УДК 004.8:519.24:614.2
DOI https://doi.org/10.32782/tnv-tech.2025.4.1.38

МЕТОДИ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ У МЕДИЧНІЙ СТАТИСТИЦІ УКРАЇНИ: АНАЛІТИЧНИЙ ОГЛЯД

Бобко Б. В. *– аспірант кафедри системного аналізу та інформаційних технологій Вінницького національного технічного університету*
ORCID ID: 0009-0004-1845-4841

Жуков С. О. *– кандидат технічних наук,*
доцент кафедри системного аналізу та інформаційних технологій Вінницького національного технічного університету
ORCID ID: 0009-0002-3443-6139

У статті розглянуто теоретичні засади застосування методів машинного навчання для виявлення аномалій у медичних статистичних даних. Проаналізовано сучасні підходи до детектування аномалій у великих масивах медичної інформації, включаючи агреговані показники охорони здоров'я та дані пацієнтів. Визначено основні типи аномалій (точкові, контекстні, колективні) та класифіковано методи виявлення аномалій (контрольовані, напівконтрольовані, неконтрольовані) з акцентом на останніх, що найчастіше застосовуються через брак розмічених даних. Розглянуто алгоритми і моделі машинного навчання, які використовуються для пошуку нетипових спостережень: статистичні, метрики близькості, кластерні, однокласові класифікатори, ансамблеві методи (наприклад, Isolation Forest) та глибокі нейронні мережі (автоенкодери). Наведено приклади використання таких методів для підвищення якості медичних даних (виявлення помилкових записів, аномальних значень показників) та для виявлення незвичних патернів, що можуть сигналізувати про нові явища (спалахи захворювань, шахрайство у сфері охорони здоров'я тощо). Узагальнення літератури показало, що використання методів машинного навчання дозволяє автоматизувати процес виявлення аномалій і значно скоротити час, який витрачається фахівцями на перевірку даних вручну. Відповідні методи успішно застосовано як за кордоном, так і в Україні, зокрема для контролю якості медичної статистики. Методи машинного навчання також демонструють потенціал у прогнозуванні тенденцій та оцінці ризиків у медичних системах, що посилює їхню роль у цифровій трансформації охорони здоров'я. Методологія машинного навчання є ефективним інструментом для знаходження аномалій у медичній статистиці, що сприяє своєчасному виявленню помилок даних та нестандартних ситуацій, підвищуючи обґрунтованість управлінських рішень у сфері охорони здоров'я.

Ключові слова: *аномалії, виявлення аномалій, медична статистика, машинне навчання, аномальні дані, охорона здоров'я.*

Bobko B. V., Zhukov S. O. Machine learning methods for anomaly detection in Ukrainian medical statistics: an analytical overview

The article examines the theoretical foundations of applying machine-learning methods to detect anomalies in medical statistical data. Contemporary approaches to anomaly detection in large-scale medical information–encompassing both aggregated health-care indicators and patient-level records – are analysed. The principal anomaly types (point, contextual and collective) are defined, and detection techniques are classified into supervised, semi-supervised and unsupervised categories, with emphasis on the latter owing to the scarcity of labelled data. The study reviews algorithms and models employed to identify atypical observations: statistical tests, distance- and density-based metrics, clustering methods, one-class classifiers, ensemble techniques (e.g., Isolation Forest) and deep neural networks (autoencoders). Illustrative applications show how these methods enhance data quality by locating erroneous entries

© Бобко Б. В., Жуков С. О., 2025
Стаття поширюється на умовах ліцензії CC BY 4.0

or abnormal indicator values and uncover unusual patterns that may signal emergent phenomena such as disease outbreaks or health-care fraud. The literature synthesis demonstrates that machine-learning-driven anomaly detection automates the identification process and markedly reduces the time experts spend on manual data verification. Such methods have been successfully implemented both internationally and in Ukraine, notably for quality control of medical statistics. Machine learning methods also demonstrate potential in forecasting trends and assessing risks in medical systems, enhancing their role in the digital transformation of healthcare. Hence, machine-learning methodology constitutes an effective instrument for detecting anomalies in medical statistics, enabling timely identification of data errors and atypical situations and thereby strengthening evidence-based decision-making in the health-care sector.

Key words: *anomalies, anomaly detection, medical statistics, machine learning, anomalous data, health-care.*

Вступ. Використання машинного навчання відкриває нові можливості для вирішення зазначених проблем. Алгоритми машинного навчання здатні виявляти складні приховані патерни у великих масивах даних і вчитися розпізнавати нетипові випадки без явних правил. Зокрема, неконтрольовані методи (без учителя) дають змогу знаходити аномалії без попередньо розмічених прикладів, що особливо актуально для виявлення нових, невідомих раніше відхилень. У світі вже накопичено значний досвід застосування таких методів у домені охорони здоров'я. Наприклад, методи кластерного аналізу та дистанційні алгоритми використовували для автоматичного пошуку помилкових записів у великих масивах даних опитувань населення (алгоритм DBSCAN успішно виявив понад 94 % аномальних записів ваги у базі з сотень тисяч анкет). Страхові медичні компанії застосовують комплексні рішення на основі ансамблів моделей та методів глибокого навчання для виявлення лікарів або установ з нетиповими показниками ресурсокористування, що може свідчити про шахрайство або перевитрати. Нейронні мережі (автоенкодери, рекурентні мережі) вже продемонстрували ефективність у детектуванні аномальних патернів у часових рядах, таких як дані моніторингу пацієнтів або епідеміологічні показники. В Україні також розпочато впровадження подібних підходів: зокрема, розроблено прототипи систем для автоматизованого аналізу результатів медичних оглядів з метою знаходження аномальних показників у документації пацієнтів. Очікується, що інтеграція таких інструментів у медичні інформаційні системи підвищить якість даних і зменшить навантаження на персонал, який займається перевіркою статистичних форм. Таким чином, актуальність теми зумовлена потребою забезпечити надійність медичної статистики та своєчасне виявлення нетипових ситуацій у сфері охорони здоров'я.

Аналіз останніх досліджень і публікацій. Фундаментальну теоретичну базу виявлення аномалій заклали класичні монографії C. C. Aggarwal [1] та V. Barnett і T. Lewis [2], які окреслили статистичні, евристичні й робастні підходи до роботи з викидами. Подальший розвиток полягає у системних оглядах V. Chandola, A. Banerjee й V. Kumar [3], V. Hodge та J. Austin [4] і M. A. F. Pimentel із колегами [5], що запропонували класифікацію алгоритмів за наявністю розмітки та типом аномалії. На емпіричну перевірку цих класифікацій спирався порівняльний аналіз M. Goldstein та S. Uchida [6]. Методологічні віхи позначили щільні алгоритми LOF авторства M. M. Breunig, H.-P. Kriegel, R. T. Ng і J. Sander [7], ансамблевий Isolation Forest від F. T. Liu, K. M. Ting та Z.-H. Zhou [8] і критерій F. E. Grubbs для одновимірних викидів [9]. У сфері високорозмірних даних значний внесок зробили B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola й R. Williamson із концепцією одноманітної опорної оболонки [10], тоді як обробка часових залежностей спирається на архітектуру S. Hochreiter та J. Schmidhuber (LSTM) [11] та її узагальнення у глибинному огляді G. Pang, C. Shen, L. Cao й A. Van Den Hengel [12]. Додатковий

практичний вимір внесла праця K. G. Mehrotra, C. K. Mohan і H. Huang із формалізацією принципів побудови алгоритмів аномалій [13].

У медичному домені статистичний алгоритм C. P. Farrington, N. J. Andrews, A. D. Beale та M. A. Catchpole став прототипом сучасних систем епідеміологічного сигналювання [14], а його поєднання з автоенкодером для COVID-19 обґрунтували H. Homayouni, I. Ray, S. Ghosh, S. Gondalia та M. G. Kahn [15]. Ефективність неконтрольованих кластерних підходів у виявленні помилок введення підтвердили A. Sau, S. Phadikar та I. Bhakta на масивах громадського здоров'я [16], тоді як автоенкодер і метод FindFPOF у реєстрі онкопацієнтів оцінили P. Röchner і F. Rothlauf [17]. Для українських реалій практичний прототип системи аномалій лабораторних показників розробив V. Ya. Drozdov [18]. Нарешті, комплекс пояснюваного неконтрольованого аналізу страхових даних, який актуалізує питання інтерпретованості, запропонувала група H. De Meulemeester, F. De Smet, J. van Dorst та співавт. [19]. У сукупності ці праці демонструють еволюцію від статистичних критеріїв до ансамблевих і глибинних методів, що утворюють надійну парадигму для підвищення достовірності та оперативності медичної статистики.

Мета дослідження. Метою роботи є систематизація методів машинного навчання для виявлення аномалій у медичній статистиці та оцінка їх придатності для української системи охорони здоров'я. Для цього виконано: узагальнення понять і класифікації аномалій та методів їх виявлення; аналіз категорій алгоритмів (з поясненням їх роботи та прикладами для медичних даних); огляд застосування методів у світі та в Україні з виокремленням переваг і проблем; надання рекомендацій щодо вибору методів для очищення даних, моніторингу та пошуку аномалій, а також визначення перспектив подальших досліджень.

Виклад основного матеріалу. Медична статистика оперує великими обсягами інформації: дані реєстрів пацієнтів, результати лабораторних досліджень, показники захворюваності і смертності, дані епіднагляду, фінансові показники системи охорони здоров'я тощо. У таких даних можливі аномалії різної природи. По-перше, це можуть бути помилки або шум – наприклад, хибно введені значення (опечатки, одиничні помилки вимірювань). Такі аномальні записи не несуть нової медичної інформації, але здатні суттєво спотворити аналіз, якщо їх вчасно не виявити. По-друге, аномалії можуть відповідати рідкісним подіям або новим явищам, які навпаки становлять практичний інтерес і потребують вивчення. Прикладом є несподіване різке збільшення випадків захворювання у певному регіоні (можливий спалах інфекції) або поява нетипового патерну у клінічних показниках пацієнта (можливе нове ускладнення). Виявлення таких аномальних патернів дає змогу своєчасно реагувати – проводити епідеміологічне розслідування, коригувати лікування тощо [3, с. 16].

Задача автоматизованого пошуку аномалій у медичних даних є складною через низку факторів. По-перше, медичні дані часто високорозмірні, гетерогенні та мають складні зв'язки (часові залежності, просторові контексти, множинні атрибути) [15, с. 289]. Наприклад, електронна медична карта пацієнта містить десятки полів різних типів; записи згруповані по пацієнтах і часу візитів, тож аномалії можуть проявлятися лише у певному контексті (наприклад, нелогічна послідовність діагнозів у межах однієї історії хвороби). По-друге, як правило, відсутні наперед розмічені приклади аномалій. Нові аномальні ситуації важко передбачити та задати вручну, тому методи з навчанням під наглядом (на розмічених даних) малоефективні [7, с. 95]. По-третє, пояснення виявленої аномалії є критично важливим для медичних фахівців [19, с. 24]. Якщо алгоритм сигналізує про підозрілий

випадок, лікарю чи аналітику потрібно розуміти, у чому саме полягає нетиповість (наприклад, «у цього пацієнта незвично низький пульс при високій температурі тіла»). Тому сучасні дослідження приділяють увагу розробці методів не лише для виявлення, а й для інтерпретації аномалій (засоби пояснення на основі правил, показників впливу атрибутів тощо).

Поняття аномалії є багатогранним, і різні автори можуть акцентувати різні його аспекти. У загальному випадку виділяють три основні типи аномалій за характером прояву [3, с. 17]:

– Точкові аномалії (point anomalies): поодинокі записи або вимірювання, які є нетиповими відносно решти даних. Це класичні викиди, коли один об'єкт значно відхиляється за своїми ознаками від інших. Наприклад, якщо серед тисяч вимірів температури тіла пацієнтів (нормально 36–39 °C) трапився запис 42 °C, то він може розглядатися як точкова аномалія (можливо, помилка або унікальний випадок). Точкові аномалії є найпростішими для розуміння і часто виявляються методами, заснованими на статистичних порогах (наприклад, правило 3σ, критерій Граббса) [9, с. 5].

– Контекстуальні аномалії (contextual anomalies): такі аномальні точки, які є нетиповими лише у певному контексті або умовах [17, с. 217]. Один і той самий показник може вважатися нормальним в одному контексті і аномальним в іншому. Контекст задається додатковими ознаками (наприклад, час, місце). Приклад: рівень захворюваності на грип 100 випадків на 100 тис. населення – загалом це звичайне значення, але якщо такий рівень зафіксовано влітку (коли зазвичай <10 випадків), то для літнього сезону це аномально високо. Тут контекст – літній період; без урахування контексту значення 100/100 тис. не виглядало б дивним. Контекстуальні аномалії також називають умовними, і для їх виявлення застосовують методи, що моделюють нормальну поведінку з урахуванням контекстних змінних (сезонність, тренди тощо).

– Колективні аномалії (collective anomalies): групи або послідовності пов'язаних між собою точок, які в сукупності формують аномальний патерн [17, с. 216]. Окремі елементи групи можуть не виглядати як аномалії самі по собі, але їх поєднання є нетиповим. Приклад: у часовому ряді відвідувань лікарень послідовність з 5 днів, коли кількість звернень зросла на 50 % від середнього, може бути колективною аномалією (ознака спалаху), хоча кожне окреме значення може лежати в межах історичних коливань. Інший приклад – набір показників пацієнта (температура, тиск, пульс), які в певній комбінації нетипові між собою, хоча кожен окремо в межах норми. Виявлення колективних аномалій потребує аналізу взаємозв'язків між даними (часових, просторових, логічних) і часто базується на методах послідовного аналізу, серій часів, методах пошуку частих патернів тощо [19, с. 23].

Зазначені типи аномалій перетинаються з поняттями «детектування викидів» та «детектування новизни». Детектування викидів (outlier detection) зазвичай має на увазі пошук точкових аномалій у вже наявному датасеті, включно з шумом [3, с. 20]. Натомість детектування новизни (novelty detection) фокусується на виявленні нових патернів, яких не було в навчальній вибірці – тобто передбачається, що навчальні дані не містять аномалій, і модель має спрацьовувати при появі принципово нового об'єкту [17, с. 219]. У практиці різниця між цими поняттями розмита, і термін «anomaly detection» часто охоплює обидва випадки.

Існує багато підходів до автоматизованого розпізнавання аномальних даних, і їх можна класифікувати за різними ознаками [3, с. 18]. Один з важливих

критеріїв – наявність навчальних міток (розмітки) у даних. За цим критерієм розрізняють три широкі категорії методів [14, с. 547]:

1. Контрольовані методи (supervised anomaly detection): вимагають розміченої вибірки – тобто даних, де кожне спостереження позначене як «нормальне» або «аномальне». Тоді задача зводиться до навчання бінарного класифікатора, який відноситиме нові об'єкти до одного з двох класів. Контрольовані підходи мало поширені в задачах пошуку аномалій, оскільки здебільшого отримати репрезентативну вибірку аномалій складно (аномальні випадки рідкісні, нові типи відхилень невідомі тощо). Ба більше, класифікатори схильні до зміщення через сильну дисбалансованість класів (аномалій значно менше). У медичних задачах прикладом контрольованого підходу може бути навчання моделі для виявлення шахрайства, де є наперед помічені випадки шахрайських дій і нормальної практики – але в більшості випадків такі мітки отримати важко [10, с. 145].

2. Напівконтрольовані методи (semi-supervised anomaly detection): передбачають, що частина даних розмічена, зазвичай тільки нормальні зразки. Модель навчається лише на нормальних даних, будуючи межу «нормальної поведінки», а потім для нових спостережень визначає, наскільки ймовірно вони вписуються в цю норму. Фактично це задача однокласової класифікації (one-class classification) – навчання алгоритму, який розпізнає тільки один (нормальний) клас, сигналізуючи, якщо об'єкт йому не належить. Багато популярних алгоритмів аномалії працюють саме у такому режимі, використовуючи лише «хороші» дані для тренування. Наприклад, якщо маємо вибірку коректних медичних записів, модель може навчитися описувати розподіл цих коректних даних, а все, що суттєво відхиляється від цього розподілу, вважати потенційною аномалією. Цей підхід добре працює для контролю якості даних, коли ми впевнені, що навчальна база очищена від явних помилок [10, с. 146].

3. Неконтрольовані методи (unsupervised anomaly detection): не потребують жодної розмітки даних. Вони безпосередньо аналізують структуру набору даних, виходячи з припущення, що аномалії є статистично рідкісними та відмінними від масиву нормальних точок [6, с. 176]. Іншими словами, такі алгоритми шукають об'єкти, які виглядають ізольовано або не вписуються в щільні кластери основної маси даних. Неконтрольовані методи є найбільш універсальними і тому найчастіше використовуються на практиці. Вони особливо цінні у випадках, коли апріорі невідомо, якими можуть бути аномалії (тобто неможливо зібрати їх приклади заздалегідь). Практично всі сучасні застосування виявлення аномалій у медичній статистиці покладаються на неконтрольовані алгоритми [7, с. 9]. Наприклад, алгоритми пошуку викидів у базах пацієнтів або часових рядах захворюваності працюють без учителя: вони самостійно виявляють записи, що мають низьку ймовірність появи згідно зі статистичною моделлю загальних даних.

Ще одна класифікація методів – за основним підходом або моделлю, яку використовує алгоритм [7, с. 11]. У літературі запропоновано різні схеми групування алгоритмів аномалій; одна з поширених підходить з точки зору сімейств алгоритмів [15, с. 290]. Наведемо узагальнену класифікацію методів виявлення аномалій, адаптовану до задач медичної статистики, у вигляді таблиці (Таблиця 1).

Як видно з Таблиці 1, для задач виявлення аномалій у медичних даних найважливішими є неконтрольовані методи, оскільки вони не потребують заздалегідь розмічених аномальних прикладів і здатні виявляти нові види відхилень. Серед них особливо популярними є алгоритми на основі близькості (LOF, k-NN), кластеризації (DBSCAN) та ансамблеві методи на кшталт Isolation Forest. Багато

досліджень відзначають високу ефективність ансамблевих підходів: так, Isolation Forest успішно застосовано для пошуку викидів у різних доменах, зокрема медичних, завдяки його властивості комбінувати результати багатьох випадкових ізоляцій і виявляти як поодинокі викиди, так і невеликі кластери аномалій [13].

Окремо варто згадати про часові ряди та просторово-часові дані, оскільки медична статистика часто представлена у вигляді динаміки показників (наприклад, щоденна захворюваність, щотижнева смертність, часові сигнали пацієнтів). Для часових рядів існують спеціалізовані методи аномального зрушення від тренду. Класичний статистичний підхід – контрольні карти і методи на кшталт CUSUM, методи Шумського-Еффронта для епіднагляду – дозволяють виявляти сплески показників [14, с. 547]. Сучасні ML-підходи використовують рекурентні нейронні мережі (LSTM) або гістерезисні методи. Так, у роботі Homayouni et al. (2021) запропоновано підхід IDEAL на основі LSTM-автоенкодера для аналізу часових послідовностей COVID-19 – алгоритм розбиває багатовимірні дані на підпослідовності, знаходить аномальні підпослідовності і візуалізує ступінь їх відхилення [15, с. 290]. Це дозволило автоматично виявляти незвичні тренди у статистиці COVID-19 для окремих штатів та пояснювати, які показники (захворювання, смертність, госпіталізації) найбільше виходять за межі очікуваних значень.

Методи виявлення аномалій вже застосовані у низці завдань, пов'язаних із медициною та охороною здоров'я. Наведемо деякі помітні напрямки:

Верифікація якості та цілісності даних. Перший рівень – це пошук грубих помилок і нелогічних записів у медичних базах. Дослідження Sau et al. (2024) на даних великого опитування населення (India Annual Health Survey) показало, що кластерний алгоритм DBSCAN може виявити переважну більшість випадкових помилок введення (наприклад, неправильно вказана вага дитини) з високою повнотою (~95 %) при прийнятній точності [16, с. 130]. Röchner & Rothlauf (2023) протестували неконтрольовані алгоритми на реальних записах ракового реєстру і виявили, що і патерновий метод (FindFPOF), і автоенкодер в 3.5 рази ефективніше знаходять неправдоподібні записи, ніж випадковий перегляд записів експертами [17, с. 127]. Зокрема, автоенкодер зміг автоматично виявити ~28 % імплозибельних (підозрілих) записів у вибірці, у той час як при ручній вибірці експерти знаходили лише ~8 %. Це підтверджує, що ML-алгоритми значно знижують трудомісткість контролю даних у реєстрах та статистичних системах охорони здоров'я. В Україні подібний підхід може бути надзвичайно корисним, враховуючи великі обсяги звітності, що зводяться до Центру медичної статистики МОЗ. Автоматизоване «очищення» медичних форм від аномальних записів (наприклад, нереальні дати, вік пацієнтів, діагнози невідповідні статі, тощо) підвищить достовірність офіційних показників і дозволить зосередити увагу експертів на дійсно важливих нестандартних випадках.

Виявлення шахрайства та аномальних практик. У системах медичного страхування і адміністрування медичних послуг актуальне завдання – пошук підозрілих відхилень у поведінці учасників (лікарів, клінік, пацієнтів), що можуть сигналізувати про неефективне витрачання коштів чи навмисне шахрайство. Класичні методи аудиту базуються на експертних правилах, але вони не встигають за складними схемами зловживань. Тому застосовуються ML-рішення для аналізу великих масивів даних про страхові випадки. Дослідження Van Leeuwen et al. (2024) описує комплексний підхід: поєднання кластеризації категоріальних ознак (через embedding-методи) з ансамблями неконтрольованих алгоритмів (Isolation Forest, One-Class SVM, Gaussian Mixture тощо) та методами пояснення SHAP для

Таблиця 1

Основні категорії методів виявлення аномалій та приклади алгоритмів

Категорія методів	Приклад алгоритму	Характеристика та примітки	Застосування в медицині
Статистичні методи (розподільчі)	Z-критерій, критерій Граббса	Припускають певний розподіл даних (нормальний тощо) і шукають точки, що мають низьку ймовірність (хвости розподілу). Працюють для простих (часто одновимірних) випадків, чутливі до вибору моделі розподілу	Контроль якості окремих показників (лабораторні аналізи – пошук позамежних значень)
Методи відстаней та густини (пошук сусідів)	k-NN outlier, Local Outlier Factor (LOF)	Базуються на метриці у просторі ознак: аномалії – це точки, які далекі від найближчих сусідів або знаходяться в областях із низькою локальною густиною даних. LOF порівнює локальну густину точки з густиною сусідів. Ефективні при не надто великих розмірностях	Виявлення нетипових пацієнтів шляхом порівняння з «схожими» пацієнтами (за віком, діагнозом тощо); пошук аномальних днів за кількістю звернень у лікарню (порівняння з близькими датами)
Кластерні методи	DBSCAN, OPTICS	Виявляють кластери даних; точки, що не входять до жодного щільного кластера, трактуються як аномалії. Параметри (радіус, мін. розмір кластера) впливають на результат. Ефективні для виявлення групових аномалій. DBSCAN показав високу точність у задачах пошуку data entry помилок	Відокремлення груп пацієнтів з рідкісними комбінаціями симптомів; виділення регіонів із нетипово низькою/високою захворюваністю на карті (якщо регіон не належить до жодного кластеру схожих регіонів)
Методи на основі класифікації (однокласові, SVM)	One-Class SVM, дерева рішень	Створюють модель, що відділяє нормальні дані від решти простору. One-Class SVM будує поверхню (гіперплощину) навколо нормальних точок, викиди знаходяться поза нею. Потребують налаштування параметрів (наприклад, відсоток допустимих викидів). Добре працюють, коли нормальний клас чітко відокремлений	Автоматична перевірка медичних записів на відповідність правилам: модель навчена на валідних записах, новий запис, що не підтримує навчений патерн, позначається для перевірки
Ансамблеві методи (дерева, ліс)	Isolation Forest, ансамбль дерев	Використовують комбінацію багатьох простих моделей (дерев) для ізоляції точок. Isolation Forest випадково розбиває дані, будуючи безліч «дерев ізоляції»; аномалії ізолюються на короткій глибині дерева. Ансамблеві алгоритми часто мало чутливі до масштабування ознак і можуть обробляти великі масиви	Виявлення пацієнтів з нетиповими показниками шляхом випадкового розбиття ознак; використовувався для виявлення аномалій у медичних опитуваннях і показав інваріантність до масштабів одиниць виміру

Продовження таблиці 1

Методи зменшення розмірності (лінійні моделі)	PCA, Robust PCA	Будують лінійну модель (напр. головні компоненти), що описує більшу частину даних. Точки, що мають великий реконструкційний промах (reconstruction error) – тобто сильно відхиляються від цієї моделі – вважаються аномаліями. Robust PCA витримує певний відсоток викидів при навчанні. Лінійні методи швидкі, але ефективні лише за лінійних залежностей	Контроль даних обстежень: PCA-модель на нормальних випадках виявляє пацієнтів, чиї результати аналізів не пояснюються жодною лінійною комбінацією типових результатів
Методи глибокого навчання	Автоенкодер, VAE, GAN	Використовують нейронні мережі для навчання складних нелінійних моделей нормальних даних. Автоенкодер навчається стискати та відновлювати дані; якщо для нового об'єкту помилка відновлення висока – він аномальний. Варіаційні автоенкодери (VAE) генерують розподіл латентних змінних, а Generative Adversarial Networks (GAN) можуть генерувати нормальні зразки, позначаючи погано генеровані як потенційні аномалії. Глибокі моделі потребують багато даних для навчання, але можуть виявляти дуже складні патерни	Виявлення аномалій на ЕКГ або медичних зображеннях: автоенкодери навчені на нормальних кардіосигналах виділяють пацієнтів з рідкісними аритміями; VAE використовували для пошуку аномальних МРТ-зображень мозку. У страхових даних глибокі моделі (разом із пояснювальними методами SHAP) допомагають знаходити підозрілі шаблони виставлення рахунків
Інші/ гібридні підходи	Rule-based + ML	Комбінують експертні правила та ML. Наприклад, спочатку відфільтровують явно некоректні дані за бізнес-правилами, а потім застосовують модель для решти. Інший приклад – інформаційно-теоретичні методи, що шукають аномалії як точки, що дають найбільшу поправку до ентропії системи. Гібридні методи можуть підвищити точність, але складні у реалізації	У медичних реєстрах часто є вбудовані правила валідації (логічні контролі). Поєднання: спочатку перевірка на такі правила (наприклад, вік пацієнта не може бути від'ємним), потім ML-модель шукає складніші аномалії у вже очищених даних

інтерпретації результатів [19, с. 21]. Цей підхід успішно виявив раніше не помічені аномальні патерни серед лікарів загальної практики у Бельгії, які полягали у надмірному призначенні певних процедур. В Україні, зі впровадженням НСЗУ (Національна служба здоров'я України) і електронних даних про медичні послуги, такі технології могли б використовуватися для моніторингу підозрілих відхилень (наприклад, значно більша кількість послуг, виписаних одним лікарем у порівнянні з колегами; нетипові комбінації призначених ліків тощо).

Епідеміологічний нагляд та виявлення спалахів. У сфері громадського здоров'я аномалії часом відповідають початку епідемій або надзвичайних ситуацій. Виявити нестандартне зростання захворюваності раніше традиційних методів – значить виграти час для реагування. Класичні алгоритми (наприклад, методи контрольних меж, алгоритм Farrington) давно застосовуються для цього, але зараз їх доповнюють ML-моделі. Так, поєднання моделей машинного навчання і динамічних систем було використано для раннього виявлення спалахів COVID-19, перевершивши за чутливістю стандартні методи епіднагляду [15, с. 293]. Методи класифікації (Random Forest, SVM) також залучаються для прогнозування та виявлення епідемічних аномалій – наприклад, при моніторингу грипу або денге [4, с. 87]. В українських реаліях аналогічні інструменти могли б використовуватися для аналізу регіональних статистичних рядів (виявлення нетипових відхилень у показниках захворюваності, смертності) з урахуванням сезонності та інших чинників, або для виявлення кластерів рідкісних захворювань (онко, орфанні хвороби) шляхом автоматизованого сканування просторових даних.

Клінічні застосування: моніторинг пацієнтів. У лікарняних інформаційних системах та пристроях спостереження (монітори в реанімації, носимі сенсори) також впроваджуються алгоритми аномалій. Вони допомагають ранньому виявленню погіршення стану пацієнта. Наприклад, алгоритм на основі One-Class SVM успішно застосовано для аналізу потоків даних з носимих датчиків, визначаючи незвичні патерни фізіологічних показників, що можуть передувати епізодам погіршення здоров'я [10, с. 147]. У кардіології автоенкодери й інші нейромережеві моделі навчаються на нормальних ЕКГ, ідентифікуючи аномальні ритми, що можуть означати аритмію або інші патології [12, с. 383]. Такі рішення інтегруються у системи підтримки прийняття рішень, попереджаючи лікарів про потенційні проблеми ще до того, як вони стануть критичними.

Отже, аналіз показує, що проблема виявлення аномалій у медичних даних інтенсивно досліджується. З'являються все нові алгоритми, здатні ефективніше і точніше виокремлювати нетипові випадки. Комплексні підходи – наприклад, поєднання традиційних статистичних правил із гнучкістю машинного навчання – дають найкращі результати у практичних системах контролю якості даних [16, с. 125]. Водночас залишається низка викликів: забезпечення інтерпретованості («прозорості») моделей для медиків, боротьба з хибними спрацюваннями (коли алгоритм позначає нормальні варіації як аномалії), адаптація моделей до зміни поведінки даних з часом (дрейф даних). Але загальний висновок дослідників одностайний: методи машинного навчання значно підсилюють можливості виявлення аномалій у медичній статистиці, роблячи цей процес більш автоматизованим, оперативним та об'єктивним [3, с. 17].

Висновки. У підсумку теоретичний аналіз засвідчив, що методологія машинного навчання створює чітку та гнучку рамку для детектування аномалій у медичній статистиці. Систематизація алгоритмічних підходів – від статистичних порогових критеріїв і методів відстаней до кластерних, однокласових та ансамблевих

моделей і глибоких нейромереж – дозволяє обирати оптимальні інструменти залежно від характеру даних та наявності розмітки. Особливо цінними в умовах обмеженої аннотації є неконтрольовані методи, тоді як ансамблеві та однокласові підходи демонструють високу точність у контрольних завданнях, а глибокі моделі сприяють виявленню складних патернів.

Практична інтеграція цих рішень в інформаційні системи охорони здоров'я України обіцяє суттєве підвищення достовірності статистики, прискорення перевірки даних і своєчасне виявлення нетипових явищ. При цьому ключовими завданнями залишаються забезпечення інтерпретованості рішень моделей, адаптація алгоритмів до національних особливостей медичних реєстрів та розвиток потокових рішень для моніторингу пацієнтів у реальному часі. Подальші дослідження мають зосередитися на оптимізації моделей під специфіку українських даних, впровадженні інтуїтивних інтерфейсів пояснення результатів та оцінці їхнього впливу на якість управлінських рішень у сфері охорони здоров'я.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Aggarwal C. C. Outlier Analysis. 2nd ed. New York: Springer, 2017. 466 p. DOI: https://doi.org/10.1007/978-3-319-47578-3

2. Barnett V., Lewis T. Outliers in Statistical Data. 3rd ed. Chichester: John Wiley & Sons, 1994. 584 p. DOI: https://doi.org/10.1002/bimj.4710370219

3. Chandola, Varun & Banerjee, Arindam & Kumar, Vipin. (2009). Anomaly Detection: A Survey. *ACM Comput. Surv*. 41. DOI: https://doi.org/10.1145/1541880.1541882

4. Hodge V., Austin J. A survey of outlier detection methodologies. *Artificial Intelligence Review*. 2004. Vol. 22, No. 2. P. 85–126. DOI: https://doi.org/10.1007/s10462-004-4304-y

5. Pimentel M. A. F., Clifton D. A., Clifton L., Tarassenko L. A review of novelty detection. *Signal Processing*. 2014. Vol. 99. P. 215–249. DOI: https://doi.org/10.1016/j.sigpro.2013.12.026

6. Goldstein M., Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*. 2016. Vol. 11, No. 4. Art. 173. DOI: https://doi.org/10.1371/journal.pone.0152173

7. Breunig M. M., Kriegel H.-P., Ng R. T., Sander J. LOF: Identifying density-based local outliers. *Proc. ACM SIGMOD International Conference on Management of Data*. 2000. P. 93–104. DOI: https://doi.org/10.1145/342009.335388

8. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest. *Proc. 2008 IEEE International Conference on Data Mining (ICDM)*. 2008. P. 413–422. DOI: https://doi.org/10.1109/ICDM.2008.17

9. Grubbs F. E. Procedures for detecting outlying observations in samples. *Technometrics*. 1969. Vol. 11, No. 1. P. 1–21. DOI: https://doi.org/10.1080/00401706.1969.10490657

10. Schölkopf B., Platt J., Shawe-Taylor J., Smola A., Williamson R. Estimating the support of a high-dimensional distribution. *Neural Computation*. 2001. Vol. 13, No. 7. P. 144–147 DOI: https://doi.org/10.1162/089976601750264965

11. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*. 1997. Vol. 9, No. 8. P. 135–160. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

12. Pang G., Shen C., Cao L., Van Den Hengel A. Deep learning for anomaly detection: A review. *ACM Computing Surveys*. 2021. Vol. 54, No. 2. P. 381–387. DOI: https://doi.org/10.1145/3439950

13. Mehrotra K. G., Mohan C. K., Huang H. 2017. Anomaly Detection Principles and Algorithms. DOI: https://doi.org/10.1007/978-3-319-67526-8

14. Farrington C. P., Andrews N. J., Beale A. D., Catchpole M. A. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the*

Royal Statistical Society. 1996. Vol. 159, No. 3. P. 547–563. DOI: https://doi.org/10.2307/2983331

15. Homayouni H., Ray I., Ghosh S., Gondalia S., Kahn M. G. Anomaly detection in COVID-19 time-series data. *SN Computer Science*. 2021. Vol. 2, No. 4. – Art. 288. DOI: https://doi.org/10.1007/s42979-021-00658-w

16. Sau A., Phadikar S., Bhakta I. Efficient detection of data entry errors in large-scale public health surveys: an unsupervised machine learning approach. *Discover Public Health*. 2024. Vol. 21. Art. 129. DOI: https://doi.org/10.1186/s12982-024-00245-3

17. Röchner P., Rothlauf F. Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. *BMC Medical Research Methodology*. 2023. Vol. 23. Art. 125. DOI: https://doi.org/10.1186/s12874-023-01946-0

18. Drozdov V. Ya. Development of a system for automated anomaly detection in medical research results: Master's thesis. Ternopil : TNTU, 2023. URL: https://elartu.tntu.edu.ua/handle/lib/44471

19. Meulemeester, Hannes & De Smet, Frank & Dorst, Johan & Derroitte, Elise & Moor, Bart. Explainable unsupervised anomaly detection for healthcare insurance data. *BMC Medical Informatics and Decision Making*. 2025. DOI: https://doi.org/10.1186/s12911-024-02823-6

REFERENCES:

1. Aggarwal C. C. (2017). Outlier Analysis. 2nd ed. New York: Springer, 466 p. DOI: https://doi.org/10.1007/978-3-319-47578-3

2. Barnett V., Lewis T. (1994). Outliers in Statistical Data. 3rd ed. Chichester: John Wiley & Sons, 584 p. DOI: https://doi.org/10.1002/bimj.4710370219

3. Chandola, Varun & Banerjee, Arindam & Kumar, Vipin. (2009). Anomaly Detection: A Survey. *ACM Comput. Surv*. 41. DOI: https://doi.org/10.1145/1541880.1541882

4. Hodge V., Austin J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*. Vol. 22, No. 2. P. 85–126. DOI: https://doi.org/10.1007/s10462-004-4304-y

5. Pimentel M. A. F., Clifton D. A., Clifton L., Tarassenko L. (2014). A review of novelty detection. *Signal Processing*. Vol. 99. P. 215–249. DOI: https://doi.org/10.1016/j.sigpro.2013.12.026

6. Goldstein M., Uchida S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*. Vol. 11, No. 4. Art. 173. DOI: https://doi.org/10.1371/journal.pone.0152173

7. Breunig M. M., Kriegel H.-P., Ng R. T., Sander J. (2000). LOF: Identifying density-based local outliers. *Proc. ACM SIGMOD International Conference on Management of Data*. P. 93–104. DOI: https://doi.org/10.1145/342009.335388

8. Liu F. T., Ting K. M., Zhou Z.-H. (2008). Isolation Forest. *Proc. 2008 IEEE International Conference on Data Mining (ICDM)*. P. 413–422. DOI: https://doi.org/10.1109/ICDM.2008.17

9. Grubbs F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*. Vol. 11, No. 1. P. 1–21. DOI: https://doi.org/10.1080/00401706.1969.10490657

10. Schölkopf B., Platt J., Shawe-Taylor J., Smola A., Williamson R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*. Vol. 13, No. 7. P. 144–147. DOI: https://doi.org/10.1162/089976601750264965

11. Hochreiter S., Schmidhuber J. (1997). Long short-term memory. *Neural Computation*. Vol. 9, No. 8. P. 135–160. DOI: https://doi.org/10.1162/neco.1997.9.8.1735

12. Pang G., Shen C., Cao L., Van Den Hengel A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys*. Vol. 54, No. 2. P. 381–387. DOI: https://doi.org/10.1145/3439950

13. Mehrotra K. G., Mohan C. K., Huang H. (2017). Anomaly Detection Principles and Algorithms. DOI: https://doi.org/10.1007/978-3-319-67526-8

14. Farrington C. P., Andrews N. J., Beale A. D., Catchpole M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society*. Vol. 159, No. 3. P. 547–563. DOI: https://doi.org/10.2307/2983331

15. Homayouni H., Ray I., Ghosh S., Gondalia S., Kahn M. G. (2021). Anomaly detection in COVID-19 time-series data. *SN Computer Science*. Vol. 2, No. 4. – Art. 288. DOI: https://doi.org/10.1007/s42979-021-00658-w

16. Sau A., Phadikar S., Bhakta I. (2024). Efficient detection of data entry errors in large-scale public health surveys: an unsupervised machine learning approach. *Discover Public Health*. Vol. 21. Art. 129. DOI: https://doi.org/10.1186/s12982-024-00245-3

17. Röchner P., Rothlauf F. (2023). Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. *BMC Medical Research Methodology*. Vol. 23. Art. 125. DOI: https://doi.org/10.1186/s12874-023-01946-0

18. Drozdov V. Ya. (2023). Development of a system for automated anomaly detection in medical research results: Master's thesis. Ternopil : TNTU, https://elartu.tntu.edu.ua/handle/lib/44471

19. Meulemeester, Hannes & De Smet, Frank & Dorst, Johan & Derroitte, Elise & Moor, Bart. (2025). Explainable unsupervised anomaly detection for healthcare insurance data. *BMC Medical Informatics and Decision Making*. DOI: https://doi.org/10.1186/s12911-024-02823-6

Дата першого надходження рукопису до видання: 20.08.2025
Дата прийнятого до друку рукопису після рецензування: 16.09.2025
Дата публікації: 30.10.2025