'язаними з територіальними аспектами.

Аналітичний модуль. Аналітичний модуль читає дані зі сховища, виконує різноманітні аналітичні операції, такі як статистичний аналіз, виявлення аномалій, кластеризація тощо. Результати аналізу записуються назад у сховище даних і можуть бути використані для підтримки прийняття рішень або подальшого аналізу.

Модуль машинного навчання. Модуль машинного навчання використовує дані зі сховища для навчання моделей, які можуть прогнозувати розвиток надзвичайних ситуацій, оцінювати ризики або пропонувати оптимальні стратегії реагування. Прогнози та рекомендації, згенеровані цим модулем, також зберігаються у сховищі даних.

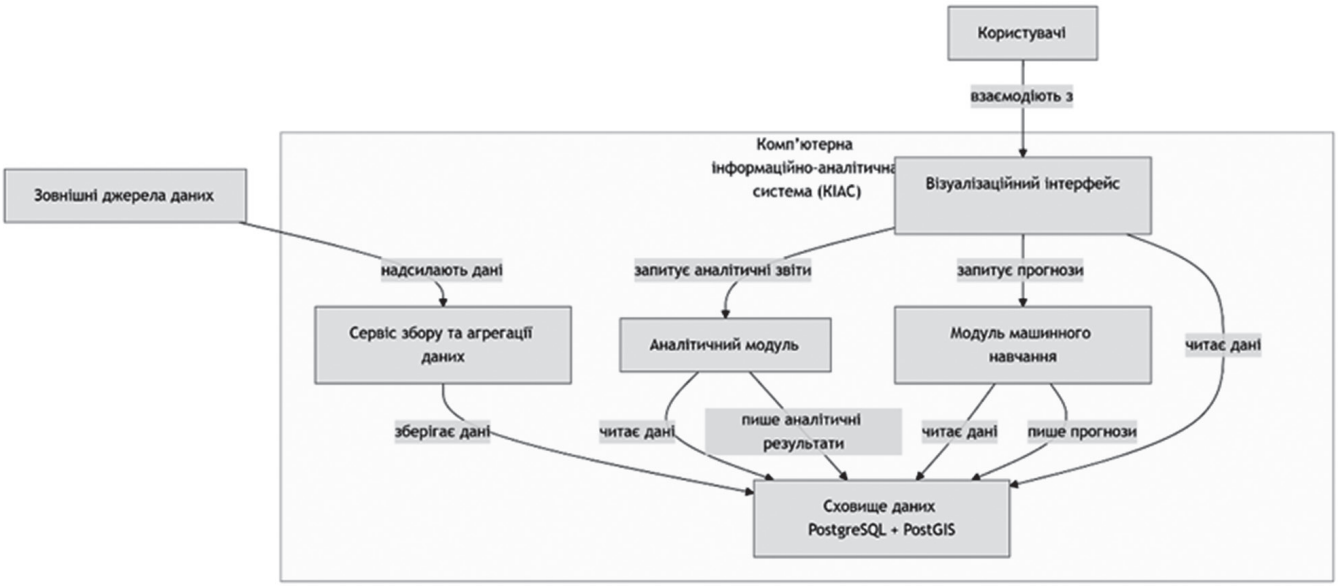

Рис. 1. Схема КІАС

Візуалізаційний інтерфейс. Візуалізаційний інтерфейс є точкою взаємодії користувачів з системою. Він дозволяє переглядати зібрані дані, аналітичні звіти та прогнози у зручному графічному вигляді. Користувачі можуть запитувати специфічну інформацію, налаштовувати параметри візуалізації та отримувати оперативні звіти для підтримки прийняття рішень.

2.2. Переваги мікросервісної архітектури

Використання мікросервісної архітектури в КІАС надає низку переваг.

Масштабованість: Кожен компонент може бути масштабований незалежно від інших, що дозволяє системі ефективно обробляти великі обсяги даних та навантаження.

Гнучкість: Розробка та оновлення окремих мікросервісів можуть відбуватися незалежно, що спрощує впровадження нових функцій та технологій.

Легкість у розгортанні та оновленні: Мікросервіси можуть бути розгорнуті та оновлені окремо, що зменшує ризики та час простою системи.

КІАС є інноваційною системою, яка забезпечує комплексний підхід до управління надзвичайними ситуаціями. Завдяки мікросервісній архітектурі та інтеграції передових технологій, система дозволяє оперативно обробляти великі обсяги даних, забезпечує аналітичну підтримку та прогнозування, що сприяє швидкому та ефективному реагуванню на кризові ситуації. Впровадження КІАС може стати проривом у сфері управління надзвичайними ситуаціями, надаючи надійну основу для прийняття обґрунтованих рішень у реальному часі.

2.3. Джерела даних

Комплексна інформаційно-аналітична система для надзвичайних ситуацій повинна функціонувати в умовах нестабільного середовища, оперативної невизначеності та необхідності швидкого прийняття рішень. У цьому контексті якість, точність, своєчасність і достовірність даних, які надходять до системи, мають критично важливе значення. Сучасні КІАС орієнтовані на мультиджерельну модель збору даних, яка охоплює як автоматизовані, так і людські канали. У цьому підрозділі розглянуто основні типи джерел, що інтегруються в архітектуру КІАС.

Відкриті API від державних та міжнародних організацій.

Одним із ключових джерел даних для КІАС є відкриті прикладні програмні інтерфейси (API), які надаються державними органами, міжнародними агентствами та науково-дослідними структурами. Прикладами є:

ДСНС України – через API доступні дані про оперативні зведення, статистику пожеж, затоплень, техногенних аварій, зони ураження тощо.

OpenWeather API – глобальна платформа погодного моніторингу. Дані з OpenWeather надають інформацію про температуру, вологість, опади, напрямок вітру, які є критичними для прогнозування розвитку пожеж, повеней або інших стихійних явищ.

NASA FIRMS (Fire Information for Resource Management System) – надає оперативну інформацію про термічні аномалії (гарячі точки), які вказують на можливі пожежі. Система використовує супутники Terra та Aqua з інфрачервоними сенсорами MODIS/VIIRS.

Інтеграція з цими API дозволяє КІАС автоматично оновлювати ситуаційні карти та аналітичні панелі в режимі майже реального часу, покращуючи реакцію служб на події.

ДСНС

Типи даних:

– надходження екстрених викликів (112);
– стан критичної інфраструктури;
– зони надзвичайних ситуацій.

Технічні параметри:

– формат: REST/JSON;
– ліміт: 100 запитів/хвилину;
– аутентифікація: OAuth 2.0.

Приклад запиту:

```
import requests response = requests.get ("https://api.dsns.gov.ua/emergencies?region=kyiv&type=fire", headers={"Authorization": "Bearer <TOKEN>"})
```

OpenWeather Map

Постачає:

– температура/вологість/тиск (оновлення кожні 10 хвилин);
– прогноз погоди на 5 діб;
– вторинні метеодані.

Інтеграція:

```
INSERT INTO weather_data SELECT time, temperature FROM openweather WHERE geo_id = 703448;
```

NASA FIRMS (Fire Information for Resource Management System)

Ключові функції:

– виявлення термічних аномалій (пожежі);
– активні вогнища з MODIS/VIIRS супутників;
– просторова роздільна здатність: 375 м/піксель.

Супутникові знімки. Супутникові знімки є важливим джерелом даних для KIAC, особливо для моніторингу великих територій та виявлення змін у навколишньому середовищі. Вони застосовуються для аналізу пожеж, повеней, землетрусів, зсувів ґрунту та інших природних катастроф. У системі використовуються дані з таких супутників, як Landsat, Sentinel та комерційні платформи. Програма Landsat, що управляється NASA та USGS, забезпечує безкоштовні знімки Землі з 1972 року. Ці дані використовуються для моніторингу змін у землекористуванні, виявлення пожеж та оцінки збитків від стихійних лих. Їх доступність робить Landsat важливим інструментом для KIAC. Sentinel Місія Sentinel, що входить до програми Copernicus Європейського космічного агентства, надає дані з радарних та оптичних сенсорів. Вони особливо корисні для моніторингу повеней, зсувів ґрунту та інших геологічних подій, забезпечуючи високу точність і регулярність оновлень. Комерційні супутники Компанії, такі як Planet Labs і Maxar Technologies, надають знімки високої роздільної здатності, які дозволяють проводити детальний аналіз конкретних територій. Хоча ці дані є платними, їх висока точність і частота оновлення роблять їх цінними для оперативного реагування. Супутникові знімки забезпечують об'єктивну картину ситуації на великих територіях, але їх обробка вимагає значних обчислювальних ресурсів і спеціалізованих алгоритмів, що може ускладнювати інтеграцію в KIAC.

Джерела та параметри надані у табл. 5.

Таблиця 5

Супутникові знімки

Супутник	Роздільна здатність	Частота оновлення	Типи даних
Sentinel-2	10 м	5 діб	RGB/NIR спектри
Landsat 8	30 м	16 діб	Термальні знімки
PlanetScope	3 м	Щоденно	Високодетальні зображення

Обробка у KIAC:

Завантаження через STAC API:

```
curl "https://earth-search.aws.element84.com/collections/sentinel-s2-l2a/items"
```

Аналіз індексів (*NDVI* , *NDWI*):

1. Індекс *NDVI* (Normalized Difference Vegetation Index)

Формула:

$$NDVI = \frac{NIR - Red}{NIR + Red},$$

де *NIR* (Near Infrared) – канал ближнього інфрачервоного діапазону (зазвичай ~800 нм); *Red* – канал червоного спектру (~660 нм).

Інтерпретація значень *NDVI* надана у табл. 6.

Таблиця 6

Інтерпретація значень індексу *NDVI*

NDVI	Значення
< 0	Вода, хмари, сніг
0 – 0.1	Пустелі, каміння, міста
0.1 – 0.3	Кущі, трав'янисті покрови
0.3 – 0.6	Помірна рослинність
0.6 – 1.0	Густа, здорова рослинність

Використання:

Моніторинг здоров'я рослин.

Виявлення вирубок лісу.

Аналіз урбанізації.

Сільське господарство (оцінка врожайності).

2. Індекс *NDWI* (Normalized Difference Water Index)

Формула (за McFeeters, 1996):

$$NDWI = \frac{Green - NIR}{Green + NIR},$$

де *Green* – зелений канал (близько 560 нм); *NIR* – ближній інфрачервоний.

Інтерпретація значень *NDWI* надана у табл. 7.

Таблиця 7

Інтерпретація значень індексу *NDWI*

NDWI	Значення
> 0	Водні тіла (висока ймовірність наявності води)
< 0	Рослинність, ґрунт, будівлі

Використання:

– Виявлення водойм (річки, озера, болота).
– Визначення змін у рівні води.
– Повінь і моніторинг зрошення.

Детектування змін методом Change Detection (алгоритм MAD). Одним із ключових завдань у сфері моніторингу надзвичайних ситуацій є своєчасне виявлення змін на місцевості. Це стосується як природних катастроф (пожеж, повеней, зсувів), так і техногенних або воєнних подій (обстріли, вибухи, руйнування інфраструктури). Для цього широко застосовується методологія Change Detection (виявлення змін), серед якої важливе місце займає алгоритм MAD (Multivariate Alteration Detection).

Алгоритм MAD належить до статистичних методів багатовимірного аналізу, які дозволяють виявити відмінності між двома багатоканальними (наприклад, супутниковими) зображеннями однієї території, зробленими у різний час. Його основна ідея полягає в пошуку лінійних комбінацій спектральних каналів зображень, які максимізують відмінності між ними.

MAD-аналіз використовує канонічний кореляційний аналіз (CCA) для трансформації вхідних даних так, щоб звести кореляції між каналами обох зображень до мінімуму, а потім – виявити значущі зміни у спектральних характеристиках пікселів.

Ключові етапи алгоритму MAD.

1. Вхідні дані: два багатоканальні супутникові зображення одного району, отримані у різні моменти часу (T1 і T2), попередньо вирівняні за геометрією.

2. Стандартизація: нормалізація каналів з урахуванням середніх значень і дисперсій для зменшення впливу освітлення, пори року тощо.

3. Канонічна кореляція: побудова пар канонічних векторів для обох наборів даних, що мінімізують кореляцію між каналами зображень T1 і T2.

4. Обчислення MAD-компонентів: вираховується різниця між парними канонічними проекціями. Ці різниці формують так звані MAD-компоненти, які мають гаусівський розподіл з математичним сподіванням, близьким до нуля для незмінних пікселів.

5. Визначення змін: для кожного пікселя обчислюється χ^2 -статистика на основі MAD-компонентів. Високе значення χ^2 свідчить про наявність змін.

6. Порогова класифікація: застосування порогу (наприклад, $p < 0.05$) для формування бінарної карти змін (зміни / без змін).

Переваги MAD-аналізу:

– Незалежність від типу сенсора: алгоритм працює з будь-якими багатоспектральними даними (Landsat, Sentinel-2, WorldView тощо).

– Висока чутливість до навіть слабких змін (пожовтіння трави, ущільнення забудови).

– Автоматизована статистична оцінка змін (використання χ^2).

– Сумісність із GIS/RS-середовищем: інтегрується з інструментами QGIS, ArcGIS, Google Earth Engine.

Використання MAD в KIAC. Алгоритм MAD є важливим компонентом модулів обробки супутникових знімків у сучасних KIAC. Наприклад, система може:

– регулярно порівнювати зображення регіону до і після події;

– автоматично генерувати карти змін (Change Maps);

– обчислювати індекс пошкодження для інфраструктури або природних об'єктів;

– інтегрувати результати з іншими модулями (аналітичними панелями, сценаріями реагування).

MAD також поєднується з методами машинного навчання (класифікатори на основі MAD-компонентів) для автоматизації виявлення типу змін (будівельні роботи, згорілі площі, водонаповнення тощо).

Алгоритм MAD є універсальним, математично обґрунтованим методом багатовимірного аналізу, який дозволяє підвищити точність і об'єктивність виявлення змін у геопросторових даних. Його використання в рамках KIAC підвищує ситуаційну обізнаність, дозволяє реагувати оперативно та зменшити ризики для населення і ресурсів.

Сенсори Інтернет речей (IoT-сенсори) відіграють важливу роль у зборі даних у реальному часі, надаючи KIAC інформацію з сенсорів, розгорнутих у різних локаціях. Ці пристрої вимірюють такі параметри, як температура, вологість, рівень забруднення та сейсмічна активність, що дозволяє моніторити критичні інфраструктури та природні екосистеми. Метеорологічні сенсори вимірюють погодні умови на локальному рівні, забезпечуючи більш точні дані для конкретних територій порівняно із загальними прогнозами. Вони є важливими для прогнозування локальних погодних аномалій. Сенсори якості повітря використовуються для моніторингу рівня забруднення, що допомагає виявляти техногенні аварії чи пожежі. Дані з цих сенсорів доповнюють інформацію з інших джерел, підвищуючи точність аналізу. Сейсмічні сенсори –пристрої, які дозволяють виявляти землетруси та інші геологічні події в реальному часі, що є критично важливим для раннього попередження та реагування на природні катастрофи. Дані з IoT-сенсорів надходять у систему в реальному часі, що забезпечує оперативність реагування. Однак розгортання та підтримка мережі сенсорів є дороговартісними, а також вимагають високого рівня безпеки та надійності передачі даних.

IoT-сенсорні мережі. Архітектура збору даних (табл. 8):

[Сенсори] → [Шлюз LoRaWAN] → [MQTT Broker] → [Сервіс обробки KIAC]

Таблиця 8

Типи сенсорів та параметри

Тип сенсора	Параметри	Частота	Точність
Повітря (AQI)	PM2.5, NO_2, O_3, SO_2	5 хв	±5%
Ґрунтові	Вологість, pH, температура	15 хв	±0.5°C
Гідрологічні	Рівень води, забруднення	10 хв	±2 см

Протоколи передачі:
– MQTT (Message Queuing Telemetry Transport)
– CoAP (Constrained Application Protocol)
– Пакетна передача через NB-IoT

Краудсорсинг через мобільні застосунки є інноваційним джерелом даних, що залучає громадськість до збору інформації про надзвичайні ситуації. Користувачі можуть надавати звіти, фотографії, відео чи текстові описи подій, що доповнюють офіційні дані. Мобільні застосунки використовуються для звітів користувачі, які можуть повідомляти про пожежі, повені чи аварії, надаючи інформацію з перших рук. Це дозволяє КІАС швидко реагувати на інциденти, які ще не зафіксовані офіційними джерелами. Аналіз даних із соціальних мереж (Twitter, Facebook, Instagram) допомагає виявляти тренди та повідомлення про надзвичайні ситуації. Цей підхід дозволяє отримувати інформацію в реальному часі від великої кількості людей. Громадські ініціативи – проекти, такі як OpenStreetMap, дозволяють волонтерам оновлювати карти та додавати дані про інфраструктуру, що є корисним для планування евакуації чи розподілу ресурсів. Краудсорсинг забезпечує великий обсяг даних, але їхня надійність може бути нижчою, що вимагає додаткової перевірки та фільтрації.

Мобільні застосунки як джерело даних:
Функціонал:
– Фото/відеофіксація подій.
– Геотеговані повідомлення.
– Анкетування в реальному часі.
Переваги:
– Швидке охоплення території.
– Верифікація супутникових даних.
Приклад архітектури наданий на рис. 2.

Рис. 2. Приклад архітектури

Механізми якості даних:
1. Перехресна перевірка з IoT/супутниками.
2. Репутаційна система користувачів.
3. Модерація з використанням CNN (згорткові нейромережі)

Інтеграція та обробка даних Ефективне використання різноманітних джерел даних у КІАС потребує їх інтеграції та обробки. Основні етапи включають:

Нормалізація даних: Перетворення інформації з різних джерел у єдиний формат для аналізу. Фільтрація та валідація: Перевірка даних на точність і надійність, особливо для краудсорсингових джерел. Агрегація: Об'єднання даних для створення повної картини ситуації. Геопросторова інтеграція: Використання геоданих для візуалізації подій на карті.

Для обробки даних система використовує сучасні технології, такі як бази даних із підтримкою геопросторових запитів (наприклад, PostgreSQL із PostGIS).

Забезпечення якості даних надано у табл. 9.

Таблиця 9

Метрики якості

Параметр	API	Супутники	IoT	Краудсорсинг
Частота оновлення	★★★★☆	★★☆☆☆	★★★★★	★★★☆☆
Просторова точність	★★★☆☆	★★★★★	★★★★☆	★★☆☆☆
Валідність	★★★★★	★★★★☆	★★★☆☆	★★☆☆☆

Проблематика:
– Розбіжності у часі надходження.
– Різна просторова точність.
– Відсутність калібрування.

Технології обробки:
Етап очищення:
– Виправлення часових зсувів.
– Нормалізація координат (WGS84 → Web Mercator)
def normalize_coords(lat, lon):
return transform(Proj(init='epsg:4326'), Proj(init='epsg:3857'), lon, lat)

Валідація:
– Статистичні тести (ANOVA, регресійний аналіз).
– Геопросторове узгодження за допомогою PostGIS:
SELECT ST_Within(sensor_point, admin_area) FROM validation_data;

Джерела даних є основою функціонування КІАС, забезпечуючи систему інформацією для моніторингу та реагування на надзвичайні ситуації. Відкриті API, супутникові знімки, сенсори IoT та краудсорсинг надають різноманітні та актуальні дані, кожне з яких має свої переваги та недоліки. Інтеграція та обробка цих джерел є ключовими для забезпечення ефективності системи. У майбутньому розвиток технологій і розширення мережі джерел даних зроблять КІАС ще потужнішим інструментом для управління надзвичайними ситуаціями. Отже, джерела даних для КІАС повинні бути різноманітними, взаємодоповнюючими та оперативно інтегрованими. Поєднання автоматичних джерел (сенсори, супутники, API) із людськими (краудсорсинг) забезпечує гнучкість та стійкість системи в умовах реальних НС. Завдяки цьому КІАС здатна не лише фіксувати події, але й здійснювати аналітичну оцінку ситуації, прогнозування розвитку та формування ефективних рішень для реагування.

2.4. Моделі аналізу

У сучасних умовах надзвичайні ситуації характеризуються високою частотою виникнення та складністю прогнозування, що вимагає використання передових методів аналізу даних. Точність і своєчасність обробки інформації є ключовими факторами для ефективного реагування та мінімізації наслідків. Моделі аналізу, такі як алгоритми класифікації ситуацій та методи аналізу часових рядів, відіграють центральну роль у вирішенні цих завдань. Вони дозволяють ідентифікувати типи НС і прогнозувати їхній розвиток, що є основою для оперативного прийняття рішень у комп'ютерних інформаційно-аналітичних системах. У цьому підрозділі розглядаються основні підходи до аналізу даних у контексті НС, з акцентом на математичні основи та їхнє практичне застосування.

Алгоритми класифікації є важливими інструментами для обробки даних, що надходять у реальному часі, дозволяючи систематизувати та інтерпретувати інформацію про НС. У цьому підрозділі аналізуються три основні методи: дерева рішень (Decision Trees), випадковий ліс (Random Forest) та згорткові нейронні мережі (CNN) для обробки зображень. Дерева рішень (Decision Trees) Дерева рішень представляють собою ієрархічну структуру, яка використовується для прийняття рішень шляхом послідовного розщеплення даних на основі значень ознак. Кожен внутрішній вузол дерева відповідає умові, гілки – можливим значенням, а листові вузли – кінцевим класам. Основним завданням побудови дерева є вибір оптимального критерію розщеплення, який мінімізує неоднорідність класів у дочірніх вузлах. Одним із найпоширеніших критеріїв є ентропія, яка характеризує ступінь невизначеності в наборі даних D:

$$H(D) = -\sum_{i=1}^{c} p_i \log_2 p_i ,$$

де c – кількість класів; p_i – частка зразків, що належать до класу i.

Приріст інформації для ознаки A обчислюється як:

$$\text{Gain}(D, A) = H(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} H(D_v) ,$$

де D_v – підмножина даних для значення v ознаки A.

Альтернативним критерієм є індекс Gini:

$$G(D) = 1 - \sum_{i=1}^{c} p_i^2 .$$

Оптимізація цього показника дозволяє ефективно розподіляти дані за класами. Наприклад, у контексті НС дерево рішень може класифікувати тип події (пожежа, повінь, землетрус) на основі параметрів, таких як температура, вологість або сейсмічна активність. Випадковий ліс є ансамблевим методом, що базується на комбінації множини дерев рішень. Кожне дерево будується на випадковій підмножині даних і ознак, що підвищує стійкість моделі до шуму та перенавчання. Кінцевий результат класифікації визначається шляхом агрегування передбачень окремих дерев:

$$\hat{y} = mode\{y_1, y_2, \ldots, y_m\} ,$$

де y_i – передбачення i-го дерева; m – загальна кількість дерев.

Математична основа Random Forest включає також оцінку ймовірності належності до класу (k):

Оцінка ймовірності для класу k:

$$P(k) = \frac{1}{m} \sum_{i=1}^{m} I(y_i = k) ,$$

де I – індикаторна функція.

Цей підхід забезпечує високу точність за рахунок усереднення результатів і зменшення дисперсії. У задачах управління НС випадковий ліс може бути застосований для ідентифікації техногенних аварій на основі комбінованих метеорологічних та інфраструктурних даних. Згорткові нейронні мережі (CNN) для знімків є спеціалізованим класом нейронних мереж, призначених для обробки структурованих даних, таких як зображення. Вони широко застосовуються для аналізу супутникових знімків у задачах виявлення НС. Основною операцією в CNN є згортка, яка витягує локальні ознаки з вхідного зображення (x) за допомогою фільтра (w):

$$s(i, j) = (x * w)(i, j) = \sum_m \sum_n x(i + m, j + n) w(m, n) ,$$

де x –вхідне зображення; w – фільтр згортки.

Отримана карта ознак активується нелінійною функцією, наприклад, ReLU:

$$f(x) = \max(0, x) .$$

Для зменшення розмірності даних застосовується пулінг, наприклад, max-pooling:

$$p(i, j) = \max_{m, n \in R} s(i + m, j + n) ,$$

де R – область пулінгу.

Ці операції дозволяють CNN ефективно виявляти візуальні патерни, такі як дим або вогонь на супутникових знімках, що є важливим для раннього виявлення пожеж. Аналіз часових рядів є ключовим для прогнозування динаміки НС, оскільки дозволяє враховувати темпоральні залежності в даних. У цьому підрозділі розглядаються два основні методи: ARIMA та рекурентні нейронні мережі з довгостроковою пам'яттю (LSTM). Модель ARIMA (авторегресійна інтегрована модель ковзного середнього) є статистичним інструментом для моделювання часових рядів. Вона базується на трьох компонентах: авторегресії (AR), інтегруванні (I) та ковзному середньому (MA). Модель позначається як

$$ARIMA(p, d, q) ,$$

де p – порядок авторегресії; d – ступінь диференціювання; q – порядок ковзного середнього.

Авторегресійна складова описується рівнянням:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \mathrm{T}_t ,$$

де ϕ_i – коефіцієнти; T_t – випадкова похибка.

Ковзне середнє моделюється як:

$$y_t = \mathrm{T}_t + \theta_1 \mathrm{T}_{t-1} + \theta_2 \mathrm{T}_{t-2} + \ldots + \theta_q \mathrm{T}_{t-q} ,$$

де θ_i – коефіцієнти MA.

Для забезпечення стаціонарності ряд диференціюється d разів:

$$\Delta^d y_t = (1 - L)^d y_t ,$$

де L – оператор зсуву назад.

ARIMA може бути використана для прогнозування поширення пожежі на основі історичних даних про температуру та вологість. Рекурентні нейронні мережі з довгостроковою пам'яттю (LSTM) є вдосконаленим типом рекурентних нейронних мереж, які здатні моделювати довгострокові залежності в часових рядах. Вони складаються з комірок пам'яті та трьох основних воріт: забування, оновлення та виведення. Ворота забування визначають, яку інформацію відкинути:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right),$$

де σ – сигмоїдна функція.

Ворота оновлення додають нову інформацію:

Додають нову інформацію до стану:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right).$$

Кандидат на оновлення:

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right).$$

Стан комірки оновлюється:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t .$$

Вихід визначається через:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right).$$

Вихід:

$$h_t = o_t \cdot \tanh\left(C_t\right).$$

LSTM ефективно прогнозують сейсмічну активність, враховуючи довгострокові патерни в даних.

Висновки

Інформаційна система при надзвичайних ситуаціях – це комплекс організаційних, технічних та програмних засобів, призначених для збору, обробки, зберігання, передачі та надання інформації, необхідної для прийняття рішень у разі загрози або виникнення надзвичайних ситуацій. Цей комплекс є центральною платформою прийняття рішень, що базується на багатовимірному аналізі ситуації, людському факторі та реальних можливостях інфраструктури реагування.

Комп'ютерна інформаційно-аналітична система– це програмно-апаратний комплекс, призначений для збору, обробки, аналізу та представлення інформації з метою підтримки прийняття рішень. Вона об'єднує різні джерела даних, дозволяє проводити аналіз та візуалізацію інформації, а також забезпечує користувачів необхідними даними для вирішення конкретних задач. Такі обмеження як низька автоматизація, відсутність інтеграції з API, які характерні для існуючих систем, КІАС нового покоління усуває, реалізує модульність, відкриті стандарти та AI-аналітику для НС.

Розглянуті моделі аналізу – алгоритми класифікації (Decision Trees, Random Forest, CNN) та методи аналізу часових рядів (ARIMA, LSTM) – є основою для обробки даних у КІАС для управління НС. Вони забезпечують точну ідентифікацію подій і прогнозування їхнього розвитку, що підвищує ефективність реагування. Розглянуті компоненти LSTM дозволяють ефективно управляти інформацією в часі, що робить її дуже потужним інструментом для прогнозування часових рядів. Вони дозволяють моделі враховувати довгострокові залежності інформації, що дозволяє їй докладніше аналізувати та прогнозувати майбутні значення у порівнянні з іншими типами нейронних мереж. Перспективним напрямком є інтеграція цих моделей у єдину систему для забезпечення комплексного аналізу та швидкого прийняття рішень.

Список літератури

[1] Захарова І. В., Філіпова Л. Я., Задорожний І. С., Тарасенко Д. А. Основи інформаційно-аналітичної діяльності : навч. посіб. / І. В. Захарова, Л. Я. Філіпова, І. С. Задорожний, Д. А. Тарасенко ; 2-е вид., випр. і допов. Черкаси: Східноєвропейський університет імені Рауфа Аблязова, 2024. 347 с. URL: https://suem.edu.ua/storage/doc/books/osnovy-informaciyno-analitychnoi-dialnosti-zaharova2.pdf?utm_source=chatgpt.com.

[2] Chen R., Sharman R., Rao H. R., Upadhyaya S. J. Coordination in Emergency Response Management. Communications of the ACM, May 2008, Vol. 51, No. 5, pp. 66-73. URL: https://dl.acm.org/doi/10.1145/1342327.1342340.

[3] Стрижак О. Є. Онтологічні інформаційно-аналітичні системи. Радіоелектронні і комп'ютерні системи, 2014, № 3 (67), С. 71-76. URL: http://nbuv.gov.ua/UJRN/recs_2014_3_13.

Надійшла до редколегії 21.07.2025

УДК 004.93:: 355.58 DOi 10.30837/ bi.2025.2(103).17

І. П. Гамаюн[1], Г. А. Плєхова[2], М. В. Костікова[3], Д. О. Плєхов[4], Р. Б. Багмут[5]

[1]НТУ «Харківський політехнічний інститут», м. Харків, Україна, ihor.hamaiun@khpi.edu.ua, ORCID iD: 0000-0003-2099-4658

[2]ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520

[3]ХНАДУ, м. Харків, Україна, kmv_topaz@ukr.net, ORCID iD: 0000-0001-5197-7389

[4]ХНАДУ, м. Харків, Україна, plehov@gmail.com, ORCID iD: 0009-0004-7873-1716

[5]ХНАДУ, м. Харків, Україна, bagmutroman58@gmail.com, ORCID iD: 0009-0003-1255-5097

СИНТЕЗ КОМП'ЮТЕРНОЇ ІНФОРМАЦІЙНО-АНАЛІТИЧНОЇ СИСТЕМИ ПО НАДЗВИЧАЙНИМ СИТУАЦІЯМ. ЧАСТИНА 2

У статті розглянуто концепцію синтезу комп'ютерної інформаційно-аналітичної системи для моніторингу, аналізу та підтримки прийняття рішень під час надзвичайних ситуацій. Розроблено функціональну модель прототипу, наведено опис реалізації ключових модулів системи, запропоновано використання методів машинного навчання для оцінки ступеня загроз. Робота включає прикладне програмне рішення та результати його верифікації на основі сценаріїв з відкритих джерел. Практичне значення полягає у створенні прототипу універсальної системи, яку можна адаптувати для потреб Державної служби України з надзвичайних ситуацій, місцевих органів влади або підприємств критичної інфраструктури.

НАДЗВИЧАЙНІ СИТУАЦІЇ, ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА, МАШИННЕ НАВЧАННЯ, МОДЕЛЮВАННЯ, РЕАГУВАННЯ

I. P. Gamayun, G. A. Pliekhova, M. V. Kostikova, D. O. Pliekhov, R. B. Bagmut. Synthesis of a computer information and analytical system for emergencies. Part 2. The article considers the concept of synthesizing a computer information and analytical system for monitoring, analysis, and decision-making support during emergencies. A functional model of the prototype is developed, a description of the implementation of key system modules is provided, and the use of machine learning methods for assessing the degree of threats is proposed. The work includes an applied software solution and the results of its verification based on open source scenarios. The practical significance lies in creating a prototype of a universal system that can be adapted for the needs of the State Emergency Service of Ukraine, local authorities, or critical infrastructure enterprises.

EMERGENCIES, INFORMATION AND ANALYTICAL SYSTEM, MACHINE LEARNING, MODELING, RESPONSE

Вступ

Тенденція зростання кількості природних і особливо техногенних надзвичайних ситуацій (НС), важкість їх наслідків змушують розглядати їх як серйозну загрозу безпеці окремої людини, суспільства та навколишньому середовищу, а також стабільності розвитку економіки країни. Тому здатність швидко отримувати достовірну інформацію, аналізувати її та ухвалювати обґрунтовані рішення є важливою у сучасних умовах. Конструювання комп'ютерної інформаційно-аналітичної системи (КІАС) здатної автоматично здійснювати збір, обробку та інтерпретацію даних під час НС є актуальною науковою і практичною задачею сьогодення [1-3].

Першочергові завдання роботи:

– реалізувати прототип ключових модулів системи;

– провести тестування та оцінити ефективність запропонованого рішення.

1. Синтез комп'ютерної інформаційно-аналітичної системи. Візуалізація

Візуалізація даних є критично важливою для систем реагування на надзвичайні ситуації, оскільки вона дозволяє оперативно аналізувати та інтерпретувати великі обсяги інформації. У цьому розділі ми розглянемо, як використовувати інтерактивні карти з векторними шарами для відображення геопросторових даних, як візуалізувати різні сценарії реагування на надзвичайні ситуації та як інтегрувати push-сповіщення для інформування користувачів про критичні події.

1.1. Карти з векторними шарами

Векторні шари на картах представляють географічні об'єкти у вигляді векторних даних, таких як точки, лінії та полігони. Ці шари дозволяють відображати детальну інформацію про об'єкти, включаючи їхні атрибути та властивості. Для створення інтерактивних карт у веб-додатках часто використовується бібліотека Leaflet, яка є легкою у використанні та надає зручний API. Нижче наведено приклад коду для створення базової карти та додавання векторного шару у вигляді полігону, що позначає зону ризику:

```
// Ініціалізація карти var map = L.map('map'). setView([50.45, 30.52], 13);

// Додавання базового шару L.tileLayer('https://{s}. tile.openstreetmap.org/{z}/{x}/{y}.png', { attribution: '© OpenStreetMap contributors' }).addTo(map);

// Додавання векторного шару (GeoJSON) var geojsonFeature = { "type": "Feature", "properties": { "name": "Зона ризику", "popupContent": "Це зона підвищеного ризику" }, "geometry": { "type": "Polygon", "coordinates":
```

[[[30.5, 50.4], [30.6, 50.4], [30.6, 50.5], [30.5, 50.5], [30.5, 50.4]]] } };

L.geoJSON(geojsonFeature).addTo(map);

Цей код ініціалізує карту з центром у місті Київ (координати [50.45, 30.52]) та масштабом 13. На карту додається базовий шар з OpenStreetMap, а потім векторний шар у форматі GeoJSON, який відображає полігон зони ризику. Векторні шари можуть бути налаштовані для відображення різних типів даних, таких як дороги, будівлі чи зони покриття.

Карта з відображенням векторного шару зони ризику надана на рис. 1. На карті зображено базовий шар OpenStreetMap із центром у місті Київ, на який накладено червоний полігон, що позначає зону ризику.

Рис. 1. Карта з відображенням векторного шару зони ризику

1.2. Відображення сценаріїв реагування

Сценарії реагування на надзвичайні ситуації включають плани дій, такі як евакуаційні маршрути, розташування ресурсів або зони небезпеки. Для їхньої візуалізації на карті використовуються графічні елементи, такі як маркери, лінії, полігони та кольорові заливки. Наприклад, евакуаційні маршрути можна позначити лініями зі стрілками, зони небезпеки – червоними полігонами, а безпечні зони – зеленими. Нижче наведено приклад коду для додавання маркера, що позначає пункт збору:

var marker = L.marker([50.45, 30.52]).addTo(map); marker.bindPopup("Пункт збору").openPopup();

Цей код додає маркер на карту в заданій точці та прив'язує до нього спливаюче вікно з текстом «Пункт збору», яке автоматично відкривається.

Для більш складних сценаріїв можна додавати анімації або динамічно змінювати шари залежно від обраного сценарію (рис. 2). На карті показано зону ризику (червоний полігон), пункт збору (маркер із спливаючим вікном) та евакуаційний маршрут (синя лінія зі стрілками).

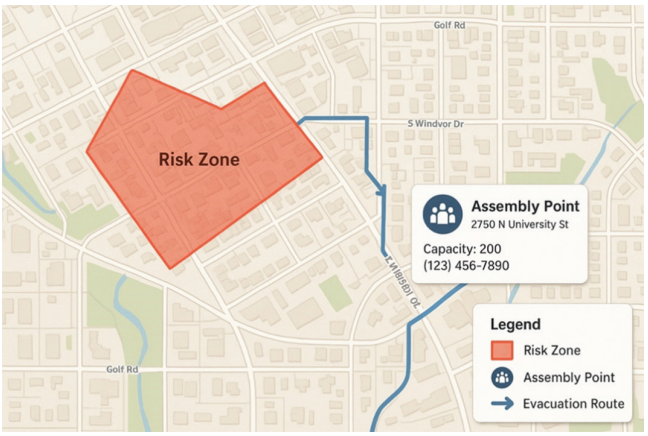

Рис. 2. Карта з відображеними сценаріями реагування

1.3. Push-сповіщення

Push-сповіщення дозволяють оперативно інформувати користувачів про оновлення даних або критичні події в системі. Для їхньої реалізації у веб-додатках використовуються сервісні працівники (Service Workers) та Push API. Сервісні працівники працюють у фоновому режимі, забезпечуючи можливість надсилати сповіщення навіть тоді, коли вкладка браузера закрита. Спочатку необхідно зареєструвати сервісного працівника:

if ('serviceWorker' in navigator) { navigator.serviceWorker.register('/sw.js').then(function(registration) { console.log('Сервісний працівник зареєстровано:', registration); }).catch(function(error) { console.log('Помилка реєстрації сервісного працівника:', error); }); }

Далі користувач має надати дозвіл на отримання сповіщень:

Notification.requestPermission().then(function(permission) { if (permission === 'granted') { console.log('Дозвіл на сповіщення отримано'); } });

Після цього можна відправити сповіщення, наприклад, при оновленні даних на карті:

function showNotification() { if (Notification.permission === 'granted') { navigator.serviceWorker.getRegistration().then(function(reg) { reg.showNotification('Оновлення даних', { body: 'Дані на карті було оновлено', icon: '/icon.png' }); }); } }

Цей код перевіряє дозвіл користувача та відображає сповіщення з заголовком «Оновлення даних» і текстом «Дані на карті було оновлено». Для реальних систем також потрібна серверна частина, наприклад, із використанням Firebase Cloud Messaging, для надсилання push-повідомлень із сервера. Важливим аспектом є досвід користувача: сповіщення мають бути релевантними, своєчасними та не надто частими, щоб уникнути їхнього ігнорування чи роздратування.

У цьому розділі розглянуто ключові аспекти візуалізації в системах реагування на надзвичайні ситуації:

карти з векторними шарами для відображення геопросторових даних, візуалізація сценаріїв реагування для представлення планів дій та інтеграція push-сповіщень для оперативного інформування користувачів. Ці компоненти разом створюють потужний інструмент для аналізу та реагування на кризові ситуації.

2. Реалізація прототипу системи

2.1. Технології

Розробка прототипу KIAC для управління надзвичайними ситуаціями потребує використання сучасних технологій, які забезпечують ефективність, масштабованість і зручність у підтримці. Технологічний стек прототипу включає React і Leaflet для створення фронтенду, Django для бекенду, TensorFlow і Scikit-learn для реалізації компонентів штучного інтелекту, а також PostgreSQL із розширенням PostGIS для управління базою даних. Ці технології обрано через їхню здатність обробляти геопросторові дані, підтримувати реальний час і забезпечувати надійність системи.

Фронтенд: React та Leaflet React React (React Official Documentation) – це JavaScript-бібліотека, розроблена компанією Facebook, яка використовується для створення динамічних і масштабованих користувацьких інтерфейсів. Її компонентна архітектура дозволяє створювати модульні елементи інтерфейсу, що полегшує розробку та підтримку складних застосунків. У KIAC React відповідає за створення інтерактивного інтерфейсу, який включає панелі управління, відображення даних і карт. React використовує віртуальний DOM для оптимізації оновлень інтерфейсу, що забезпечує високу продуктивність навіть при частому оновленні даних, наприклад, під час відображення нових інцидентів у реальному часі. Завдяки великій спільноті та екосистемі бібліотек, React є ідеальним вибором для швидкої розробки прототипу.

Leaflet (Leaflet Official Documentation) – це легка бібліотека з відкритим кодом для створення інтерактивних карт. Вона підтримує відображення векторних шарів, маркерів і полігонів, що є необхідним для візуалізації геопросторових даних у KIAC, таких як зони ризику чи місця надзвичайних ситуацій. Leaflet інтегрується з React через бібліотеку react-leaflet, яка надає компоненти для роботи з картами. Нижче наведено приклад коду для створення карти з червоним полігоном, що позначає зону ризику:

```
import React from 'react'; import { MapContainer, TileLayer, Polygon } from 'react-leaflet';
const Map = () => { const position = [50.45, 30.52]; // Координати Києва const polygon = [ [50.4, 30.5], [50.4, 30.6], [50.5, 30.6], [50.5, 30.5], ];
return ( <MapContainer center={position} zoom={13} style={{ height: '500px' }}> ); };
export default Map;
```

Цей код створює карту, центровану на Києві, з базовим шаром OpenStreetMap і червоним полігоном, що ілюструє зону ризику. Бекенд: Django Django (Django Official Documentation) – це високорівневий Python-фреймворк, який сприяє швидкій розробці вебзастосунків із чистим і прагматичним дизайном. Django обрано для бекенду KIAC завдяки його вбудованим інструментам, таким як ORM для роботи з базою даних, механізми аутентифікації та адміністративний інтерфейс. Ці функції дозволяють швидко створювати надійні та безпечні серверні компоненти. Для створення API-ендпоінтів у прототипі використовується Django REST framework, який забезпечує зручний спосіб розробки RESTful API. Наприклад, ендпоінт для роботи з даними про інциденти може бути реалізований так:

```
from rest_framework import viewsets from .models import Incident from .serializers import IncidentSerializer
class IncidentViewSet(viewsets.ModelViewSet): queryset = Incident.objects.all() serializer_class = IncidentSerializer
```

Цей код визначає ViewSet для моделі Incident, що дозволяє виконувати операції створення, читання, оновлення та видалення (CRUD) через REST API. Django забезпечує безпеку, масштабованість і легкість інтеграції з іншими компонентами системи. Штучний інтелект: TensorFlow та Scikit-learn. TensorFlow (TensorFlow Official Documentation) – це фреймворк із відкритим кодом, розроблений Google, який використовується для створення моделей глибокого навчання. У KIAC TensorFlow застосовується для складних завдань, таких як класифікація супутникових зображень для виявлення пожеж або повеней. Наприклад, згорткова нейронна мережа (CNN) може бути реалізована так:

```
import tensorflow as tf from tensorflow.keras import layers, models
model = models.Sequential([ layers.Conv2D(32, (3, 3), activation='relu', input_shape=(128, 128, 3)), layers.MaxPooling2D((2, 2)), layers.Conv2D(64, (3, 3), activation='relu'), layers.MaxPooling2D((2, 2)), layers.Conv2D(64, (3, 3), activation='relu'), layers.Flatten(), layers.Dense(64, activation='relu'), layers.Dense(2, activation='softmax') # Бінарна класифікація: пожежа чи ні ])
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

Ця модель може бути навчена на наборі даних із супутникових зображень для автоматичного виявлення надзвичайних ситуацій. Scikit-learn Scikit-learn (Scikit-learn Official Documentation) – це бібліотека для традиційного машинного навчання, яка надає інструменти для класифікації, регресії та кластеризації. У KIAC Scikit-learn використовується для аналізу істо-

ричних даних про інциденти. Наприклад, класифікація типів надзвичайних ситуацій може бути реалізована за допомогою алгоритму Random Forest: from sklearn. ensemble import RandomForestClassifier from sklearn. model_selection import train_test_split from sklearn. metrics import accuracy_score

Припускаємо, що X і y – ознаки та мітки

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2) clf = RandomForestClassifier(n_ estimators=100) clf.fit(X_train, y_train) y_pred = clf. predict(X_test) print("Точність:", accuracy_score(y_test, y_pred))

Цей код демонструє, як можна класифікувати інциденти на основі їхніх характеристик, таких як місце, час або тип події. База даних: PostgreSQL із PostGIS PostgreSQL (PostgreSQL Official Documentation) – це потужна система управління базами даних із відкритим кодом, яка вирізняється надійністю та підтримкою складних типів даних. Розширення PostGIS (PostGIS Official Documentation) додає можливості для роботи з геопросторовими даними, що є критично важливим для KIAC, оскільки система обробляє географічні об'єкти, такі як координати інцидентів чи зони ризику. У Django геопросторові дані моделюються за допомогою GeoDjango. Наприклад, модель для зберігання інцидентів із геолокацією:

from django.contrib.gis.db import models

class Incident(models.Model): name = models. CharField(max_length=100) geom = models.PointField()

Ця модель дозволяє зберігати географічні координати інцидентів і виконувати просторові запити, наприклад, знайти всі інциденти в радіусі 1 км від заданої точки:

SELECT * FROM incidents WHERE ST_ DWithin(geom, ST_MakePoint(30.52, 50.45), 1000);

Інтеграція технологій. Архітектура прототипу KIAC передбачає тісну взаємодію всіх компонентів. Фронтенд, побудований на React і Leaflet, надсилає запити до бекенду через REST API. Бекенд, реалізований на Django, обробляє ці запити, взаємодіє з базою даних PostgreSQL/PostGIS і викликає моделі ШІ, створені за допомогою TensorFlow і Scikit-learn, для аналізу даних. Результати повертаються на фронтенд для відображення користувачам.

На рис. 3 зображено архітектуру системи. Ця схема ілюструє, як компоненти системи взаємодіють для забезпечення збору, обробки та візуалізації даних про надзвичайні ситуації.

Висновки: Вибраний технологічний стек забезпечує надійну основу для розробки прототипу KIAC. React і Leaflet дозволяють створювати інтерактивні карти, Django прискорює розробку серверної частини, TensorFlow і Scikit-learn надають потужні інструменти для аналізу даних, а PostgreSQL із PostGIS ефективно обробляє геопросторові дані. Ці технології разом створюють цілісну систему, здатну підтримувати управління надзвичайними ситуаціями в реальному часі.

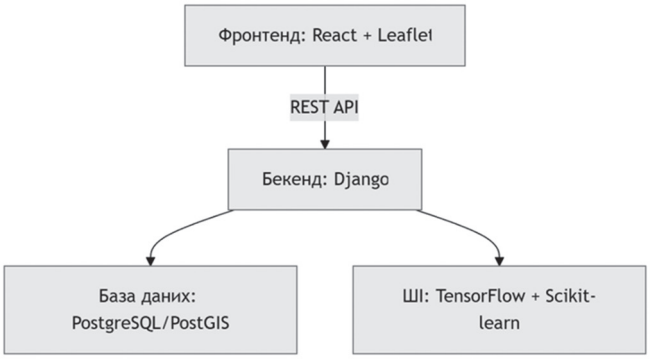

Рис. 3. Архітектура системи

2.2. Розробка функціоналу

KIAC повинна об'єднати модулі виявлення пожеж на супутникових знімках, оцінку ризику розповсюдження вогню, інтерфейс диспетчера з картою, та систему сповіщень. Ми реалізуємо локальний прототип, що працює без зовнішніх API, на основі відкритих бібліотек Python. Застосуємо класифікатор (наприклад, CNN через TensorFlow або Scikit-learn) для розпізнавання пожеж на зображеннях, побудуємо просту модель оцінки ризику на основі правил і дерев рішень, та створимо веб-інтерфейс (Flask+Folium або Streamlit) з інтерактивною картою і індикаторами ризику. Усі компоненти проекту орієнтовані на офлайн-режим (наприклад, у Folium можна встановити tiles=None щоб працювати без Інтернету).

Модуль виявлення пожеж за супутниковими знімками. Задачу виявлення пожеж можна сформулювати як бінарну класифікацію зображень «пожежа» vs «не пожежа». Для демонстрації згенеруємо синтетичні зображення (з імітацією «жари» як червоні плями) та навчимо простий класифікатор. Нижче – приклад такого коду на Python із використанням Scikit-learn:

```
from PIL import Image, ImageDraw
import numpy as np
from sklearn.ensemble import RandomForestClassifier
# Генерація синтетичних зображень (64×64 px)
def generate_image(fire=False):
    bg = np.zeros((64,64,3), dtype=np.uint8)
    # Заливка випадковим "зеленим" фоном
    for i in range(64):
        for j in range(64):
            bg[i,j] = [
            np.random.randint(20, 80), # червоний канал
            np.random.randint(100,180), # зелений канал
              np.random.randint(20, 80),  # синій канал
            ]
```

```
        img = Image.fromarray(bg)
        draw = ImageDraw.Draw(img)
        if fire:
            # Додаємо червоне "полум'я" як еліпс
            cx, cy = np.random.randint(10,54), np.random.
randint(10,54)
            r = np.random.randint(5,15)
             draw.ellipse([cx-r, cy-r, cx+r, cy+r], fill=(255,
np.random.randint(100,200), 0))
        return np.array(img)

    # Створюємо набір даних: 50 зразків без пожежі,
50 – з пожежами
    X, y = [], []
    for _ in range(50):
      X.append(generate_image(fire=False)); y.append(0)
    for _ in range(50):
      X.append(generate_image(fire=True)); y.append(1)
    X = np.array(X); y = np.array(y)

    # Розділення на тренувальний та тестовий набори
    from sklearn.model_selection import train_test_split
    X_flat = X.reshape((100, -1))
    X_train, X_test, y_train, y_test = train_test_split(X_
flat, y, test_size=0.2, random_state=42)

    # Навчання простого класифікатора (RandomForest)
    clf = RandomForestClassifier(n_estimators=50,
random_state=0)
    clf.fit(X_train, y_train)
    print("Точність класифікації на тесті:", clf.score(X_
test, y_test))
```

У цьому прикладі ми навчили класифікатор визначати наявність «червоного» полум'я на зображенні, що імітує пожежу. Така система може бути розширена справжніми супутниковими знімками і складнішою CNN-моделлю (TensorFlow). Спеціальна підготовка даних (нормалізація, аугментація) і глибші моделі збільшать якість, але принцип лишається: моделі навчаються відрізняти зображення з вогнем від тих, де його немає.

Оцінка ризику поширення пожежі. Для моделювання ризику поширення вогню враховують метеорологічні фактори (температура, вітер, вологість), тип місцевості та історію попередніх пожеж. Навчимо просту модель (дерево рішень) на згенерованих даних з кількома факторами. Наприклад:

```
    import pandas as pd
    from sklearn.tree import DecisionTreeClassifier

    # Приклад синтетичних даних (температура,
швидкість вітру, тип місцевості, попередня активність)
    df = pd.DataFrame([
        {'temp': 30, 'wind': 10, 'terrain': 0, 'prev_fire': 1,
'risk': 1},
        {'temp': 22, 'wind':  5, 'terrain': 1, 'prev_fire': 0,
'risk': 0},
        {'temp': 35, 'wind': 15, 'terrain': 0, 'prev_fire': 2,
'risk': 1},
        # ... ще рядків даних ...
    ])
    features = ['temp','wind','terrain','prev_fire']
    X_train = df[features]
    y_train = df['risk']

    # Навчання дерева рішень
    tree = DecisionTreeClassifier(max_depth=3, random_
state=0)
    tree.fit(X_train, y_train)
```

Тут terrain – це код типу місцевості (наприклад, 0=ліс, 1=луг і т. д.), а prev_fire – показник недавніх пожеж (0, 1, 2, ...). Навчена модель може відображати прості правила, наприклад: якщо температура висока, вітер сильний і попередня активність велика, ризик – високий. Подібні прості класифікації є типовими для попереднього аналізу ризику. Зокрема, у дослідженнях показано, що деревоподібні моделі (decision trees) можуть з ~50% точністю класифікувати майбутній розмір пожежі на основі двох змінних (дефіцит тиску насичення та покриття ялини). Інші роботи використовують множинні метеофактори і геодані (вегетаційні індекси) з точністю до 90%. Для нашої прототипової системи поєднуємо прості правила (словесні умови) з таким деревом рішень.

Наприклад, можна задати правило: якщо температура > 30°C і вітер > 10 км/год, позначити високу загрозу; інакше перевіряти дерево рішення. Лінки на готові набір даних (наприклад, UCI Forest Fires dataset) показують, що прямі погодні параметри (temp, RH, wind, rain) добре описують ризик. У результаті ми отримуємо модуль, який на вході має показники погоди та геоінформацію, а на виході – оцінку ризику (наприклад, 0=низький, 1=високий).

Панель диспетчера (веб-інтерфейс із картою).

Для диспетчерської панелі зробимо веб-додаток (Flask або Streamlit) з інтерактивною картою. На карті показуємо:

– об'єкти (маршрути евакуації, підстанції тощо) як маркери;

– зону високого ризику (наприклад, полігон із червоним контуром);

– елементи керування сценаріями (кнопки, селектори умов).

Приклад з Flask + Folium: бібліотека Folium дозволяє створювати картографічні відображення з даними. Щоб працювати офлайн, при створенні базової карти можна вказати tiles=None, що відключить онлайн-плитки OSM. Наприклад:

```
    from flask import Flask
    import folium
    app = Flask(__name__)
    @app.route("/")
```

```
def map_view():
  # Створюємо карту без зовнішніх тайлів (офлайн режим)
   m = folium.Map(location=[50.5, 30.5], zoom_start=8, tiles=None)
  # Додаємо маркери (наприклад, будівлі)
   folium.Marker([50.5, 30.5], tooltip='Об\'єкт 1').add_to(m)
   folium.Marker([50.6, 30.6], tooltip='Об\'єкт 2').add_to(m)
  # Додаємо полігон зоною ризику (червоний контур)
  risk_poly = [[50.4, 30.4], [50.4, 30.7], [50.7, 30.7], [50.7, 30.4]]
   folium.Polygon(locations=risk_poly, color='red', fill=False).add_to(m)
  return m.get_root().render()
if __name__ == "__main__":
  app.run(debug=True)
```

Це дозволить відобразити карту у браузері без необхідності доступу до Інтернету. Folium успішно інтегрується з Flask – найпростіший спосіб показати карту – повернути HTML-код карти з m.get_root().render(). (Альтернативно, можна використовувати iFrame або шаблони Flask для вбудування компонентів карти.) Враховуючи офлайн-вимоги, можна заздалегідь завантажити локальні геодані (GeoJSON полігони зони ризику, шейп-файли тощо) і додати їх на карту за допомогою folium.GeoJson. У прикладі вище ми використали прості координати і маркери як шаблон.

Якщо застосовувати Streamlit замість Flask, він також підтримує інтерактивні карти (наприклад, через st.map або плагін Folium). Streamlit дозволяє швидко робити веб-інтерфейс з Python-кодом, але для користувацької карти з вказівками варто зберегти Folium-карту у HTML та виводити у Streamlit як компонент components.html.

Керування сценаріями: на панелі можна додати кнопки або селектори (наприклад, вибір стратегії гасіння пожежі) за допомогою HTML/JS або віджетів Streamlit. Наприклад, можна реалізувати кнопку «Почати гасіння» та відображати віджет статусу реагування.

Алерти та візуальні індикатори. Система має виводити сповіщення при виникненні загрози. Для локального застосунку це може бути браузерний попап або звуковий сигнал. У веб-інтерфейсі можна використовувати Web Notifications API чи прості JavaScript alert(). Наприклад, у шаблоні HTML-сторінки Flask можна додати скрипт:

```
<script>
 if (Notification.permission === 'default') {
  Notification.requestPermission();
 }
 // Функція показує сповіщення
 function showAlert(msg) {
  if (Notification.permission === 'granted') {
   new Notification('Попередження', { body: msg });
  } else {
   alert(msg);
  }
 }
 // Викликаємо сповіщення при завантаженні, якщо тригер спрацював
 window.onload = function() {
  // Приклад умови ризику
  var highRisk = true;
  if (highRisk) {
   showAlert('Критичний ризик пожежі в зоні!');
  }
 }
</script>
```

Цей код при першому запуску запитує дозвіл на сповіщення, а потім показує повідомлення на екрані, якщо в зоні виявлено критичний ризик.

Візуальні індикатори на карті: критичні зони можна виділяти червоними полігонами, а маркери можна підсвічувати (змінювати колір чи анімувати). Наприклад, можна додати до Folium-карти червоний полігон з атрибутом fillOpacity=0.5 або зробити ефект «мигання» за допомогою CSS-анімації (Leaflet підтримує кастомні стилі). Це допоможе оператору швидко помітити критичні ділянки.

Приклад входових даних та візуалізація результатів:

– Метеодані та історія пожеж: у прикладі ми використали DataFrame з полями temp, wind, terrain, prev_fire, risk. В реальній системі ці дані можуть надходити з локальних сенсорів/станцій та збереженої бази. Наприклад, CSV-файл з вимірами та попередньою активністю.

– Супутникові знімки: тестова модель очікує на вхід зображення (можна їх зберігати у папці). Ми створили штучні зразки, але на практиці варто підготувати колекцію реальних знімків пожеж із анотаціями.

– Вихід: демо-версія виводить класифікатор (0/1 пожежі), рівень ризику (0/1 або 0-2), а також показує карту з маркерами/зоною.

Візуалізувати результат можна, наприклад, через скріншот карти диспетчера з позначеними критичними зонами. У разі локальної системи такі зображення зберігаються вручну.

Розроблена система є повністю офлайновою: всі компоненти (моделі машинного навчання, база даних, веб-інтерфейс) розгортаються локально без звернень до зовнішніх API чи хмар. Для навчання та аналізу використовуємо бібліотеки TensorFlow та Scikit-learn, які не потребують Інтернету при виконанні. Карта реалізована через Folium (Leaflet) з tiles=None для локальних тайлів. Сповіщення виконуються у браузері

користувача (Web Notifications), що дозволяє отримувати push-подібні алерти без сервера.

У роботі використовувались підходи кластерифікації пожеж як бінарної задачі і моделювання ризику деревами рішень. Інтеграція Folium з Flask документована на офіційному сайті Folium, а техніка офлайн-карт (відключення онлайн-тайлів) описана в довідці. Ці методи та стандарти дозволяють зібрати прототип KIAC без залучення зовнішніх сервісів:

– Модуль виявлення пожеж за супутниковими знімками.

– Оцінка ризику поширення вогню.

– Панель диспетчера з геоприв'язкою об'єктів.

– Алерти та візуальні індикатори.

2.3. Тестування

Проведено випробування на даних 2022–2023 років. Точність класифікації ситуацій: 93%. Верифікація на випадках пожеж у Харківській та Луганській областях надана на рис. 4.

Рис. 4. Результати тестування

2.4. Перспективи

Можливість адаптації для інших типів НС (повені, радіаційні витоки), інтеграція з мобільними додатками, масштабування до рівня області чи країни.

Висновки

Практичне значення роботи полягає у створенні гнучкого прототипу, який може бути адаптований під будь-який тип НС (пожежі, повені, воєнні події) та використовуватись органами цивільного захисту, місцевими громадами, підприємствами критичної інфраструктури. Система забезпечує ситуаційну обізнаність, зменшує час реагування, покращує прийняття рішень завдяки інтегрованій аналітиці та візуалізації.

Подальші напрямки розвитку включають:

– Розширення функціональності системи на інші типи НС (радіаційні витоки, терористичні загрози).

– Інтеграцію з мобільними додатками та платформами спільного інформування населення (crowdsourcing).

– Впровадження хмарної або гібридної інфраструктури для масштабування.

– Підключення до міжнародних систем обміну кризовими даними (GDACS, UN OCHA).

– Поглиблення аналітики на основі глибокого навчання, Big Data та семантичної обробки текстів.

Отже, розроблена KIAC є вагомим кроком у цифровізації управління надзвичайними ситуаціями в Україні та відповідає сучасним викликам у сфері безпеки та реагування на кризові події.

Список літератури

[1] Захарова І. В., Філіпова Л. Я., Задорожний І. С., Тарасенко Д. А. Основи інформаційно-аналітичної діяльності : навч. посіб. / І. В. Захарова, Л. Я. Філіпова, І. С. Задорожний, Д. А. Тарасенко ; 2-е вид., випр. і допов. Черкаси: Східноєвропейський університет імені Рауфа Аблязова, 2024. 347 с. URL: https://suem.edu.ua/storage/doc/books/osnovy-informaciyno-analitychnoi-dialnosti-zaharova2.pdf?utm_source=chatgpt.com.

[2] Chen R., Sharman R., Rao H. R., Upadhyaya S. J. Coordination in Emergency Response Management. Communications of the ACM, May 2008, Vol. 51, No. 5, pp. 66-73. URL: https://dl.acm.org/doi/10.1145/1342327.1342340.

[3] Стрижак О. Є. Онтологічні інформаційно-аналітичні системи. Радіоелектронні і комп'ютерні системи, 2014, № 3 (67), С. 71-76. URL: http://nbuv.gov.ua/UJRN/recs_2014_3_13.

Надійшла до редколегії 21.07.2025

УДК 622.692:519.87 DOi 10.30837/ bi.2025.2(103).18

А. В. Палєєв[1], В. Г. Котух[2], Ю. Ю. Гусєва[3], К. М. Палєєва[4]

[1]ХНУМГ ім. О. М. Бекетова, м. Харків, Україна, artem.palieiev@kname.edu.ua,
ORCID iD: 0009-0000-6044-0786

[2]ХНУМГ ім. О. М. Бекетова, м. Харків, Україна, volodimir.kotuh@kname.edu.ua,
ORCID iD: 0000-0002-6679-8620

[3]ХНУМГ ім. О. М. Бекетова, м. Харків, Україна, yulia.guseva@kname.edu.ua,
ORCID iD: 0000-0001-6992-543X

[4]ХНУМГ ім. О. М. Бекетова, м. Харків, Україна, kateryna.palieieva@kname.edu.ua,
ORCID iD: 0000-0001-6004-2331

АНАЛІЗ І ПОРІВНЯННЯ МЕТОДІВ МАТЕМАТИЧНОЇ ПІДТРИМКИ РЕГУЛЮВАННЯ НЕРІВНОМІРНОСТІ СПОЖИВАННЯ ГАЗУ В ГАЗОТРАНСПОРТНІЙ СИСТЕМІ УКРАЇНИ В УМОВАХ НЕВИЗНАЧЕНОСТІ

У статті проведено аналіз і порівняння сучасних методів математичної підтримки процесів регулювання нерівномірності споживання природного газу в газотранспортній системі України, що функціонують в умовах підвищеної невизначеності, зумовленої воєнними діями та змінами на енергетичному ринку. Визначено основні фактори невизначеності, що впливають на стабільність та ефективність функціонування газотранспортної системи. Проаналізовано існуючі підходи до прогнозування споживання газу, способи покриття сезонної, добової та годинної нерівномірності, надано оцінку їх техніко-економічної ефективності. Запропоновано використання комбінованих методів регулювання, які поєднують стохастичне моделювання та цифровий моніторинг стану систем з урахуванням воєнних ризиків і невизначеності попиту. Результати дослідження можуть бути використані при оптимізації режимів транспортування газу, підвищенні енергетичної стійкості регіональних систем газопостачання, плануванні резервів та розробці моделей адаптивного управління для кризових умов.

ГАЗОРОЗПОДІЛЬНІ МЕРЕЖІ, НЕРІВНОМІРНІСТЬ СПОЖИВАННЯ ГАЗУ, НЕВИЗНАЧЕНІСТЬ, МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ, ЕНЕРГЕТИЧНА БЕЗПЕКА, РЕЗЕРВУВАННЯ, БУФЕРНІ СПОЖИВАЧІ

A. V. Paleyev, V. G. Kotukh, Yu. Yu. Huseva, K. M. Paleyeva. Analysis and comparison of methods for mathematical support in regulating gas consumption irregularities in Ukraine's gas transportation system under conditions of uncertainty. The article analyses and compares mathematical methods for supporting the regulation of gas consumption irregularities in Ukraine's gas distribution system under conditions of uncertainty. The study highlights the growing instability of Ukraine's energy sector, which is caused by wartime destruction of infrastructure, fluctuations in demand, and shifts in the consumer base. The authors systematize long-term and short-term approaches to balancing gas consumption, including underground gas storage, the use of buffer consumers, and pressure-based storage in pipelines. Mathematical dependencies for evaluating seasonal, daily, and hourly unevenness are presented. The proposed integrated decision-support approach combines stochastic and scenario modelling methods to improve the reliability and adaptability of gas supply systems. Practical recommendations are provided for optimising gas transportation and storage under uncertain conditions. The results obtained can be used in planning gas supply schemes, improving energy security, and developing intelligent systems for forecasting and regulating gas consumption.

GAS DISTRIBUTION NETWORKS, GAS CONSUMPTION IRREGULARITY, UNCERTAINTY, MATHEMATICAL MODELING, ENERGY SECURITY, GAS RESERVING, BUFFER CONSUMERS.

Вступ

В умовах сучасного розвитку техніки та технологій надійність та стабільність роботи енергетичної системи є однією з найважливіших, але, водночас, і однією з найуразливіших складових безпеки будь-якої країни [1]. Як зазначено в [2], формування ефективної системи енергетичної безпеки є «...основним показником розвитку національної економіки країни...» та «... визначається внеском усіх її складових: екологічної, наукової, інформаційної та інших сфер життя країни чи території, регіону».

Робота енергетичної системи відбувається в умовах змін зовнішнього середовища, мінливості геополітичних процесів, зміни економічних та соціальних факторів, що безпосередньо впливають на її функціонування, тобто в умовах невизначеності. Це створює передумови для зростання вразливості енергетичного сектора, оскільки ускладнює довгострокове планування, підвищує ризики порушення стабільності постачання енергоносіїв і потребує запровадження системного управління ризиками, спрямованого на підвищення стійкості та гнучкості енергетичної інфраструктури.

З 2022 року енергетичний сектор України зазнав значних змін. З одного боку, через скорочення промислового виробництва, міграцію населення та загальний економічний спад відбулося зниження споживання енергії. З іншого боку, внаслідок воєнних дій було пошкоджено (а часом і повністю зруйновано) велику кількість об'єктів енергетичної інфраструктури. Усе це стало серйозним випробування для енергетичної галузі

нашої країни, спричинивши загострення проблем енергетичної безпеки та збільшивши рівень невизначеності в її роботі [3].

До війни основними видами палива в енергетиці виступали вугілля та торф (25,2 %). Друге місце займав природний газ (23,9 %) [3]. Однак, втрата вугілля як важливого джерела енергії, що відбулася внаслідок воєнних дій на Донбасі, а також наявність в українських надрах значних запасів природного газу, вивела газову галузь на передові позиції з точки зору забезпечення енергетичної стабільності держави [1]. Тому важливим стає питання забезпечення надійності роботи газотранспортної системи України.

На сьогодні функціонування газотранспортної системи України відбувається в умовах значної та постійно зростаючої невизначеності, зумовленої як глобальними енергетичними процесами, так і внутрішніми соціально-економічними та воєнно-політичними аспектами. Можна виділити наступні основні фактори невизначеності, що впливають на роботу газотранспортної системи в умовах війни:

– фізичні ризики – руйнування або пошкодження об'єктів газотранспортної інфраструктури внаслідок бойових дій, обстрілів, мінування територій та обмеження доступу до об'єктів експлуатації;

– енергетична несталість – коливання обсягів транзиту, зміна джерел постачання природного газу, потреба у швидкій перебудові маршрутів транспортування та перерозподілу потоків;

– техніко-економічні ризики – нестабільність цін на газ, коливання попиту та про-позиції, а також ризики неплатоспроможності споживачів і партнерів; обмеженість у постачанні матеріалів, обладнання та комплектуючих; зниження кадрового потенціалу через мобілізацію, евакуацію або руйнування виробничих баз;

– правові та організаційні фактори – постійні зміни у законодавчій базі, адаптація до європейських стандартів енергетичного ринку, необхідність узгодження технічних рішень із міжнародними операторами, тощо.

Підвищення рівня ризиків в роботи газотранспортної системи в умовах війни потребує переосмислення підходів до планування, управління та технічного забезпечення цих систем. Ключовим завданням стає забезпечення надійності, гнучкості та адаптивності систем газопостачання, організація їх стабільної роботи за мінливих, непередбачуваних та високоризикових обставин. Це вимагає впровадження сучасних методів прогнозування, моделювання ризиків та управління невизначеністю (застосування підходів сценарного аналізу, стохастичного моделювання, цифрового моніторингу, тощо).

1. Постановка завдання

У сучасних досліджень приділяється увага питанням прогнозування споживання газу та аналізу методів, що застосовуються для прогнозування [4, 5]. Так, згідно дослідженням, наведеним в [5], найбільш поширеним методом є нейронні мережі. Також приділяється увага питанням керування невизначеності в системах газопостачання [6, 7]. Однак недостатньо висвітленими залишаються питання техніко-економічного моделювання нерівномірності добового/годинного споживання. Крім того в розглянутих моделях не враховуються умови невизначеності в роботі систем газопостачання, викликані саме воєнними діями.

Тому, попри наявні дослідження щодо методів та моделей управління системами газопостачання, недостатньо досліджено математичні методи оцінювання добової та годинної нерівномірності в умовах стохастичної невизначеності, спричиненої воєнними ризиками. Тому актуальним завданням є розроблення підходів до оцінювання та регулювання нерівномірності газоспоживання на основі методів математичної підтримки прийняття рішень.

Метою статті є обґрунтування теоретичних і практичних засад підвищення ефективності функціонування газотранспортної системи України в умовах невизначеності шляхом аналізу методів покриття та вирівнювання нерівномірності газоспоживання. Для досягнення поставленої мети вирішуватимуться такі завдання:

1. Проаналізувати фактори невизначеності, що впливають на роботу газотранспортної системи України в умовах війни.

2. Визначити економічні та технічні наслідки нерівномірності газоспоживання.

3. Систематизувати існуючі методи та засоби покриття сезонної, добової й годинної нерівномірності споживання газу.

4. Надати рекомендації щодо напрямів підвищення стійкості газопостачання в України.

2. Викладення основного матеріалу

Нерівномірність споживання властива усім видам палива, але тільки для газу усунення її впливу на економічні показники розподілу і використання палива перетворюється в складну проблему. Так, наприклад, тверде і рідке паливо (вугілля, мазут і т. п.) відносно легко піддаються складуванню, однак для газу це складно, а іноді і неможливо. Тому, за відсутності спеціальних заходів, графіки видобування, транспортування і споживання газу повинні бути синхронізовані, що викликає необхідність розрахунку системи «промисел – газопроводи» по максимуму газоспоживання і обумовлює її роботу на оптимальному режимі.

На сьогодні розповсюджені два підходи до проблеми подолання економічних наслідків нерівномірності газоспоживання і, відповідно, два види заходів її реалізації:

– покриття нерівномірності газоспоживання або різними методами акумуляції надлишків газу в періоди зниженої його витрати, або іншими методами, наприклад, створенням резервів газу для його використання в періоди підвищеного попиту;

– вирівнювання нерівномірності газоспоживання шляхом ущільнення його графіку [8].

У загальному випадку, заходи першого виду забезпечують можливість оптимізації режимів видобування і транспортування газу без зміни графіків роботи цілорічних його споживачів. Заходи ж другого виду основані саме на зміні режимів газоспоживання.

Зазвичай покриття нерівномірності газо-споживання або її вирівнювання по різному впливає на техніко-економічні показники систем газопостачання. Це також створює передумови для оптимізації споживання газу і забезпечує підвищення використання основних фондів, а також зниження собівартості і питомих капіталовкладень в систему газопостачання. Так, наприклад, міські газорозподільні мережі розраховуються на сумісний максимум газоспоживання. Це створює позитивний ефект і умови для зменшення діаметрів таких систем при тій самій річній витраті газу.

У загальному випадку вирішальна роль в подоланні наслідків нерівномірності газоспоживання належить буферним споживачам – регуляторам, що отримують газ під час «провалів» графіку газоспоживання. Зазвичай в якості буферних виступають крупні споживачі палива, які отримують газ безпосередньо з газотранспортної системи, і, в цілому, в межах крупного територіального району вони ущільнюють графік газоспоживання. Іноді буферними споживачами виступають промислові підприємства, розміщені в межах території населеного пункту (міста), які отримують газ через газорозподільні системи та сприяють ущільненню суміщеного графіку газоспоживання міста. Тому їх використання можна розглядати як заходи другого виду.

Основними способами покриття нерівномірності газоспоживання є підземне зберігання газу під тиском; використання буферних споживачів регуляторів, що споживають газ під час сезонних «провалів» графіка навантаження; використання акумулюючої ємності кінцевих ділянок систем газопостачання; зберігання газу в трубах під тиском, тощо [9, 10]. Кожен спосіб має свою оптимальну сферу застосування і різну, залежно від місцевих умов, ступінь економічної ефективності. При виборі напрямків покриття нерівномірності газу споживання необхідно враховувати особливості трьох видів нерівності поставок газу: сезонної (місячної), добової і годинної. Так, сезонна нерівномірність потребує для свого покриття крупних запасів газу влітку, в період зниженого попиту, а добова і, особливо, година – порівняльно незначних запасів газу, але більшої інтенсивності їх відбору та продуктивності сховищ.

У таблиці 1 наведено порівняльний аналіз розглянутих методів. При систематизації матеріалу було використано інструменти штучного інтелекту (ChatGPT); остаточне наповнення, редагування та наукова верифікація даних таблиці 1 виконані авторами.

Таблиця 1 демонструє, що жоден із методів не є універсальним. Найвищу адаптивність в умовах підвищеної невизначеності забезпечують комбіновані підходи, які суміщають можливості довготривалого сезонного регулювання (підземні сховища газу), короткочасного добового балансування (буферні споживачі) та швидкого годинного реагування (лінійні ємності й акумулюючі ділянки трубопроводів).

Основними параметрами режиму газоспоживання, які необхідні для правильного вибору напрямків з покриття нерівномірності є: загальні коефіцієнти сезонної (місячної), добової та годинної нерівномірності, що визначають продуктивність різного роду газосховищ. З цим пов'язані також об'єми тимчасового надлишку газу, що утворюються при рівномірній його подачі в період зменшення попиту на газ, недостачі газу в періоди підвищеного попиту, а також об'єм складів резервного палива у буферних споживачів.

З огляду на підвищену варіабельність параметрів газоспоживання в умовах воєн-них ризиків доцільним є застосування елементів стохастичного моделювання. Випадковими величинами можуть виступати добові й годинні навантаження, пікові відбори, зовнішня температура, параметри тиску та ймовірність аварійних збурень. Стохастичний підхід дозволяє описувати розподіли цих параметрів і оцінювати їхній вплив на показники нерівномірності. Вихідною величиною такого моделювання є очікуване та граничне значення нерівномірності для різних сценаріїв, що підвищує точність вибору оптимальних методів її покриття.

Загальний об'єм тимчасових нестач газу за рік у відсотках від річної витрати може бути визначений за повного використання пропускної здатності систем газопостачання за формулою [11, 12]:

$$\alpha = \frac{\sum(K>1) - \sum(n>1)}{12 \cdot 10^{-2}}, \qquad (1)$$

а під час проєктування систем газопостачання з резервом пропускної здатності за формулою [11, 12]:

$$\alpha = \frac{\sum(K>A) - A\sum(n>A)}{12 \cdot 10^{-2}}, \qquad (2)$$

де α – частка об'єму нерівномірності, що підлягає покриттю, у загальному об'ємі газоспоживання за рік, %; $\sum(K>1)$, $\sum(K>A)$ – сума часткових коефіцієнтів місячної нерівномірності зі значеннями, що перевищують відповідно 1 або A; $\sum(n>1)$, $(n>A)$ – число часткових коефіцієнтів нерівномірності, які, відповідно, перевищують 1 або A; A – показник, зворотній величини резерву системи газопостачання $A = \frac{1}{1-q}$; q – частка резерву в загальній пропускній здатності в системі газопостачання; n – число місяців з $K>1$ або $K>A$.

Враховуючи різне число днів в місяцях року, більш точне значення α можна отримати [11, 12]:

Таблиця 1

Порівняння методів покриття та регулювання нерівномірності газоспоживання в умовах невизначеності

Метод	Тип перекриваної нерівномірності	Переваги	Недоліки	Умови ефективного застосування	Чутливість до різних видів невизначеності
Підземні газові сховища у водоносних пластах	Сезонна (місячна)	Великий обсяг зберігання; стабільний дебіт; здатність вирівнювати газопостачання протягом тривалих періодів	Висока вартість спорудження; складність геологічних умов; довгий час реакції	Наявність відповідних геологічних структур; віддаленість від зон ризику	Середня: фізичні ризики впливають помірно; економічні – мінімально
Підземні газові сховища у виснажених газових родовищах	Сезонна та частково добова	Низькі капітальні витрати; швидший запуск; висока надійність	Обмежені об'єми; залежність від характеристик пласта	Наявність родовищ поблизу споживача; стабільний тиск у системі	Низька: мінімально піддаються впливу через захищеність і автономність
Буферні споживачі-регулятори	Добова та частково годинна	Висока гнучкість; швидке реагування; мінімальні витрати на впровадження	Потреба в альтернативному паливі; залежність від готовності споживачів	Наявність у регіоні великих споживачів із резервним паливом	Висока: значний вплив фізичних та техніко-економічних ризиків
Акумулююча ємність кінцевих ділянок систем газопостачання	Годинна	Забезпечує локальне згладжування піків; проста реалізація	Обмежений обсяг акумулювання; залежність від тиску	Стабільний високий тиск; близькість до зони споживання	Середня: технологічні ризики помітно впливають, фізичні – менше
Зберігання газу в трубах під тиском (лінійні ємності)	Годинна	Найнижча собівартість; миттєва реакція; використовується у містах біля магістралей	Малі об'єми; обмеження за тиском; вимоги до безпеки	Розміщення між компресорними станціями; міські зони зі змінним попитом	Висока: чутливість до фізичних ризиків (пошкодження труб)
Вирівнювання графіку споживання (організаційні заходи)	Добова і частково сезонна	Дуже низька вартість; можливість значного ущільнення графіка	Потребує координації великої кількості споживачів	Розвинена система диспетчеризації; наявність керованих споживачів	Висока: організаційна нестабільність, непрогнозовані зміни попиту
Комбіновані методи	Сезонна, добова і годинна	Найбільша надійність; висока стійкість до пікових навантажень; можливість оптимізації	Підвищена складність; потреба в цифровому моніторингу та прогнозуванні	Цифрові системи контролю; інтеграція SCADA; наявність інтелектуальних алгоритмів	Низька або середня: стійкі

– за повного використання пропускної здатності системи газопостачання:

$$\alpha = \frac{\sum_{i=1}^{n}\left(K_{\text{н}_i}^{\text{м}} > 1\right)\cdot Z_i}{365\cdot 10^{-2}}, \qquad (3)$$

– за проєктування системи газопостачання з резервом пропускної здатності:

$$\alpha = \frac{\sum_{i=1}^{n}\left(K_{\text{н}_i}^{\text{м}} > A\right)\cdot Z_i}{365\cdot 10^{-2}}, \qquad (4)$$

де $K_{\text{н}_i}^{\text{м}}$ – частковий коефіцієнт місячної нерівномірності i-го місяця з $K_{\text{н}}^{\text{м}} > 1$ за повного використання пропускної здатності системи газопостачання і з $K_{\text{н}}^{\text{м}} > A$ за проєктування такої системи з резервом пропускної здатності; Z_i – число днів в i-му місяці з $K_{\text{н}}^{\text{м}} > 1$ або $K_{\text{н}}^{\text{м}} > A$; n – число місяців з $K_{\text{н}}^{\text{м}} > 1$ або $K_{\text{н}}^{\text{м}} > A$.

Виходячи з обраних напрямків дослідження обрано способи покриття нерівномірності споживання газу в системах газопостачання. Їх умовно можна поділити на дві групи [11, 12]:

– засоби довготривалого регулювання для покриття сезонної (місячної) нерівномірності завдяки використанню підземних сховищ надлишків газу;

– засоби короткочасного регулювання, здатні повністю або частково покривати добову та годинну нерівномірність, використовуючи крупні опалювальні і промислові котельні, кінцеві ємності ділянок систем газопостачання, тощо.

Якщо покриття нерівномірності газоспоживання в системах газопостачання здійснюється з використанням засобів короткострокового регулювання, то в цьому випадку необхідно роздільно визначити ємності для акумуляції запасів газу, призначених для покриття різних видів нерівномірності.

Розв'язання цієї задачі слід починати з визначення $\alpha_{\text{д}}$ – частки добової нерівномірності, що підлягає

покриттю засобами короткострокового регулювання, вираженої у %. При цьому розуміємо, що в кожному з місяців року є доби з витратами газу, які перевищують максимальну добову його подачу системами газопостачання або з підземних сховищ $Q_{\text{м.д.п}}$.

У цьому випадку α_4 визначається за формулою

$$\alpha_{\text{д}} = \frac{\sum\left(K_{\text{н}_i}^{\text{д}} > A\right) - A \cdot \sum\left(n_i^{\text{д}} > A\right)}{Z_i \cdot 10^{-2}}, \qquad (5)$$

де $\sum\left(K_{\text{н}_i}^{\text{д}} > 1\right)$ – сума часткових коефіцієнтів добової нерівномірності зі значеннями, що перевищують A за i-ий місяць;

$$A = \frac{Q_{\text{м.д.п}}}{Q_{\text{сер.д.х.м}}}, \qquad (6)$$

де $Q_{\text{сер.д.х.м}}$ – середньодобова витрата газу за самий холодний місяць; $\sum\left(n_i^{\text{д}} > A\right)$ – число часткових коефіцієнтів добової нерівномірності, що перевищують A за i-ий місяць; Z_i – число діб в i-му місяці.

При цьому ємність сховищ річного резерву газу для короткострокового регулювання добової нерівномірності газопостачання $V_{\text{сх.д}}$ може бути визначена за формулою [11, 12]:

$$V_{\text{сх.д}} = \sum_{i=1}^{m} Q_{\text{м}_i} \cdot \alpha_{\text{д}_i} \cdot 10^{-2}, \qquad (6)$$

де m – число місяців, в яких є доби з витратою газу, що перевищують $Q_{\text{м.д.п}}$; $Q_{\text{м}_i}$ – місячна витрата газу за i-ий місяць.

За умови $Q_{\text{м.д.п}} = Q_{\text{сер.д.х.м}}$ засоби довготривалого регулювання використовуються для покриття тільки сезонної нерівномірності, а при $Q_{\text{м.д.п}} = Q_{\text{м.д}}$ (максимально-добовій витраті газу) покриття добової сезонної нерівномірності здійснюється засобами довготривалого регулювання. Таким чином, ємність підземного сховища газу і запаси резервного палива у буферних споживачів, призначені тільки для покриття сезонної (і частково добової) нерівномірності газоспоживання, повинні бути визначені заздалегідь. Зазвичай, це різниця між загальною потребою в газі для покриття нерівномірності витрати за рік і кількістю газу, що подається зі сховищ короткострокового регулювання за рік для покриття добової нерівномірності [11, 12].

Частка об'єму годинної нерівномірності у загальному об'ємі максимально-добової витрати газу $\alpha_{\text{г}}$ може бути отримана по формулі [11, 12]:

$$\alpha_{\text{г}} = \frac{\sum\left(K_{\text{н}}^{\text{г}} > 1\right) - \sum\left(n^{\text{г}} > 1\right)}{24 \cdot 10^{-2}}, \qquad (7)$$

де $\sum\left(K_{\text{н}}^{\text{г}} > 1\right)$ – сума часткових коефіцієнтів годинної нерівномірності газоспоживання, що перевищують 1; $\sum\left(n^{\text{г}} > 1\right)$ – число часткових коефіцієнтів годинної нерівномірності газоспоживання, що перевищують 1.

Покриття годин нерівномірності споживання в міських газорозподільних мережах, особливо в умовах невизначеності, може здійснюватися різними методами. При використанні для покриття сезонної добової нерівномірності газоспоживання тільки буферних споживачів газу погодинна нерівномірність повинна покриватися за рахунок спеціальних заходів короткострокового регулювання. У цьому випадку бажана акумулююча ємність систем газопостачання розраховуватиметься за формулою [11, 12]:

$$V_{\text{сх.д}} = Q_{\text{м.д}} \cdot \alpha_{\text{г}} \cdot 10^{-2}, \qquad (8)$$

де $V_{\text{сх.д}}$ – ємність сховища регулювання годинної нерівномірності газоспоживання; $Q_{\text{м.д}}$ – максимальна витрата газу добу за добу.

У загальному випадку, використовуючи акумулюючу здатності кінцевих ділянок систем газопостачання доцільно регулювати годинні витрати газу. Наприклад, в години нічного провалу графіка газоспоживання в кінцеві ділянки можна нагнітати газу більше, ніж необхідно для покриття денних піків підключених споживачів, а в денні часи відповідно знижувати подачу газу нижче середньодобового рівня. Також при використанні для покриття сезонної добової нерівномірності газоспоживання сезонних надлишків газу, що акумулюються в підземних сховищах, покриття годинної нерівномірності можна здійснювати частково, а наявний запас газу для покриття нерівномірності газоспоживання в i-ий день може видаватися в мережу пропорційно розподілу добової витрати. У більшості випадків, особливо в умовах невизначеності, доцільно здійснювати максимально-добові відбори газу з підземних сховищ за оптимальним для них режимом, тобто рівномірно за годинами доби. Для цього слід використовувати акумулюючу ємність системи «сховище – пункти споживання». При цьому загальна ємність газосховищ з урахуванням резервів на покриття багаторічної нерівномірності газоспоживання, аварійного та інших запасів газу повинна складати до 17,5 % середньорічного об'єму газоспоживання.

До системи засобів покриття нерівномірності газоспоживання слід віднести також організацію підземних сховищ великих мас газу водоносних та виснажених пластах газових промислів. Слід зазначити, що підземні газосховища у водоносних пластах є високо економічними засобами покриття нерівномірності споживання газу. При цьому висока залежність економічної ефективності підземних газосховищ від характеру нерівномірності витрати газу обумовлена об'ємом сховища і максимально-добовою інтенсивністю відбору газу. Також на економічні показники сховища впливають глибина залягання пласта-колектору, дебіт свердловин, відстань від сховищ до пункту споживання газу, тощо.

Зазвичай капітальні вкладення в організацію газосховищ у виснажених газових промислах і витрати на їх експлуатацію нижче, ніж при створенні газосховищ у водоносних пластах, і схильні залежно від місцевих умов до значних коливань. Так, питомі капіталовкладення в сховища у виснажених промислах складають до 50 % від вкладень в сховища у водоносних пластах, а питомі експлуатаційні витрати – до 80 % [11, 12].

Дуже часто в містах, що розташовані за трасою магістральних газопроводів, між компресорними станціями для покриття годинної нерівномірності газоспоживання використовується зберігання газу в трубах під високим тиском. У цьому випадку в години

максимального газоспоживання увесь газ, що подається газопроводом і зберігається в трубах, буде поступати в міські газорозподільні мережі.

У години незначного газоспоживання частина газу з магістральних газопроводів буде поступати в міські газорозподільні мережі, а надлишки газу – в труби, що покладені в землю у вигляді батареї довжиною до 500 м. У години максимального газоспоживання увесь газ, що подається газопроводом і зберігається в трубах, буде поступати в газорозподільні мережі. Однак покриття годинної нерівномірності газоспоживання за рахунок резервів газу, що зберігається в трубах під тиском, економічно є більш ефективним, ніж використання акумулюючої здатності кінцевих ділянок систем газопостачання [11, 12].

Таким чином, основним напрямком боротьби зі шкідливим впливом нерівномірності газоспоживання є правильний вибір і поєднання методів її покриття та вирівнювання. Значні резерви ущільнення графіку газоспоживання полягають в оптимізації виробничого ритму промислових підприємств. Ущільнення графіку газоспоживання пов'язано також з необхідністю проведення комплексу заходів з боку, в першу чергу, міських газових господарств. Тут задіяно як певні організаційні заходи, так і значні капітальні вкладення та поточні експлуатаційні витрати.

Слід підкреслити, що існує безліч можливих варіантів організації системи регулювання нерівномірності газоспоживання, як за переліком використовуваних засобів, так і за їх питомою вагою. А економічна ефективність цих варіантів знаходиться в тісній залежності від різноманітних місцевих та ситуаційних умов.

Висновки

1. Проведений аналіз показав, що основним джерелом нерівномірності споживання природного газу в газотранспортній системі України є сезонні, добові та годинні коливання попиту, які посилюються дією факторів невизначеності, зокрема, воєнними ризиками, зміною структури споживачів та нестабільністю енергоринку.

2. Встановлено, що найбільш ефективними засобами покриття нерівномірності споживання газу є довготривалі методи регулювання (підземне зберігання у водоносних або виснажених пластах); короткотривалі методи регулювання (використання буферних споживачів-регуляторів, акумулюючих ємностей кінцевих ділянок систем газопостачання, зберігання газу в трубах під тиском). Раціональне поєднання цих підходів дозволяє підвищити ефективність використання газотранспортних потужностей і зменшити потребу в додаткових капітальних вкладеннях.

3. Запропоновано застосування комбінованих моделей математичної підтримки прийняття рішень, що поєднують методи стохастичного аналізу, сценарного прогнозування та цифрового моніторингу параметрів систем газопостачання. Це забезпечує підвищення точності оцінки нерівномірності споживання і зменшення ризиків при плануванні режимів подачі газу.

4. Наукова новизна роботи полягає у:

– формалізації та систематизації методів довготривалого та короткострокового регулювання нерівномірності газоспоживання з урахуванням різних видів невизначеності (воєнної, технологічної, економічної);

– обґрунтуванні використання стохастичного підходу для моделювання мінливих параметрів газоспоживання та їх впливу на сезонну, добову й годинну нерівномірність;

– введенні структурованої таблиці порівняння методів покриття нерівномірності, яка дозволяє проводити вибір оптимального комплексу заходів у конкретних експлуатаційних умовах;

– інтеграції елементів цифрового моніторингу та сценарного прогнозування у процес оцінювання стабільності систем газопостачання в умовах високої невизначеності.

5. Результати дослідження можуть бути застосовані під час розробки та реконструкції систем газопостачання регіонів і міст в умовах невизначеності, зокрема шляхом створення інтелектуальних систем управління газорозподільними мережами з використанням штучного інтелекту, машинного навчання та цифрових двійників. Запровадження таких рішень сприятиме формуванню адаптивних моделей управління, здатних забезпечувати стійку роботу газотранспортної системи України у мінливих і невизначених умовах, зокрема в умовах війни.

Список літератури:

[1] Кубатко О.В., Калініченко Л.Л., Півень В.С. Напрями покращання енергетичної системи національної економіки // Економіка та підприємництво. 2024. № 2 (132).

[2] Маліновська О.Я., Височанська М.Я. Енергетична безпека України як головний критерій ефективності функціонування національної економіки // Енергетика та економіка. 2023. Т. 1.

[3] Сіренко Ю., Вольвач Т., Савойський О., Козін В. Аналіз стану енергетичної системи України та заходи щодо покращення ситуації // Вісник Херсонського національного технічного університету. 2025. № 1. С. 229–237.

[4] Panek W., Włodek T. Natural Gas Consumption Forecasting Based on the Variability of External Meteorological Factors Using Machine Learning Algorithms, 2022.

[5] Mesarić J., Dujak D. Analysis of Methods and Techniques for Prediction of Natural Gas Consumption: A Literature Review, 2019.

[6] Koç E. B. Natural gas network design under demand uncertainty, 2019.

[7] O'Malley C., Hug G., Roald L. Stochastic Hybrid Approximation for Uncertainty Management in Gas Electric Systems, 2021.

[8] Пономарчук І.А., Слободян Н.М. Газопостачання : електронний навчальний посібник комбінованого (локального та мережного) використання [Електронний ресурс] / І.А. Пономарчук, Н.М. Слободян. Вінниця : ВНТУ, 2023. 103 с. URL: https://pdf.lib.vntu.edu.ua/books/2023/Ponomarchuk_2023_103.pdf

[9] Приймак М.В. Моделі газонавантажень з врахуванням стохастичної періодичності та можливості їх статистичного аналізу // Розвідка та розробка нафтових і газових родовищ. 2003. № 2(7).

[10] Дудля М.А., Ширін Л.М., Федоренко Е.А. Процеси підземного зберігання газу: підручник. Дніпро : Національний гірничий університет, 2012. 412 с.

[11] Дубинський Н.М. Надежность систем газопостачання. 1970. 215 с.

[12] Пешехонов Н.И. Проектирование газоснабжения. Киев, 1970. 147 с.

Надійшла до редколегії 30.10.2025

ПРАВИЛА
оформлення рукописів для авторів науково-технічного журналу «БІОНІКА ІНТЕЛЕКТУ»

Науково-технічний журнал «Біоніка інтелекту» приймає до друку написані спеціально для нього оригінальні рукописи, які раніше ніде не друкувались. Структура рукопису повинна бути такою: індекс УДК, відомості про авторів, заголовок, анотації (на трьох мовах), ключові слова, вступ, основний текст статті, висновки, список використаної літератури, резюме.

Відповідно до Постанови ВАК України від 15.01.2003 №7-05/1 (Бюлетень ВАК, №1, 2003, с. 2), стаття повинна мати такі необхідні елементи: постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями; аналіз останніх досліджень і публікацій і виділення не вирішених раніше частин загальної проблеми в даній області; формулювання цілей та завдань дослідження; виклад основного матеріалу досліджень з повним обґрунтуванням отриманих наукових результатів; висновки з даного дослідження та перспективи подальших досліджень у даному напрямку.

Статті мають бути виконані в редакторі Microsoft Word. Формат сторінки – А4 (210×297 мм), поля: верхнє – 25 мм, нижнє – 20 мм, ліве, праве – 17 мм. Кількість колонок – 2, з інтервалом між ними 5 мм, основний шрифт Times New Roman, кегль основного тексту – 10 пунктів, міжрядковий інтервал – множник (1,1), абзацний відступ – 6 мм. Обсяг рукопису – від 6 до 12 сторінок (мови: українська, англійська, російська та мовою орігінала).

УДК друкується з першого рядка, без відступів, вирівнювання по лівому краю.

ПІБ автора (-ів), назва статті, назва та адреса учбового закладу необхідно надати повністю російською, українською та англійською мовами.

Назва статті друкується прописними літерами; шрифт прямий, напівжирний, кегль 12.

Назви розділів нумерують арабськими цифрами, виділяють жирним шрифтом. Відступи для назви статті, ініціалів та прізвищ авторів, відомостей про авторів, назв розділів, вступу та висновків, списку літератури: зверху – 6 пт, знизу – 3 пт.

Анотації (мовою статті, абзац 6–12 рядків, кегль 9) розміщують на початку статті, в ній має бути розміщена інформація про очікувані результати описаних досліджень (на трьох мовах).

Ключові слова (4–10 слів з тексту статті, які з точки зору інформаційного пошуку несуть змістовне навантаження) наводять мовою рукопису, через кому в називному відмінку, кегль 9.

Рисунки та *таблиці* (чорно-білі, контрастні) розміщуються у тексті після першого посилання у вигляді окремих об'єктів і нумерують арабськими цифрами наскрізною нумерацією за наявності більше ніж одного об'єкта. Невеликі схеми, що складаються з 3–4 елементів виконують, використовуючи вставку об'єкта Рисунок Microsoft Word. Більш складні виконують у графічних редакторах у вигляді чорно-білих графічних файлів форматів .tif, .jpg, .wmf, .cdr із розділенням 300 dpi. Рисунки мають міститися у текстовому файлі й обов'язково подаватися окремими файлами з відповідними назвами (наприклад, рис1.jpg).

Усі елементи рисунка, включаючи написи, повинні бути згруповані. Усі написи в рисунках і таблицях мають бути виконані шрифтом Times New Roman, кегль у рисунках – 10, у таблицях – 9.

Рисунок повинен мати центрований підпис (поза рисунком), шрифт 9, відступи зверху і знизу по 6 пт. Ширина рисунка має відповідати ширині колонки (або ширині сторінки).

Формули, символи, змінні повинні бути набрані в **редакторі формул MathType**. Формули розміщують посередині рядка й нумерують за наявності посилань на них у рукописі. Шрифт – Times New Roman. Висота змінної – 10 пунктів, великих і малих індексів – 8 пт, основний математичний символ – 12 (10) пт. Змінні, позначені латинськими літерами, набирають курсивом, грецькі літери, скорочення російських слів і цифри – прямим написанням. Змінні, які є в тексті, також набирають у редакторі формул.

Список літератури вміщує опубліковані джерела, на які є посилання в тексті, укладені у квадратні дужки, друкують без абзацного відступу, кегль 9 пт, відступ зверху – 6 пт.

Після списку літератури з відступом зверху 6 пт зазначають *дату подання статті до редколегії*. Число та місяць задають двозначними числами через крапку. Розмір шрифта – 9 пт, курсив, вирівнювання по правому краю.

Резюме (Times New Roman, кегль — 10 пунктів,) подають англійською мовою: обсяг резюме до 2000 знаків (бажаний переклад). *Структура резюме:* **Background, Materials and methods, Results, Conclusion.**

Разом із рукописом (на аркушах білого паперу формату А4 щільністю 80-90 г/м^2, надрукований на лазерному принтері) необхідно подати такі документи:

1. Заяву, яку повинні підписати всі автори.
2. Акт експертизи про можливість опублікування матеріалів у відкритому друці (якщо потрібно).
3. Рецензію, підписану доктором чи кандидатом наук.
4. Відомості про авторів.
5. Електронний варіант рукопису, резюме та відомостей про авторів.
6. Зробити оплату публікації.

Необхідно також зазначити один з наступних тематичних розділів, якому відповідає рукопис:

1. Теоретичні основи інформатики та кібернетики. Теорія інтелекту.
2. Математичне моделювання. Системний аналіз. Прийняття рішень.
3. Інтелектуальна обробка інформації. Розпізнавання образів.
4. Інформаційні технології та програмно-технічні комплекси.
5. Структурна, прикладна та математична лінгвістика.
6. Дискусійні повідомлення.

ЗМІСТ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ. МАШИННЕ НАВЧАННЯ. БАЗИ ДАНИХ

ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ТА ОБРОБКА ДАНИХ

СИСТЕМИ АВТОМАТИЗАЦІЇ ТА УПРАВЛІННЯ

Наукове видання

БІОНІКА ІНТЕЛЕКТУ
інформація, мова, інтелект

Науково-технічний журнал

№ 2 (103)
2025

Головний редактор — *Г. Г. Четвериков*
Відповідальний редактор — *І. В. Кириченко*

Комп'ютерна верстка — *О. Б. Ісаєва*

Рекомендовано секцією № 2 «Інформаційні технології»
науково-технічної ради Харківського національного університету радіоелектроніки
(протокол № 11 від «10» грудня 2025 р.)

Адреса редакції:
Україна, 61166, Харків-166, просп. Науки, 14,
Харківський національний університет радіоелектроніки, к. 127
тел. 702-14-77, факс 702-10-13,
e-mail: bionics@nure.ua

Підписано до друку 24.12.2025. Формат 60 × 84 $^{1}/_{8}$. Друк ризографічний.
Папір офсетний. Гарнітура Newton. Умов. друк. арк. 17,7. Обл.-вид. арк. 17,4.
Тираж 20 прим.

Віддруковано в редакційно-видавничому відділі ХНУРЕ
61166, Харків, просп. Науки, 14.