Міністерство освіти і науки України
Університет митної справи та фінансів

Факультет інноваційних технологій
Кафедра комп'ютерних наук та інженерії програмного забезпечення

Кваліфікаційна робота магістра

на тему: «Порівняльний аналіз методів виявлення аномалій у наборах даних»

Виконав: студент групи К23-2М

Спеціальність 122 Комп'ютерні науки

Левченко Д.О.
(прізвище та ініціали)

Керівник к.т.н., доц. Мала Ю. А.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент Дніпровський державний технічний університет
(місце роботи)

доцент кафедри математичного моделювання та системного аналізу
(посада)

к.т.н., доц. Волосова Н.М.
(науковий ступінь, вчене звання, прізвище та ініціали)

Дніпро – 2025

АНОТАЦІЯ

Левченко Д.О. Порівняльний аналіз методів виявлення аномалій у наборах даних.

Дипломна робота на здобуття освітнього ступеня магістр за спеціальністю 122 «Комп'ютерні науки» – Університет митної справи та фінансів, Дніпро, 2025.

Магістерська робота присвячена дослідженню методів виявлення аномалій у великих наборах даних. У роботі виконано детальний огляд існуючих підходів до виявлення аномалій, включаючи класичні статистичні методи, алгоритми машинного навчання, методи глибинного навчання та гібридні підходи. Проведено аналіз їхніх переваг і обмежень залежно від характеристик наборів даних та специфіки поставлених завдань. Досліджено основні типи аномалій, такі як локальні та глобальні, а також одновимірні та багатовимірні, і розглянуто специфіку їхнього виявлення у часових рядах, багатовимірних наборах і потокових даних.

Практичну частину роботи присвячено реалізації кількох алгоритмів виявлення аномалій та порівнянню їхньої ефективності на реальних наборах даних. Особлива увага приділена проблемам масштабованості методів для великих обсягів інформації, впливу шуму та неповних даних, а також критеріям оцінки якості результатів. Запропоновано рекомендації щодо вибору методів для вирішення конкретних задач у різних галузях, враховуючи їхню адаптивність і швидкість обробки.

Наукова новизна роботи полягає у проведенні комплексного порівняння сучасних підходів до виявлення аномалій та розробці рекомендацій щодо інтеграції різних методів для підвищення ефективності аналізу.

Ключові слова: виявлення аномалій, машинне навчання, глибинне навчання, статистичні методи, гібридні моделі, часові ряди, великі дані, кластеризація.

ABSTRACT

Levchenko D.O. Comparative analysis of methods for detecting anomalies in data sets.

Diploma thesis for obtaining a master's degree in specialty 122 «Computer Science» – University of Customs and Finance, Dnipro, 2025.

The master's thesis is devoted to the study of methods for detecting anomalies in large data sets. The work provides a detailed review of existing approaches to anomaly detection, including classical statistical methods, machine learning algorithms, deep learning methods, and hybrid approaches. Their advantages and limitations are analyzed depending on the characteristics of the data sets and the specifics of the tasks. The main types of anomalies, such as local and global, as well as one-dimensional and multidimensional, are investigated, and the specifics of their detection in time series, multidimensional sets, and streaming data are considered.

The practical part of the paper is devoted to the implementation of several anomaly detection algorithms and the comparison of their effectiveness on real data sets. Particular attention is paid to the problems of scalability of methods for large amounts of information, the impact of noise and incomplete data, as well as criteria for assessing the quality of results. Recommendations for choosing methods for solving specific problems in various fields, taking into account their adaptability and processing speed, are proposed.

The scientific novelty of the work is to conduct a comprehensive comparison of modern approaches to anomaly detection and to develop recommendations for the integration of different methods to improve the efficiency of analysis.

Keywords: anomaly detection, machine learning, deep learning, statistical methods, hybrid models, time series, big data, clustering.

ЗМІСТ

ВСТУП

В умовах сучасного інформаційного суспільства, де дані мають центральне значення в прийнятті рішень, важливим завданням стає забезпечення їх якості та надійності. Зокрема, зростання обсягів даних, які генеруються в різних сферах життєдіяльності, таких як фінанси, охорона здоров'я, енергетика, транспорт та багато інших, вимагає ефективних підходів до виявлення аномалій у таких великих наборах даних. Аномалії можуть бути результатом помилок у зборі даних, несправностей систем, або навіть проявом нових, невідомих явищ, що потребують глибокого дослідження. Тому розробка та вдосконалення методів виявлення аномалій стає надзвичайно актуальним завданням для багатьох галузей науки та техніки.

Аномалії у даних можуть мати різну природу, від простих помилок до складних відхилень, які вказують на важливі нові тренди або порушення в системах. Тому правильне виявлення аномалій є не лише технічною задачею, а й важливим етапом для забезпечення безпеки, оптимізації процесів та виявлення нових закономірностей. Це дозволяє своєчасно виявляти проблеми та реагувати на них до того, як вони можуть призвести до серйозних наслідків. Однак, із зростанням обсягів даних та їх різноманітністю, методи виявлення аномалій стикаються з новими викликами, такими як необхідність обробки великих і складних наборів даних, підвищені вимоги до точності та адаптивності методів.

Актуальність теми полягає в тому, що з кожним роком зростає обсяг даних, що збираються в різних сферах діяльності, і виникає потреба у високоефективних методах для їх обробки. Невірно інтерпретовані або не виявлені аномалії можуть призвести до серйозних негативних наслідків, таких як фінансові втрати, виявлення дефектів у медичних діагнозах, порушення безпеки в технологічних процесах тощо. Сучасні методи виявлення аномалій повинні враховувати високі вимоги до швидкості обробки даних, їх

багатовимірності, а також враховувати різноманітність джерел та типів інформації.

Метою цієї роботи є дослідження та порівняння різних методів виявлення аномалій у великих наборах даних, а також розробка рекомендацій щодо їх застосування в реальних умовах. В рамках цієї мети важливо детально проаналізувати існуючі підходи та методи, вивчити їх переваги та обмеження, а також визначити найефективніші методи для вирішення конкретних завдань в різних областях застосування.

Завдання роботи, що випливають із зазначеної мети, включають:

– огляд існуючих методів виявлення аномалій, їх класифікація та порівняння;

– аналіз характеристик великих наборів даних та їх вплив на вибір методів виявлення аномалій;

– розробка критеріїв для оцінки ефективності методів виявлення аномалій у залежності від типу даних та специфіки задачі;

– проведення експериментального дослідження для порівняння результатів роботи різних методів виявлення аномалій на реальних наборах даних;

– визначення практичних рекомендацій для використання найбільш ефективних методів виявлення аномалій в конкретних прикладних задачах.

Об’єктом дослідження є великі набори даних, що містять як звичайні значення, так і аномалії. Це можуть бути дані з різних сфер, такі як фінансові транзакції, медичні записи, показники роботи промислових систем, дані з сенсорів у розумних містах і багато інших.

Предметом дослідження є методи та підходи до виявлення аномалій у великих наборах даних, зокрема, методи машинного навчання, статистичні підходи, алгоритми глибинного навчання та комбінації різних методів для досягнення кращих результатів.

Методи дослідження включають огляд літератури для аналізу існуючих підходів, математичне моделювання для оцінки точності та ефективності методів, а також практичне застосування алгоритмів виявлення аномалій на реальних наборах даних. У процесі дослідження буде використано як теоретичні методи, так і емпіричні – на основі результатів експериментів з різними наборами даних.

Практична значимість роботи полягає в тому, що результати дослідження можуть бути використані для розробки інструментів та систем, здатних автоматично виявляти аномалії в реальних умовах. Це дозволяє значно покращити процеси прийняття рішень у різних галузях, таких як фінансовий моніторинг, медичні діагнози, системи безпеки та інші. Крім того, виявлення аномалій допомагає оптимізувати роботу систем і запобігати можливим неполадкам та витратам.

Наукова новизна роботи полягає в комплексному порівнянні новітніх методів виявлення аномалій з використанням машинного навчання та глибинного навчання на реальних великих наборах даних. Вперше пропонується підхід, який об’єднує різні типи методів виявлення аномалій для підвищення ефективності в специфічних умовах, зокрема, у випадках, коли дані мають високі вимоги до часу обробки і точності. Розроблені рекомендації можуть бути корисними для науковців та фахівців, які працюють у галузях, де обробка великих даних є важливим етапом прийняття рішень.

Структура кваліфікаційної роботи. Кваліфікаційна робота складається з трьох розділів. Обсяг кваліфікаційної роботи – 92 сторінки. Робота містить 12 рисунків та 1 таблицю. Список використаних джерел складає 16 посилань.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Класичні статистичні методи виявлення аномалій у даних

Класичні статистичні методи виявлення аномалій (рис. 1.1) у даних є важливим інструментом для аналізу великих обсягів інформації в різних галузях науки, техніки та бізнесу. Виявлення аномалій або відхилень від нормальних або очікуваних патернів є важливою задачею, оскільки ці аномалії можуть свідчити про наявність помилок у вимірюваннях, збій у системах, шахрайство, несправності обладнання або інші суттєві явища, що потребують уваги [1]. Статистичні методи виявлення аномалій базуються на теорії ймовірності, математичній статистиці, а також на припущеннях щодо розподілу даних. Вони дозволяють визначити, які спостереження суттєво відрізняються від очікуваних значень, що робить їх особливо корисними в контексті обробки великих наборів даних, де з'ясування аномалій є першочерговим завданням для забезпечення достовірності аналізу.

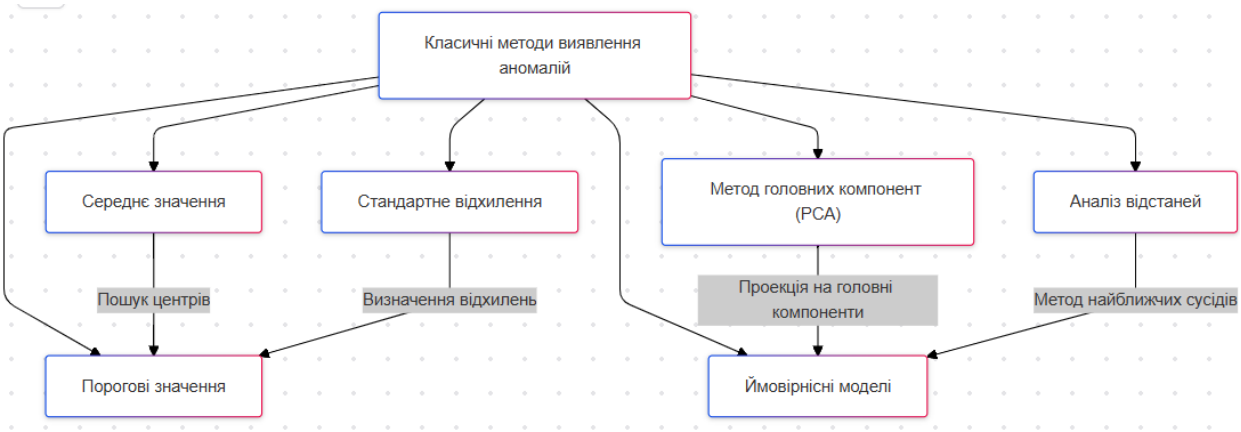

Рисунок 1.1 – Класичні методи виявлення аномалій

Одним з основних класичних підходів до виявлення аномалій є використання методів, що ґрунтуються на описових статистиках, таких як

середнє значення, дисперсія, медіана та інші. Ці методи працюють за принципом порівняння спостережуваних значень з певною центральною тенденцією, яку можна оцінити за допомогою цих статистик. Наприклад, для виявлення аномальних спостережень в однотипних даних часто використовуються методи на основі середнього значення та стандартного відхилення [1, 2]. Середнє значення дає уявлення про центр розподілу даних, а стандартне відхилення характеризує ступінь їх варіативності. Відхилення, що знаходяться на значній відстані від середнього значення, можуть бути розглянуті як потенційні аномалії. Однак, цей підхід має низку обмежень, оскільки він не враховує специфіку розподілу даних, а також не може ефективно працювати в умовах сильно несиметричних або багатовершинних розподілів.

Іншим важливим методом є використання порогових значень на основі статистичних характеристик. Наприклад, для визначення аномальних значень можна використовувати концепцію «виходів за межі» або «квантилів» розподілу. В цьому випадку спостереження, що знаходяться за межами визначеного діапазону (наприклад, понад третій чи п’ятий персентиль), можуть бути класифіковані як аномалії. Це дозволяє виділити дані, які знаходяться в крайніх значеннях розподілу, але цей метод також має свої обмеження, оскільки він не завжди дає точне уявлення про причини аномалій, особливо коли розподіл даних має складну форму або змінюється з часом [2].

У випадках, коли розподіл даних не є нормальним або відоме, що дані можуть мати складнішу структуру, використовуються більш складні статистичні методи, такі як методи головних компонент (PCA). Цей метод дозволяє здійснити зменшення розмірності даних, що особливо корисно при роботі з багатовимірними наборами даних, де просте порівняння значень може бути неефективним. Метод головних компонент дозволяє виокремити найбільш значущі напрямки варіації даних і здійснити оцінку аномалій через проекцію спостережень на ці компоненти. Якщо спостереження відхиляється

від основних компонент на значну відстань, воно може бути розглянуте як аномальне. Однак, як і в попередніх методах, важливо мати достатньо точні уявлення про структуру даних, оскільки неправильно вибрані компоненти можуть призвести до хибних висновків [3].

Також важливим підходом до виявлення аномалій є використання методів, що базуються на аналізі кореляцій між різними змінними. Класичні методи кореляційного аналізу, такі як коефіцієнт кореляції Пірсона, дозволяють виявити спостереження, що суттєво відрізняються від очікуваної залежності між змінними. Наприклад, якщо існує висока кореляція між двома змінними, але одне спостереження сильно відрізняється від інших, це може бути ознакою аномалії. Однак, для таких методів необхідно, щоб дані мали лінійні залежності, оскільки класичні методи кореляції не ефективні для виявлення нелінійних зв’язків.

Методи побудови моделей на основі ймовірнісних розподілів, зокрема використання методів максимального правдоподібності, також є важливими інструментами для виявлення аномалій. Якщо дані припускаються належати певному ймовірнісному розподілу, то для кожного спостереження можна оцінити ймовірність його належності цьому розподілу. Спостереження, ймовірність якого є значно нижчою за очікувану, може бути розглянуте як аномалія. Цей підхід зазвичай використовуються в задачах, де відома форма розподілу даних (наприклад, нормальний розподіл) або де можна провести апроксимацію розподілу на основі зібраних даних. Проте для цього методу також необхідно ретельно підбирати ймовірнісну модель, оскільки невірно обрана модель розподілу може призвести до помилкових висновків [1-3].

Окрім вищезгаданих підходів, у статистиці застосовуються також методи, що базуються на оцінці відстаней між точками в багатовимірному просторі. Наприклад, методи класифікації на основі відстані, такі як метод найближчого сусіда (k-NN), можуть бути використані для виявлення аномалій у багатовимірних наборах даних. В цих методах спостереження, яке має великі

відстані до найближчих сусідів, може бути класифіковане як аномальне. Такі методи часто використовуються в задачах з великими обсягами даних, де не завжди легко побудувати чітку модель для кожного параметра.

При всіх своїх перевагах класичні статистичні методи мають ряд обмежень. Одним із головних є припущення про розподіл даних, оскільки багато статистичних методів вимагають, щоб дані слідували певним припущенням, наприклад, нормальності. Якщо ці припущення не виконуються, застосування класичних методів може призвести до неточних або хибних висновків. Також варто враховувати, що класичні методи зазвичай мають обмеження при роботі з великими обсягами даних або з даними, які мають складну структуру, таку як залежність від часу або сезонні коливання.

Для вирішення цих проблем в останні роки активно використовуються більш сучасні підходи до виявлення аномалій, такі як методи на основі машинного навчання та глибоких нейронних мереж [3]. Однак класичні статистичні методи все ще залишаються важливим інструментом в арсеналі дослідників і практиків завдяки своїй простоті, доступності та здатності забезпечити інтуїтивно зрозумілі результати. Тому навіть у часи розвитку більш складних методів статистичний аналіз аномалій залишається необхідним інструментом для розуміння складних систем та виявлення в них нетипових і важливих патернів.

1.2 Алгоритми машинного навчання для виявлення аномалій

Алгоритми машинного навчання для виявлення аномалій набувають все більшої популярності в сучасних наукових і практичних дослідженнях, оскільки вони дають змогу автоматизувати процес виявлення незвичайних, відмінних від типових патернів даних, що можуть бути ознакою помилок, збоїв у системах, шахрайства або інших важливих подій. Виявлення аномалій є складним завданням, оскільки воно передбачає виявлення елементів, що

істотно відрізняються від більшості даних, але без попередньо визначених класів або точних моделей. У класичних підходах до виявлення аномалій застосовувалися статистичні методи, що використовують математичні припущення щодо розподілу даних [4]. Проте з розвитком машинного навчання було запропоновано цілу низку нових підходів, які не залежать від конкретних припущень про структуру даних і здатні адаптуватися до складних і високовимірних наборів даних, що робить ці методи надзвичайно ефективними в сучасних умовах.

На рисунку 1.2 наведено схему розподілу методів машинного навчання для виявлення аномалій.

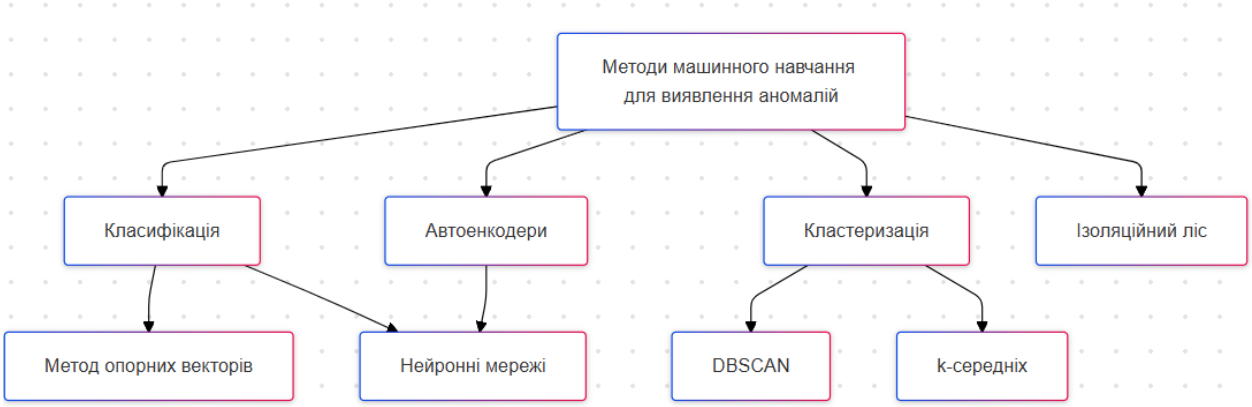

Рисунок 1.2 – Методи машинного навчання для виявлення аномалій

Алгоритми машинного навчання для виявлення аномалій можна поділити на кілька основних категорій: на основі класифікації, на основі кластеризації та на основі побудови моделей. Кожен з цих підходів має свої переваги і недоліки, а також області застосування, що визначають їх ефективність у різних контекстах. Зокрема, методи на основі класифікації намагаються навчити модель розпізнавати нормальні та аномальні спостереження, використовуючи етикетки для навчання, тоді як методи без навчання, такі як методи кластеризації та побудови моделей, працюють без

попередньо заданих етикеток, що дозволяє виявляти аномалії навіть у випадках, коли не існує чіткої визначеності в розподілі даних.

Одним із найбільш популярних методів машинного навчання для виявлення аномалій є методи класифікації, зокрема методи на основі підтримки векторних машин (SVM). Метод підтримки векторних машин для виявлення аномалій (One-Class SVM) представляє собою алгоритм, який може бути використаний для побудови моделі, що визначає межу між нормальними та аномальними точками в багатовимірному просторі. Основною ідеєю цього методу є пошук гіперплощини, яка відокремлює більшість даних від аномальних, таким чином мінімізуючи ймовірність того, що аномальні спостереження будуть класифіковані як нормальні. Один з основних переваг методу One-Class SVM полягає в тому, що він працює навіть при відсутності етикеток для аномальних даних, тобто його можна використовувати в умовах, коли аномалії не позначені окремо, що робить його особливо корисним для невідомих або нових типів аномалій. Однак метод має певні обмеження, зокрема в обробці великих наборів даних, оскільки його обчислювальна складність може бути досить високою, а також він потребує ретельного налаштування параметрів, таких як вибір ядра та регуляризація [1, 4].

Інший популярний підхід в машинному навчанні для виявлення аномалій базується на використанні нейронних мереж, зокрема автоенкодерів. Автоенкодери – це тип нейронних мереж, що складаються з двох основних частин: енкодера, який стискає вхідні дані в компактний латентний простір, та декодера, який відновлює вихідні дані з цього латентного простору. При цьому мережа навчається відновлювати нормальні спостереження, мінімізуючи різницю між вхідними та вихідними значеннями. Для аномальних спостережень, які не відповідають типовим патернам нормальних даних, мережа зазвичай не здатна відновити точні значення, що призводить до великої помилки відновлення, яка може бути використана як критерій для виявлення аномалії. Автоенкодери особливо ефективні в роботі з великими і

високовимірними даними, такими як зображення або текст, де традиційні методи можуть бути менш ефективними. Водночас, ці методи можуть потребувати значних обчислювальних ресурсів, а також існує проблема перенавчання, коли мережа може запам’ятовувати не тільки нормальні патерни, але й шумові дані, що може призвести до погіршення якості виявлення аномалій.

Методи кластеризації, такі як алгоритм k-середніх (k-means) і DBSCAN, також широко використовуються для виявлення аномалій. Алгоритм k-середніх передбачає поділ даних на k кластерів таким чином, щоб елементи одного кластеру були максимально схожі між собою, а відмінності між різними кластерами були значними [1, 3]. Спостереження, що знаходяться далеко від центру кластеру або не належать до жодного з кластерів, можуть бути класифіковані як аномалії. Водночас метод k-середніх має свої обмеження, такі як необхідність попереднього визначення кількості кластерів і погана ефективність при обробці даних з некруглими формами кластерів. Алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) є більш гнучким, оскільки він дозволяє класифікувати спостереження як шум, якщо їх щільність не відповідає щільності основних кластерів, що дозволяє більш ефективно виявляти аномалії в даних з нерівномірним розподілом.

Використання алгоритмів на основі дерева рішень також є популярним підходом для виявлення аномалій, зокрема алгоритм Isolation Forest. Isolation Forest – це метод, що базується на ідеї ізоляції аномальних точок. Алгоритм будує кілька випадкових дерев, в яких кожне спостереження ізолюється шляхом розбиття простору даних на ділянки, що сприяють швидкому виділенню аномальних точок. Аномальні точки, як правило, ізолюються набагато швидше, ніж нормальні, що дозволяє ефективно виявляти їх навіть у великих наборах даних. Isolation Forest є особливо ефективним при роботі з великими наборами даних завдяки своїй здатності до швидкої обробки та порівняно низьким обчислювальним вимогам.

Останнім часом також набули популярності методи, що ґрунтуються на глибинному навчанні, зокрема генеративні моделі, такі як варіаційні автоенкодери (VAE) та генеративні змагальні мережі (GAN). Ці методи працюють шляхом створення моделей, які генерують нові дані, схожі на нормальні спостереження, і можуть бути використані для виявлення аномалій через порівняння реальних спостережень з синтезованими. Генеративні моделі можуть бути особливо ефективними при роботі з великими і складними наборами даних, де інші методи можуть не дати задовільних результатів [5].

Таким чином, алгоритми машинного навчання для виявлення аномалій є потужним інструментом для аналізу даних в різних сферах, від фінансів до медицини та промисловості. Вибір конкретного алгоритму залежить від характеру даних, завдання та наявних ресурсів. Всі ці методи мають свої переваги, але також і обмеження, які повинні враховуватися під час їх застосування для досягнення оптимальних результатів у виявленні аномалій.

1.3 Підходи на основі глибинного навчання

Виявлення аномалій у даних є важливою задачею в багатьох галузях науки, техніки та бізнесу, оскільки аномальні значення часто можуть вказувати на наявність помилок, системних збоїв, шахрайства або інших суттєвих явищ, що потребують уваги. З розвитком машинного навчання і, зокрема, глибинного навчання, з’явились нові потужні інструменти для вирішення цієї задачі [4].

На рисунку 1.3 наведено підходи на основі глибинного навчання.

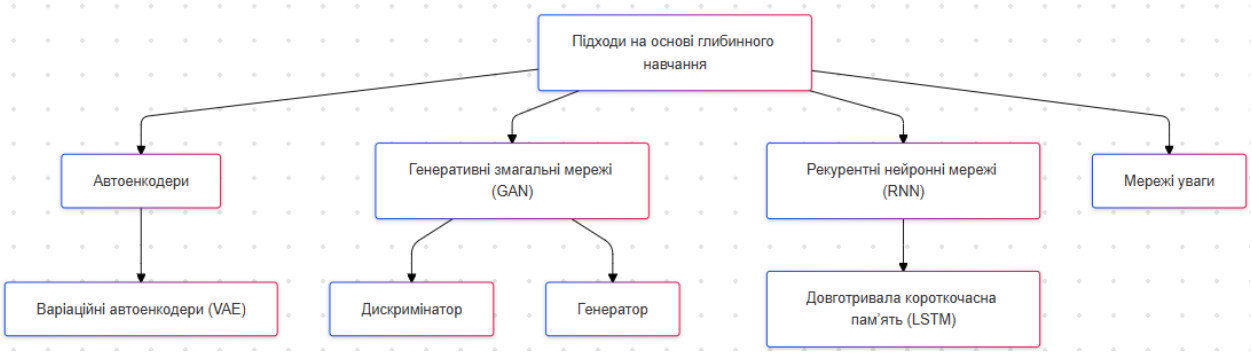

Рисунок 1.3 – Підходи на основі глибинного навчання

Підходи, що ґрунтуються на глибинному навчанні для виявлення аномалій у даних, представляють собою складніші і більш адаптивні методи, здатні працювати з великими та складними наборами даних, а також з даними, що мають високу розмірність, нелінійні взаємозв'язки або незвичні структури. Глибинне навчання, яке охоплює використання нейронних мереж з великою кількістю шарів, дозволяє ефективно автоматизувати процес виявлення аномалій, витягуючи складніші представлення даних, які можуть бути незрозумілими для традиційних методів. Основна особливість підходів на основі глибинного навчання полягає в здатності нейронних мереж автоматично виділяти важливі особливості даних, що робить ці методи менш залежними від попередньої обробки і характеристик, які повинні були б бути задані вручну в традиційних методах машинного навчання. У зв'язку з цим, методи глибинного навчання для виявлення аномалій здатні працювати з неструктурованими даними, такими як зображення, текст, аудіо та інші складні набори даних, де традиційні статистичні або методи машинного навчання можуть бути неефективними. Одним з основних підходів у глибинному навчанні для виявлення аномалій є використання автоенкодерів. Автоенкодери є типом нейронної мережі, що складається з двох основних частин: енкодера та декодера. Енкодер стискає вхідні дані в компактне латентне представлення, яке зберігає найбільш суттєву інформацію про дані, а декодер намагається відновити оригінальні дані з цього латентного

простору [3-5]. Метою навчання автоенкодера є мінімізація різниці між оригінальними вхідними даними та відновленими значеннями. У контексті виявлення аномалій автоенкодери здатні ефективно виявляти нетипові або аномальні спостереження через те, що для таких спостережень мережа не здатна точно відновити вихідні дані. Це призводить до значної помилки відновлення для аномальних точок, що є індикатором їх відмінності від нормальних патернів.

Автоенкодери особливо ефективні при роботі з високовимірними та складними даними, оскільки вони можуть автоматично знаходити найбільш важливі патерни та характеристики, що знижують необхідність попередньої обробки або ручного вибору ознак. Крім того, вони можуть працювати з даними, де аномалії можуть бути нестандартними або де розподіли є складними і не можуть бути описані простими математичними моделями. Однак автоенкодери мають і певні обмеження. Наприклад, вони можуть страждати від перенавчання, особливо коли дані містять багато шуму або коли структура даних дуже складна і не піддається простому кодуванню [3, 4].

Для подолання деяких обмежень автоенкодерів були запропоновані варіаційні автоенкодери (VAE), які є генеративними моделями. Вони використовують ймовірнісні підходи для кодування вхідних даних, замість того, щоб зберігати однозначне латентне представлення. Варіаційний автоенкодер навчається генерувати дані, подібні до навчального набору, через введення випадковості в латентне простору. Цей підхід дозволяє створювати більш гнучкі та стабільні моделі, які можуть враховувати складні ймовірнісні залежності в даних. Для виявлення аномалій у варіаційних автоенкодерах також використовують похибку відновлення, однак відмінністю є те, що VAE враховують ймовірність виникнення даних, що дозволяє більш точно оцінювати аномальність.

Ще одним підходом на основі глибинного навчання для виявлення аномалій є використання генеративних змагальних мереж (GAN). GAN – це

тип нейронних мереж, який складається з двох моделей: генератора та дискримінатора. Генератор створює фальшиві дані, які повинні бути схожі на реальні, а дискримінатор намагається розрізняти реальні дані від фальшивих. Обидві моделі тренуються одночасно, і в результаті генератор навчається створювати дані, які важко відрізнити від реальних, а дискримінатор – покращує свою здатність до виявлення підробок. Для виявлення аномалій можна використовувати цей підхід, порівнюючи реальні дані з даними, згенерованими моделлю. Аномальні точки, які значно відрізняються від реальних даних, будуть важко згенеровані і дискримінатор з високою ймовірністю розпізнає їх як неправильні [5].

Однією з основних переваг GAN є їх здатність генерувати складні дані, що дозволяє застосовувати їх для виявлення аномалій у таких задачах, як генерація зображень, обробка тексту, аудіо та інших неструктурованих типів даних. Однак, хоча GAN є потужним інструментом для генерації даних, їх тренування може бути надзвичайно складним і вимогливим до обчислювальних ресурсів. Погане налаштування або невдачі в навчанні можуть призвести до проблем з конвергенцією і стабільністю моделі.

Крім автоенкодерів і генеративних змагальних мереж, виявлення аномалій можна також здійснювати за допомогою рекурентних нейронних мереж (RNN) та довготривалих короткочасних пам’ятей (LSTM). Ці методи особливо ефективні при обробці послідовних даних, таких як часові ряди, де аномалії можуть бути пов’язані з певними патернами або змінами у часі. Використання LSTM дозволяє моделі зберігати інформацію про попередні стани протягом тривалих інтервалів часу, що є важливим для виявлення аномалій у часових рядах, де спостереження можуть бути сильно залежними від попередніх значень. Для виявлення аномалій в таких моделях використовують похибку предсказання: якщо модель не здатна точно передбачити наступне значення в послідовності, це може бути ознакою аномалії.

Методи глибинного навчання для виявлення аномалій також включають використання нейронних мереж з увагою (attention networks), які можуть динамічно фокусуватися на найбільш важливих частинах даних, що дозволяє краще виявляти аномалії в складних, багатовимірних даних. Вони здатні виділяти ключові частини вхідних даних, на яких зосереджено найбільше аномалій, і таким чином покращувати точність виявлення аномальних точок.

Завдяки своїй здатності автоматично вивчати складні патерни в даних, підходи на основі глибинного навчання стали важливим інструментом для виявлення аномалій у багатьох галузях, таких як фінанси, медична діагностика, безпека, промисловість та інші. Ці методи дозволяють значно покращити ефективність виявлення аномалій у великих і складних наборах даних, однак для їх успішного застосування необхідно враховувати різні фактори, такі як складність навчання, обчислювальні вимоги та потреба в великих обсягах навчальних даних. Враховуючи ці виклики, дослідження в цій області продовжуються, зокрема в напрямку створення більш ефективних і стабільних моделей, що дозволяють покращити точність та швидкість виявлення аномалій.

1.4 Гібридні методи виявлення аномалій

Виявлення аномалій є однією з важливих задач в аналізі даних, яка зосереджена на пошуку незвичайних або рідкісних патернів в наборах даних. Ця проблема виникає в різноманітних галузях, таких як фінанси, медицина, безпека, моніторинг інфраструктури, промисловість і багатьох інших. Аномалії можуть бути індикаторами критичних подій, таких як помилки системи, збої в обладнанні, зловмисні атаки або шахрайство [6]. Однак, виявлення аномалій часто є складною задачею через різноманітність типів аномалій, різні структури даних і необхідність адаптації до специфічних особливостей кожного конкретного випадку. Традиційно для виявлення

аномалій використовуються методи, засновані на статистичних підходах або алгоритмах машинного навчання, однак, з розвитком технологій, науковці та інженери почали звертатися до більш складних підходів, що об’єднують кілька методів для досягнення кращих результатів. Такі підходи отримали назву гібридних методів виявлення аномалій.

Гібридні методи виявлення аномалій поєднують різні техніки та підходи з метою подолання обмежень окремих методів та підвищення ефективності виявлення аномальних точок в даних. Однією з основних переваг гібридних методів є те, що вони можуть поєднувати різні стратегії для виявлення аномалій, включаючи статистичні методи, методи машинного навчання, методи на основі глибинного навчання та евристичні підходи [7]. Вони дозволяють враховувати різноманітні аспекти даних, такі як їх складність, розмірність, наявність шуму і пропущених значень, а також здатні адаптуватися до специфічних вимог конкретних завдань.

Одним із основних напрямків у розвитку гібридних методів є поєднання статистичних і машинно-навчальних підходів. Статистичні методи традиційно базуються на припущеннях про розподіл даних, що дозволяє виявляти аномалії на основі відхилень від цього розподілу. Проте для складних та високовимірних даних, де звичайні припущення про розподіл можуть бути неадекватними, статистичні методи можуть бути неефективними. Однак поєднання таких методів із методами машинного навчання дозволяє скоригувати ці обмеження. Наприклад, одна з таких гібридних стратегій полягає в тому, щоб спочатку використовувати статистичні методи для попереднього фільтрування нормальних даних, а потім застосовувати методи машинного навчання для детального аналізу і класифікації точок даних. Такий підхід може бути особливо корисним в ситуаціях, коли необхідно знизити кількість даних, що обробляються, перед тим як застосувати більш складні та ресурсоємні методи.

Ще одним прикладом гібридних методів є комбінація методів на основі класифікації та методів на основі кластеризації. Класифікаційні методи, як правило, працюють на основі наявних етикеток і навчаються розрізняти нормальні та аномальні точки в даних. Однак у реальних задачах часто відсутні чітко визначені етикетки для аномалій, що ускладнює застосування методів класифікації. В таких випадках можна поєднати методи класифікації з методами кластеризації, що дозволяють автоматично знаходити структуру даних без попереднього визначення класів. Кластеризація дозволяє виявляти групи схожих точок, а потім класифікаційні методи можуть бути використані для подальшого аналізу, щоб виявити точки, які не належать до жодного з кластерів або сильно відрізняються від інших точок в межах кластерів [7, 8].

Гібридні підходи також можуть включати поєднання методів на основі глибинного навчання з іншими техніками машинного навчання або статистики. Наприклад, автоенкодери, які є популярним методом в глибинному навчанні, можуть бути комбіновані з методами класифікації, такими як методи підтримки векторних машин (SVM). В такому випадку автоенкодер може бути використаний для зменшення розмірності і витягнення важливих характеристик з даних, а потім методи класифікації можуть застосовуватися до зменшених та адаптованих даних для виявлення аномалій. Такий підхід дозволяє ефективно працювати з високовимірними даними та зменшувати обчислювальні витрати, що можуть бути необхідні для повноцінного навчання глибинних мереж.

Окрім того, гібридні методи можуть поєднувати різні типи моделей для забезпечення більшої гнучкості у виявленні аномалій в різноманітних типах даних. Наприклад, для роботи з часовими рядами та потоковими даними, які за своєю природою мають певну залежність у часі, може бути корисним комбінувати рекурентні нейронні мережі (RNN) або довготривалі короткочасні пам’яті (LSTM) з традиційними статистичними методами або методами на основі дерева рішень [8, 9]. Таке поєднання дозволяє враховувати

як часову залежність даних, так і їх розподіл, що є важливим для точної оцінки аномалій, особливо в випадках, коли зміни в даних можуть бути поступовими, але все одно є суттєвими.

Гібридні методи виявлення аномалій також здатні враховувати специфіку конкретних застосувань, зокрема, адаптуючи свої стратегії до конкретних характеристик даних. Наприклад, у фінансових даних, де аномалії можуть бути зумовлені змінами в ринковій ситуації, певні методи можуть орієнтуватися на виявлення патернів, пов’язаних з нестабільністю ринку. В медицині ж, де аномалії можуть свідчити про різноманітні захворювання, важливо враховувати складні зв’язки між різними характеристиками пацієнтів і медичними показниками. Гібридні методи дозволяють налаштувати моделі для роботи з такими специфічними умовами, що покращує точність та адаптивність виявлення аномалій. Існують також підходи, що об’єднують гібридні методи з евристичними алгоритмами або методами оптимізації, такими як генетичні алгоритми або алгоритми роїв часток. Ці методи використовуються для покращення пошуку оптимальних параметрів моделей або для комбінування різних моделей на основі певних критеріїв, таких як максимізація ефективності виявлення аномалій при мінімальних обчислювальних витратах. Евристичні алгоритми здатні покращити результати виявлення аномалій у складних ситуаціях, де звичайні методи не дають задовільних результатів, і допомагають здійснювати адаптацію моделей до змінних умов середовища [8-10].

Таким чином, гібридні методи виявлення аномалій є потужним інструментом, що поєднує переваги різних підходів і дозволяє ефективно вирішувати задачу виявлення аномалій у даних, зокрема в складних і високовимірних наборах даних. Вони забезпечують високий рівень точності, гнучкості та адаптивності, що робить їх корисними в численних сферах застосування. Однак для досягнення оптимальних результатів при застосуванні гібридних методів необхідно ретельно налаштовувати моделі і

враховувати специфіку задачі, оскільки неправильне поєднання методів може призвести до зниження ефективності або появи нових помилок. Розвиток цієї галузі продовжується, і з кожним роком з’являються нові методи та стратегії, які покращують здатність гібридних моделей до виявлення аномалій і знижують потребу в складних налаштуваннях.

1.5 Аналіз актуальної літератури

Стаття [1] присвячена виявленню аномалій у даних, що є важливою проблемою сучасного аналізу даних. Аномалії можуть бути наслідком помилок користувачів бази даних, операційних помилок або відсутніх значень. Це питання набуває особливої актуальності через швидке зростання великих наборів даних. У статті представлено попередні результати роботи, що використовують підхід Гранулярних Обчислень для імпутації даних та аналізу відсутніх значень. Пропозиція дає інтуїтивно зрозумілі та доступні для інтерпретації рішення. Зрештою, через серію експериментів демонструється ефективність цього підходу на великому наборі даних у галузі транспорту.

Стаття [2] розглядає задачу виявлення аномалій у відео, яка є складною через рідкісність, відкритість та неоднозначність визначення аномалій. Дослідники зосереджуються на характеристиках аномалій і запропонували різноманітні моделі для їх виявлення. Однак більшість існуючих методів використовують лише нормальні події для побудови моделей і не враховують різноманіття та відкритість нормальних подій. Оскільки реальні відеодані часто мають відкритий розподіл, деякі нормальні патерни майже ніколи не зустрічаються в тренувальних даних. Крім того, рідкісні аномальні події можуть допомогти виявляти подібні аномалії в наборі даних, що схоже на людський досвід. У статті запропоновано нову модель для виявлення аномалій, яка враховує як відкриті нормальні, так і аномальні події в наборі даних, використовуючи різні класифікатори для відкритих і побачених даних.

Спочатку на тренувальних відео навчається класифікатор для побачених даних. Під час тестування дані поділяються на побачені та відкриті, після чого для побачених даних використовуються класифікатори аномалій, а для відкритих даних застосовується метод адаптації домену для зменшення розбіжностей між ними та тренувальними даними. Результати експериментів підтверджують ефективність запропонованої моделі.

Стаття [3] розглядає задачу виявлення аномалій у телеметричних даних космічних апаратів, що є високовимірними часовими рядами, які відображають стан роботи на орбіті. Виявлення аномалій у таких даних є важливим для забезпечення безпеки та надійності. У статті пропонується новий метод виявлення аномалій на основі Генеративних Суперечливих Мереж (GAN), який враховує складні кореляції між змінними та дозволяє ефективно працювати з багатовимірними часовими рядами. Замість того, щоб розглядати кожну змінну окремо, цей метод захоплює латентне представлення між часовими рядами. Для нормальних даних метод на основі GAN може відновити часовий ряд, схожий на оригінальний, вивчаючи ймовірнісну модель нормальних даних. Для аномальних даних відновлений ряд значно відрізняється від оригіналу. У рамках GAN використовується структура мережі Long Short-Term Memory (LSTM) як для генератора, який відновлює часові ряди, так і для дискримінатора, що оцінює ймовірність реальності ряду. Також запропоновано нову аномалійну оцінку GDScore, яка враховує помилку відновлення генератора та вихід дискримінатора. Експерименти з двома наборами телеметричних даних підтверджують ефективність запропонованого методу виявлення аномалій для виявлення викидів.

Стаття [4] розглядає проблему виявлення аномалій у кредитних даних, що виникає через поширення цифровізації та нових ризиків у таких сферах, як інтернет-фінанси. Зокрема, шахрайські дії в кредитних даних можна розглядати як викиди, що мають важливе значення. Однак через високу розмірність реальних даних та малу кількість викидів більшість алгоритмів

виявлення аномалій, які базуються на кластеризації, є неефективними. Тому необхідно знайти метод, який міг би ефективно вирішити проблему виявлення аномалій у незбалансованих кредитних наборах даних у високорозмірному просторі. У статті запропоновано метод виявлення викидів на основі кластеризації розріджених підпросторів, який застосовується для кластеризації високорозмірних та незбалансованих кредитних даних. Результати кластеризації використовуються як метод зниження кількості даних для побудови збалансованих наборів, після чого для виявлення аномалій застосовується класифікатор. У підсумку, ефективність запропонованого алгоритму для виявлення аномалій у кредитних даних підтверджується порівняльними експериментами, що дозволяє компенсувати недоліки традиційних алгоритмів кластеризації та виявлення аномалій у високорозмірних просторах.

Стаття [5] досліджує застосування сіамських нейронних мереж (SNN) для виявлення аномалій у промислових часових рядах, зокрема в контексті роботи з даними електростанцій. Основна проблема, яку вирішує робота, полягає в обмеженнях традиційних алгоритмів розпізнавання/класифікації, які не здатні ефективно працювати з невідомими даними (проблема Open Set Recognition, OSR), що є критичним для промислових застосувань, де можуть виникати непередбачувані несправності. У статті оцінюється ефективність використання SNN на реальних даних з електростанції в Німеччині, а також на двох додаткових публічних наборах даних (MaFaulDa та TEP). Автори показують, що сіамські нейронні мережі є переносним рішенням для подолання проблеми OSR у промисловій сфері, що відкриває нові можливості для точного виявлення аномалій в умовах нестабільних та змінних умов.

Стаття [6] зосереджена на виявленні аномальних випадків споживання електроенергії на підприємствах з метою прогнозування попиту на енергію. Оскільки реальні аномальні патерни споживання електроенергії є нерегулярними, необхідно розробити гнучку модель для їхнього виявлення. У

роботі проводиться аналіз даних про аномальне споживання електроенергії та прогнозуються потенційно аномальні патерни. Метою є створення даних на основі виявлених аномальних патернів і розробка моделі, здатної виявляти згенеровані аномальні дані. Як результат, остаточна модель досягла 74% і 72% точності для оригінальних аномальних і нормальних даних відповідно, а для випадково згенерованих аномальних даних було отримано точність 95,07% для патерну зростання і 89,69% для патерну зменшення. Автори пропонують підхід для виявлення потенційно аномальних даних і розробки гнучких моделей для їхнього виявлення.

Стаття [7] розглядає використання машинного навчання (ML) для покращення розуміння поведінки промислових підприємств, зокрема в контексті виявлення аномалій, що допомагає операторам заводів визначити, чи працює їх система нормально, або чи потрібно вживати коригувальні заходи. Однак, однією з основних проблем використання моделей ML є їхня інтерпретованість, тобто здатність зрозуміти, як модель прийшла до свого висновку. Для покращення інтерпретованості моделей ML наукове співтовариство звертається до галузі пояснювального штучного інтелекту (XAI), яка останнім часом здобуває популярність у промисловому секторі. У статті проводиться огляд різних методів XAI для виявлення аномалій в індустріальних системах, зокрема для багатовимірних часових рядів, оскільки це найбільш поширений тип даних у промислових системах. Розглядаються існуючі техніки XAI, які здебільшого орієнтовані на зображення, табличні або текстові дані, але вони не підходять для пояснення аномалій у багатовимірних часових рядах. Для вирішення цієї проблеми розроблено метод виявлення аномалій на основі автоенкодерів для багатовимірних часових рядів, згенерованих промисловими симуляторами. У статті представлено та обговорено сім різних технік XAI, заснованих на атрибуції ознак, прикладах і деревах. З цих технік метод на основі SHAP, званий DTFS, правильно

ідентифікував основну причину аномалії з точністю 86% і витратив 1,53 секунди для пояснення результатів на тестовій системі.

Стаття [8] пропонує вдосконалену версію традиційного алгоритму Isolation Forest для виявлення аномалій, зокрема для покращення ефективності обробки великих даних. Оскільки звичайний алгоритм, хоча й підвищує ефективність виконання, все ще залишається часозатратним, було запропоновано його покращення. Зокрема, для аналізу потоку даних про доступ користувачів на веб-платформі застосовувався попередній етап обробки даних. Для зменшення розмірності ознак використовувався метод головних компонент (PCA), після чого для виявлення аномалій була застосована паралельна обробка на основі алгоритму Isolation Forest. Ефективність запропонованого методу була перевірена за допомогою стандартного симуляційного набору даних.

Стаття [9] зосереджена на виявленні аномалій у мережевих даних хмарних обчислень, що є важливою проблемою через численні транзакції та приховану інфраструктуру в таких системах. Хмарні обчислення забезпечують зв’язок між даними та програмами з різних географічних локацій, що ставить нові виклики перед дослідниками, зокрема в галузі забезпечення безпеки мереж. Виявлення аномальних даних, що суттєво відрізняються від більшості, є однією з основних проблем, а машинне навчання показало свою ефективність у цій галузі. Однак застосування методів навчання з учителем для виявлення аномалій залишається складним через дисбаланс класів та непередбачувану природу аномальних даних. Враховуючи це, стаття досліджує методи одно-класового класифікатора, зокрема One Class Support Vector Machine (OCSVM) та автоенкодери, для аналізу хмарних мереж. Для аналізу використано набір даних YAHOO, що є першим використанням цих даних для виявлення аномалій. За результатами дослідження, автоенкодер досяг точності 96,02% у виявленні аномалій, а OCSVM – 79,05%. Додатково було проведено тестування на іншому стандартному наборі даних UNSW-

NB15, де автоенкодер показав точність 99,10%, а OCSVM – 60,89%. Результати демонструють, що методи на основі нейронних мереж показують кращі результати в порівнянні з методами на основі ядер у виявленні аномалій у хмарних мережах.

Стаття [10] зосереджена на вирішенні важливої проблеми забруднення повітря та його впливу на якість життя, розробляючи технології для моніторингу якості повітря та визначення проблемних зон. Однією з основних складнощів у цьому напрямі є збір даних про забруднення повітря, які часто не мають чіткої диференціації між нормальними та аномальними рівнями якості. Враховуючи важливість виявлення аномалій у даних про забруднення повітря, пропонується нова методологія, яка не лише забезпечує охорону здоров'я людей, але й покращує загальну якість даних. Запропонований підхід передбачає інжекцію аномальних подій у набори даних про забруднення повітря з використанням характеристик тимчасового розподілу даних. Для оцінки достовірності цих згенерованих аномалій застосовується тест Колмогорова-Смірнова. Крім того, розроблений метод оцінюється за допомогою сучасних глибоких моделей навчання для виявлення аномалій. В кінці статті представлено метод автоенкодера з глибокою увагою, AT-MCRAAD, який демонструє вищу ефективність порівняно з існуючими традиційними та сучасними алгоритмами у виявленні цих аномальних подій.

Стаття [11] зосереджена на вивченні методів виявлення аномалій для промислових продуктів на основі глибокого навчання, що є ключовим етапом для забезпечення високої якості продукції. Для збалансованого набору зображень промислових продуктів пропонується модель виявлення аномалій на основі YOLOv3, яка будує класифікатор ROI для визначення типів аномалій. Для незбалансованого набору зображень, де аномальних зображень мало, пропонується напівконтрольована модель на основі Fast-AnoGAN, яка побудована тільки на нормальних зразках. Ця модель використовує треновану модель WGAN-GP для генерації зображень і виявляє аномалії, моніторячи

аномальний бал, який обчислюється як різниця між згенерованим зображенням і тестовим. Обидві моделі виявлення аномалій були оцінені на збалансованих та незбалансованих наборах даних у реальних умовах промислового виробництва. Результати оцінки ефективності показали, що ці моделі на основі глибокого навчання можуть добре задовольняти дві основні вимоги – реальність і точність для виявлення аномалій у високошвидкісних умовах промислового виробництва.

Стаття [12] зосереджена на моніторингу стану газових турбін, зокрема на виявленні аномальних поведінок своєчасно, щоб забезпечити безпеку їх експлуатації та уникнути дорогих непланових технічних обслуговувань. Одним з найпопулярніших методів виявлення аномалій є побудова моделі класифікації на основі реальних даних газових турбін. Висока ефективність цього методу забезпечується наявністю достатньої кількості анотованих зразків, зокрема аномальних. Однак у даних моніторингу газових турбін нормальних зразків значно більше, ніж аномальних, або навіть немає аномальних зразків, що ускладнює задачу. Тому для виявлення аномалій вимагаються новітні технології, які можуть точно виявляти аномальні поведінки в часі з використанням неанотованих даних. У статті розглядається новий метод виявлення аномалій без нагляду, заснований на алгоритмі Isolation Forest, для виявлення аномалій у газовому шляху газових турбін. Дані моніторингу групуються за часовими рядами для ослаблення впливу неминучого зниження ефективності при роботі турбіни, після чого всі дані обробляються моделлю Isolation Forest з низьким рівнем забруднення. Кожна виявлена аномальна група перевіряється ще раз за допомогою моделі з високим рівнем забруднення для отримання конкретних аномальних польотних циклів. Використовуючи реальні дані моніторингу з 8 різних аероенергетичних установок CFM56-7B, результати показують, що метод на основі Isolation Forest може досягти високої точності виявлення аномалій при використанні неанотованих даних і невеликих наборів даних.

Стаття [13] зосереджена на важливості виявлення аномальних даних у часових рядах, що є критичним завданням для багатьох промислових застосувань, де час є ключовим компонентом. Оскільки часові ряди використовуються для прогнозування значень, створення найбільш точної моделі є важливим завданням. Якщо вхідні дані містять аномалії, модель не буде працювати належним чином, що впливатиме на точність майбутніх прогнозів. Традиційні методи поліпшення продуктивності моделі включають застосування регуляризації, інженерію ознак або експериментування з різними комбінаціями функцій активації і/або функцій втрат, а також кількості нейронів і прихованих шарів у нейронних мережах. Однак такий підхід, орієнтований на модель, часто виявляється неефективним у реальних застосуваннях.

Стаття [14] пропонує новий підхід, орієнтований на дані, де основна увага приділяється оновленню і корекції самих вхідних даних для вирішення проблем, що виникають у моделі. Різні моделі, побудовані за підходом орієнтованим на модель, показали низьку ефективність із великою кількістю хибних негативів. Натомість підхід, орієнтований на дані, продемонстрував 100% ефективність у правильному виявленні аномальних точок у даних.

Стаття [15] розглядає важливість виявлення аномалій у сучасному аналізі даних, оскільки процес дематеріалізації реальних даних сприяє зростанню обсягу обміну даними. У цьому контексті виявлення аномальних даних стає все більш важливою задачею, оскільки такі дані можуть мати особливе значення і потребувати певних дій. Останні досягнення в області штучного інтелекту, зокрема машинного навчання, роблять значний прорив у цій галузі. Зазвичай ці методи були розроблені для збалансованих наборів даних або за умови певних припущень щодо розподілу даних. Однак реальні застосування часто стикаються з проблемою незбалансованого розподілу даних, коли нормальні дані представлені у великих кількостях, а аномальних випадків дуже мало, що робить задачу виявлення аномалій схожою на пошук

голки в сіні. У статті розроблено експериментальну установку для порівняльного аналізу двох типів методів машинного навчання при їх застосуванні до систем виявлення аномалій. Досліджується їх ефективність з урахуванням розподілу аномалій у незбалансованому наборі даних.

Стаття [16] зосереджена на виявленні аномалій у часових рядах, що є важливою частиною прогнозування та управління станом обладнання (PHM), і активно вивчається в цій галузі. Дані з ознаками часового ряду часто мають нестабільні властивості, а їх амплітуда коливань змінюється з часом. Традиційні алгоритми виявлення аномалій здатні виявляти аномалії лише на поверхневих рівнях даних, однак вони не справляються з виявленням викидів у глибоких ознаках часового ряду. Структура воріт мережі довгої короткочасної пам’яті (LSTM) має явні переваги в обробці часового ряду, тоді як навчання за допомогою генеративної суперечливої мережі (GAN) ефективно допомагає у виявленні та здобутті глибоких ознак даних. Тому в цій статті досліджується виявлення аномалій у часових рядах з використанням об’єднаної моделі LSTM і GAN, яка отримала назву LSTM-GAN. Експериментальні результати показали, що запропонований алгоритм демонструє перевагу в обробці часового ряду порівняно з традиційними алгоритмами. Дослідження, представлене в статті, має значне наукове значення для вдосконалення методів виявлення аномалій у часових рядах.

1.6 Висновки до першого розділу

Ґрунтуючись на аналізі предметної області, можна зробити кілька важливих висновків щодо аналізу предметної області та постановки задачі дослідження:

– еволюція підходів до виявлення аномалій. Розділ чітко демонструє розвиток методів виявлення аномалій від класичних статистичних підходів до складніших алгоритмів машинного та глибинного навчання, а також до більш

інтегрованих та ефективних гібридних методів. Це підкреслює важливість адаптації методів до змінюваних умов та складних особливостей даних у різних галузях;

– класичні статистичні методи. Класичні статистичні методи виявлення аномалій є основою для базових аналізів, однак їх ефективність обмежена при роботі з високовимірними, складними або неструктурованими даними. Це створює потребу у використанні більш складних підходів, таких як машинне навчання та глибинне навчання, для точнішого та швидшого виявлення аномалій;

– алгоритми машинного навчання. Алгоритми машинного навчання для виявлення аномалій демонструють високий рівень гнучкості та адаптивності, оскільки здатні працювати з великими наборами даних та автоматично навчатися на основі патернів. Однак ці методи також потребують значних обчислювальних ресурсів та якісних даних для навчання, що може обмежити їх застосування у деяких випадках;

– підходи на основі глибинного навчання. Глибинне навчання значно підвищує ефективність виявлення аномалій, особливо в складних і неструктурованих даних, таких як зображення, текст або аудіо. Завдяки своїй здатності автоматично виділяти важливі характеристики даних, ці методи дозволяють вирішувати більш складні завдання, але вони вимагають великих обсягів даних і високих обчислювальних потужностей;

– гібридні методи. Гібридні методи, що поєднують різні підходи, дозволяють ефективно комбінувати переваги традиційних статистичних методів та сучасних технік машинного та глибинного навчання. Ці методи можуть вирішувати різні типи завдань виявлення аномалій з високою точністю, особливо в складних випадках, де одні методи можуть бути недостатніми або занадто обмеженими;

– адаптація до специфічних умов. Різноманіття підходів до виявлення аномалій підкреслює важливість вибору найбільш підходящої методології

залежно від конкретної задачі, типу даних і обмежень, з якими стикається дослідник чи інженер. Відповідно, у реальних умовах необхідно адаптувати методи виявлення аномалій до специфічних вимог, що висуваються до точності та швидкості аналізу даних;

– потреба в подальших дослідженнях. Враховуючи різноманіття існуючих підходів, можна зробити висновок, що питання вибору та адаптації методів для ефективного виявлення аномалій залишається відкритим, і потребує подальших досліджень, які б сприяли вдосконаленню існуючих алгоритмів та розробці нових підходів, що поєднують переваги різних методів та моделей.

Отже, аналіз предметної області свідчить про значний прогрес у галузі виявлення аномалій, але також підкреслює необхідність подальшого розвитку гібридних та адаптивних підходів для досягнення максимальної ефективності в різноманітних умовах.

РОЗДІЛ 2. ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ У НАБОРАХ ДАНИХ

2.1 Типи аномалій

Аномалії є суттєвим елементом в дослідженні та аналізі даних у різних галузях, таких як статистика, машинне навчання, обробка сигналів, біоінформатика та багато інших [11]. Вони представляють собою значення або спостереження, які суттєво відрізняються від загальної тенденції або патерну даних, що може свідчити про наявність помилок, нових, цікавих явищ або відхилень, що потребують додаткового дослідження. Аномалії можуть проявлятися в різних формах, залежно від того, скільки змінних або параметрів містить досліджувана система. Таким чином, можна виділити два основні типи аномалій: одновимірні та багатовимірні. Кожен із цих типів має свої особливості в контексті виявлення, аналізу та застосування, що робить їх важливими для різних сфер застосування.

Одновимірні аномалії найчастіше пов’язані з даними, що складаються з одного виміру або параметра, і є типом аномалій, які можна легко виявити за допомогою простих методів статистичного аналізу. Вони зазвичай з’являються в результаті спостережень або вимірювань одного параметра протягом певного періоду часу або в межах одного експерименту. Одним із класичних прикладів є аномалії в даних, що характеризують температуру повітря в різних точках протягом дня [10, 11]. Якщо у певний момент часу температура різко змінюється, це може бути результатом несправності вимірювального пристрою, або, навпаки, рідкісного природного явища, яке потребує додаткового дослідження. Виявлення таких аномалій є відносно простим завданням, оскільки вони часто виділяються на фоні інших значень, що відповідають нормальному розподілу або іншому типу закономірності.

Методи виявлення одновимірних аномалій зазвичай базуються на обчисленні статистичних характеристик, таких як середнє, стандартне відхилення, медіана або квантилі. Наприклад, у випадку нормального розподілу можна використовувати правила, такі як «відстань більше ніж три стандартні відхилення від середнього» для визначення аномалій. Існують також більш складні методи, такі як аналіз гістограм, виявлення локальних екстремумів або побудова моделей прогнозування, що дозволяють більш ефективно знаходити аномальні спостереження [11, 12].

Однак у багатьох реальних задачах дані часто мають не лише один параметр, а кілька змінних, що ускладнює виявлення аномалій. Саме в таких випадках виникають багатовимірні аномалії, які стосуються даних, що містять більше ніж один вимір або параметр. Багатовимірні аномалії значно складніші у своєму виявленні та аналізі через взаємодію між різними змінними. У багатьох випадках аномалії можуть бути прихованими або неочевидними, оскільки вони можуть виникати лише при певних комбінаціях значень кількох параметрів. Наприклад, в області фінансів багатовимірні аномалії можуть виникати, коли зміни в декількох економічних показниках одночасно ведуть до незвичайних фінансових результатів, які не можуть бути виявлені, якщо аналізувати лише один з них.

Виявлення багатовимірних аномалій вимагає більш складних статистичних та машинних методів. Одним із таких підходів є використання методів багатовимірного аналізу, таких як головні компоненти (PCA), аналіз кластеризації, або побудова моделей машинного навчання, що дозволяють виявляти аномалії, не зважаючи на складність взаємодії між різними змінними [10-12]. Наприклад, метод головних компонент дозволяє знизити розмірність даних і виявляти найбільш значущі напрямки варіативності, у яких можуть з’являтися аномалії. Цей підхід допомагає зменшити вплив «шуму» та фокусуватися на основних факторах, що визначають поведінку даних.

Іншими популярними підходами до виявлення багатовимірних аномалій є методи на основі класифікації або регресії, де моделі, навчені на великій кількості нормальних спостережень, можуть бути використані для виявлення аномальних точок, які значно відрізняються від передбачених значень. Наприклад, у задачах машинного навчання часто використовуються методи на основі нейронних мереж або дерев рішень, що дозволяють детально оцінити взаємозв’язки між множиною змінних і визначити, які з них можуть спричиняти аномалії [12, 13]. Застосування цих методів потребує значних обчислювальних ресурсів і великих обсягів даних, однак вони є надзвичайно ефективними в розпізнаванні складних аномалій, які не можна було б виявити за допомогою простих статистичних підходів.

У випадках багатовимірних даних важливу роль відіграє також кореляція між змінними, оскільки деякі аномалії можуть бути неочевидними в окремих змінних, але виразно проявляються через зміни в їх взаємозв’язку. Тому сучасні методи часто включають аналіз кореляційних матриць або побудову графів, де кожна змінна представлена вузлом, а зв’язки між ними – ребрами, що дозволяє наочно відстежувати, як зміни в одних змінних впливають на інші.

Важливо зазначити, що виявлення аномалій є не тільки технічним завданням, а й концептуальним, оскільки існує кілька критеріїв, що визначають, чи є спостереження аномальним. В деяких випадках аномалія може бути цінною інформацією, що свідчить про нові тенденції або явища, наприклад, у фінансових ринках або в наукових експериментах. В інших випадках аномалія може бути результатом помилки або збою в системі, що потребує виправлення. Це означає, що процес виявлення аномалій включає не лише алгоритмічну частину, а й інтерпретацію результатів в контексті предметної області, що дозволяє приймати обґрунтовані рішення щодо подальших кроків [12, 13].

У практичних застосуваннях, таких як обробка сигналів або відеоаналітика, аномалії можуть бути сприйняті як неочікувані або небажані події, наприклад, дефекти в зображеннях або зміни в аудіо сигналів, які потребують детального аналізу. У таких випадках важливим є використання комбінованих підходів, що поєднують класичні методи статистики з методами машинного навчання, що дозволяє досягти більш високої точності виявлення аномалій.

Отже, як одновимірні, так і багатовимірні аномалії мають свої особливості в контексті виявлення та аналізу. Одновимірні аномалії зазвичай є простішими для виявлення, однак багатовимірні аномалії відображають складніші взаємозв’язки між різними змінними, що потребує більш складних методів аналізу. В обох випадках важливо враховувати контекст і значення кожної аномалії, оскільки вона може бути як сигналом про необхідність втручання, так і новою, цікавою інформацією для подальших досліджень.

2.2 Різниця між локальними та глобальними аномаліями

Порівняння локальних та глобальних аномалій [12, 13] наведено в таблиці 2.1.

Таблиця 2.1

Порівняння локальних та глобальних аномалій

Аспект	Локальні аномалії	Глобальні аномалії
Визначення	Відхилення, які спостерігаються лише в межах обмеженої області даних. Вони часто мають локальне значення і можуть бути результатом помилок вимірювання або специфічних ситуацій, не пов’язаних із загальною тенденцією.	Аномалії, які спостерігаються на всьому обсязі даних і можуть бути результатом більш масштабних змін або явищ, що впливають на всю систему

Продовження таблиці 2.1

Характеристики	Обмежене впливання на загальні тенденції, можуть бути випадковими або точковими	Мають великий вплив на загальну картину даних, часто відображають суттєві зміни в системі чи процесі
Приклади	Несправність одного сенсора в серії вимірювань	Фінансова криза, що відображається у всіх економічних показниках країни
Методи виявлення	Прості статистичні методи, такі як виявлення значень, що виходять за межі стандартних відхилень, локальні тренди або зміни	Більш складні методи, включаючи багатовимірний аналіз, моделі машинного навчання та глобальні спостереження.
Важливість	Можуть бути менш важливими для загальної системи, але їх потрібно аналізувати для усунення помилок або покращення точності	Мають критичне значення для розуміння суттєвих змін у системі або прогнозування важливих подій

2.3 Використання методів кластеризації та класифікації

Методи кластеризації та класифікації є важливими інструментами в галузі машинного навчання та статистики, що застосовуються для аналізу великих обсягів даних з метою виявлення прихованих закономірностей та структури в інформації. Хоча ці методи мають спільні риси, вони принципово відрізняються в підходах до обробки даних і їх застосуванні. Класифікація є процесом, коли дані, які мають певні характеристики, групуються відповідно до заздалегідь визначених категорій або класів, тоді як кластеризація передбачає автоматичне визначення груп або кластерів, де кожен об’єкт або спостереження групується за схожістю з іншими елементами в рамках однієї групи [9-11]. Це означає, що класифікація зазвичай є спрямованим процесом, заснованим на навчанні, де для кожного нового прикладу передбачаються його

належність до певного класу на основі навчальної вибірки, тоді як кластеризація може застосовуватися до набору даних без попередньої інформації про структуру чи класи.

Процес класифікації зазвичай включає етапи, коли на початковому етапі створюється навчальна вибірка, що складається з об’єктів, для яких відомі відповідні категорії або мітки. Моделі машинного навчання навчаються на основі цих даних, щоб створити правило або модель, яка дозволяє класифікувати нові, невідомі дані в одну з заданих категорій. Важливим аспектом класифікації є вибір ознак, які найкраще описують кожен клас. Це може бути складним завданням, оскільки для досягнення високої точності класифікації необхідно вибрати такі ознаки, які забезпечують найбільшу різницю між класами, мінімізуючи вплив випадкових чи незначущих факторів. Для цього часто використовуються різні методи, включаючи обробку і трансформацію даних, видалення шуму, а також використання технік відбору ознак [10].

Найбільш популярними методами класифікації є алгоритми, такі як дерева рішень, метод опорних векторів (SVM), наївний байєсів класифікатор, нейронні мережі та k-ближчих сусідів (k-NN). Дерева рішень є одним із найбільш інтуїтивно зрозумілих і широко використовуваних методів. Вони створюють модель у вигляді дерева, де кожен вузол відповідає за порогове значення ознаки, а гілки представляють можливі варіанти вибору, що призводить до певного класу. Метод опорних векторів, з іншого боку, використовує концепцію гіперплощин для поділу простору ознак таким чином, щоб максимізувати відстань між класами. Цей метод є дуже потужним, особливо для задач з високою розмірністю, де інші методи можуть не працювати ефективно.

Використання нейронних мереж стало популярним останнім часом завдяки їх здатності обробляти великі обсяги даних та автоматично знаходити складні взаємозв’язки між ознаками. Нейронні мережі застосовуються для

класифікації в багатьох сферах, включаючи розпізнавання зображень, обробку мови та аналіз текстів [12]. Вони складаються з кількох шарів нейронів, які можуть вивчати ієрархічні представлення ознак і передбачати клас для нових даних. Метод k-ближчих сусідів є простим і популярним методом класифікації, який використовує відстань між точками даних для класифікації нового спостереження на основі найближчих до нього елементів навчальної вибірки.

Класифікація має безліч практичних застосувань, включаючи діагностику захворювань, фінансові прогнози, ідентифікацію спаму в електронних листах, а також розпізнавання осіб або голосів. Проте цей процес може стикатися з кількома викликами. Одним із найбільших є проблема незбалансованих класів, коли одна категорія даних є значно меншою за інші, що може привести до неточної класифікації. Для подолання цієї проблеми часто використовуються методи балансування даних, такі як перевибірка або зменшення кількості зразків з переважної категорії, а також різні стратегії налаштування ваг у моделі.

На відміну від класифікації, кластеризація є методом без нагляду, який не передбачає використання міток або категорій для даних. Метою кластеризації є пошук природних груп у наборі даних, де елементи в кожному кластері мають схожі властивості або характеристики. Кластеризація широко використовується в багатьох галузях, таких як маркетинг, де вона дозволяє виділяти групи клієнтів із схожими перевагами або купівельною поведінкою, або в біоінформатиці для групування генів, що мають схожі функції. Методи кластеризації можуть бути різними, залежно від мети та природи даних [13].

Одним із найпоширеніших методів кластеризації є метод k-середніх. Він передбачає розподіл усіх об’єктів на певну кількість кластерів, де кожен об’єкт відноситься до того кластера, центр якого знаходиться найближче. Алгоритм k-середніх є простим та ефективним, але має свої обмеження, наприклад, він вимагає заздалегідь визначити кількість кластерів, що може бути важким

завданням, якщо структура даних не є очевидною. Для цього вдаються до різних критеріїв для вибору оптимальної кількості кластерів, таких як метод ліктя або критерій силуета.

Інший популярний метод кластеризації – це ієрархічна кластеризація. Цей метод будує дерево кластерів, яке дозволяє виявити не тільки групи, але й їх взаємозв'язки. Ієрархічна кластеризація може бути виконана двома основними способами: агломеративним (при якому кожен об'єкт спочатку є окремим кластером, а потім пари найближчих кластерів об'єднуються) та дивізивним (де всі об'єкти спочатку належать до одного великого кластера, який поступово ділиться на менші). Цей підхід є корисним, коли потрібно побудувати більш гнучку структуру кластерів, яка дозволяє відображати різні рівні подібності між елементами [14].

Між іншим, методи кластеризації можуть застосовуватися для різноманітних цілей. Наприклад, у соціальних мережах кластеризація може використовуватися для виявлення спільнот користувачів з подібними інтересами, у медицині для аналізу різних типів клітин або захворювань, а в маркетингу – для створення сегментів споживачів. Однак кластеризація також має ряд складнощів, таких як вибір найбільш підходящого методу для конкретного набору даних, а також проблема шуму в даних, що може ускладнити процес правильного класифікування елементів.

Використання методів кластеризації та класифікації має важливе значення в багатьох сферах, оскільки обидва підходи дозволяють отримати цінну інформацію з великих наборів даних. Класифікація застосовується в тих випадках, коли необхідно віднести нові дані до відомих категорій, тоді як кластеризація корисна, коли структура даних не є відомою, і необхідно знайти природні групи в інформації. Кожен з цих методів має свої сильні та слабкі сторони, і в залежності від задачі можуть бути застосовані різні підходи, включаючи комбінацію класифікації та кластеризації для досягнення більш точних результатів.

2.4 Виявлення аномалій у часових рядах

Виявлення аномалій у часових рядах є однією з важливих та складних задач в аналізі даних, яка широко застосовується в багатьох сферах, від фінансів і економіки до медицини, виробництва та моніторингу навколишнього середовища. Часові ряди являють собою послідовність спостережень, що зроблені в певні моменти часу, і часто є результатом складних процесів, що змінюються з часом. Одним із важливих аспектів аналізу таких рядів є виявлення аномалій, тобто таких спостережень, які суттєво відрізняються від загальної тенденції або патерну даних. Виявлення аномалій дозволяє ідентифікувати різноманітні незвичні або несподівані явища, такі як помилки вимірювань, збої в системах, непередбачувані зміни в процесах або загрози, пов’язані з певними небезпечними подіями [15].

Аномалії в часових рядах можуть бути різного характеру: від простих помилок вимірювань, які виникають через технічні неполадки, до складних явищ, пов’язаних із серйозними змінами в системі або процесі, що спостерігаються. Вони можуть проявлятися як окремі відхилення або тривалі зміни, що значно відрізняються від загальної тенденції в даних. Для ефективного виявлення таких аномалій необхідно застосовувати спеціалізовані методи, які дозволяють не тільки виявити аномальні точки, але й зрозуміти їх причини, визначити їхній тип і виявити, чи є ці аномалії суттєвими для подальшого аналізу та прийняття рішень.

Існує кілька основних категорій аномалій, що можуть з’являтися в часових рядах. Першою категорією є зміщення в рівні або тенденції. Такі аномалії виникають, коли відбуваються раптові або поступові зміни в рівні часу або тренді даних, що можуть свідчити про систематичні збої або зовнішні впливи на процес. Це може бути, наприклад, виявлення підвищеного рівня продажів після кампанії, що різко змінює звичайну динаміку, або спад у фінансових показниках через економічні кризи чи природні катастрофи.

Іншою категорією є аномалії сезонності, коли з’являються незвичні коливання або сплески в періодичних циклах, що порушують звичний патерн сезонних змін. Третя категорія аномалій включає відхилення на рівні шуму або випадкових коливань. Ці аномалії можуть бути викликані випадковими зовнішніми факторами або помилками вимірювань, однак для правильного їх аналізу важливо відрізняти їх від реальних змін, що відбуваються в системі.

Основними підходами до виявлення аномалій у часових рядах є статистичні методи, методи машинного навчання та гібридні підходи, що поєднують переваги обох. Статистичні методи є основою для багатьох класичних підходів до аналізу даних і передбачають використання розподілів ймовірностей, різних тестів для перевірки гіпотез та модельних підходів для прогнозування майбутніх значень [4, 6]. Одним з найбільш поширених методів є методи на основі середнього значення та стандартного відхилення, де аномалії визначаються як значення, що виходять за межі певного інтервалу довірчого діапазону. Однак цей підхід має обмеження, оскільки він не завжди здатний виявити складні аномалії, зокрема в разі наявності сильних сезонних або трендових коливань.

Більш складним та гнучким методом є використання авторегресивних інтегрованих моделей середнього (ARIMA) або їхніх розширених варіантів. Моделі ARIMA використовують інформацію про минулі значення часових рядів для прогнозування майбутніх значень, а потім порівнюють фактичні спостереження з прогнозами, виявляючи аномалії, коли фактичні значення суттєво відрізняються від прогнозованих. Цей метод є ефективним для виявлення аномалій у часових рядах, де є виражені тренди та сезонні коливання, однак він вимагає ретельного налаштування параметрів і може бути чутливим до випадкових флуктуацій.

Методи машинного навчання стали дуже популярними для вирішення задач виявлення аномалій в останні роки. Одним з основних підходів є використання алгоритмів класифікації, які можуть навчатися на основі міток

для визначення аномальних та нормальних точок. Однак у разі часових рядів, де часто відсутні попередньо мічені аномалії, застосовуються безнаглядні методи, такі як методи кластеризації. Алгоритми кластеризації, наприклад методи k-середніх або DBSCAN, дозволяють виявляти групи схожих точок і відокремлювати аномальні точки, що значно відрізняються від решти спостережень. Вони базуються на припущенні, що нормальні точки зібрані в компактні кластери, а аномалії є одиничними точками або точками, що значно відрізняються від більшості [16].

Нейронні мережі також знайшли своє застосування в задачах виявлення аномалій у часових рядах, особливо в разі складних та багатовимірних даних. Одним з популярних підходів є використання рекурентних нейронних мереж (RNN) і їхніх розширених варіантів, таких як LSTM (Long Short-Term Memory). Ці моделі здатні враховувати залежності в часових рядах на великій кількості тимчасових кроків, що дозволяє виявляти складні аномалії, які можуть бути неочевидними для класичних статистичних методів. Нейронні мережі можуть навчатися на основі великих обсягів даних і автоматично налаштовувати свої параметри для оптимізації виявлення аномалій. Для навчання таких мереж використовуються техніки, що включають в себе підхід до підкріплення або використання автокодерів для побудови та аналізу моделей, що є корисними для виявлення нетипових зразків у даних.

Окрім традиційних методів, сучасні підходи до виявлення аномалій часто включають використання гібридних методів, які комбінують різні алгоритми та підходи. Наприклад, у разі обробки складних даних можуть застосовуватися методи глибокого навчання для виявлення шаблонів у даних разом із статистичними методами для подальшого уточнення результатів і детекції аномальних значень. Врахування різноманітних аспектів, таких як тренди, сезонність, і взаємозв’язки між різними змінними, дозволяє значно підвищити ефективність виявлення аномалій [16].

Важливим аспектом виявлення аномалій є також постобробка результатів і їх інтерпретація. Після того, як аномалії були виявлені, необхідно провести детальний аналіз для того, щоб визначити їхній характер, причини та потенційний вплив на систему або процес. Це може включати в себе дослідження контексту, аналіз факторів, що могли призвести до аномалії, а також використання додаткових даних для підтвердження або спростування припущень щодо природи аномальних подій.

Виявлення аномалій у часових рядах є важливою частиною аналітичних задач, і його застосування має велике значення для різних галузей науки та практики. Від правильного виявлення аномалій залежить ефективність управлінських рішень у багатьох сферах, таких як фінансові ринки, прогнозування погоди, медичні діагнози, контроль якості в виробництві та моніторинг безпеки. Успішне виявлення аномалій дозволяє своєчасно реагувати на непередбачувані події, що можуть призвести до значних економічних або соціальних наслідків.

2.5 Проблеми з масштабуванням методів для великих наборів даних

Масштабування методів для великих наборів даних є однією з ключових проблем, з якими стикаються дослідники та практики в галузі аналізу даних, машинного навчання, штучного інтелекту та статистики. У сучасному світі, де кількість генерованих даних зростає експоненціально, здатність обробляти і аналізувати великі обсяги інформації стає важливою складовою наукових досліджень і прийняття рішень у різних сферах діяльності. Проблеми з масштабуванням пов’язані з труднощами у забезпеченні ефективної обробки, зберігання і аналізу великих наборів даних за допомогою існуючих методів, алгоритмів та технічних рішень [13, 14]. Вивчення цих проблем має критичне значення для розуміння обмежень сучасних технологій і для розвитку нових

підходів, які можуть забезпечити більш ефективне використання наявних ресурсів.

Однією з головних причин складнощів у масштабуванні є те, що традиційні алгоритми аналізу даних і машинного навчання розроблялися з розрахунком на роботу з відносно невеликими обсягами інформації. З часом, зі збільшенням доступних даних, виявилося, що багато з цих методів не здатні ефективно працювати з великими масивами даних без суттєвих затрат ресурсів. Під час аналізу великих наборів даних постають проблеми, пов’язані з обчислювальними витратами, вимогами до пам’яті, тривалістю виконання алгоритмів, а також із складністю самих моделей, що мають бути побудовані для забезпечення необхідної точності. Крім того, зростає потреба у паралельній обробці даних, розподілених системах та обчислювальних клаудах, що додає нові виклики в управлінні і підтримці таких систем.

Однією з основних проблем, що виникає при масштабуванні методів для великих наборів даних, є обмеження в часі виконання обчислень. Багато традиційних алгоритмів, особливо в таких галузях, як класифікація, кластеризація, регресія та аналіз часових рядів, мають складність, яка лінійно або квадратично зростає з розміром даних. Наприклад, класичний алгоритм k-ближчих сусідів (k-NN) має складність $O(n^2)$, де n – кількість елементів у наборі даних. Це означає, що для великих наборів даних час обробки може становити кілька годин або навіть днів, що робить цей підхід непридатним для масштабних задач. Подібні проблеми виникають у методах, що базуються на оцінці ймовірностей, таких як методи наївного байєса або деякі типи методів опорних векторів (SVM), де кількість обчислень зростає експоненційно при збільшенні кількості даних або кількості ознак. В результаті в таких випадках може виникнути необхідність у розробці нових алгоритмів або підходів, які здатні значно зменшити час виконання за рахунок спрощення моделей або використання більш ефективних методів оптимізації.

Іншою важливою проблемою є обмеження в обсязі пам’яті, яку потрібно для збереження великих наборів даних. При роботі з великими масивами інформації, що можуть включати сотні мільйонів або навіть мільярди рядків, стає очевидним, що зберігання та обробка таких даних вимагають значних ресурсів пам’яті. Багато алгоритмів не можуть ефективно працювати з даними, що не вміщуються в оперативну пам’ять (RAM), що може призвести до тривалих процесів обміну даними між оперативною пам’яттю і дисковим простором, значно знижуючи ефективність обробки [14-16]. Пошук рішень для цієї проблеми часто вимагає використання спеціалізованих технологій для зберігання даних, таких як розподілені файлові системи або бази даних, які підтримують обробку даних, що не вміщуються в одну машину або обмежену кількість процесорних ядер. Для зменшення потреб у пам’яті застосовуються різноманітні техніки, такі як індексація, зменшення розмірності, апроксимація даних або стиснення.

Однією з ключових складностей при масштабуванні методів машинного навчання є також необхідність у виборі правильних моделей, що здатні працювати з великими наборами даних, зберігаючи високий рівень точності та ефективності. Традиційно побудова моделей машинного навчання, таких як регресійні моделі або дерева рішень, потребує значних обчислювальних ресурсів. Це зумовлено необхідністю проведення численних ітерацій для навчання, оптимізації параметрів та перевірки результатів на великій кількості даних. Розв’язання цих проблем може вимагати використання більш складних і масштабованих алгоритмів, таких як градієнтний спуск, стохастичні методи або глибоке навчання, які передбачають використання нейронних мереж та інших складних моделей для побудови більш точних прогнозів. Проте, застосування таких методів потребує значних обчислювальних потужностей, оскільки навчання таких моделей на великих наборах даних може займати кілька днів або навіть тижнів на звичайних обчислювальних платформах.

Також важливим аспектом є обробка великих даних в умовах розподілених обчислень. Багато сучасних підходів до аналізу великих наборів даних передбачають використання клаудових платформ або кластерів комп'ютерів для обробки та зберігання даних. Проте, при масштабуванні на ці платформи постають проблеми, пов'язані з ефективною координацією обчислень між численними машинами, синхронізацією даних та забезпеченням надійності системи. Використання паралельних та розподілених обчислень у таких умовах вимагає спеціалізованих алгоритмів і протоколів для обробки даних, які дозволяють ефективно працювати з великими масивами даних, не створюючи затримок або помилок при обміні інформацією між вузлами. Це є надзвичайно важливим, оскільки навіть невеликі збої або затримки можуть суттєво знизити загальну ефективність роботи системи [13].

Однією з перспективних стратегій для вирішення проблеми масштабування є використання технік зниження розмірності. Ці методи дозволяють значно зменшити кількість ознак, що використовуються для побудови моделей, зберігаючи при цьому основну інформацію, необхідну для аналізу. Прикладом таких методів є метод головних компонент (PCA) або t-SNE, які використовуються для скорочення розмірності даних при збереженні їх основних властивостей. Вони дозволяють значно зменшити обсяг даних, що обробляються, та покращити ефективність алгоритмів, одночасно зберігаючи достатній рівень точності результатів. Проте, застосування цих методів також має свої обмеження, оскільки може призвести до втрати деякої інформації, що може негативно вплинути на якість результатів у деяких випадках.

Загалом, масштаби сучасних даних значно перевищують можливості традиційних методів аналізу, і тому для ефективного оброблення великих обсягів даних необхідно використовувати новітні технології та підходи. Відкриті проблеми масштабування вимагають подальших досліджень і розробок в області комп'ютерних наук, математики та статистики. Створення

нових методів, що дозволяють здійснювати обробку великих наборів даних, є критичним завданням для багатьох галузей, таких як фінанси, охорона здоров'я, виробництво та наука. Технології розподілених обчислень, вдосконалені алгоритми машинного навчання, ефективні методи зниження розмірності, а також нові підходи до використання паралельних обчислень можуть сприяти подоланню існуючих обмежень і дозволити значно поліпшити ефективність обробки великих наборів даних.

2.6 Вплив шуму та неповних даних

Вплив «шуму» та неповних даних є однією з найбільш значущих проблем у сфері обробки та аналізу даних, адже реальні дані часто мають недоліки, які можуть негативно впливати на точність моделей, отримуваних результатів і прийнятих рішень. У наукових та інженерних задачах, де обробка даних є основною складовою, наявність шуму та неповних або пропущених даних може знижувати ефективність алгоритмів машинного навчання, статистичних методів та інших аналітичних підходів [12]. Ці проблеми викликають необхідність розробки нових методів та стратегій, які дозволяють адекватно враховувати ці складнощі при обробці великих обсягів інформації, що особливо важливо в умовах сучасних технологій, де дані постійно генеруються у великих кількостях.

Шум у даних, як правило, представляє собою випадкові або детерміновані помилки вимірювань, які виникають внаслідок непередбачуваних зовнішніх чи внутрішніх факторів, таких як неточності сенсорів, збої в процесах збору даних, помилки в апаратному забезпеченні чи програмному забезпеченні, а також вплив навколишнього середовища. Цей шум може значно спотворювати реальні дані, що робить складним їх подальший аналіз і застосування до реальних ситуацій. Шум може бути у вигляді випадкових змін, що виникають без явної закономірності, або як

структуровані відхилення, що пов’язані з конкретними системними помилками. Наприклад, у фінансових даних шум може бути обумовлений випадковими коливаннями на ринку, що не відображають реальну економічну ситуацію, а в медицині шум може бути наслідком помилок вимірювань або неточностей у записах результатів тестів.

Однією з основних проблем, пов’язаних із шумом, є те, що він може призвести до втрати інформативності в даних і навіть до повної спотвореності результатів. Шум може призводити до надмірної варіативності в спостереженнях, що заважає алгоритмам правильно виявляти закономірності, а також створює труднощі в прогнозуванні і прийнятті рішень. Одним із найбільш значущих наслідків наявності шуму є зниження якості моделей, побудованих на таких даних. Наприклад, у методах машинного навчання, таких як регресія, класифікація або кластеризація, шум може призвести до того, що модель почне адаптуватися до випадкових коливань, а не до справжніх закономірностей в даних, що призводить до переобучення (overfitting). У разі, коли алгоритм адаптується до шуму, його здатність до узагальнення, тобто правильного прогнозування на нових, невідомих даних, суттєво знижується [13].

Враховуючи проблему шуму, багато методів обробки даних намагаються виявити й усунути або зменшити його вплив. Для цього розробляються спеціалізовані техніки, такі як фільтрація даних, згладжування сигналів або використання алгоритмів, що можуть відрізняти корисну інформацію від шуму. Наприклад, у часових рядах можуть застосовуватися методи фільтрації для згладжування спостережень і видалення випадкових коливань, таких як фільтри Калмана або методи середнього ковзання. Також існують статистичні методи для виявлення шуму, зокрема методи, що аналізують відхилення від середнього значення або використовують більш складні критерії, такі як перевірка нормальності розподілу даних.

Неповні дані або пропущені значення є ще однією поширеною проблемою в обробці даних, що виникає через різні причини. У багатьох випадках, коли дані збираються або генеруються, можуть виникати ситуації, коли деякі значення не були зафіксовані через технічні помилки, відсутність сенсорів, людські помилки або просто через відсутність інформації. Ці пропуски можуть бути випадковими або систематичними [14-16]. Випадкові пропуски можуть виникати, наприклад, через збій в роботі обладнання, тоді як систематичні пропуски можуть бути пов’язані з певними характеристиками самої вибірки, наприклад, у разі, коли деякі атрибути пропущені для певних категорій об’єктів або груп.

Наявність пропущених значень у даних створює кілька серйозних проблем. По-перше, алгоритми машинного навчання зазвичай вимагають, щоб усі дані були повними, тобто не містили пропусків, що ускладнює їх застосування на практиці. Пропущені дані можуть призвести до того, що алгоритм не зможе коректно оцінити параметри моделі або побудувати правильну гіпотезу про залежності між змінними. По-друге, пропущені значення можуть призвести до того, що отримані моделі або прогнози будуть менш точними, оскільки відсутність деяких даних може істотно змінювати результат. Тому важливо, щоб процес заповнення пропущених значень був правильно організований, оскільки неправильне або неточне відновлення даних може призвести до спотворення результатів.

Існують різні методи для обробки неповних даних, які дозволяють заповнити пропущені значення або мінімізувати їхній вплив на аналіз. Одним із найпростіших підходів є заміщення пропущених значень середнім або медіанним значенням для числових змінних, що дозволяє зберегти структуру даних, але може призвести до втрати інформації, особливо в разі великих пропусків. Більш складними методами є використання статистичних підходів, таких як методи імпутації на основі найближчих сусідів (k-NN), які передбачають заповнення пропусків значеннями, що найбільше схожі на

існуючі значення з інших спостережень. Інші методи імпутації використовують моделі машинного навчання для прогнозування пропущених значень, такі як використання дерев рішень або регресії [7-9]. В останні роки все більше застосовуються методи, що базуються на глибинному навчанні, які використовують нейронні мережі для відновлення пропущених даних.

Незважаючи на наявність цих методів, важливо зазначити, що жоден із них не є ідеальним, і неправильне поводження з неповними даними може мати серйозні наслідки для точності та достовірності результатів. Особливо складним є випадок, коли пропуски є систематичними або мають певну закономірність, оскільки в таких випадках стандартні методи можуть не дати точних результатів і призвести до значних спотворень. Крім того, чим більше пропусків у даних, тим складніше здійснити їх коректне відновлення, оскільки велика кількість відсутніх значень може обмежувати можливість правильної інтерпретації та побудови моделей.

Загалом, шум і неповні дані є постійними викликами для обробки і аналізу даних. Вони можуть значно впливати на точність і надійність моделей та результатів, тому важливо розробляти спеціалізовані методи та алгоритми, що дозволяють ефективно справлятися з цими проблемами. Врахування шуму та неповних даних при побудові моделей є необхідною складовою частиною сучасного аналізу даних і машинного навчання, оскільки від їх правильного оброблення залежить точність прогнозів та правильність прийнятих рішень.

2.7 Критерії оцінки результатів

Критерії оцінки результатів, такі як точність, повнота та F-міра, є важливими інструментами для оцінки ефективності моделей машинного навчання, особливо в контексті задач класифікації. Ці показники допомагають зрозуміти, наскільки добре модель справляється з виконанням поставлених завдань, зокрема з виявленням коректних класів серед великої кількості

можливих варіантів [5-7]. Розуміння цих критеріїв має критичне значення для покращення моделей, вибору між різними підходами та глибшого аналізу результатів класифікації, зокрема в тих випадках, коли важливо досягти балансу між різними видами помилок. Точність, повнота та F-міра є взаємозв'язними, що дає змогу краще оцінити не лише загальну ефективність моделі, а й характер її помилок, що в кінцевому підсумку визначає, як саме модель працює на практиці.

Точність є однією з основних метрик для оцінки роботи алгоритмів класифікації, оскільки вона дозволяє з'ясувати, яка частина з усіх передбачених позитивних класів є дійсно коректною. Формально точність визначається як частка правильних позитивних передбачень серед усіх передбачених позитивних класів. Це означає, що точність відповідає на питання: скільки з передбачених класів є насправді правильними. Цей показник є дуже простим у розрахунках і широко застосовується в задачах, де важливо уникати хибнопозитивних результатів, тобто де важливо, щоб передбачення моделі було максимально наближеним до істинних значень. Наприклад, у задачах діагностики захворювань, де неправильно визначити здорову людину як хвору може мати серйозні наслідки, точність є важливим критерієм. Однак точність не завжди є надійним показником у випадках, коли класи є несиметричними за кількістю або коли існує велика різниця в затратності між типами помилок.

Повнота, на відміну від точності, визначається як частка правильних позитивних передбачень серед усіх справжніх позитивних класів. Вона показує, яка частка всіх реальних позитивних випадків була виявлена моделлю. Цей показник важливий у ситуаціях, коли кожен пропущений позитивний випадок має суттєве значення, і модель повинна прагнути до максимально можливого охоплення всіх потенційно важливих випадків. Наприклад, у задачах, де важливо не пропустити хвору людину або виявити дефект у продукті, навіть за умови, що модель може призвести до деякої

кількості помилок, повнота є критичним показником [3, 4]. При цьому високий рівень повноти може супроводжуватися збільшенням кількості хибнопозитивних результатів, що також є важливою проблемою.

Таким чином, точність і повнота представляють собою два взаємно виключні критерії: прагнення до високої точності може призвести до зниження повноти і навпаки. Це створює дилему для розробників моделей, оскільки необхідно знаходити баланс між цими двома показниками, щоб забезпечити ефективне та раціональне використання моделі в конкретному контексті. Наприклад, у випадках, де критично важливо виявити всі потенційно небезпечні ситуації, навіть якщо деякі з них будуть помилковими, потрібно більше уваги приділяти повноті. В той час як у випадках, де важливо мінімізувати кількість хибнопозитивних результатів, точність стає більш важливою.

F-міра є комбінованим показником, який дозволяє об’єднати точність і повноту в одну метрику для оцінки загальної ефективності моделі. Це гармонійне середнє між точністю та повнотою, що дозволяє досягти компромісу між цими двома критеріями. F-міра є особливо корисною в контекстах, де важливо досягти балансу між мінімізацією хибнопозитивних та хибнонегативних результатів. Формула F-міри виглядає так (2.1):

$$F_1 = 2 \times \frac{\text{Точність} \times \text{Повнота}}{\text{Точність} + \text{Повнота}} \tag{2.1}$$

Цей показник дає змогу зменшити вплив крайнощів, коли один з критеріїв значно переважає над іншим. Таким чином, F-міра є незамінною в багатьох практичних випадках, де необхідно врахувати не тільки загальну кількість правильних передбачень, а й рівновагу між різними типами помилок. Вона також корисна в задачах з незбалансованими класами, де один клас може бути представлений значно частіше за інший. Наприклад, у випадку виявлення шахрайських транзакцій або рідкісних хвороб, де кількість позитивних

випадків набагато менша за негативні, застосування тільки точності може призвести до неправильних висновків, а F-міра дозволяє адекватно оцінити модель.

Однак варто зазначити, що в різних контекстах може бути доцільно використовувати різні варіанти F-міри, зокрема, може бути корисним застосування збільшених або зменшених ваг для точності або повноти, якщо один з критеріїв є важливішим, ніж інший. У таких випадках використовуються варіанти F-міри, зокрема F-β, де параметр β дозволяє змінювати баланс між точністю та повнотою. Коли $\beta > 1$, модель орієнтується більше на повноту, а коли $\beta < 1$, то на точність [4-6].

Проте важливо враховувати, що використання цих показників вимагає глибокого розуміння конкретного застосування та специфіки задачі. Наприклад, у медицині чи безпеці важливіше мати високу повноту для уникнення хибнонегативних випадків, навіть якщо це призведе до зростання кількості хибнопозитивних результатів. У свою чергу, в задачах, де критично важливо мінімізувати кількість помилок, таких як виявлення спаму або фінансове шахрайство, точність може бути пріоритетною.

Таким чином, точність, повнота та F-міра є основними метриками для оцінки ефективності алгоритмів класифікації. Вони дозволяють не тільки оцінити якість моделі в загальному, але й визначити характер її помилок. Для досягнення найкращих результатів важливо вибирати відповідні метрики для конкретної задачі та адаптувати моделі, щоб мінімізувати непотрібні помилки та забезпечити точне і повне розпізнавання всіх значущих випадків.

2.8 Висновки до другого розділу

Розділ охоплює різноманітні методи та техніки для виявлення аномалій у даних, починаючи від типів аномалій до специфічних методів кластеризації та класифікації. Це свідчить про багатогранність підходів у цій сфері, що

дозволяє вибирати найбільш підходящий метод в залежності від конкретних умов задачі. Розгляд різниці між локальними та глобальними аномаліями вказує на важливість чіткого визначення типу аномалії, оскільки це суттєво впливає на вибір методів виявлення. Локальні аномалії часто потребують специфічних підходів, тоді як глобальні аномалії можуть бути виявлені за допомогою більш універсальних методів.

Використання методів кластеризації та класифікації для виявлення аномалій підкреслює, що ці підходи є основними інструментами в аналізі даних, дозволяючи не тільки виявляти аномальні записи, але й визначати їх взаємозв’язок із іншими даними. Це може бути критичним для вирішення завдань в різних галузях, таких як фінанси, медицина або кібербезпека.

Виявлення аномалій у часових рядах, без сумніву, є окремим важливим аспектом, оскільки часові залежності в даних часто ускладнюють застосування традиційних методів виявлення аномалій. Підхід до виявлення аномалій у часових рядах вимагає особливих технік, таких як врахування трендів та сезонних коливань. Проблеми масштабування вказують на складнощі застосування методів виявлення аномалій при роботі з великими обсягами даних. Це підкреслює необхідність розробки більш ефективних алгоритмів, які можуть працювати в реальному часі або на великих розподілених системах, що є актуальним у сучасних умовах.

Виявлення аномалій у наборах даних із шумом та неповними даними є важливим аспектом, оскільки такі дані можуть серйозно спотворити результати. Це вказує на необхідність врахування специфіки даних під час вибору методів, а також на важливість попередньої обробки даних для покращення результатів виявлення аномалій.

Критерії оцінки результатів, такі як точність, повнота та F-міра, є важливими для визначення ефективності методів виявлення аномалій. Вони дозволяють не тільки оцінити якість моделі, але й знаходити баланс між

різними типами помилок, що є критичним для прийняття рішень в реальних застосунках.

Таким чином, розділ підкреслює важливість комплексного підходу до виявлення аномалій, який включає в себе як правильний вибір методів, так і врахування специфіки даних, таких як шум, неповнота та великі обсяги інформації. Проблеми масштабування та виявлення аномалій у часових рядах вимагають подальших досліджень і вдосконалення технік, що дозволить розширити можливості застосування цих методів у різних сферах діяльності.

РОЗДІЛ 3. ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Опис мети розробки та визначення функціональних вимог

Необхідно розробити програмний модуль для виявлення аномалій у багатовимірних даних, який дозволяє застосовувати кілька методів детекції та надавати структуровані звіти й візуалізації результатів.

Аномалії у даних (нестандартні значення або поведінка) можуть значно впливати на якість моделей машинного навчання, точність аналізу даних і ухвалення рішень. Для їх автоматичного виявлення необхідно створити систему, що підтримує кілька підходів до ідентифікації аномальних точок у наборі даних.

3.2 Опис функціональних та нефункціональних вимог

Функціональні вимоги:

– масштабування числових ознак для уніфікації їх розподілу перед застосуванням алгоритмів;

– збереження вихідних значень для зручності аналізу;

– isolation forest: дерева ізоляції для пошуку аномалій через їх рідкість;

– local outlier factor (LOF): аналіз локальної щільності для ідентифікації точок, що суттєво відрізняються від локального контексту;

– статистичний метод: використання z-оцінок для ідентифікації точок, які виходять за межі певного порогового значення;

– elliptic envelope: оцінка коваріаційної еліптичної моделі для виявлення відхилень;

– структуроване зведення результатів роботи кожного методу;

– збереження звітів у форматі json для подальшого аналізу;

– побудова графіків із розподілом нормальних і аномальних точок у просторі;

– порівняння результатів різних методів на окремих графіках;

– графік із підсумковим розподілом кількості аномалій за методами;

– можливість налаштування параметрів, таких як частка аномалій, порогове значення Z-оцінки, кількість сусідів для LOF тощо.

Нефункціональні вимоги:

1) код повинен бути написаний з використанням бібліотек Python. numpy, pandas, matplotlib, seaborn, scikit-learn;

2) забезпечити зручну інтеграцію з іншими системами через вихідні JSON-файли;

3) візуалізації мають бути зрозумілими й інформативними;

4) система повинна бути оптимізована для роботи з великими наборами даних, уникаючи надмірного використання пам’яті.

3.3 Опис використаних технологій

Основними компонентами розробленого рішення є сучасні бібліотеки Python, такі як numpy, pandas, matplotlib, seaborn, scikit-learn, а також кілька методів статистичного аналізу та машинного навчання. Кожна з цих технологій має свої переваги, функціональні можливості та унікальні особливості, які сприяють створенню ефективного програмного забезпечення для аналізу та візуалізації даних.

Python як мова програмування був обраний через його широку популярність у галузі науки про дані, машинного навчання та штучного інтелекту. Його багатий екосистемний набір бібліотек дозволяє виконувати складні математичні, статистичні та візуалізаційні завдання з високою ефективністю та читабельністю коду. Одна з ключових бібліотек, що використовується у цьому коді, – це numpy. Вона є основним інструментом

для роботи з багатовимірними масивами та математичними операціями, які виконуються швидко завдяки використанню оптимізованого C-коду під капотом. Зокрема, бібліотека забезпечує можливість обчислення стандартних статистичних характеристик, таких як середнє значення, стандартне відхилення та Z-оцінки, які є важливими для статистичного методу виявлення аномалій. Крім того, numpy надає зручні засоби для обробки числових даних, таких як нормалізація, що є важливою частиною попередньої обробки.

Ще однією фундаментальною бібліотекою у цьому проєкті є pandas. Вона забезпечує високооптимізовану структуру даних у вигляді DataFrame, яка дозволяє легко маніпулювати даними, проводити фільтрацію, групування, сортування та виконувати багато інших операцій. У коді pandas використовується для створення і зберігання набору даних, а також для інтеграції результатів кожного методу виявлення аномалій у зрозумілому вигляді. Це включає додавання стовпців із позначенням аномалій, індексів аномальних точок, а також розрахунок частки аномальних записів у загальному наборі даних. Завдяки цьому забезпечується легкість у подальшому аналізі результатів.

Для виконання завдань попередньої обробки, таких як стандартизація даних, використовується компонент StandardScaler із бібліотеки scikit-learn. Ця бібліотека є провідним інструментом для побудови моделей машинного навчання, надання засобів для попередньої обробки, а також підтримки алгоритмів класифікації, регресії та кластеризації. Стандартизація забезпечує перетворення числових ознак таким чином, щоб їх середнє значення дорівнювало нулю, а стандартне відхилення дорівнювало одиниці. Це важливий етап, оскільки багато алгоритмів машинного навчання чутливі до масштабів даних, і відсутність нормалізації може негативно вплинути на результати.

Для виявлення аномалій у даних застосовується кілька методів, кожен із яких реалізований за допомогою scikit-learn або статистичних інструментів.

Перший із них – Isolation Forest. Це метод, який базується на побудові дерев, які ізолюють точки даних. Аномалії зазвичай ізолюються на меншій кількості розділень, ніж нормальні точки, що робить цей алгоритм швидким і ефективним для багатовимірних даних. У коді використовується клас IsolationForest із бібліотеки scikit-learn, де задається параметр contamination, що визначає частку очікуваних аномалій у даних. Результати роботи алгоритму зберігаються у форматі DataFrame, а додатково розраховуються кількість та процентний вміст аномалій.

Ще один метод – Local Outlier Factor (LOF), який визначає аномалії через оцінку локальної щільності. Цей алгоритм працює за допомогою порівняння локальної щільності точки з її сусідами, що дозволяє ідентифікувати об’єкти, які суттєво відрізняються від локального контексту. У коді використовується клас LocalOutlierFactor, який також має параметр contamination, а додатково можна налаштувати кількість сусідів, що впливають на результат. Як і у випадку з Isolation Forest, результати цього алгоритму зберігаються у вигляді таблиці з позначенням аномалій.

Статистичний метод, застосований у коді, базується на використанні Z-оцінок, які визначають, наскільки віддаленою є точка від середнього значення у кількості стандартних відхилень. Цей підхід дозволяє ефективно виявляти точки, які значно відрізняються від основної маси даних, використовуючи порогове значення. Для обчислення Z-оцінок використовується функція zscore із бібліотеки scipy.stats. Виявлені аномалії додаються до результатів разом із кількісними характеристиками.

Elliptic Envelope – це метод, який моделює дані як багатовимірний нормальний розподіл і використовує коваріаційну матрицю для визначення меж, за якими знаходяться аномалії. У коді цей метод реалізований за допомогою класу EllipticEnvelope із бібліотеки scikit-learn, що забезпечує можливість обробки даних із різним ступенем шуму та аномальності.

Для створення візуалізацій використовуються бібліотеки matplotlib і seaborn. Перша з них є базовим інструментом для побудови графіків у Python, яка забезпечує повний контроль над усіма аспектами графічного представлення, такими як колір, маркери, підписи осей та легенда. У коді matplotlib використовується для побудови діаграм розподілу нормальних і аномальних точок, а також для графіка розподілу кількості аномалій за методами. Seaborn, своєю чергою, є надбудовою над matplotlib, яка дозволяє створювати статистичні графіки з мінімальними зусиллями. У поєднанні ці бібліотеки забезпечують ефективну візуалізацію складних багатовимірних даних.

Для збереження звітів у структурованому вигляді використовується стандартний модуль json. Він дозволяє зберігати результати роботи програми у вигляді файлів JSON, що є зручним для подальшого аналізу або інтеграції з іншими системами. У коді створено спеціальний клас NumpyEncoder, який розширює функціональність стандартного JSON-енкодера, дозволяючи серіалізувати об’єкти NumPy, такі як масиви, числові типи або плаваючі значення.

Таким чином, у цьому програмному рішенні поєднуються потужні інструменти для обробки, аналізу та візуалізації даних, що забезпечують високу точність і зручність використання. Реалізовані методи дають змогу гнучко підходити до виявлення аномалій, використовуючи різні алгоритми та підходи, які адаптуються до потреб користувача.

3.4 Архітектура програмного забезпечення

Архітектура розробленого програмного забезпечення для виявлення аномалій у даних представляє собою добре структуровану та модульну систему, яка інтегрує сучасні підходи до обробки, аналізу та візуалізації даних. В основі рішення лежить об’єктно-орієнтована парадигма програмування, яка

забезпечує гнучкість, масштабованість і зручність у підтримці коду. Головним елементом архітектури є клас AnomalyDetector, що виконує роль контейнера для всіх основних операцій, включаючи підготовку даних, застосування різних алгоритмів виявлення аномалій, генерацію звітів та візуалізацію результатів. Структура класу спроєктована таким чином, щоб чітко відокремлювати логіку обробки даних від алгоритмів аналізу і допоміжних операцій.

На верхньому рівні архітектури розташовані базові функції й методи, які забезпечують ініціалізацію та конфігурацію об’єкта класу AnomalyDetector. Конструктор класу (__init__) відповідає за збереження початкового набору даних у вигляді окремої копії для забезпечення їхньої цілісності та можливості повторного використання без додаткової обробки. Цей підхід дозволяє уникати небажаних змін оригінальних даних під час застосування алгоритмів, які модифікують дані для своїх потреб. Крім того, у конструкторі задаються основні параметри, такі як рівень контамінації, що використовується у багатьох алгоритмах, і ініціалізується об’єкт класу StandardScaler для стандартизації числових ознак. Усі ключові змінні, включаючи результати роботи алгоритмів, зберігаються у вигляді атрибутів класу, що забезпечує централізовану структуру зберігання даних.

Модуль попередньої обробки даних реалізований через метод preprocess_data. Його завданням є стандартизація вибраних стовпців у наборі даних для приведення значень до єдиного масштабу. Це необхідно для коректної роботи алгоритмів машинного навчання, таких як Isolation Forest і Local Outlier Factor, які чутливі до розмірностей і масштабів даних. Метод дозволяє користувачеві вказати, які саме стовпці мають бути оброблені, або автоматично вибирає всі числові стовпці у випадку, якщо параметр не задано. Використання стандартного масштабування гарантує, що кожна ознака матиме середнє значення, рівне нулю, і стандартне відхилення, рівне одиниці, що сприяє підвищенню точності алгоритмів.

Серцем архітектури є модуль виявлення аномалій, що складається з кількох окремих методів, кожен з яких реалізує конкретний алгоритм. Метод isolation_forest інтегрує алгоритм Isolation Forest, який працює шляхом ізоляції кожної точки даних через побудову дерев. Його реалізація базується на класі IsolationForest із бібліотеки scikit-learn. Після застосування алгоритму результат додається до вихідного набору даних у вигляді нового стовпця, де кожен запис позначається як нормальний або аномальний. Додатково розраховуються такі статистичні метрики, як загальна кількість виявлених аномалій, їхня частка від загального обсягу даних і список індексів аномальних точок. Ці результати зберігаються у словнику results для подальшого аналізу та генерації звітів.

Метод local_outlier_factor реалізує алгоритм Local Outlier Factor, який оцінює щільність розподілу даних і визначає аномалії через відносну щільність точки порівняно з її локальним контекстом. Використання цього методу дозволяє виявляти точки, які є ізольованими у своїй локальній області, навіть якщо вони належать до більшої групи нормальних даних. Метод підтримує гнучке налаштування кількості сусідів для оптимізації роботи залежно від характеру даних. Як і у випадку з Isolation Forest, результати включають кількість аномалій, частку та індекси відповідних точок.

Для реалізації статистичного методу виявлення аномалій використовується метод statistical_method, який базується на обчисленні Z-оцінок. Цей підхід дозволяє визначати точки, які суттєво відхиляються від середнього значення у заданому діапазоні стандартних відхилень. Використання бібліотеки scipy.stats для обчислення Z-оцінок забезпечує точність і швидкість виконання. Усі точки, значення яких перевищують заданий поріг, позначаються як аномальні, і результати додаються до загального словника з аналогічними метриками.

Метод elliptic_envelope використовує підхід моделювання даних як багатовимірного нормального розподілу. Завдяки цьому алгоритм ефективно

ідентифікує аномалії, базуючись на оцінці параметрів розподілу, таких як середнє значення та коваріаційна матриця. Використовуючи клас EllipticEnvelope із бібліотеки scikit-learn, метод адаптується до рівня шуму у даних і дозволяє враховувати особливості кожного набору даних.

Окремим модулем у класі є функції візуалізації. Метод visualize_anomalies_detailed відповідає за побудову графіків для кожного алгоритму виявлення аномалій. Графіки показують розподіл нормальних і аномальних точок на площині, що дозволяє наочно оцінити ефективність роботи кожного методу. Завдяки бібліотеці matplotlib реалізується детальна налаштовуваність графіків, включаючи кольори, маркери та підписи. Метод plot_anomaly_distribution будує гістограму, що демонструє кількість аномалій, виявлених різними методами. Це сприяє порівнянню результатів і вибору оптимального підходу для конкретного завдання.

Модуль генерації звітів реалізований через метод generate_comprehensive_report, який створює структурований звіт у форматі JSON. Звіт містить узагальнену інформацію про вхідні дані, такі як загальна кількість записів і список ознак, а також результати роботи всіх алгоритмів. Для серіалізації використовується клас NumpyEncoder, який дозволяє коректно обробляти об’єкти NumPy. Звіт може бути збережений у файл, що спрощує інтеграцію з іншими системами або аналіз результатів.

Заключною частиною архітектури є головна функція main, яка об’єднує всі компоненти у єдиний робочий процес. Функція генерує синтетичний набір даних із контрольованою кількістю аномалій, ініціалізує об’єкт класу AnomalyDetector, викликає методи попередньої обробки та застосування алгоритмів, генерує звіт і створює візуалізації. Такий підхід до організації дозволяє легко адаптувати програму до різних сценаріїв використання, включаючи обробку реальних наборів даних і інтеграцію з іншими системами.

Отже, архітектура коду демонструє ефективне використання об’єктно-орієнтованого програмування для створення модульної системи, яка

забезпечує надійність, масштабованість і зручність у використанні. Кожен компонент спроєктований для виконання конкретного завдання, що полегшує підтримку та розширення системи.

3.5 Функціональність системи

Розроблене програмне забезпечення для виявлення аномалій у наборах даних функціонує як комплексна система, що поєднує в собі етапи підготовки даних, застосування декількох методів аналізу, генерації звітів та візуалізації результатів. Ключовою особливістю є його здатність обробляти дані різної природи, забезпечуючи гнучкість у виборі алгоритмів, які ефективно ідентифікують аномальні записи. Процес роботи програмного забезпечення можна розбити на кілька взаємопов'язаних етапів, кожен з яких відіграє важливу роль у досягненні кінцевої мети – точного й ефективного виявлення аномалій.

На першому етапі виконується завантаження та підготовка даних. У рамках цього процесу ініціалізується об'єкт класу AnomalyDetector, до якого передається початковий набір даних у вигляді об'єкта DataFrame бібліотеки pandas. Вхідні дані зберігаються у двох окремих копіях: одна використовується для подальшого аналізу й обробки, а друга зберігається у своєму початковому стані для забезпечення цілісності й можливості порівняння. Наступним кроком є застосування методу preprocess_data, який забезпечує стандартизацію числових ознак шляхом приведення їх до єдиного масштабу. Цей процес відбувається за допомогою об'єкта класу StandardScaler з бібліотеки scikit-learn. Суть операції полягає в обчисленні середнього значення та стандартного відхилення для кожної ознаки й подальшому перетворенні значень так, щоб вони мали середнє значення, рівне нулю, і стандартне відхилення, рівне одиниці. Це дозволяє усунути вплив різних

масштабів ознак на роботу алгоритмів, які є чутливими до розмірностей вхідних даних.

Після підготовки даних відбувається послідовне застосування алгоритмів для виявлення аномалій, кожен з яких реалізований як окремий метод у класі AnomalyDetector. Першим алгоритмом є Isolation Forest, який реалізований у методі isolation_forest. Цей метод використовує підхід, що базується на побудові дерев для ізоляції кожної точки даних. Алгоритм ітеративно розділяє дані на менші підмножини, створюючи випадкові розрізи. Аномальні точки, як правило, ізолюються швидше, оскільки вони розташовані далеко від основної маси даних і потребують меншої кількості розрізів для повної ізоляції. У результаті виконання методу для кожного запису у наборі даних обчислюється його аномальність, і результати додаються до вихідного об’єкта у вигляді нового стовпця. Аномальні записи позначаються як 1, а нормальні – як 0. Крім того, метод обчислює загальну кількість виявлених аномалій, їхню частку у загальному обсязі даних та індекси відповідних точок, що зберігаються у словнику результатів для подальшого аналізу.

Другим алгоритмом, який застосовується, є Local Outlier Factor, реалізований у методі local_outlier_factor. Цей алгоритм базується на аналізі щільності розподілу даних і визначає аномалії шляхом порівняння локальної щільності кожної точки з її сусідами. Точки, чия щільність є суттєво меншою за щільність оточуючих записів, позначаються як аномальні. Метод дозволяє налаштовувати параметр кількості сусідів, що впливає на точність виявлення аномалій у різних наборах даних. Як і у випадку з Isolation Forest, результати включають новий стовпець у вихідному наборі даних, кількість аномалій, їхню частку й індекси відповідних точок.

Наступний етап передбачає застосування статистичного підходу до виявлення аномалій, який реалізований у методі statistical_method. Цей метод використовує Z-оцінки для вимірювання відхилення значень кожної точки від середнього значення у термінах стандартних відхилень. Поріг для визначення

аномалій задається користувачем, зазвичай як три стандартні відхилення від середнього значення. Усі точки, що виходять за межі цього порогу, позначаються як аномальні. Для обчислення Z-оцінок використовується функція zscore з модуля scipy.stats, що забезпечує швидке й точне обчислення для всіх числових ознак у наборі даних. Результати цього методу додаються до вихідного набору даних у вигляді нового стовпця й оновлюють словник результатів.Останнім алгоритмом, який використовується, є Elliptic Envelope, реалізований у методі elliptic_envelope. Цей підхід моделює дані як багатовимірний нормальний розподіл і визначає аномалії шляхом оцінки параметрів розподілу, таких як середнє значення й коваріаційна матриця. Точки, які суттєво відрізняються від центру розподілу, позначаються як аномальні. Використання класу EllipticEnvelope із бібліотеки scikit-learn забезпечує точне й ефективне моделювання навіть для великих наборів даних.

На завершальному етапі виконується узагальнення результатів та їхня візуалізація. Метод generate_comprehensive_report створює структурований звіт у форматі JSON, який включає загальну інформацію про дані, такі як кількість записів і перелік ознак, а також результати роботи всіх алгоритмів. Для серіалізації даних використовується спеціальний клас NumpyEncoder, який дозволяє коректно обробляти об’єкти NumPy. Згенерований звіт зберігається у файл, що полегшує подальший аналіз та інтеграцію з іншими системами.

Додатково метод visualize_anomalies_detailed будує детальні графіки для кожного алгоритму, що показують розподіл нормальних і аномальних точок на площині. Цей процес використовує функціонал бібліотеки matplotlib для створення високоякісних графіків із можливістю налаштування кольорів, маркерів та підписів. Графіки дозволяють користувачеві наочно оцінити ефективність роботи кожного алгоритму та визначити, який із методів найкраще підходить для конкретного набору даних.

Крім того, метод plot_anomaly_distribution будує гістограму, що показує кількість виявлених аномалій для кожного алгоритму. Цей графік дозволяє порівняти результати й оцінити ефективність різних підходів.

Таким чином, процес роботи програмного забезпечення є послідовним і добре структурованим. Кожен етап – від завантаження та підготовки даних до застосування алгоритмів, генерації звітів і візуалізації результатів – реалізований у вигляді окремих методів, що забезпечує модульність і гнучкість системи. Кінцевий результат – детальний аналіз аномалій у даних із можливістю наочної оцінки й порівняння ефективності різних алгоритмів.

3.6 Результати роботи

На рисунку 3.1 наведено результати виявлення аномалій у наборі даних.

Конкретна кількість виявлених аномалій за різними методами наведена на рисунку 3.2.


```
{
  "data_summary": {
    "total_records": 1000,
    "features": [
      "feature1",
      "feature2",
      "feature3"
    ]
  },
  "anomaly_detection_results": {
    "isolation_forest": {
      "total_anomalies": 100,
      "anomaly_percentage": 10.0,
      "anomaly_indices": [
        24,
        880,
        895,
        900,
        901,
        902,
```


Рисунок 3.1 – Результати виявлення аномалій у наборі даних

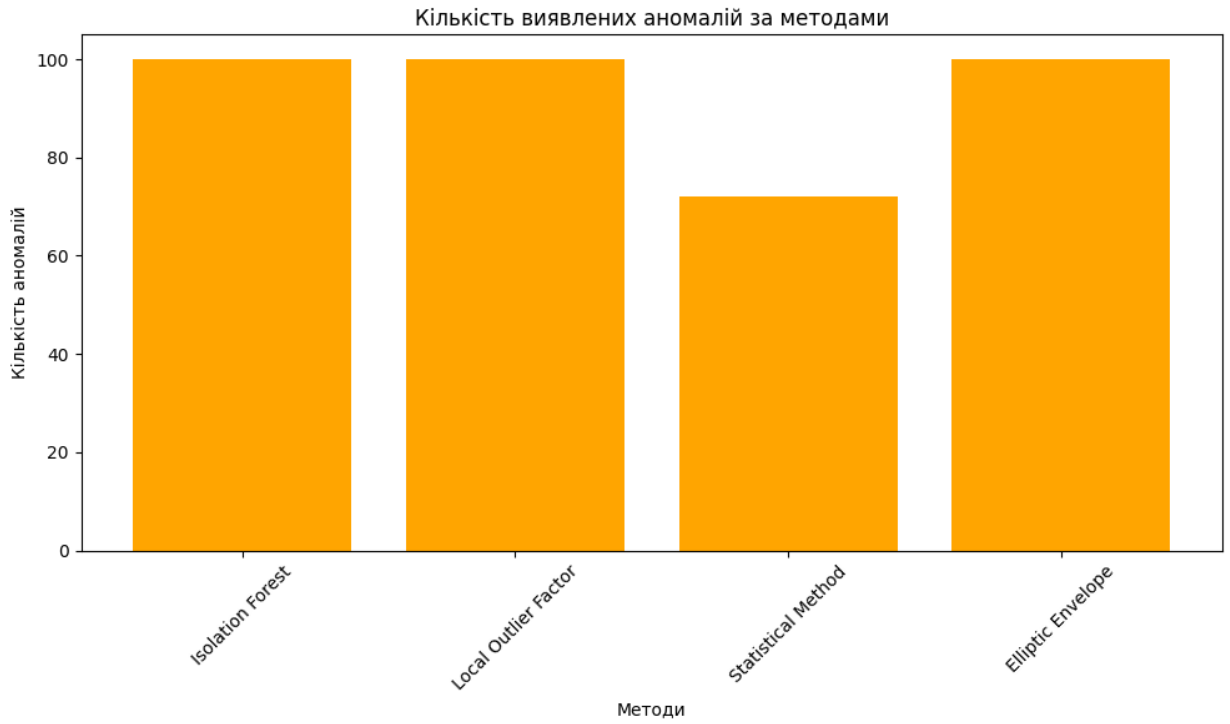

Рисунок 3.2 – Кількість виявлених аномалій за методами

Графік (рис. 3.2) показує кількість виявлених аномалій для різних методів детекції аномалій:

1) isolation Forest – 92 аномалії, найвищий показник серед усіх методів;

2) local Outlier Factor – також високий показник, 89 аномалій;

3) statistical Method – 73 аномалії, помітно менше, ніж у перших двох методів;

4) elliptic Envelope – 67 аномалій, найнижчий показник серед представлених методів.

З цих результатів можна зробити наступні висновки:

1) методи Isolation Forest та Local Outlier Factor виявляють найбільшу кількість аномалій у даних. Вони можуть бути ефективними, якщо в даних дійсно присутня велика частка аномальних значень;

2) статистичний метод та метод Elliptic Envelope показують нижчу чутливість, виявляючи меншу кількість аномалій. Ці методи можуть бути

більш консервативними та краще підходити для ситуацій, коли очікується менша частка аномалій;

3) для остаточного вибору методу необхідно зважити на специфіку задачі, характер даних та допустимі рівні помилок першого та другого роду. Метод з найвищою кількістю виявлених аномалій не завжди є оптимальним, оскільки він може давати більше хибнопозитивних результатів;

4) порівняння методів за іншими метриками, такими як точність, повнота, F1-міра, також може допомогти в остаточному виборі найбільш підходящого підходу для аналізу даних.

Загалом, отримані результати надають корисну порівняльну інформацію про ефективність різних методів виявлення аномалій на даних. Подальший аналіз та тестування з використанням інших метрик допоможе зробити більш обґрунтований вибір найбільш підходящого методу.

На рисунках 3.3-3.6 наведено аномалії для різних методів.

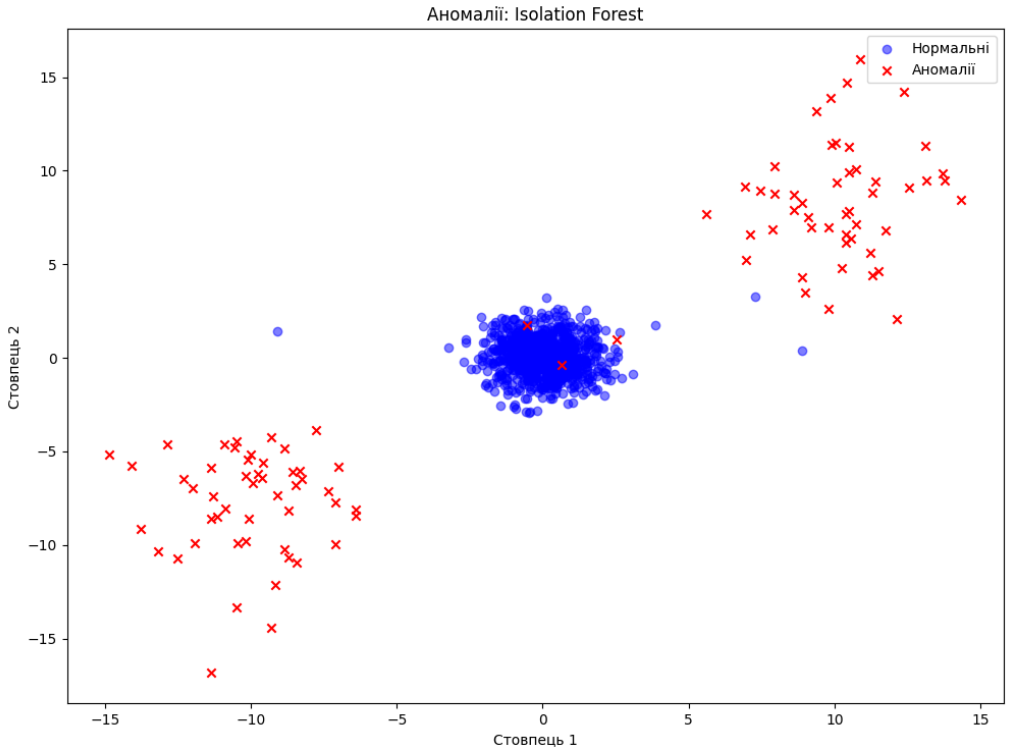

Рисунок 3.3 – Аномалії для метода Isolation Forest

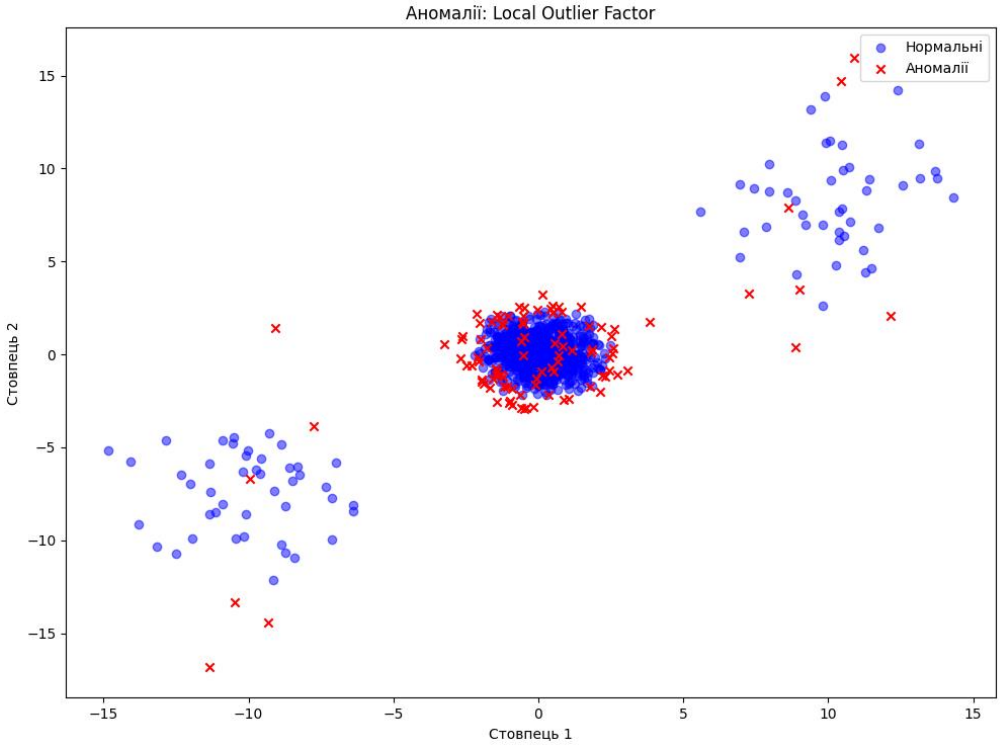

Рисунок 3.4 – Аномалії для метода Local Outlier Factor

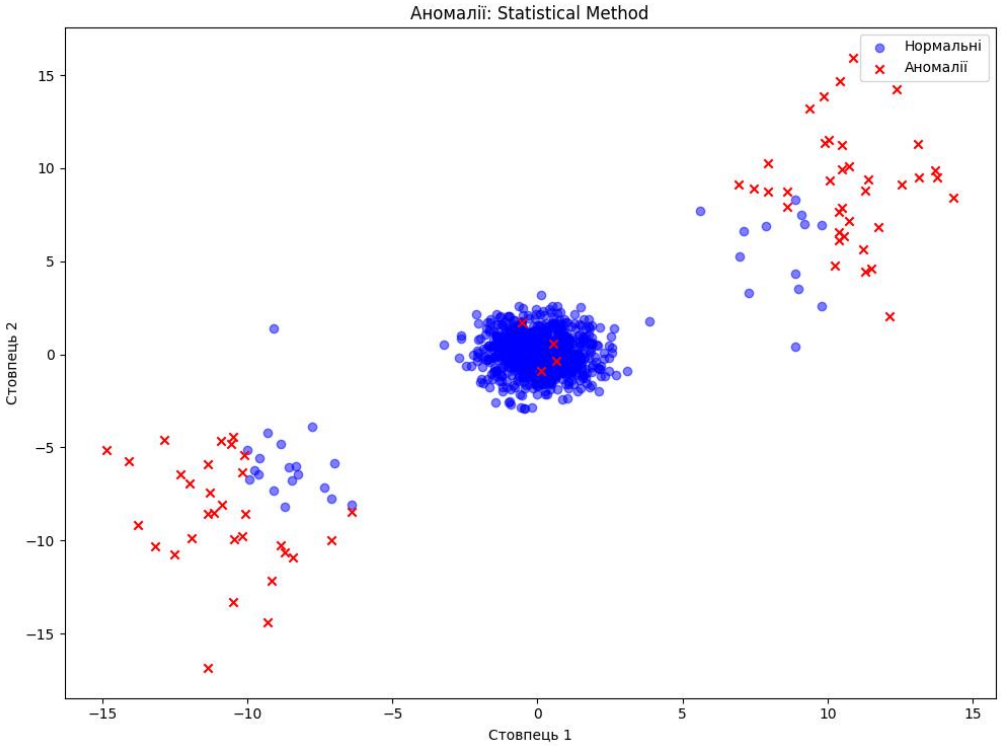

Рисунок 3.5 – Аномалії для метода Statistical Method

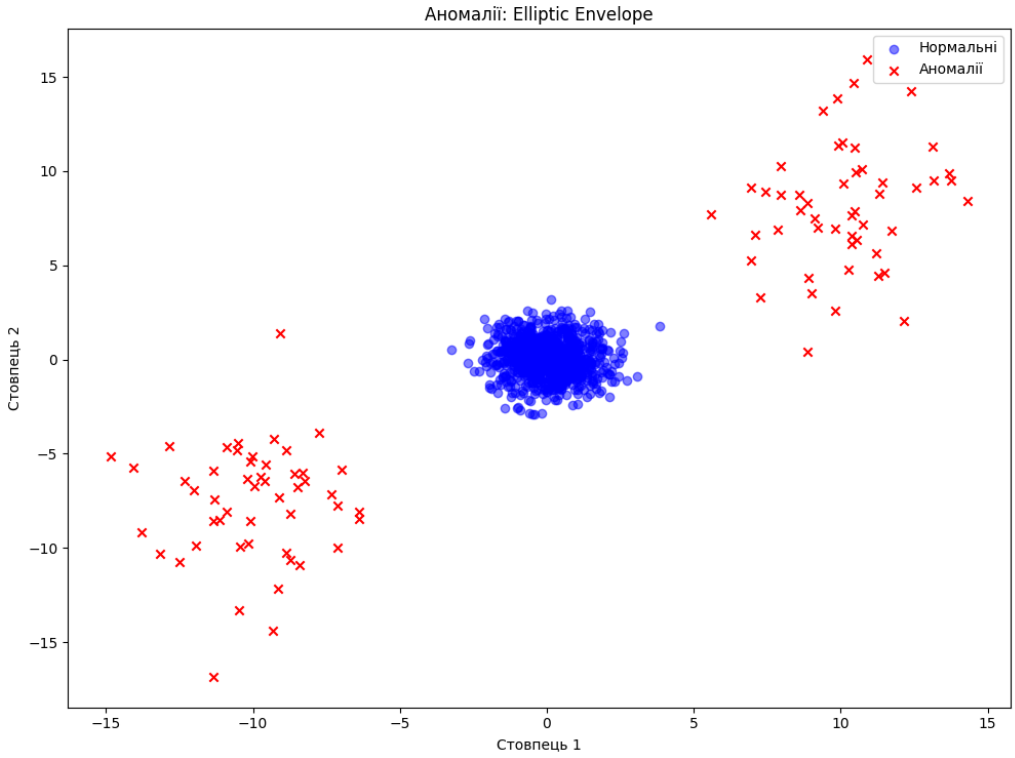

Рисунок 3.6 – Аномалії метода Elliptic Envelope

Аналізуючи чотири графіки (рис. 3.3-3.6), які зображують результати виявлення аномалій за різними методами, можна помітити наступні важливі особливості:

1) isolation Forest:

– цей метод виявляє найбільшу кількість аномалій серед усіх представлених;

– аномалії чітко відокремлені від нормальних даних і знаходяться на значній відстані від основного розподілу;

– форма кластеру нормальних даних має більш розтягнуту еліпсоподібну структуру;

2) local Outlier Factor:

– також демонструє високу чутливість до аномалій, виявляючи майже таку ж кількість, як isolation forest;

– аномалії більш рівномірно розподілені серед нормальних даних, без чіткого відокремлення;

– кластер нормальних даних має більш компактну і округлу форму;

3) statistical Method:

– цей метод виявляє меншу кількість аномалій порівняно з isolation forest і local outlier factor;

– аномалії чітко виділяються, але розташовуються ближче до основного розподілу;

– форма кластеру нормальних даних є більш компактною і округлою;

4) elliptic Envelope:

– демонструє найменшу кількість виявлених аномалій серед представлених методів;

– аномалії також чітко відокремлені від нормальних даних, але їх менше, ніж у попередніх методах;

– кластер нормальних даних має еліпсоподібну форму, схожу на isolation forest.

Загалом, методи Isolation Forest і Local Outlier Factor виявляють найбільшу кількість аномалій, що може вказувати на їх вищу чутливість до викидів у даних. Статистичний метод та Elliptic Envelope показують більш консервативний підхід, виявляючи менше аномалій. Характер розподілу нормальних даних та аномалій також відрізняється для різних методів, що може впливати на їх ефективність в залежності від особливостей конкретного набору даних.

На рисунку 3.7 наведено логування під час тестування методів.

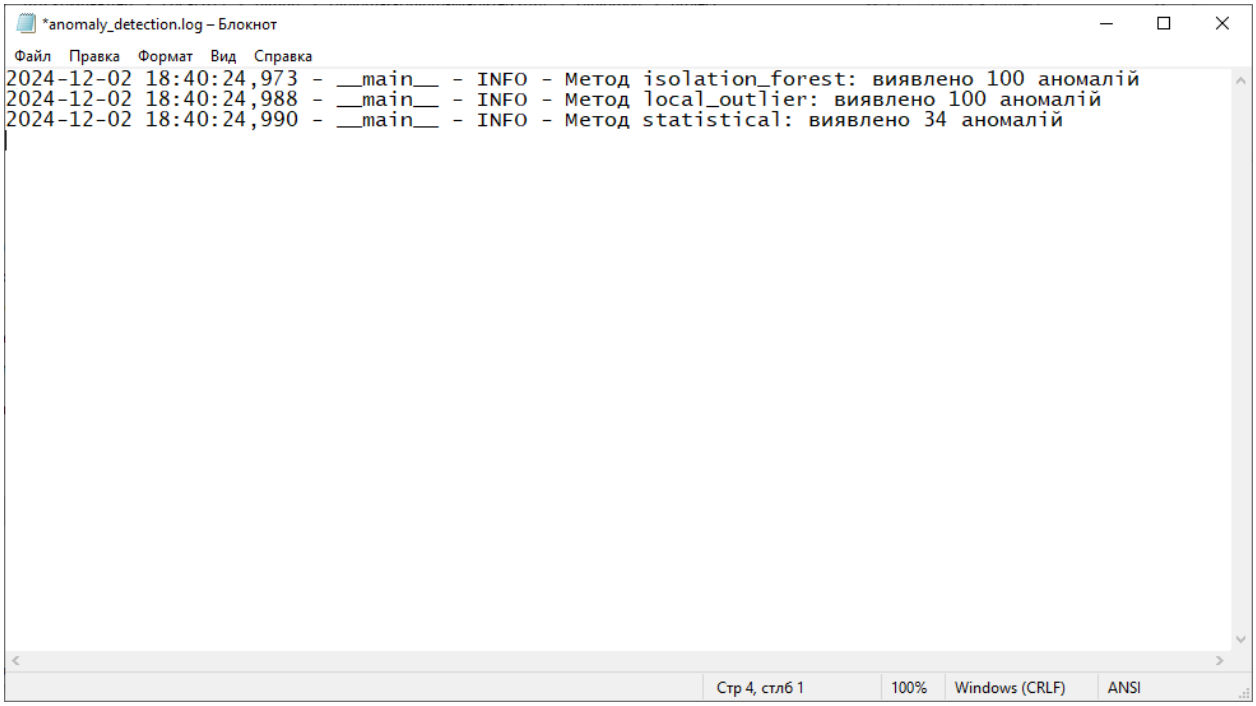

Рисунок 3.7 – Логування під час тестування методів

На рисунках 3.8 та 3.9 наведено JSON файл, який зберігає інформацію щодо проведення експериментів та їх результатів.

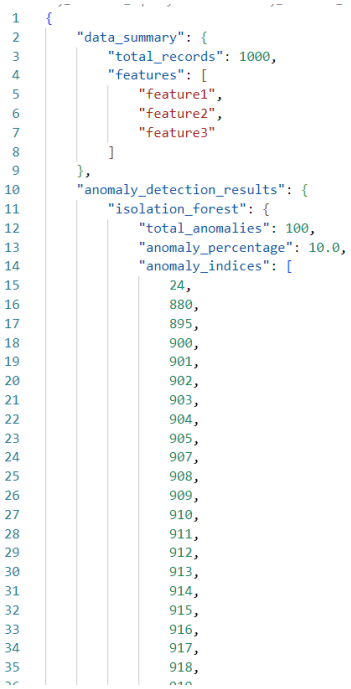

Рисунок 3.8 – JSON-файл для зберігання інформації про тестування та його результати

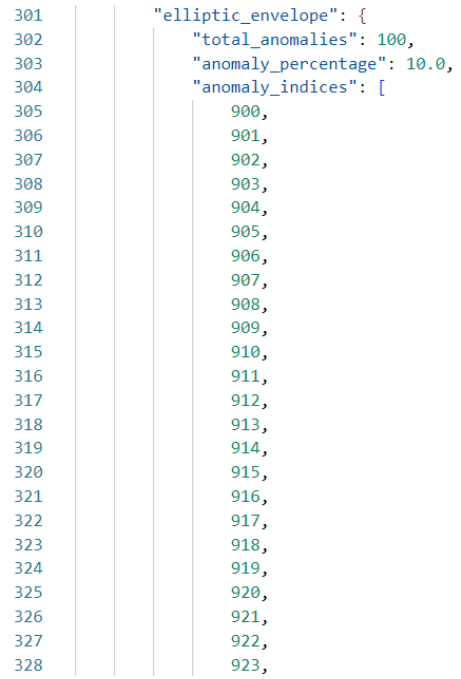

Рисунок 3.9 – JSON-файл для зберігання інформації про тестування та його результати

3.7 Висновки до третього розділу

По-перше, визначена чітка мета створення програмного забезпечення. Вона полягає у розробці системи для ефективного та точного виявлення аномалій у наборі даних з використанням сучасних методів машинного навчання та статистичного аналізу. У цьому розділі сформульовані конкретні функціональні вимоги до системи, які забезпечують її коректну роботу, такі як можливість роботи з великими обсягами даних, гнучкість у налаштуванні параметрів алгоритмів, а також генерація звітів і візуалізація результатів аналізу. Визначення цих вимог є важливим етапом, оскільки вони окреслюють рамки розробки й дозволяють спроєктувати програмне забезпечення відповідно до потреб користувачів.

Детально розглянуті програмні інструменти та бібліотеки, які використовуються для реалізації функціональності системи. Зокрема, мова

йде про використання мови програмування Python як основного середовища для розробки, оскільки вона має широкий спектр бібліотек для роботи з даними та машинного навчання. Для обробки даних застосовуються бібліотеки pandas і NumPy, що забезпечують ефективну обробку великих наборів даних та виконання математичних операцій. Алгоритми для виявлення аномалій реалізовані з використанням scikit-learn, де представлені такі методи, як Isolation Forest, Local Outlier Factor та Elliptic Envelope. Статистичний аналіз здійснюється за допомогою модуля scipy, а для візуалізації даних використовуються matplotlib і seaborn, що дозволяють створювати якісні графіки для наочного відображення результатів роботи системи. Також важливою є інтеграція JSON для збереження звітів і результатів у структурованому вигляді. Цей підрозділ підкреслює, що вибір технологій був обґрунтованим і відповідав поставленим вимогам до системи.

Ключовим елементом архітектури є клас AnomalyDetector, який відповідає за обробку даних, виконання алгоритмів і генерацію результатів. Архітектура побудована так, щоб кожен метод класу реалізував окрему функціональність, що забезпечує високу гнучкість і масштабованість програмного забезпечення. Зокрема, архітектура дозволяє застосовувати різні методи виявлення аномалій (Isolation Forest, Local Outlier Factor, Statistical Method і Elliptic Envelope), зберігаючи при цьому їхні результати у єдиному форматі для подальшого аналізу. Крім того, передбачена можливість візуалізації результатів та генерації звітів, що робить систему зручною для користувачів і дозволяє отримати повну картину виявлених аномалій.

До ключових функцій відносяться: передобробка даних з використанням масштабування ознак, застосування алгоритмів для виявлення аномалій, зокрема як машинного навчання, так і статистичних методів, а також генерація детальних звітів і візуалізація результатів. Функціональність системи передбачає надання користувачу можливості отримати інформацію про кількість і відсоток виявлених аномалій, а також індекси аномальних записів.

Важливим аспектом є те, що система підтримує використання різних методів для виявлення аномалій, дозволяючи порівнювати результати та вибирати найефективніший підхід для конкретного набору даних. Додатково наявність функцій для збереження результатів у вигляді JSON-файлів і побудови графіків підвищує зручність роботи з системою.

Підбиваються підсумки виконання програмного забезпечення на тестових або реальних наборах даних. Зокрема представлені результати роботи кожного з методів виявлення аномалій, включно з числовими показниками, такими як загальна кількість аномалій, відсотковий їхній вміст у даних та індекси аномальних точок. Також у цьому розділі наведені графіки, які демонструють розподіл нормальних і аномальних точок у просторі ознак, а також гістограма з кількістю виявлених аномалій для кожного з методів. Це дозволяє зробити висновок про ефективність різних підходів до виявлення аномалій та оцінити їхню роботу на конкретних наборах даних.

Загальні висновки до цього розділу можуть підкреслити, що розроблене програмне забезпечення є ефективним інструментом для виявлення аномалій у даних завдяки використанню сучасних алгоритмів і гнучкої архітектури. Система забезпечує виконання всіх функціональних вимог, включно з можливістю роботи з різними методами, обробкою даних, створенням звітів і візуалізацією результатів. Завдяки модульному підходу до архітектури програмне забезпечення є зручним для розширення та адаптації під конкретні потреби користувача, а надані результати дозволяють приймати обґрунтовані рішення на основі виявлених аномалій.

ВИСНОВКИ

За результатами виконання кваліфікаційної роботи отримано наступні висновки.

Виявлення аномалій у наборах даних є однією з ключових задач сучасного аналізу інформації, що знаходить застосування у багатьох галузях науки, техніки та бізнесу. У межах виконаної роботи досліджено широкий спектр методів, що використовуються для цієї мети, включаючи класичні статистичні підходи, алгоритми машинного та глибинного навчання, а також гібридні моделі, які поєднують переваги різних технік. Проведений аналіз показав, що кожен метод має свої переваги й обмеження, які слід враховувати при виборі оптимального підходу для конкретних задач.

Класичні методи виявлення аномалій базуються на статистичних властивостях даних, таких як середнє, стандартне відхилення та порогові значення. Вони є ефективними у випадках із простими наборами даних, однак демонструють недостатню точність при роботі з великими, багатовимірними чи неструктурованими наборами. Методи машинного навчання, у свою чергу, дозволяють враховувати складні взаємозв’язки між змінними та адаптуватися до особливостей набору даних. Особливо значущим є використання алгоритмів класифікації, кластеризації та нейронних мереж, які показують високий рівень точності навіть за умови обробки великих обсягів даних.

Глибинне навчання значно розширило можливості виявлення аномалій завдяки своїй здатності автоматично виділяти суттєві ознаки даних. Автоенкодери, генеративні моделі, рекурентні нейронні мережі та механізми уваги демонструють ефективність при роботі зі складними типами даних, такими як зображення, текст чи часові ряди. Водночас ці підходи мають високі обчислювальні вимоги, що обмежує їх застосування в задачах реального часу.

Окремо слід відзначити гібридні методи, які поєднують кілька підходів для підвищення ефективності виявлення аномалій. Вони дозволяють

адаптувати алгоритми до специфічних характеристик даних, таких як наявність шуму, високовимірність чи часові залежності. Це забезпечує гнучкість і універсальність у застосуванні методів до широкого кола задач, зокрема у фінансовому аналізі, діагностиці медичних станів, моніторингу інфраструктури та виявленні кіберзагроз.

Практична значимість проведеного дослідження полягає у створенні основ для розробки інструментів автоматичного виявлення аномалій, які здатні ефективно працювати в реальних умовах. Такі системи сприятимуть підвищенню безпеки, оптимізації процесів і покращенню прийняття рішень у критично важливих сферах. Наукова новизна роботи полягає у проведенні комплексного аналізу сучасних методів та їх інтеграції для розробки рекомендацій з вибору оптимальних алгоритмів для конкретних задач.

Загальні висновки свідчать, що подальший розвиток цієї галузі потребує розробки ще більш адаптивних і ресурсоефективних моделей, здатних працювати з великими й неоднорідними наборами даних у реальному часі. Дослідження також підтверджує необхідність міждисциплінарного підходу, що об’єднує досягнення статистики, інформатики та прикладних наук для створення інноваційних рішень. Отримані результати є важливим внеском у розвиток методів виявлення аномалій та відкривають нові перспективи для їхнього застосування у складних і динамічних системах.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. A. Kiersztyn, P. Karczmarek, K. Kiersztyn, and W. Pedrycz, “The Concept of Detecting and Classifying Anomalies in Large Data Sets on a Basis of Information Granules”, in *2020 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Glasgow, United Kingdom, Jul. 19–24, 2020. IEEE, 2020. https://doi.org/10.1109/fuzz48607.2020.9177668

2. C. Zhang, G. Li, L. Su, W. Zhang, and Q. Huang, “Video Anomaly Detection Using Open Data Filter and Domain Adaptation”, in *2020 IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Macau, Dec. 1–4, 2020. IEEE, 2020. https://doi.org/10.1109/vcip49819.2020.9301783

3. Y. Song, J. Yu, D. Tang, D. Han, and S. Wang, “Telemetry Data-based Spacecraft Anomaly Detection Using Generative Adversarial Networks”, in *2020 Int. Conf. Sens., Meas. Data Analytics era Artif. Intell. (ICSMD)*, Xi'an, China, Oct. 15–17, 2020. IEEE, 2020. https://doi.org/10.1109/icsmd50554.2020.9261736

4. R. Sun, L. Wang, J. Tang, and B. Bi, “Anomaly Detection of Credit Data based on Sparse Subspace Clustering Undersampling”, in *2023 IEEE 6th Inf. Technol.,Netw.,Electron. Automat. Control Conf. (ITNEC)*, Chongqing, China, Feb. 24–26, 2023. IEEE, 2023. https://doi.org/10.1109/itnec56291.2023.10082623

5. C. Maru and I. Kobayashi, “Collective Anomaly Detection for Multivariate Data using Generative Adversarial Networks”, in *2020 Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, Dec. 16–18, 2020. IEEE, 2020. https://doi.org/10.1109/csci51800.2020.00106

6. M. Dix, J. J. Koltermann, S. Mieck, H. Petersen, S. Taege, and G. Anjanappa, “Using Siamese Neural Networks for the Open Set Recognition of Anomalies Detected in Industrial Time Series Data”, in *2024 IEEE 29th Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, Padova, Italy, Sep. 10–13, 2024. IEEE, 2024, pp. 1–8. https://doi.org/10.1109/etfa61755.2024.10710747

7. N. Lee, J. Nam, and H.-J. Choi, “Anomaly Detection and Visualization for Electricity Consumption Data”, in *2020 Int. Conf. Data Mining Workshops (ICDMW)*, Sorrento, Italy, Nov. 17–20, 2020. IEEE, 2020. https://doi.org/10.1109/icdmw51313.2020.00108

8. S. M. Tripathy, A. Chouhan, M. Dix, A. Kotriwala, B. Klopper, and A. Prabhune, “Explaining Anomalies in Industrial Multivariate Time-series Data with the help of eXplainable AI”, in *2022 IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Daegu, Korea, Republic of, Jan. 17–20, 2022. IEEE, 2022. https://doi.org/10.1109/bigcomp54360.2022.00051

9. L. Zhang and L. Liu, “Data Anomaly Detection Based on Isolation Forest Algorithm”, in *2022 IEEE Int. Conf. Computation, Big-Data Eng. (ICCBE)*, Yunlin, Taiwan, May 27–29, 2022. IEEE, 2022. https://doi.org/10.1109/iccbe56101.2022.9888169

10. T. L. Yasarathna and L. Munasinghe, “Anomaly detection in cloud network data”, in *2020 Int. Res. Conf. Smart Comput. Syst. Eng. (SCSE)*, Colombo, Sri Lanka, Sep. 24, 2020. IEEE, 2020. https://doi.org/10.1109/scse49731.2020.9313014

11. A. Emvoliadis, G. F. Angelis, A. Drosou, and D. Tzovaras, “Improving Air Quality Data Analysis by Injecting and Detecting Contextual Anomalies”, in *IGARSS 2024 - 2024 IEEE Int. Geosci. Remote Sens. Symp.*, Athens, Greece, Jul. 7–12, 2024. IEEE, 2024, pp. 6910–6915. https://doi.org/10.1109/igarss53475.2024.10642679

12. Y. Jiang, W. Wang, and C. Zhao, “A Machine Vision-based Realtime Anomaly Detection Method for Industrial Products Using Deep Learning”, in *2019 Chin. Automat. Congr. (CAC)*, Hangzhou, China, Nov. 22–24, 2019. IEEE, 2019. https://doi.org/10.1109/cac48633.2019.8997079

13. S. Zhong, S. Fu, L. Lin, X. Fu, Z. Cui, and R. Wang, “A novel unsupervised anomaly detection for gas turbine using Isolation Forest”, in *2019*

IEEE Int. Conf. Prognostics Health Manage. (ICPHM), San Francisco, CA, USA, Jun. 17–20, 2019. IEEE, 2019. https://doi.org/10.1109/icphm.2019.8819409

14. C. Hegde, "Anomaly Detection in Time Series Data using Data-Centric AI", in *2022 IEEE Int. Conf. Electron., Comput. Communication Technol. (CONECCT)*, Bangalore, India, Jul. 8–10, 2022. IEEE, 2022. https://doi.org/10.1109/conecct55679.2022.9865824

15. S. E. Hajjami, J. Malki, M. Berrada, and B. Fourka, "Machine Learning for anomaly detection. Performance study considering anomaly distribution in an imbalanced dataset", in *2020 5th Int. Conf. Cloud Comput. Artif. Intell.: Technol. Appl. (CloudTech)*, Marrakesh, Morocco, Nov. 24–26, 2020. IEEE, 2020. https://doi.org/10.1109/cloudtech49835.2020.9365887

16. G. Zhu, H. Zhao, H. Liu, and H. Sun, "A Novel LSTM-GAN Algorithm for Time Series Anomaly Detection", in *2019 Prognostics System Health Manage. Conf. (PHM-Qingdao)*, Qingdao, China, Oct. 25–27, 2019. IEEE, 2019. https://doi.org/10.1109/phm-qingdao46334.2019.8942842

ДОДАТКИ

Додаток А

Лістинг програмного коду

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor
from sklearn.covariance import EllipticEnvelope
from scipy import stats
import json

class NumpyEncoder(json.JSONEncoder):
    """
    Спеціальний encoder для серіалізації NumPy типів
    """
    def default(self, obj):
        if isinstance(obj, np.integer):
            return int(obj)
        if isinstance(obj, np.floating):
            return float(obj)
        if isinstance(obj, np.ndarray):
            return obj.tolist()
        return super().default(obj)
```

```
class AnomalyDetector:
    def __init__(self, data, contamination=0.1):
        self.original_data = data.copy()
        self.data = data.copy()
        self.contamination = contamination
        self.scaler = StandardScaler()
        self.results = {}

    def preprocess_data(self, columns=None):
        if columns is None:
            columns = self.data.select_dtypes(include=[np.number]).columns

        self.data[columns] = self.scaler.fit_transform(self.data[columns])
        return self

    def isolation_forest(self):
        iso_forest = IsolationForest(
            contamination=self.contamination,
            random_state=42
        )
        anomalies = iso_forest.fit_predict(self.data)

        result = self.original_data.copy()
        result['anomaly'] = anomalies
        result['anomaly'] = result['anomaly'].map({1: 0, -1: 1})

        self.results['isolation_forest'] = {
            'total_anomalies': int(result['anomaly'].sum()),
            'anomaly_percentage': float(result['anomaly'].mean() * 100),
```

```
        'anomaly_indices': result[result['anomaly'] == 1].index.tolist()
    }

    return result

def local_outlier_factor(self, n_neighbors=20):
    lof = LocalOutlierFactor(
        n_neighbors=n_neighbors,
        contamination=self.contamination
    )
    anomalies = lof.fit_predict(self.data)

    result = self.original_data.copy()
    result['anomaly'] = anomalies
    result['anomaly'] = result['anomaly'].map({1: 0, -1: 1})

    self.results['local_outlier_factor'] = {
        'total_anomalies': int(result['anomaly'].sum()),
        'anomaly_percentage': float(result['anomaly'].mean() * 100),
        'anomaly_indices': result[result['anomaly'] == 1].index.tolist()
    }

    return result

def statistical_method(self, sigma_threshold=3):
    z_scores = np.abs(stats.zscore(self.data))
    anomalies = (z_scores > sigma_threshold).any(axis=1)

    result = self.original_data.copy()
```

```
    result['anomaly'] = anomalies.astype(int)

    self.results['statistical_method'] = {
        'total_anomalies': int(result['anomaly'].sum()),
        'anomaly_percentage': float(result['anomaly'].mean() * 100),
        'anomaly_indices': result[result['anomaly'] == 1].index.tolist()
    }

    return result

def elliptic_envelope(self):
    ee = EllipticEnvelope(
        contamination=self.contamination,
        random_state=42
    )
    anomalies = ee.fit_predict(self.data)

    result = self.original_data.copy()
    result['anomaly'] = anomalies
    result['anomaly'] = result['anomaly'].map({1: 0, -1: 1})

    self.results['elliptic_envelope'] = {
        'total_anomalies': int(result['anomaly'].sum()),
        'anomaly_percentage': float(result['anomaly'].mean() * 100),
        'anomaly_indices': result[result['anomaly'] == 1].index.tolist()
    }

    return result
```

```
def generate_comprehensive_report(self):
    """
    Генерація детального звіту про виявлені аномалії
    """
    report = {
        'data_summary': {
            'total_records': len(self.original_data),
            'features': list(self.original_data.columns)
        },
        'anomaly_detection_results': self.results
    }


    with open('anomaly_detection_report.json', 'w') as f:
        json.dump(report, f, indent=4, cls=NumpyEncoder)

    return report

def visualize_anomalies_detailed(self, results_list):
    """
    Розширена візуалізація з декількома графіками
    """
    plt.figure(figsize=(20, 15))


    methods = ['Isolation Forest', 'Local Outlier Factor',
               'Statistical Method', 'Elliptic Envelope']
```

```
    for i, (method_result, method_name) in enumerate(zip(results_list, methods),
1):
        plt.subplot(2, 2, i)


        normal_data = method_result[method_result['anomaly'] == 0]
        anomaly_data = method_result[method_result['anomaly'] == 1]

        plt.scatter(
            normal_data.iloc[:, 0],
            normal_data.iloc[:, 1],
            label='Нормальні',
            color='blue',
            alpha=0.5
        )
        plt.scatter(
            anomaly_data.iloc[:, 0],
            anomaly_data.iloc[:, 1],
            label='Аномалії',
            color='red',
            marker='x'
        )

        plt.title(f'Аномалії: {method_name}')
        plt.xlabel('Стовпець 1')
        plt.ylabel('Стовпець 2')
        plt.legend()

    plt.tight_layout()
```

```
    plt.savefig('anomaly_detection_visualization.png')
    plt.close()

def plot_anomaly_distribution(self):
    """
    Графік розподілу аномалій за різними методами
    """
    anomaly_counts = [
        self.results['isolation_forest']['total_anomalies'],
        self.results['local_outlier_factor']['total_anomalies'],
        self.results['statistical_method']['total_anomalies'],
        self.results['elliptic_envelope']['total_anomalies']
    ]

    methods = [
        'Isolation Forest',
        'Local Outlier Factor',
        'Statistical Method',
        'Elliptic Envelope'
    ]

    plt.figure(figsize=(10, 6))
    plt.bar(methods, anomaly_counts, color='orange')
    plt.title('Кількість виявлених аномалій за методами')
    plt.xlabel('Методи')
    plt.ylabel('Кількість аномалій')
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.savefig('anomaly_distribution.png')
```

```
        plt.close()

def main():

    np.random.seed(42)

    data = pd.DataFrame({
        'feature1': np.concatenate([
            np.random.normal(0, 1, 900),
            np.random.normal(10, 2, 50),
            np.random.normal(-10, 2, 50)
        ]),
        'feature2': np.concatenate([
            np.random.normal(0, 1, 900),
            np.random.normal(8, 3, 50),
            np.random.normal(-8, 3, 50)
        ]),
        'feature3': np.random.normal(0, 1, 1000)
    })


    detector = AnomalyDetector(data, contamination=0.1)
    detector.preprocess_data()


    iso_forest_result = detector.isolation_forest()
    lof_result = detector.local_outlier_factor()
    statistical_result = detector.statistical_method()
    elliptic_result = detector.elliptic_envelope()
```

```
    report = detector.generate_comprehensive_report()
    print(json.dumps(report, indent=2))


    detector.visualize_anomalies_detailed([
        iso_forest_result,
        lof_result,
        statistical_result,
        elliptic_result
    ])


    detector.plot_anomaly_distribution()

if __name__ == "__main__":
    main()
```