Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«ПОРІВНЯЛЬНИЙ АНАЛІЗ ПІДХОДІВ ДО ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ НА ОСНОВІ МОДЕЛЕЙ ARIMA, XGBOOST ТА LSTM»**

Виконала: студентка 4-го року навчання
освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Мельник Катерина Андріївна

Керівник: Крюкова Г. В.
кандидат фіз.-мат. наук, ст. викладач

Рецензент:

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК_____
(підпис)

«_____» _____ 20___р.

Київ - 2025

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

ЗАТВЕРДЖУЮ
Зав.кафедри математики,
доцент, кандидат фіз.-мат. наук
_____ *Чорней Р.К.*
(підпис)
“____”_____ 2025

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
для кваліфікаційної роботи
студентці 4-го курсу, факультету інформатики
Мельник Катерині Андріївні

Тема: «Порівняльний аналіз підходів до виявлення аномалій у часових рядах на основі моделей ARIMA, XGBoost та LSTM»

Зміст кваліфікаційної роботи:

Анотація
Вступ
1. Основні відомості про часові ряди та їх аномалії
2. Класифікації методів виявлення аномалій
3. Опис і аналіз обраних методів визначення аномалій у часових рядах
4. Реалізація і порівняння методів
Висновки
Список літератури

Дата видачі “____”_____ 2025 Керівник _____
(підпис)
Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2025 р.

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Отримання теми кваліфікаційної роботи.	29.09.2025			
2.	Ознайомлення з темою кваліфікаційної роботи.	10.10.2025			
3.	Розробка плану та структури роботи.	20.10.2025			
4.	Робота з науковою літературою, опис основних означень. Написання вступу та анотації.	30.11.2025			
5.	Аналіз загальновідомих методів визначення аномалій і їх класифікацій.	21.02.2026			
6.	Реалізація і порівняння обраних методів на реальних даних.	18.03.2026			
7.	Робота над текстовим оформленням теоретичної частини та одержаних результатів.	11.04.2026			
8.	Попередній аналіз кваліфікаційної роботи. Виправлення помилок.	01.05.2026			
9.	Попередній захист кваліфікаційної роботи.	11.05.2026			
10.	Захист кваліфікаційної роботи.	29.05.2026			

Науковий керівник _____ *(ПІБ)*

Виконавець кваліфікаційної роботи _____ *(ПІБ)*

Зміст

Анотація **5**

Вступ **6**

Актуальність 6

Мета і завдання дослідження 7

1 Часові ряди. Основні означення та характеристики **8**

1.1 Визначення часового ряду та його складових 8

1.2 Числові характеристики часових рядів та стаціонарність 11

1.3 Типи часових рядів та структура вхідних даних 13

1.4 Аномалії і їх типи 14

1.5 Приклади часових рядів 16

2 Класифікація методів виявлення аномалій **19**

2.1 Математична постановка задачі виявлення аномалій 19

2.2 Класифікація за типом вхідних даних 20

2.3 Класифікація за методологією 21

3 Методи визначення аномалій у часових рядах **25**

3.1 Статистичний метод визначення аномалій. ARIMA 26

3.2 Метод визначення аномалій з використанням машинного навчання. XGBoost 28

3.3 Метод визначення аномалій з використанням нейронних мереж. LSTM 30

4 Практичне дослідження і порівняння методів **34**

4.1 Метрики та критерії оцінювання 34

4.2 Вибір експериментальних даних, їх налаштування та опис процедури порівняння методів 35

4.3 Детекція аномалій та порівняння результатів методів 38

Висновки **42**

Список літератури **44**

Анотація

Робота присвячена порівняльному аналізу методів виявлення аномалій в одновимірних часових рядах. Розглянуто теоретичні основи аналізу часових рядів, класифікацію аномалій та підходи до їх виявлення. Для практичного дослідження обрано три методи — ARIMA, XGBoost та LSTM, — які представляють статистичний підхід, машинне навчання та нейронні мережі відповідно. Методи реалізовано та протестовано на реальних даних з відкритого бенчмарку NAB за метриками F2-score, ROC-AUC та часом обчислення. Результати свідчать про те, що статистичний метод ARIMA забезпечує рівень детекції, порівнянний із складнішими підходами, залишаючись при цьому значно простішим у застосуванні та інтерпретації. XGBoost демонструє найкраще поєднання точності та обчислювальної ефективності. Загалом результати вказують на те, що для одновимірних рядів із точковими аномаліями складніші моделі не гарантують суттєвих переваг.

Ключові слова: часовий ряд, виявлення аномалій, ARIMA, XGBoost, LSTM, машинне навчання, нейронні мережі.

Вступ

Актуальність

У сучасних інформаційних системах часові ряди є одним із найпоширеніших форматів даних: вони формуються в результаті безперервних вимірювань параметрів технічних, біологічних та соціально-економічних процесів. Розвиток сенсорних мереж та платформ збору даних суттєво збільшив обсяги таких послідовностей, що, своєю чергою, підвищує потребу в їх автоматизованому аналізі.

Виявлення аномалій у часових рядах є однією з ключових задач цього аналізу, оскільки нетипові або екстремальні значення можуть сигналізувати про події, критично важливі для найрізноманітніших систем. Наприклад, відхилення у серцевому ритмі можуть свідчити про небезпечний стан організму, незвична активність у мережевому трафіку — про можливу кібератаку, а аномальні коливання сейсмологічних показників можуть попереджати про потенційну загрозу землетрусу.

Попри значну кількість існуючих підходів, універсального методу для виявлення аномалій не існує. Ефективність алгоритмів значною мірою залежить від природи даних, типу аномалій, а також структури самого часового ряду. Це ускладнює вибір оптимального методу для розв'язання конкретної прикладної задачі.

До того ж на сьогодні сфера машинного навчання пропонує велику кількість підходів до виявлення аномалій — від класичних статистичних моделей до складних архітектур глибоких нейронних мереж. Водночас постає питання, чи завжди складніші моделі забезпечують кращі результати. Особливо цікаво й доцільно перевірити це припущення для простих одновимірних часових рядів, де класичні статистичні методи можуть виявитися достатньо ефективними при значно менших обчислювальних витратах.

Таким чином, актуальність роботи полягає у дослідженні співвідношення між складністю алгоритму та якістю отриманих результатів. У роботі здійснюється порівняння методів машинного навчання та нейронних мереж із класичними статистичними підходами на прикладі унівариативних часових рядів. Такий підхід дозволить оцінити доцільність використання складних

моделей та визначити ситуації, у яких простіші методи забезпечують достатній рівень ефективності без надмірних обчислювальних витрат.

Мета і завдання дослідження

Метою кваліфікаційної роботи є проведення порівняльного аналізу ключових методів виявлення аномалій у часових рядах та оцінка їхньої ефективності, надійності та придатності на конкретних наборах даних, що містять специфічні аномалії, на основі експериментальної реалізації.

Об'єктом дослідження є процес виявлення аномалій у часових рядах.

Предметом дослідження є методи, алгоритми та метрики, які використовуються для виявлення та порівняльного оцінювання аномалій у часових рядах.

Робота складається зі вступу та чотирьох розділів, які містять підрозділи, а також висновків і списку використаних джерел.

Перший розділ містить теоретичні основи аналізу часових рядів: формальні означення, основні числові характеристики, а також класифікацію аномалій.

Другий розділ присвячено огляду основних підходів до класифікації методів виявлення аномалій у часових рядах.

У третьому розділі наведено детальний аналіз обраних представників трьох напрямів: статистичних методів, алгоритмів машинного навчання та нейронних мереж, що формує теоретичне підґрунтя для їх практичного застосування та порівняння.

Четвертий розділ присвячено практичному дослідженню: описується вибір експериментальних даних, процедура порівняння методів та аналізуються отримані результати.

1 Часові ряди. Основні означення та характеристики

1.1 Визначення часового ряду та його складових

Часовий ряд — це впорядкована за часом послідовність числових спостережень, кожне з яких позначається як X_t і відповідає певному моменту часу t. Формально часовий ряд подають у вигляді множини $\{X_t, t \in T\}$, де T — індексована множина моментів часу, що може бути дискретною або неперервною. У більшості практичних задач, зокрема при виявленні аномалій, аналізують саме дискретні часові ряди, у яких значення фіксуються через рівні проміжки часу.[1] Такий формат дає змогу чітко відстежувати зміни динаміки та застосовувати широкий спектр статистичних і машинних методів.

З математичної точки зору часовий ряд розглядають як одну з реалізацій випадкового процесу. Змінна X_t у цьому випадку є випадковою величиною, а сам часовий ряд — спостережуваною траєкторією відповідного процесу.[1] Наприклад, вимірювання рівня електроспоживання щогодини утворюють дискретний часовий ряд, тоді як зміна температури протягом доби теоретично може описуватися неперервним процесом. На практиці навіть такі дані зазвичай подають у дискретній формі шляхом періодичних вимірювань або усереднення значень.

Для кращого розуміння структури часового ряду його часто розглядають як поєднання кількох базових компонентів. Такий підхід дозволяє відокремити систематичні закономірності від випадкових коливань. До основних компонентів належать тренд, сезонність, циклічність і залишкова складова, які в сукупності визначають поведінку ряду.[2]

Тренд описує довгостроковий напрям зміни ряду — висхідний, спадний або відносно стабільний. Він відображає неперіодичні зміни, зумовлені фундаментальними факторами, такими як розвиток технологій, економічне зростання або зміна соціально-демографічних характеристик. Прикладом тренду може бути поступове зростання індексу фондового ринку або збільшення кількості користувачів інтернету у світі протягом останніх десятиліть.

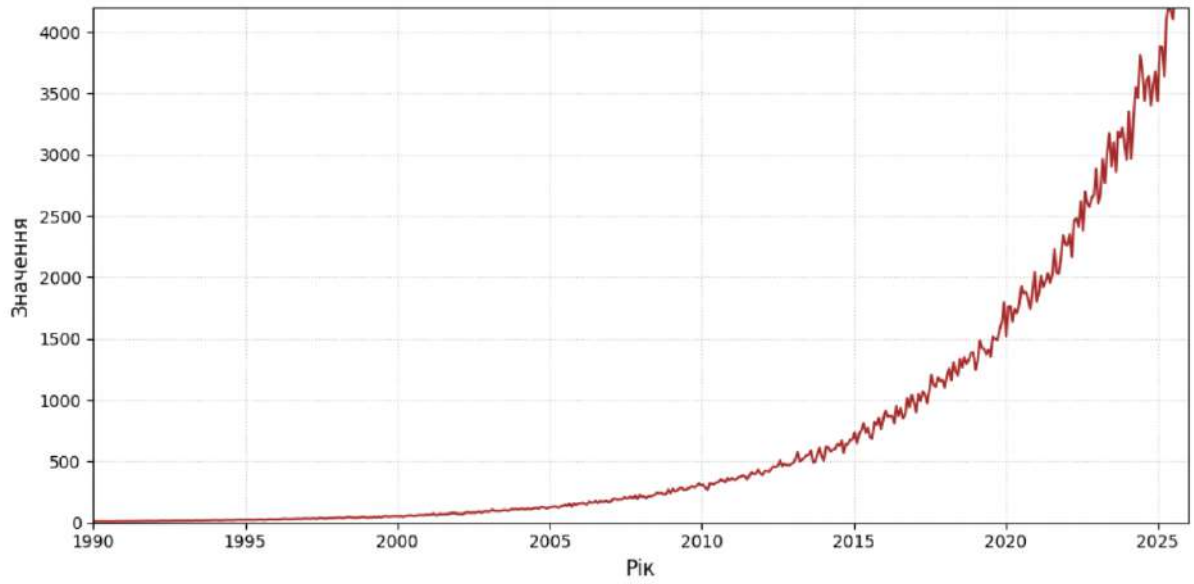

Рис. 1: Приклад тренду в часовому ряді

Сезонність — це регулярні коливання, які повторюються з фіксованою періодичністю: щотижня, щомісяця чи щороку. Такі коливання зумовлені передбачуваними календарними або кліматичними чинниками. Наприклад, попит на споживчі товари може різко зростати у святкові періоди, а обсяг продажів морозива стабільно збільшується в літні місяці.

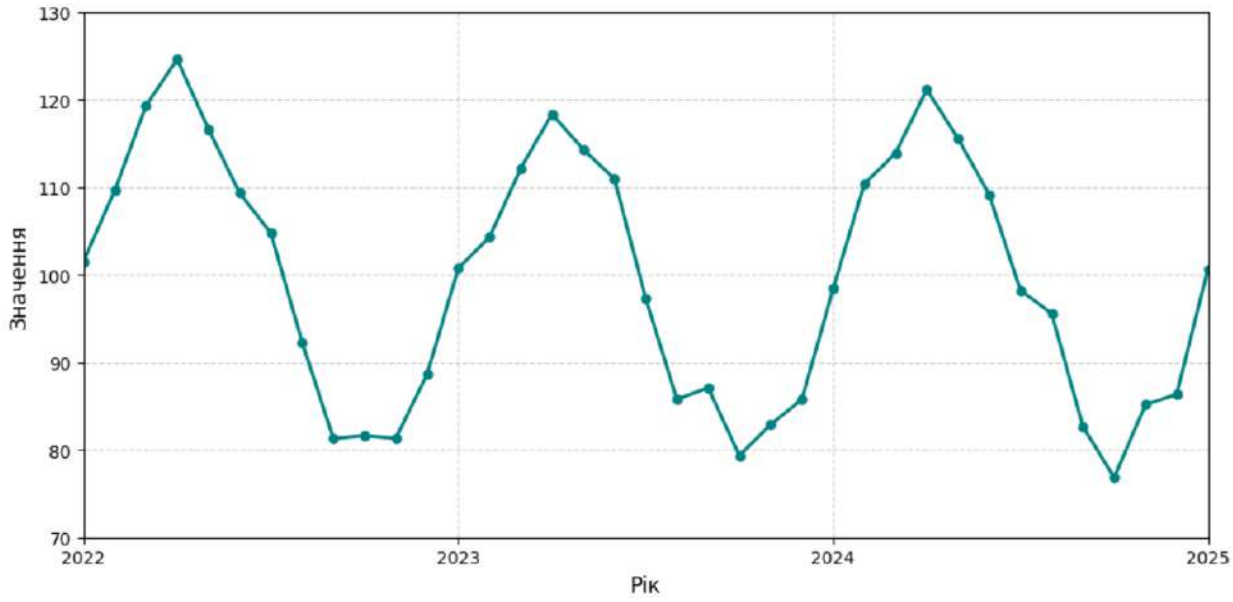

Рис. 2: Приклад сезонних коливань у часовому ряді

Циклічність характеризує коливання з довшою та нерегулярною тривалістю, які можуть охоплювати кілька років. На відміну від сезонності, циклічні

зміни не мають чітко визначеної періодичності та не прив'язані до календарних дат — вони зазвичай пов'язані з економічними циклами або іншими макроекономічними процесами. Типовим прикладом є чергування економічних підйомів і спадів.

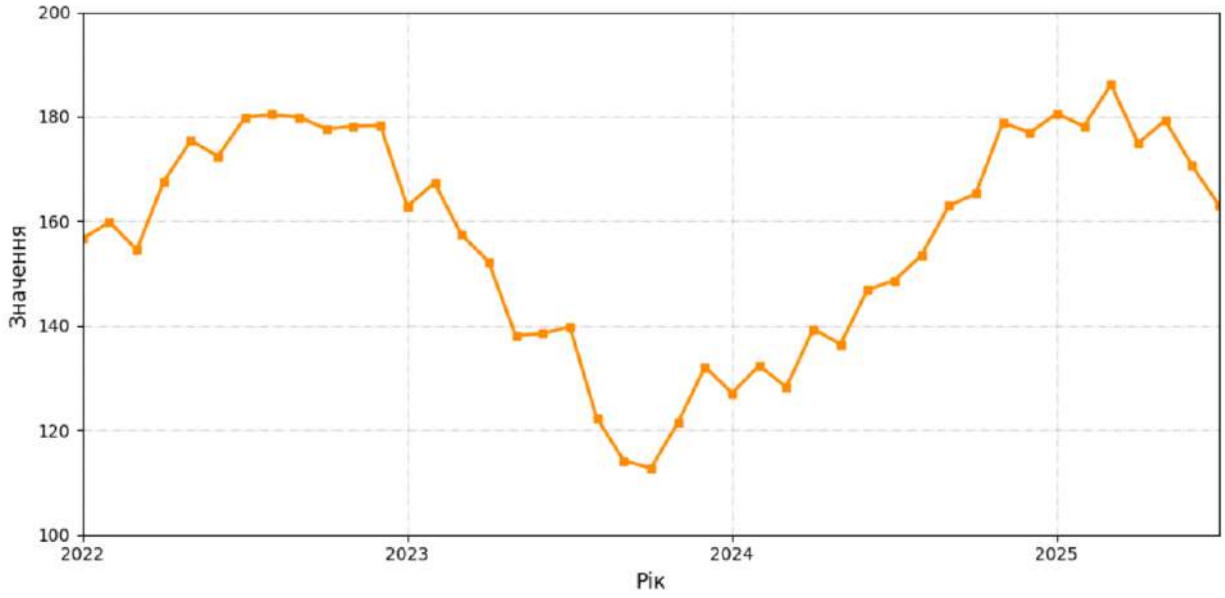

Рис. 3: Приклад несезонної циклічності у часовому ряді

Залишкова складова включає випадкові коливання, які не можуть бути пояснені трендом, сезонністю або циклічністю. Вона відображає випадковий шум, спричинений впливом непередбачуваних факторів — погодних умов, соціально-економічних подій або інших зовнішніх чинників. Варто зазначити, що саме у залишковій частині ряду найчастіше проявляються аномалії, що становлять особливий інтерес для задач моніторингу та аналізу.

Структурні компоненти можуть взаємодіяти між собою по-різному, що визначає вибір моделі декомпозиції. Адитивна модель застосовується тоді, коли амплітуда сезонних і нерегулярних коливань не залежить від рівня тренду[3]:

$$Y_t = T_t + S_t + C_t + R_t, \tag{1}$$

де Y_t — спостереження; T_t — тренд; S_t — сезонна компонента; C_t — циклічна компонента; R_t — залишки.

Таку модель використовують для рядів зі сталою амплітудою сезонності, наприклад у даних про середньодобове споживання електроенергії.

Мультиплікативна модель припускає, що амплітуда сезонних і випадкових коливань зростає разом із трендом[3]:

$$Y_t = T_t \times S_t \times C_t \times R_t. \tag{2}$$

Подібна структура властива, наприклад, місячному обсягу онлайн-продажів, де відхилення збільшуються разом із загальним зростанням ринку.

1.2 Числові характеристики часових рядів та стаціонарність

Для формального опису поведінки часового ряду використовують статистичні характеристики, які визначають типовий стан процесу. У задачах виявлення аномалій вони відіграють ключову роль: задають межі нормальної поведінки системи, відхилення від яких може свідчити про аномалію. Крім того, зазначені характеристики лежать в основі багатьох класичних моделей часових рядів, що використовуються для моделювання залежностей між спостереженнями. Розуміння цих властивостей є важливим також для подальшого застосування методів машинного навчання та нейронних мереж, оскільки вони дозволяють оцінювати структуру даних та коректно інтерпретувати результати роботи алгоритмів.

Математичне сподівання та дисперсія. Математичне сподівання μ_t характеризує середнє значення процесу в момент часу t[1] :

$$\mu_t = E(x_t) = \int_{-\infty}^{\infty} x\, f_t(x)\, dx, \tag{3}$$

де $f_t(x)$ — щільність розподілу ймовірностей у момент t.

Дисперсія σ_t^2 відображає розкид значень навколо середнього:

$$\sigma_t^2 = E\big[(x_t - \mu_t)^2\big]. \tag{4}$$

Знання математичного сподівання та дисперсії дозволяє кількісно описати типовий стан процесу. Будь-яке стійке відхилення фактичних значень від μ_t або різка зміна розкиду відносно σ_t^2 можуть слугувати сигналом про аномальну поведінку системи. Саме тому ці характеристики використовуються

як базис для формування критеріїв аномальності в деяких простих статистичних методах виявлення відхилень.

Автоковаріаційна та автокореляційна функції. Автоковаріаційна функція $\gamma(s,t)$ вимірює лінійну залежність між значеннями ряду в різні моменти часу[1] :

$$\gamma(s,t) = \operatorname{cov}(x_s, x_t) = E\big[(x_s - \mu_s)(x_t - \mu_t)\big]. \tag{5}$$

При $s = t$ автоковаріація збігається з дисперсією: $\gamma(t,t) = \sigma_t^2$.

Автокореляційна функція (АКФ) є нормованою мірою цієї залежності і набуває значень у межах $[-1,\, 1]$:

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\,\gamma(t,t)}}. \tag{6}$$

Часткова автокореляційна функція (ЧАКФ) визначає ступінь кореляції між x_t та x_{t-k} за умови, що вплив усіх проміжних значень усунено:

$$\phi_{kk} = \operatorname{corr}(x_t,\, x_{t-k} \mid x_{t-1}, \ldots, x_{t-k+1}). \tag{7}$$

Тут k позначає *лаг* — кількість кроків у часі між двома порівнюваними спостереженнями. Аналіз значень АКФ та ЧАКФ на різних лагах дозволяє охарактеризувати внутрішню структуру ряду.

Якщо АКФ залишається значущою на великих лагах, це свідчить про те, що поточні значення ряду суттєво залежать від спостережень, віддалених у часі. Такий ряд має виражену «пам'ять»: його динаміка є інерційною та передбачуваною, що, з одного боку, полегшує побудову прогнозів, а з іншого — може маскувати аномалії, оскільки відхилення поглинаються загальною тенденцією. Якщо ж АКФ швидко спадає до нуля вже на малих лагах, залежність між спостереженнями є слабкою. Такий ряд поводиться ближче до випадкового шуму: його важче прогнозувати, проте локальні відхилення виявляються помітніше на тлі загального процесу.

ЧАКФ, своєю чергою, дозволяє визначити, на яких саме лагах існує прямий зв'язок між спостереженнями, виключаючи вплив проміжних значень. Це робить її зручним інструментом для ідентифікації структури авторегресійних залежностей, що детальніше розглянемо у наступних розділах.

Стаціонарність. Стаціонарність — це властивість часового ряду, за якої його основні статистичні характеристики не змінюються з часом.[1] Формально стаціонарний процес задовольняє таким умовам:

- сталість математичного сподівання: $E(x_t) = \mu = \text{const}$;
- сталість дисперсії: $E[(x_t - \mu)^2] = \sigma^2 = \text{const}$;
- автоковаріація залежить лише від часового розриву h, а не від конкретного моменту часу: $\text{cov}(x_t,\, x_{t+h}) = \gamma(h) = \text{const}$.

Стаціонарність є необхідною умовою для коректного застосування більшості статистичних моделей аналізу часових рядів. Якщо ряд нестаціонарний — наприклад, містить тренд або зростаючу дисперсію, — його попередньо трансформують, зокрема шляхом диференціювання:

$$x'_t = x_t - x_{t-1}. \tag{8}$$

Це перетворення усуває лінійний тренд і повертає ряд до стану зі сталим середнім.

З точки зору виявлення аномалій стаціонарність відіграє принципову роль: вона дозволяє визначити стабільні статистичні межі нормальної поведінки процесу. Якщо ряд нестаціонарний, то середнє або дисперсія можуть природно змінюватися з часом, що суттєво ускладнює розмежування між закономірною зміною динаміки та справжньою аномалією.

1.3 Типи часових рядів та структура вхідних даних

Розуміння природи вхідних даних є важливою передумовою побудови алгоритмів виявлення аномалій. За кількістю параметрів, що аналізуються одночасно, виділяють два основні типи часових рядів:

- Одновимірні часові ряди, де кожне спостереження містить один скалярний показник.
- Багатовимірні часові ряди, у яких кожне спостереження представлено у вигляді вектора, матриці або тензора.[3]

Розмірність даних безпосередньо впливає на вибір методів виявлення аномалій. Часові ряди є послідовними даними: спостереження впорядковані за часом, і кожне поточне значення може залежати від попередніх. Саме ця властивість відрізняє їх від точкових даних, де кожен запис розглядається незалежно, та визначає специфіку підходів до аналізу й виявлення відхилень.

1.4 Аномалії і їх типи

Перед побудовою системи виявлення аномалій важливо визначити, що саме вважати аномалією в межах досліджуваного набору даних. У загальному випадку аномалією називають окреме спостереження або послідовність спостережень, які суттєво відрізняються від типової поведінки даних. Як правило, такі спостереження становлять лише невелику частину всієї вибірки.

При визначенні аномалій важливо не плутати їх із близькими за змістом поняттями. *Шум* виникає внаслідок випадкових помилок вимірювання або неточностей у даних і зазвичай не має аналітичної цінності — на відміну від аномалій, які відображають реальні відхилення у поведінці системи. На практиці межа між шумом та аномалією не завжди очевидна: у зашумлених даних аномальні спостереження можуть маскуватися випадковими коливаннями. Ще одним поняттям, яке варто відрізняти від аномалій, є *нові закономірності* (англ. *novelties*) — патерни, які раніше не спостерігалися в даних.[4] Їхня відмінність від аномалій полягає в тому, що після першого виявлення вони можуть поступово стати нормальною частиною процесу.

Залежно від структури та характеру прояву, аномалії поділяють на три основні категорії[4] :

1. **Точкові аномалії.** Це найпростіший тип: окремий екземпляр даних розглядається як аномальний відносно решти вибірки. Точка x_t вважається аномальною, якщо її значення суттєво відрізняється від сусідніх спостережень.

   Приклад: кредитна транзакція на суму, що значно перевищує типові витрати користувача.

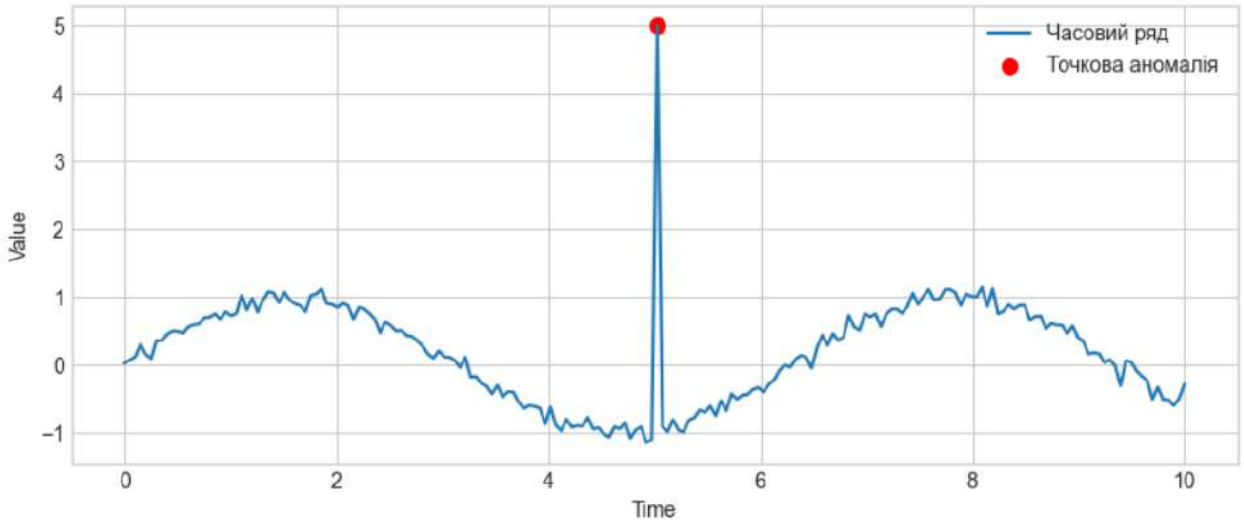

Рис. 4: Приклад точкової аномалії

2. **Контекстуальні аномалії.** Спостереження вважається аномальним лише в певному контексті, хоча за інших умов воно може бути цілком нормальним. Для їх виявлення аналізують два типи атрибутів: контекстуальні (час, місце розташування) та поведінкові (значення показника).

   Приклад: температура повітря 35°C є нормою влітку, але вважається аномалією для зимового періоду.

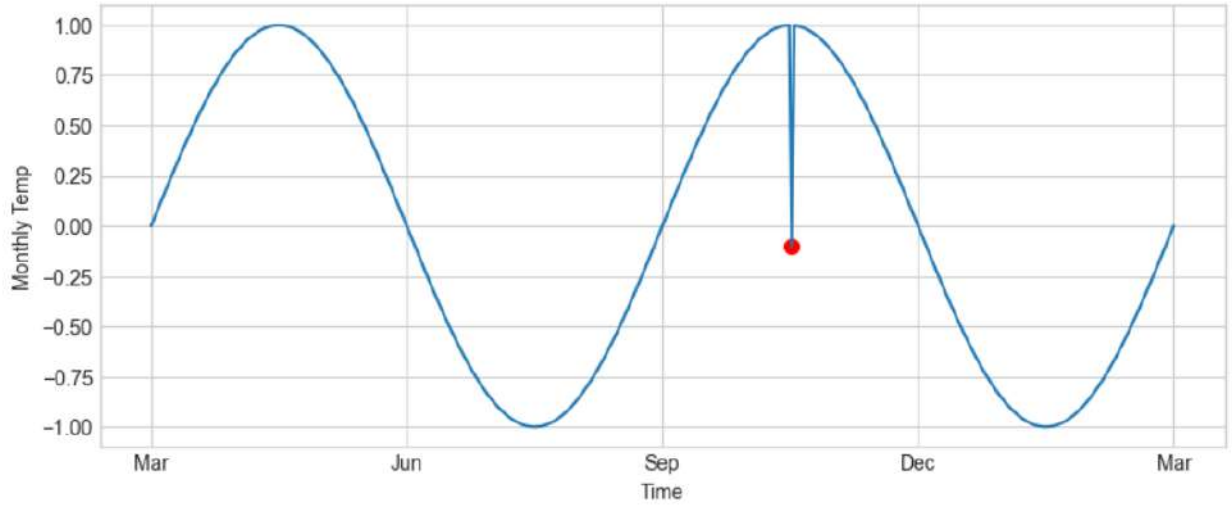

Рис. 5: Приклад контекстуальної аномалії

3. **Колективні аномалії.** Окремі точки послідовності можуть виглядати нормально, але їхня сукупність або порядок появи є аномальними. Цей

тип можливий лише в даних, де існують взаємозв'язки між спостереженнями.

Приклад: на електрокардіограмі занадто тривале збереження низького значення сигналу може свідчити про порушення ритму, хоча саме по собі це значення є допустимим.

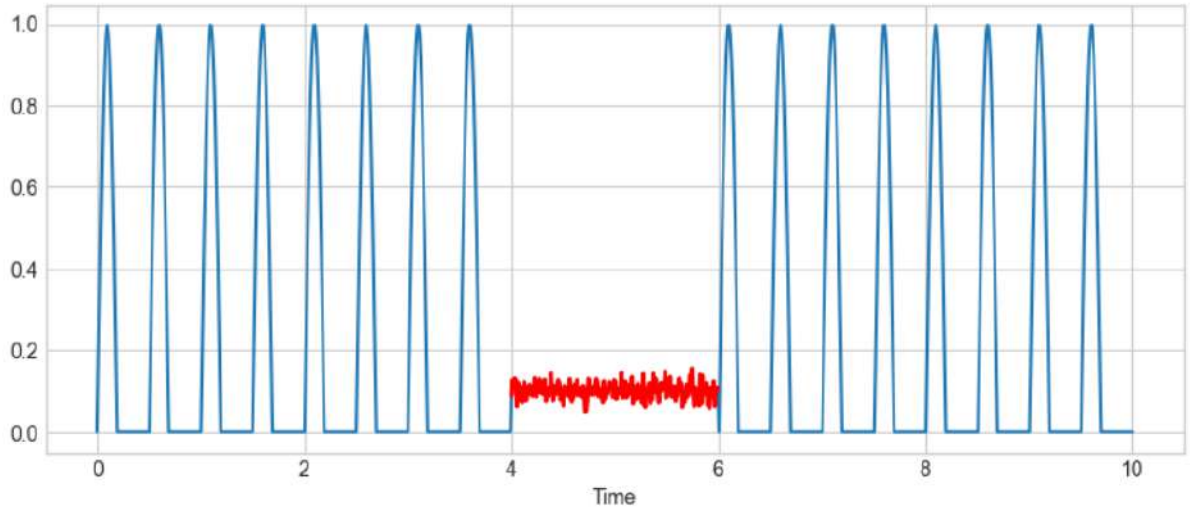

Рис. 6: Приклад колективної аномалії

Розуміння типу аномалій, які необхідно виявляти, є важливим для вибору відповідних методів аналізу. Алгоритми, що добре працюють для пошуку точкових аномалій, не завжди здатні ефективно виявляти контекстуальні або колективні відхилення, особливо в часових рядах, де важливу роль відіграє структура послідовності.

1.5 Приклади часових рядів

Часові ряди широко використовуються у різних сферах, зокрема у фінансовому аналізі, маркетингових дослідженнях, моніторингу технічних систем та наукових дослідженнях. Характер поведінки таких рядів може суттєво відрізнятися залежно від природи даних. Для ілюстрації розглянемо два приклади часових рядів із різною структурою.

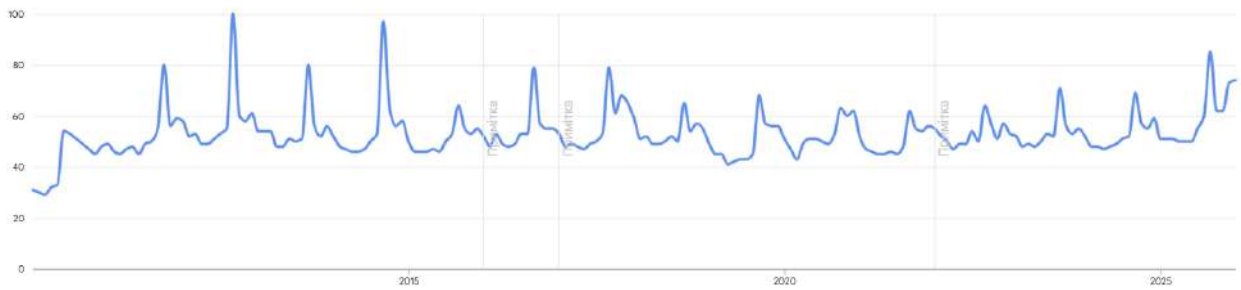

Рис. 7: Динаміка популярності пошукового запиту «iPhone» у Google за останні 15 років[5]

На даному графіку спостерігається виражена сезонна структура. Значення ряду регулярно зростають у вересні, що пов'язано з презентаціями нових моделей пристроїв. Таким чином формується чітка річна сезонність.

Рис. 8: Динаміка курсу акцій компанії Apple Inc. за останні 5 років[6]

На відміну від попереднього прикладу, фінансовий часовий ряд має іншу структуру. Для нього характерний виражений висхідний тренд та значні короткострокові коливання. Сезонні ефекти тут менш помітні, натомість добре спостерігаються локальні падіння та періоди швидкого зростання, які можуть бути пов'язані з економічними подіями, фінансовими звітами або змінами на ринку.

Наведені приклади показують, що часові ряди можуть мати різну структуру та динаміку: одні характеризуються регулярною сезонністю, інші —

вираженим трендом і значними випадковими коливаннями. У таких даних відхилення від типової поведінки не завжди легко помітити лише за допомогою візуального аналізу, особливо коли аномалії маскуються загальною динамікою процесу.

Саме тому для їх виявлення застосовують спеціалізовані методи аналізу, які дозволяють формально оцінювати ступінь відхилення спостережень від нормальної поведінки системи. У наступному розділі розглянемо математичну постановку задачі виявлення аномалій та основні підходи до класифікації відповідних методів.

2 Класифікація методів виявлення аномалій

2.1 Математична постановка задачі виявлення аномалій

У науковій літературі виявлення аномалій часто описують різними термінами, такими як детектування подій, виявлення порушень або ідентифікація викидів. Незважаючи на відмінності в термінології, всі ці підходи мають спільну мету — виявлення рідкісних спостережень, поведінка яких суттєво відрізняється від основної маси даних.

Формально задачу виявлення аномалій у часовому ряді $T = \{T_1, T_2, \ldots, T_m\}$ описують через функцію оцінювання, яка кожній точці T_i присвоює показник аномальності $s_i \in \mathbb{R}$. Отримана послідовність оцінок $S = \{s_1, s_2, \ldots, s_m\}$ задовольняє таку умову: для будь-яких двох точок T_i та T_j, якщо $s_i > s_j$, то T_i є більш аномальною, ніж T_j. Для прийняття остаточного рішення використовується порогове значення $\delta \in \mathbb{R}$: точка T_i вважається аномальною, якщо $s_i > \delta$, і нормальною в іншому випадку. Таким чином, задача зводиться до двох кроків: обчислення показника аномальності та вибору відповідного порогу.[7]

Варто зазначити, що виявлення аномалій не є класичною задачею бінарної класифікації. У більшості практичних задач аномалії становлять дуже малу частку набору даних, часто менше ніж 1%. У таких умовах традиційні класифікатори можуть демонструвати високу формальну точність, просто відносячи всі спостереження до нормального класу. Це робить задачу виявлення аномалій значно складнішою, оскільки модель повинна знаходити рідкісні та нетипові патерни в умовах сильного дисбалансу класів.

Ефективність виявлення аномалій залежить від правильного вибору методу, який визначається такими чинниками:

- **природа даних:** аналіз статичних наборів даних відрізняється від роботи з часовими рядами, де кожне спостереження має часову мітку;
- **розмірність даних:** в одновимірних рядах аномалія проявляється у відхиленні окремого значення, тоді як у багатовимірних — може виражатися через порушення взаємозв'язків між ознаками;

- **наявність розмітки:** визначає можливість використання навчання з учителем, напівкерованого навчання або методів без учителя;
- **тип аномалій:** різні алгоритми по-різному справляються з точковими, контекстуальними або колективними відхиленнями.

Для часових рядів задача додатково ускладнюється тим, що спостереження не є незалежними: кожне значення пов'язане з попередніми. Тому раптові зміни тренду, амплітуди або структури коливань можуть свідчити про аномальну подію. У таких випадках «ненормальність» визначається не лише абсолютним значенням показника, а й моментом часу, у якому воно виникає.

Таким чином, формальна постановка задачі дозволяє узагальнено описати процес пошуку відхилень незалежно від конкретного алгоритму. Проте на практиці існує велика кількість методів, що відрізняються як за типом даних, так і за математичними принципами аналізу. У наступних підрозділах розглянемо методи відповідно до типу доступної вхідної інформації та їхньої математичної методології.

2.2 Класифікація за типом вхідних даних

Вибір алгоритму для пошуку відхилень значною мірою залежить від того, наскільки повною є попередня інформація про структуру досліджуваного набору даних. Залежно від наявності або відсутності міток, що визначають приналежність кожного елемента до певного класу, зазвичай виділяють три основні підходи: навчання з учителем, напівкероване навчання та навчання без учителя[3].

Навчання з учителем базується на використанні тренувального набору даних, у якому кожен об'єкт заздалегідь позначений як нормальний або аномальний. Основна ідея полягає у побудові класифікаційної моделі, що навчається розрізняти характерні ознаки обох класів для подальшого аналізу нових спостережень. Такий підхід теоретично здатен забезпечувати високу точність, проте на практиці стикається з двома суттєвими труднощами. По-перше, аномальних прикладів зазвичай значно менше, ніж нормальних, що створює критичний дисбаланс класів. По-друге, отримати марковані дані для рідкісних подій часто дуже складно. Прикладами застосування є виявлення

шахрайських транзакцій із кредитними картками або фільтрація небажаної електронної пошти.

Напівкероване навчання передбачає, що для навчання доступна інформація лише про нормальну поведінку системи. Модель формує опис «норми» та використовує його як еталон, а будь-яке відхилення від нього в тестових даних інтерпретується як потенційна аномалія. Оскільки метод не потребує прикладів аномалій, він здатен виявляти навіть абсолютно нові типи відхилень, що раніше не траплялися. Такий підхід доцільно застосовувати там, де аварійні ситуації є надзвичайно рідкісними та їх складно змоделювати заздалегідь — наприклад, при діагностиці несправностей космічних апаратів, де аномальна подія фактично означає аварію.

Навчання без учителя є найбільш універсальним сценарієм, який не потребує жодної розмітки або попередніх знань про структуру набору даних. Алгоритм самостійно аналізує дані, виходячи з припущення, що нормальні стани трапляються значно частіше, тоді як аномалії є рідкісними та помітно відрізняються від загального розподілу. Якщо ж аномалії в наборі трапляються занадто часто, метод може генерувати значну кількість хибних тривог. Типовим прикладом є виявлення вторгнень у комп'ютерні мережі, де ручна розмітка даних є практично неможливою через їхній обсяг, а уявлення про норму може постійно змінюватися.

Варто також зазначити, що в літературі описано техніки, які навчаються виключно на аномальних прикладах. Однак такий підхід використовується рідко, оскільки сформувати навчальний набір, що охоплює всі можливі типи аномальної поведінки, на практиці вкрай складно[4].

Таким чином, тип доступної інформації про дані визначає загальну стратегію моделювання. Проте навіть у межах одного підходу математичний метод виявлення аномалій може суттєво відрізнятися. У наступному підрозділі розглянемо класифікацію методів на основі їхньої математичної методології.

2.3 Класифікація за методологією

Класифікація за методологією дозволяє структурувати алгоритми виявлення аномалій відповідно до математичного принципу, на якому ґрунтується про-

цес пошуку відхилень. Такий підхід пояснює, яким саме чином модель ідентифікує аномальні спостереження: через аналіз імовірнісних характеристик даних, вимірювання відстаней між об'єктами або визначення їхньої приналежності до певних структурних груп. У науковій літературі виділяють декілька основних методологічних підходів, серед яких найбільш поширеними є класифікаційні методи, методи на основі найближчого сусіда, кластерні та статистичні методи[4].

Класифікаційні методи базуються на побудові моделі, здатної відрізняти нормальні спостереження від аномальних на основі навчальних даних. На етапі навчання формується класифікатор, який використовує набір ознак для розмежування різних класів, після чого нові спостереження класифікуються як нормальні або аномальні.Ключовим є припущення про те, що у просторі ознак існує межа, яка дозволяє відокремити нормальні дані від відхилень.

У межах цього підходу зазвичай розрізняють багатокласову та однокласову класифікацію. У першому випадку модель навчається розрізняти декілька типів нормальної поведінки, а об'єкти, що не належать до жодного з відомих класів, розглядаються як аномальні. У другому випадку модель формує межу навколо нормальних спостережень, а всі точки поза цією областю інтерпретуються як відхилення. До поширених алгоритмів цього класу належать дерева рішень, метод опорних векторів (SVM, OC-SVM) та алгоритм XGBoost.

Перевагою класифікаційних методів є висока швидкість роботи на етапі застосування моделі та здатність ефективно працювати у просторах високої розмірності. Водночас їх використання суттєво обмежується необхідністю якісно розмічених навчальних даних для обох класів — що на практиці буває складно забезпечити, особливо для аномального класу.

Методи на основі найближчого сусіда ґрунтуються на ідеї, що нормальні спостереження, як правило, розташовані поруч із подібними об'єктами, тоді як аномалії є ізольованими у просторі ознак. Для реалізації такого підходу необхідно визначити міру подібності або відстані між об'єктами — наприклад, Евклідову відстань для безперервних атрибутів. Показник аномальності об'єкта зазвичай обчислюється через відстань до його k-го найближчого сусіда або через локальну густину оточення. Типовими представниками цього класу є алгоритм k-NN та метод локального фактора викиду (LOF — Local

Outlier Factor), який оцінює ступінь ізольованості кожної точки відносно її безпосереднього оточення.

Перевагою таких методів є відсутність потреби у припущеннях щодо імовірнісного розподілу даних та у розмічених прикладах — вони повністю спираються на структуру самого набору. Це робить їх придатними для широкого кола задач без попередніх знань про природу аномалій.

Проте методи цього класу мають суттєвий недолік — високу обчислювальну складність на етапі тестування. Для кожного нового спостереження необхідно обчислювати відстані до великої кількості точок, що може бути критичним при роботі з великими наборами даних. Крім того, зі зростанням кількості вимірів обчислення відстаней між точками втрачають здатність розрізняти нормальні та аномальні спостереження, що знижує якість виявлення відхилень.

Кластерні методи базуються на ідеї групування схожих об'єктів у кластери. У такому випадку аномаліями вважаються об'єкти, які або не належать до жодного кластера, або знаходяться далеко від центру найближчого скупчення. У деяких ситуаціях аномалії можуть формувати власні невеликі або розріджені групи, що дозволяє виявляти системні відхилення як побічний результат процесу кластеризації. Серед алгоритмів цього класу широко застосовуються k-means, DBSCAN та Isolation Forest.

Серед переваг варто виділити можливість роботи без розмітки даних, а також відносно швидку фазу тестування — нові спостереження порівнюються лише з невеликою кількістю центрів кластерів. Основним недоліком є залежність результатів від якості обраного алгоритму кластеризації: якщо він не здатен коректно відобразити структуру нормальних даних, аномалії можуть залишитися непоміченими.

Статистичні методи ґрунтуються на побудові імовірнісної моделі нормальної поведінки системи. Спостереження, що з погляду цієї моделі мають низьку ймовірність виникнення, розглядаються як потенційні аномалії. Такі методи можуть бути параметричними — коли форма розподілу задана заздалегідь і оцінюються лише його параметри — або непараметричними, коли модель будується без явних припущень щодо розподілу. До поширених представників належать моделі ARIMA, ARMA, AR, MA.

Перевагою статистичних підходів є наявність математично обґрунтованого критерію прийняття рішення, що забезпечує інтерпретованість результатів. За умови правильно підібраної моделі вони здатні ефективно виявляти відхилення навіть без розмічених даних. Проте варто зазначити, що статистичні методи ефективні переважно при невисокій розмірності даних — зі зростанням кількості вимірів підбір адекватної статистичної моделі суттєво ускладнюється. Крім того, якщо реальний розподіл даних суттєво відрізняється від прийнятих припущень, якість виявлення аномалій може помітно знизитися.

Таким чином, кожен із розглянутих підходів має свої сильні сторони та обмеження. Вибір конкретного методу залежить від властивостей набору даних, доступності розмітки, обчислювальних ресурсів та характеру відхилень, які необхідно виявити. Різноманітність підходів свідчить про відсутність універсального алгоритму, здатного ефективно працювати в усіх можливих сценаріях.

3 Методи визначення аномалій у часових рядах

Як було показано у попередньому розділі, методи виявлення аномалій можуть класифікуватися за різними критеріями: за типом вхідних даних, наявністю розмітки або математичною методологією пошуку відхилень. Проте для практичного аналізу та вибору конкретних алгоритмів доцільно також розглядати ці методи з точки зору їхньої складності та підходу до моделювання даних.

У такому випадку більшість існуючих методів виявлення аномалій у часових рядах можна умовно поділити на три основні групи:

- *статистичні методи;*
- *класичні методи машинного навчання;*
- *методи на основі нейронних мереж.*

Ключова різниця між цими підходами полягає у припущеннях щодо процесу генерації даних. Статистичні методи зазвичай виходять із того, що поведінку системи можна описати за допомогою певної математичної моделі або ймовірнісного розподілу, а спостереження з низькою ймовірністю появи розглядаються як аномальні. Методи машинного навчання, натомість, орієнтовані на пошук закономірностей без явного припущення про форму розподілу — вони аналізують структуру даних і на основі виявлених залежностей формують модель нормальної поведінки. Межа між статистичними методами та методами машинного навчання часто є менш очевидною, тоді як нейромережеві підходи вирізняються використанням спеціалізованих архітектур і зазвичай становлять окремий клас.

У наступних підрозділах розглядається по одному представнику кожної з груп. Усі три методи належать до прогнозуючих підходів: кожна модель навчається на нормальних даних, формує уявлення про очікувану поведінку ряду, а спостереження, що суттєво відхиляються від прогнозу, позначаються як потенційні аномалії. Це забезпечує спільну основу для їх подальшого по-

рівняння — умови детекції та метрики оцінювання залишаються однаковими для всіх методів.

3.1 Статистичний метод визначення аномалій. ARIMA

Одним із найбільш поширених статистичних підходів до аналізу часових рядів є модель ARIMA (Autoregressive Integrated Moving Average — авторегресивна інтегрована модель ковзного середнього). Вона широко застосовується для моделювання одновимірних часових рядів і завдяки здатності працювати з нестаціонарними даними часто використовується як базовий інструмент у задачах прогнозування та виявлення аномалій.

В основі ARIMA лежать два простіших підходи, які вона об'єднує.[8] *Авторегресійна модель* $AR(p)$ описує поточне значення ряду як лінійну комбінацію його попередніх p значень:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t, \tag{9}$$

де c — константа, φ_i — коефіцієнти моделі, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ — випадкова похибка.

Модель ковзного середнього $MA(q)$, натомість, описує поточне значення через q попередніх похибок прогнозування:

$$X_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \tag{10}$$

де μ — середнє значення ряду, θ_j — коефіцієнти моделі.

Поєднання цих двох підходів утворює *модель* $ARMA(p, q)$:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{11}$$

Проте суттєвим обмеженням моделей AR, MA та ARMA є вимога стаціонарності ряду. У реальних даних ця умова часто порушується через наявність тренду або сезонних змін. Для усунення цього обмеження модель ARMA розширюється до ARIMA(p, d, q) шляхом введення операції диференціювання. Це перетворення усуває лінійний тренд і стабілізує середнє значення ряду.

За потреби операцію повторюють — параметр d визначає, скільки разів вона застосовується.

Таким чином, кожен із трьох параметрів моделі відповідає за окремий аспект структури ряду:

- p — порядок авторегресійної складової: визначає, значення яких саме попередніх моментів часу враховуються при формуванні поточного прогнозу;
- d — кількість диференціювань, необхідних для досягнення стаціонарності; при $d = 0$ ряд вже стаціонарний, при $d = 1$ усувається лінійний тренд;
- q — визначає, скільки попередніх похибок прогнозування враховує модель, це дозволяє компенсувати короткострокові випадкові збурення.

Для ручного підбору p та q зазвичай аналізують графіки АКФ та ЧАКФ: характер їх спадання дозволяє оцінити можливі значення цих параметрів. Втім, на практиці підбір часто здійснюється автоматично — шляхом перебору комбінацій параметрів з оцінкою кожної за величиною прогнозної похибки, наприклад методом перехресної перевірки.

У задачах виявлення аномалій ARIMA використовується для побудови прогнозу нормальної поведінки ряду.[8] Модель навчається на ділянці даних без аномалій, після чого для кожного нового спостереження обчислюється залишок між фактичним та прогнозованим значенням:

$$e_t = X_t - \hat{X}_t, \tag{12}$$

де $\hat{X}_t$ — значення, передбачене моделлю. Якщо поведінка ряду відповідає вивченій нормі, залишки залишаються малими та випадковими — близькими до білого шуму. Якщо ж у ряді з'являється аномалія, модель не може її передбачити, і залишок різко зростає. Саме тому залишки порівнюють із заданим порогом: якщо відхилення перевищує його — спостереження позначається як потенційна аномалія. Для визначення порогу часто використовують Z-оцінку, що нормує залишки відносно їхнього середнього та стандартного відхилення. Втім, підходів до вибору порогу існує значно більше — від простих правил на основі квантилів до складніших статистичних методів.

Попри практичну ефективність, модель має як сильні сторони, так і обмеження, які варто враховувати при її застосуванні. До переваг ARIMA належать математична інтерпретованість та відносна простота реалізації. Модель не потребує розміченого навчального набору — достатньо лише спостережень нормальної поведінки — що відповідає сценарію напівкерованого навчання. Це робить її практичним інструментом у задачах, де аномальні приклади недоступні.

Водночас модель має суттєві обмеження. По-перше, ARIMA здатна моделювати лише лінійні залежності, що може бути недостатнім для складних нелінійних процесів. По-друге, якість прогнозів залежить від коректно підібраних параметрів (p, d, q): невдалий вибір може призвести або до пропущених аномалій — якщо модель занадто гнучко підлаштовується під дані, — або до надмірної кількості хибних спрацювань.

3.2 Метод визначення аномалій з використанням машинного навчання. XGBoost

Extreme Gradient Boosting (XGBoost) — це алгоритм машинного навчання, що належить до класу методів ансамблювання на основі дерев рішень. На відміну від статистичних моделей, які спираються на припущення про розподіл даних, XGBoost будує прогноз шляхом послідовного додавання дерев, кожне з яких намагається виправити помилки попереднього. Такий підхід дозволяє моделювати складні нелінійні залежності без жодних припущень про форму розподілу вхідних даних.

В основі алгоритму лежить ідея градієнтного бустингу: замість того щоб навчати одну потужну модель, будується ансамбль із K послідовних дерев, де кожне нове дерево наближається до залишків попереднього прогнозу.[9] Фінальний прогноз для спостереження x_i є сумою передбачень усіх дерев:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{13}$$

де $\mathcal{F}$ — простір регресійних дерев.

Навчання моделі полягає у мінімізації цільової функції, яка поєднує фун-

кцію втрат та член регуляризації:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{14}$$

де $l(\hat{y}_i, y_i)$ — похибка між прогнозом та реальним значенням, а $\Omega(f_k)$ — штраф за складність кожного дерева, що запобігає перенавчанню.

Складність дерева визначається через кількість його листків T та їхні ваги w:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2. \tag{15}$$

Параметри γ та λ контролюють, наскільки штрафується кожне нове розгалуження дерева — що більші їхні значення, то простіша і узагальненіша модель.

Оскільки цю цільову функцію неможливо мінімізувати безпосередньо, на кожному кроці навчання використовується наближення Тейлора другого порядку, що дозволяє ефективно оптимізувати модель і враховувати як напрям, так і кривизну поверхні похибок:

$$\mathcal{L}^{(t)} \approx \sum_i \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{16}$$

де $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ — похідна першого порядку функції втрат, а $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ — похідна другого порядку. Саме використання похідної другого порядку відрізняє XGBoost від звичайного градієнтного бустингу і забезпечує точнішу оптимізацію.

Для застосування XGBoost до одновимірного часового ряду $\{X_t\}$ необхідно перетворити послідовність у набір ознак. Це реалізується за допомогою методу ковзного вікна: для кожного моменту часу t формується вектор із w попередніх значень, який використовується як вхід, а X_t виступає цільовою змінною. Таким чином задача прогнозування часового ряду перетворюється на звичайну задачу регресії, яку XGBoost здатен розв'язати безпосередньо.[10]

Логіка виявлення аномалій аналогічна до описаної для ARIMA: модель навчається на нормальних даних, прогнозує очікувані значення, а спостереження, для яких відхилення між прогнозом та реальним значенням перевищує заданий поріг, позначаються як потенційні аномалії.

Попри практичну ефективність, метод має як переваги, так і обмеження. До переваг належать висока точність на складних нелінійних даних, вбудований механізм регуляризації, що знижує ризик перенавчання, а також здатність автоматично обробляти пропущені значення. Крім того, XGBoost підтримує паралельне обчислення при побудові дерев, що суттєво прискорює тренування порівняно зі звичайним градієнтним бустингом.

Водночас для роботи з часовими рядами алгоритм потребує попередньої трансформації, що вимагає вибору розміру вікна. Неправильно обраний розмір може призвести до втрати важливих часових залежностей або, навпаки, до включення нерелевантної інформації.

3.3 Метод визначення аномалій з використанням нейронних мереж. LSTM

Після значних успіхів нейронних мереж у задачах комп'ютерного зору та обробки природної мови інтерес до їх застосування в аналізі часових рядів суттєво зріс. На відміну від статистичних методів, нейронні мережі не потребують явних припущень щодо форми розподілу даних — натомість вони навчаються знаходити складні нелінійні залежності безпосередньо з прикладів. Однією з найпоширеніших архітектур для роботи з послідовностями є мережа довгої короткочасної пам'яті — Long Short-Term Memory (LSTM).[11]

LSTM належить до класу рекурентних нейронних мереж - RNN. Ідея рекурентних мереж полягає в тому, що на кожному кроці обробки послідовності мережа отримує не лише поточне значення, а й інформацію про попередній стан. Формально вихід звичайного нейрона записується як:

$$y_t = \phi(x_t^T w_x + b), \tag{17}$$

де x_t — вхідний вектор, w_x — ваговий вектор, b — зміщення, $\phi(\cdot)$ — нелінійна функція активації.

У рекурентній мережі до цієї формули додається залежність від попереднього стану:

$$y_t = \phi(x_t^T w_x + y_{t-1}^T w_y + b). \tag{18}$$

Проте звичайні RNN погано навчаються на довгих послідовностях через

проблему зникнення градієнта: при зворотному поширенні помилки через багато кроків градієнт стає настільки малим, що ранні залежності фактично перестають враховуватися. Саме для розв'язання цієї проблеми була запропонована архітектура LSTM.

Ключова відмінність LSTM від звичайної RNN полягає в наявності спеціального механізму управління інформацією — системи воріт. Кожна LSTM-комірка підтримує два вектори стану: короткострокову пам'ять h_t (прихований стан) та довгострокову пам'ять c_t (стан комірки). Потік інформації між ними регулюється трьома типами воріт:

- ворота забуття f_t — визначають, яка частина попереднього стану довгострокової пам'яті буде збережена:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f); \tag{19}$$

- вхідні ворота i_t — визначають, яка нова інформація буде записана до комірки; паралельно формується кандидат нового стану пам'яті $\tilde{c}_t$:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{20}$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c); \tag{21}$$

- вихідні ворота o_t — визначають, яка частина оновленого стану комірки передається на вихід:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o). \tag{22}$$

Довгострокова пам'ять оновлюється на основі зваженого поєднання поточного стану та нової інформації від вхідних воріт:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{23}$$

де $\odot$ — покомпонентне множення.

Вихідний прихований стан h_t, який передається на наступний крок і використовується як результат роботи комірки, обчислюється фільтруванням оновленої довгострокової пам'яті через вихідні ворота:

$$h_t = o_t \odot \tanh(c_t). \tag{24}$$

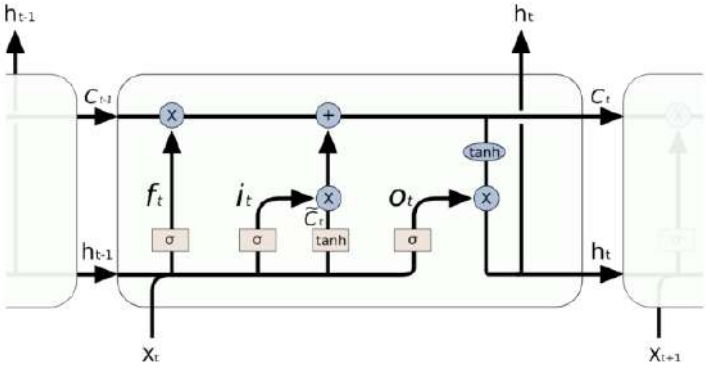

Рис. 9: Структура LSTM-комірки[11]

Завдяки такій структурі LSTM здатна зберігати важливу інформацію протягом довгих часових інтервалів і водночас відфільтровувати короткочасний шум.

У задачах виявлення аномалій LSTM може застосовуватися двома основними способами: як модель прогнозування або як частина автокодувальника.[11] У цій роботі будемо розглядати підхід на основі прогнозування.

Логіка виявлення аномалій аналогічна до описаної для попередніх методів: модель навчається на нормальних даних, після чого для кожного нового спостереження обчислюється відхилення між прогнозом та реальним значенням. Якщо похибка перевищує заданий поріг — спостереження позначається як аномалія.

Серед переваг LSTM варто відзначити здатність моделювати довгострокові нелінійні залежності у часових рядах — те, з чим статистичні моделі та класичні алгоритми машинного навчання справляються гірше. Модель не потребує явних припущень про структуру даних і може ефективно застосовуватися без розмічених аномальних прикладів. Крім того, LSTM природно враховує часову впорядкованість спостережень, що є важливою властивістю при роботі з послідовними даними.

Водночас LSTM потребує значно більших обчислювальних ресурсів порівняно зі статистичними методами та класичними алгоритмами машинного навчання: навчання займає більше часу, а велика кількість параметрів ускладнює підбір гіперпараметрів. Через послідовний характер обробки даних модель погано піддається паралелізації, що може бути критичним для систем реального часу. Нарешті, внутрішні механізми LSTM важко піддаю-

ться інтерпретації — зрозуміти, чому модель позначила певне спостереження як аномальне, значно складніше, ніж у статистичних підходах.

4 Практичне дослідження і порівняння методів

4.1 Метрики та критерії оцінювання

Для порівняння методів виявлення аномалій було обрано три показники: F2-score, ROC-AUC та час обчислення. Розглянемо кожен із них детальніше.

F2-score належить до сімейства F_β-метрик і будується на основі двох базових показників бінарної класифікації — точності та повноти.

Точність (P, Precision) показує, яка частка спостережень, позначених моделлю як аномальні, є дійсно аномальними:

$$P = \frac{TP}{TP + FP}, \tag{25}$$

де TP — кількість правильно виявлених аномалій, FP — кількість нормальних спостережень, помилково позначених як аномальні.

Повнота (R, Recall) показує, яка частка реальних аномалій була виявлена моделлю:

$$R = \frac{TP}{TP + FN}, \tag{26}$$

де FN — кількість пропущених аномалій.

Ці два показники є взаємозалежними: підвищення точності, як правило, веде до зниження повноти і навпаки. F2-score дозволяє збалансувати їх з акцентом на повноту[12]:

$$F_2 = \frac{5 \cdot P \cdot R}{4 \cdot P + R}. \tag{27}$$

У задачі виявлення аномалій пропуск реальної аномалії зазвичай є більш критичним, ніж хибне спрацювання — саме тому повнота важливіша за точність. F2-score враховує це, надаючи повноті вчетверо більшу вагу порівняно з точністю. Значення метрики знаходиться в межах [0, 1], де 1 відповідає ідеальному виявленню всіх аномалій.

Проте F2-score має суттєве обмеження: її значення залежить від конкретного порогу, що використовується для бінаризації прогнозів. Різні пороги дають різні результати, тому F2-score характеризує якість детекції лише за умови вдало підібраного порогу.

ROC-AUC використовується як доповнення до F2-score саме для усунення цього обмеження. Ця метрика оцінює здатність моделі правильно ранжу-

вати спостереження за ступенем аномальності незалежно від обраного порогу.

ROC-крива будується шляхом варіювання порогового значення і відображає співвідношення між часткою істинно позитивних (TPR) та хибно позитивних результатів (FPR):

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}. \tag{28}$$

ROC-AUC є площею під цією кривою:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(t)\big)\, dt. \tag{29}$$

Інтерпретація ROC-AUC є інтуїтивною: це ймовірність того, що випадково обрана аномальна точка отримає вищий показник аномальності, ніж випадково обрана нормальна. Значення 0.5 відповідає випадковому класифікатору, 1.0 — ідеальному ранжуванню.

Таким чином, F2-score і ROC-AUC доповнюють одна одну: перша оцінює якість детекції за конкретного порогу, друга — загальну розділювальну здатність моделі незалежно від нього.

Час обчислення вимірює сумарний час від початку навчання моделі до отримання прогнозів на тестовій вибірці в секундах. Цей показник відображає обчислювальну складність методу і є важливим критерієм при оцінці його практичної придатності — особливо в умовах, де швидкість реагування є критичною.

4.2 Вибір експериментальних даних, їх налаштування та опис процедури порівняння методів

Для експериментального дослідження було обрано три набори даних з бенчмарку Numenta Anomaly Benchmark [13] — стандартного зібрання реальних часових рядів з вручну розміченими аномаліями, що широко використовується для порівняння методів детектування відхилень. Датасети охоплюють два різних домени — мережеву інфраструктуру та дорожній трафік, — завдяки чому можемо оцінити поведінку методів на даних різної природи. Додатковим критерієм відбору була структура кожного ряду: перші 40% спостережень у кожному датасеті не містять розмічених аномалій, тому їх можна

безпосередньо використати як тренувальну вибірку без додаткової фільтрації.

- *EC2 Latency* — часовий ряд латентності запитів до сервера AWS EC2. Виміри здійснюються кожні 5 хвилин; аномалії відповідають системному збою сервера, при якому латентність різко зростає.

- *Occupancy T4013* —дані про рівень завантаженості дороги з датчика T4013, розташованого у зоні Twin Cities Metro Area, штат Міннесота. Показник відображає частку часу, протягом якої датчик фіксує присутність транспортного засобу над ним, і вимірюється з інтервалом 5 хвилин. Аномалії у цьому ряді відповідають нетиповим відхиленням рівня завантаженості — наприклад, унаслідок дорожньо-транспортних пригод або дорожніх робіт.

- *Speed T4013* — дані про середню швидкість транспортного потоку з того самого датчика T4013 з тим самим інтервалом вимірювань. Аномалії відповідають різким змінам швидкості внаслідок дорожніх інцидентів.

Дані розбивались у пропорції 40% на тренувальну частину та 60% на тестову. Тренувальна частина береться з початку ряду — саме там аномалії відсутні, — тому моделі навчаються виключно на нормальній поведінці системи.

Перед застосуванням методів усі дані пройшли стандартизацію за допомогою StandardScaler:

$$x_{\text{std}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{30}$$

де μ_{train} та σ_{train} — середнє та стандартне відхилення, обчислені на тренувальній частині. Це забезпечує методологічну однорідність експерименту та покращує збіжність нейронної мережі при навчанні. Параметри обчислюються лише на тренувальній вибірці, що запобігає витоку інформації з тестової частини.

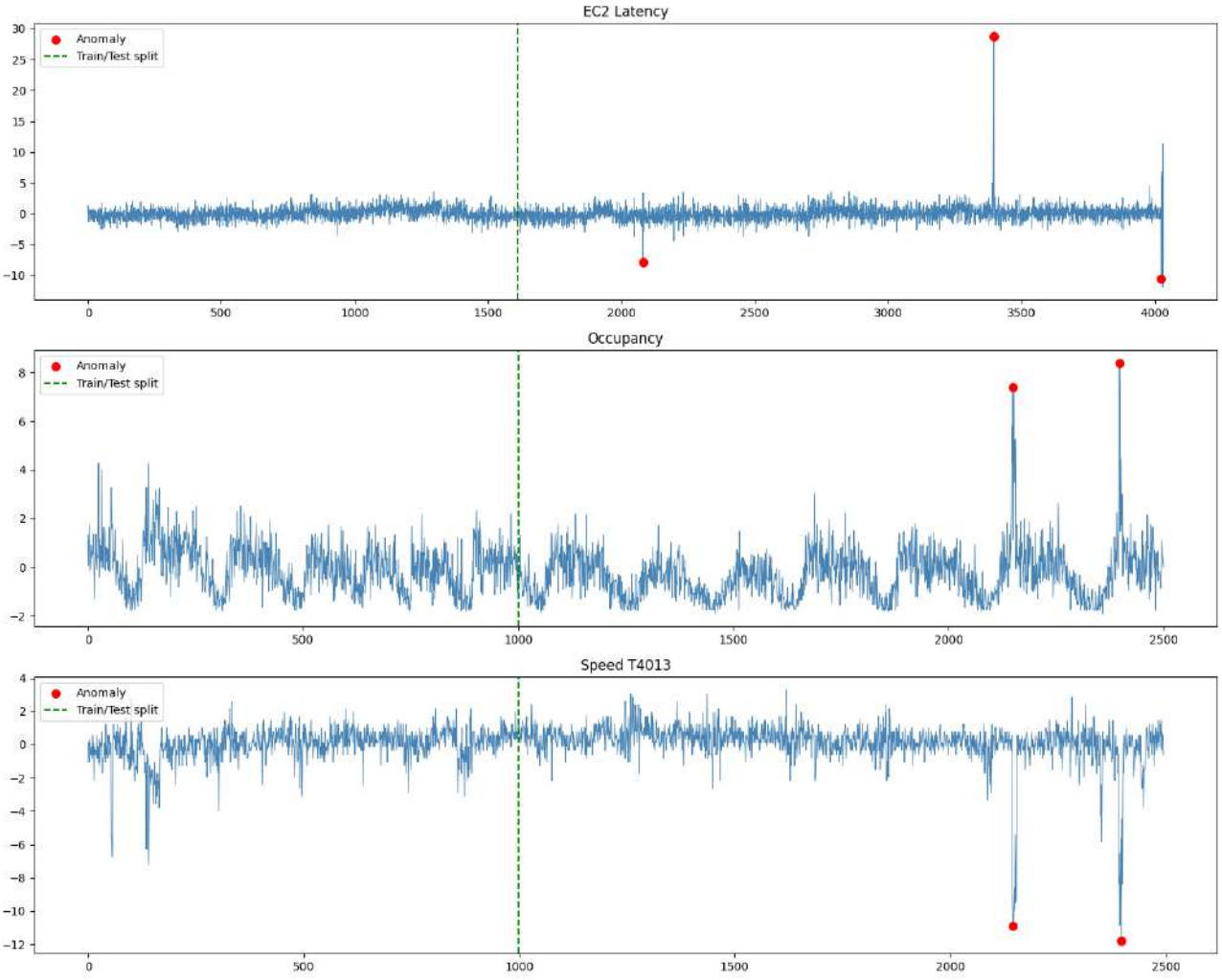

Рис. 10: Візуалізація трьох досліджуваних часових рядів з позначеними аномаліями

Оскільки всі три методи належать до прогнозуючих підходів, для них застосовувалась єдина процедура детекції аномалій:

1. Навчання моделі на тренувальній частині без аномалій.
2. Формування прогнозу $\hat{x}_t$ для кожної точки тестової вибірки.
3. Обчислення абсолютної похибки прогнозу: $e_t = |x_t - \hat{x}_t|$.
4. Визначення порогу за правилом $k\sigma$: $T = \mu_e + k \cdot \sigma_e$, де k підбирається індивідуально для кожного датасету.[12]
5. Бінаризація: точка t позначається як аномалія, якщо $e_t > T$.

Усі методи отримують однакові вхідні дані, використовують однаковий алгоритм формування показників аномальності та однаковий метод визначення порогу. Час обчислення фіксується як сумарний час від початку навчання моделі до отримання прогнозів на тестовій вибірці. Після завершення процедури обчислюються метрики F2-score та ROC-AUC.

4.3 Детекція аномалій та порівняння результатів методів

Параметри моделі ARIMA підбирались автоматично за критерієм AIC шляхом перебору комбінацій (p, d, q). Для EC2 Latency підібрано порядок (6, 1, 3), для Occupancy — (5, 0, 4), для Speed T4013 — (3, 1, 4). Ненульове значення d для перших двох рядів вказує на їх нестаціонарність, тоді як Occupancy диференціювання не потребував.

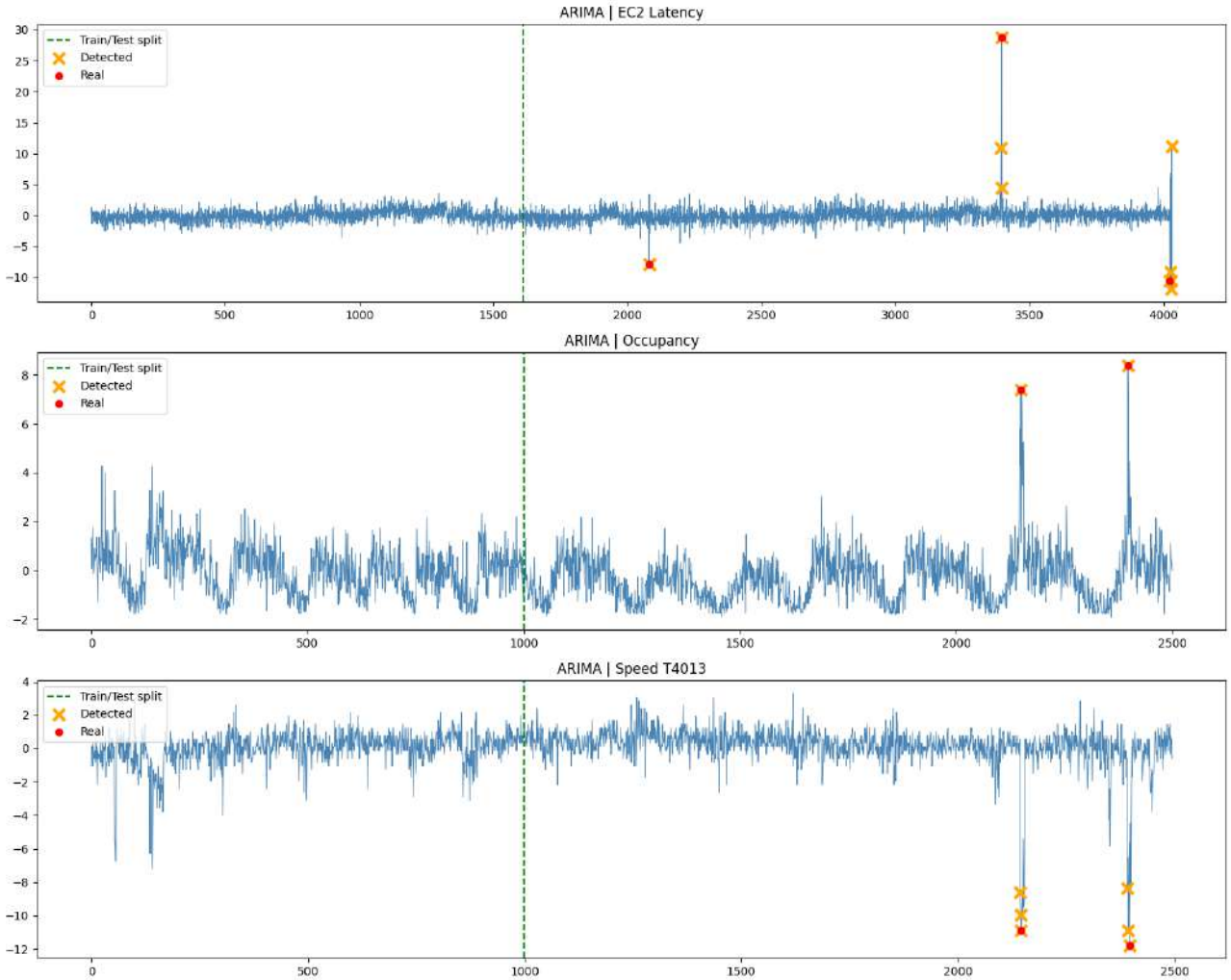

Рис. 11: Результати детектування методом ARIMA для трьох датасетів

Для моделі XGBoost ключовим параметром є розмір ковзного вікна,що визначає кількість попередніх значень, які використовуються як ознаки для прогнозу. Вікно підбиралось з діапазону $[10, 100]$ за критерієм мінімальної MSE на валідаційній частині тренувальних даних: для EC2 Latency отримано значення 50, для Occupancy — 30, для Speed T4013 — 60. Інші параметри моделі: `n_estimators=1000`, `max_depth=3`, `learning_rate=0.05`, раннє зупинення при відсутності покращення протягом 20 ітерацій.

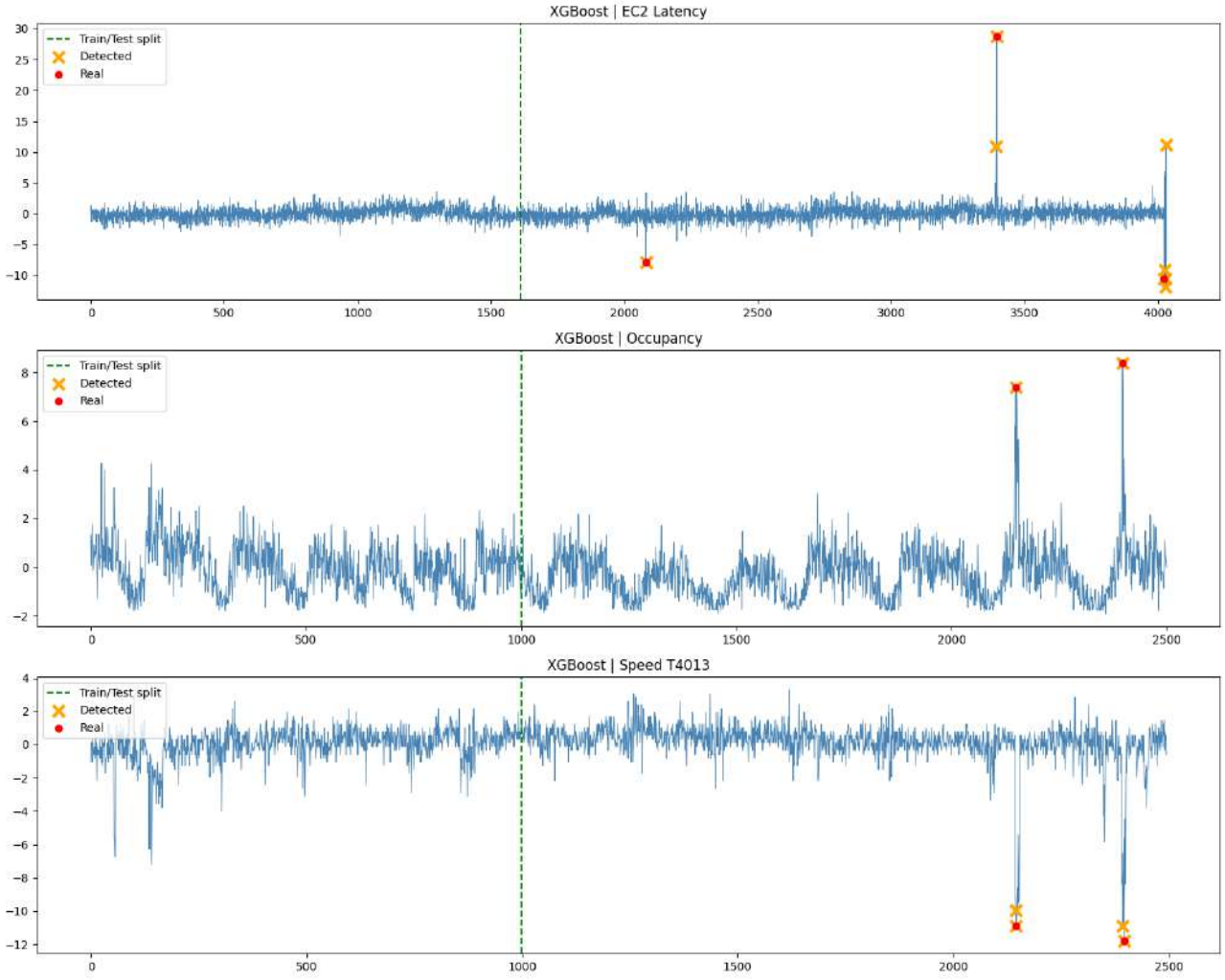

Рис. 12: Результати детектування методом XGBoost для трьох датасетів

Для LSTM використовувалась фіксована архітектура: два рекурентних шари LSTM(64) та LSTM(32), за якими йдуть Dense(16) з активацією ReLU та вихідний Dense(1). Модель навчалась з оптимізатором Adam, функцією втрат MSE, розміром батчу 16 та раннім зупиненням з `patience=10` при максимальній кількості епох 100.

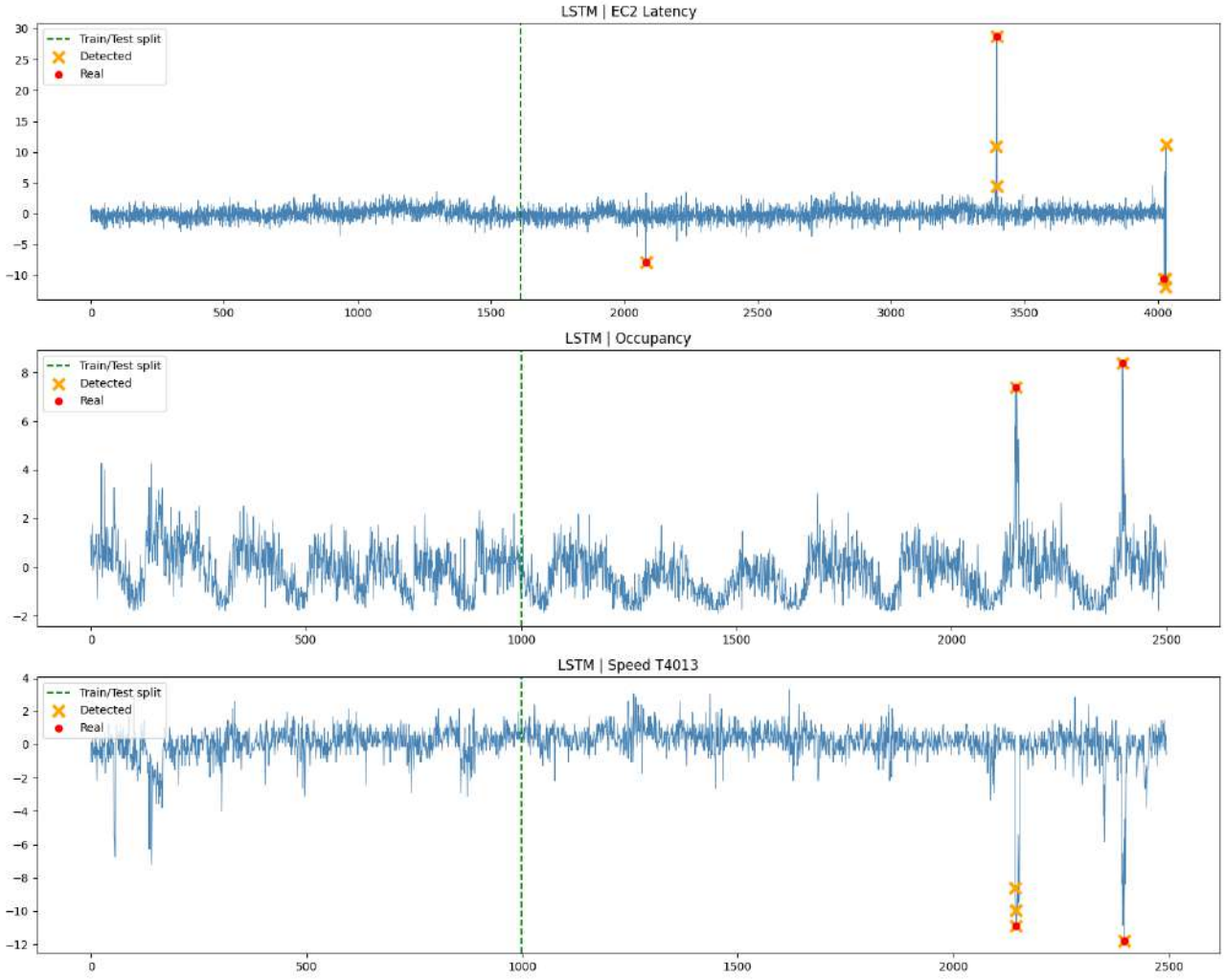

Рис. 13: Результати детектування методом LSTM для трьох датасетів

Порівняння результатів за метриками

Табл. 1: Значення F2-score для досліджуваних методів

Датасет	ARIMA	XGBoost	LSTM
EC2 Latency	0.7143	0.7500	0.7500
Occupancy	1.0000	1.0000	1.0000
Speed T4013	0.7143	0.8333	0.7692

Табл. 2: Значення ROC-AUC для досліджуваних методів

Датасет	ARIMA	XGBoost	LSTM
EC2 Latency	0.9988	0.9990	0.9990
Occupancy	1.0000	1.0000	1.0000
Speed T4013	0.9973	0.9993	0.9997

Табл. 3: Час обчислення в секундах для досліджуваних методів

Датасет	ARIMA	XGBoost	LSTM
EC2 Latency	3.75	0.34	52.42
Occupancy	3.72	0.13	24.07
Speed T4013	4.91	0.07	24.32

Повнота у всіх трьох методів на кожному датасеті становить 1.0 — жодна розмічена аномалія не була пропущена. Відмінності у F2-score зумовлені виключно різною кількістю хибних спрацювань.

На датасеті Occupancy всі методи досягають F2 = 1.0. На EC2 Latency XGBoost та LSTM отримують 0.750, тоді як ARIMA — 0.714. Найбільша різниця спостерігається на Speed T4013: XGBoost — 0.833, LSTM — 0.769, ARIMA — 0.714.

Значення ROC-AUC для всіх методів перевищує 0.997 на кожному датасеті, що свідчить про те, що похибки прогнозу на аномальних точках стабільно перевищують похибки на нормальних — усі три моделі формують якісні оцінки аномальності незалежно від обраного порогу.

Час обчислення суттєво відрізняється між методами. XGBoost є найшвидшим — менше секунди на кожному датасеті. ARIMA потребує від 3.7 до 4.9 секунди. LSTM витрачає найбільше часу: від 24 до 52 секунд.

Висновки

У роботі досліджено методи виявлення аномалій в одновимірних часових рядах: розглянуто теоретичну базу, описано обрані алгоритми та проведено їх практичне порівняння.

У теоретичній частині систематизовано основні поняття аналізу часових рядів та розглянуто класифікацію методів виявлення аномалій за наявністю розмітки і математичною методологією. Для кожного з трьох обраних класів: статистичних методів, машинного навчання та нейронних мереж, описано по одному представнику та проаналізовано його переваги й обмеження.

У практичній частині методи ARIMA, XGBoost та LSTM реалізовано на реальних часових рядах з відкритого бенчмарку та протестовано в єдиних умовах експерименту. За результатами експерименту можна зробити такі висновки:

1. Усі три методи забезпечують високу якість виявлення аномалій: статистичний метод ARIMA за показниками детектування не поступається складнішим алгоритмам машинного навчання та нейронним мережам.

2. XGBoost показав найкраще поєднання точності та швидкості: при схожих показниках детектування він витрачав на обчислення в рази менше часу, ніж ARIMA та LSTM.

3. Результати також показують, що складніша модель не гарантує кращого виявлення аномалій — принаймні для одновимірних рядів з точковими відхиленнями ARIMA показала себе не гірше за LSTM, залишаючись при цьому значно простішою у застосуванні та інтерпретації.

Загалом вибір методу варто розглядати не лише через призму метрик точності, а й з урахуванням обчислювальних витрат та інтерпретованості моделі, що особливо важливо в прикладних задачах.

Подальше дослідження може бути спрямоване на перевірку поведінки методів на рядах з контекстуальними та колективними аномаліями. Окремий інтерес становить розширення аналізу на багатовимірні часові ряди та систематизація підходів до виявлення аномалій у цьому випадку — за наявності

взаємозв'язків між ознаками складніші архітектури можуть демонструвати помітніші переваги.

Список літератури

[1] Shumway R.H., Stoffer D.S. *Time Series Analysis and Its Applications: With R Examples.* — 3rd ed. — Springer, 2011.

[2] Furizal, Ma'arif A., Kariyamin K., Firdaus A.A., Wijaya S.A., Nakib A.M., Ningrum A.F. *Understanding Time Series Forecasting: A Fundamental Study* // Buletin Ilmiah Sarjana Teknik Elektro. — 2025. — Vol. 7, No. 3. — P. 554–571.

[3] Shaukat K., Alam T.M., Luo S. et al. *A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives* // Advances in Intelligent Systems and Computing. — Springer, 2021.

[4] Chandola V., Banerjee A., Kumar V. *Anomaly Detection: A Survey* // ACM Computing Surveys. — 2009. — Vol. 41, No. 3. — P. 1–58.

[5] Google Trends. Динаміка пошукового інтересу до запиту «iPhone» (2011–2026). — Режим доступу: `https://trends.google.com/explore?q=%2Fm%2F027lnzs&date=2011-05-14%202026-05-14&geo=UA`

[6] Yahoo Finance. Apple Inc. (2021–2026). — Режим доступу: `https://finance.yahoo.com/quote/AAPL`

[7] Schmidl S., Wenig P., Papenbrock T. *Anomaly Detection in Time Series: A Comprehensive Evaluation* // Proc. VLDB Endow. — 2022. — Vol. 15, No. 9. — P. 1779–1797.

[8] Moschini G., Houssou R., Bovay J., Robert-Nicoud S. *Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model* // Engineering Proceedings. — 2021. — Vol. 5, No. 1. — P. 56.

[9] Zhang L., Bian W., Qu W., Tuo L., Wang Y. *Time Series Forecast of Sales Volume Based on XGBoost* // Journal of Physics: Conference Series. — 2021. — Vol. 1873. — Art. 012067.

[10] Zhao Q., Wen X., Huang B., Wang J., Fang J. *Optimised Extreme Gradient Boosting Model for Short Term Electric Load Demand Forecasting of Regional Grid System* // Scientific Reports. — 2022. — Vol. 12. — Art. 19282.

[11] Bowie M., Begoli E., Park B., Bopaiah J. *Towards an LSTM-based Approach for Detection of Temporally Anomalous Data in Medical Datasets* // Proceedings of the ICIQ. — 2017.

[12] Komadina A., Martinić M., Groš S., Mihajlović Ž. *Comparing Threshold Selection Methods for Network Anomaly Detection* // IEEE Access. — 2024. — Vol. 12. — P. 124944–124972.

[13] Ahmad S., Lavin A., Purdy S., Agha Z. Numenta Anomaly Benchmark. — Режим доступу: `https://github.com/numenta/NAB`

Додаток А. Реалізація методів

Код реалізації доступний за посиланням: https://colab.research.google.com/drive/1rOid541gQGlFXw3X2nCoPnTldkByKsPR?usp=sharing

У процесі підготовки кваліфікаційної роботи було використано Claude (Sonet 4.6) для стилістичного та граматичного редагування тексту. Усі запропоновані формулювання були критично проаналізовані, перевірені та за потреби відредаговані автором. Остаточний зміст роботи, інтерпретація результатів і висновки є авторськими.