DOI: https://doi.org/10.26565/2304-6201-2022-53-03

УДК 519.6:681.518.2:004.62:519.23

Методи виявлення викидів в пробних вибірках при управлінні процесами в системах за станом

О.Ю. Лихач, М.Л. Угрюмов , Д.О. Шевченко, С.І. Шматков

Лихач Олег Юрійович — *студент Харківського національного університету імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail: lykhach2018@gmail.com;* **https://orcid.org/0000-0002-4598-2912**.

Угрюмов Михайло Леонідович — *д.т.н., проф., професор кафедри ТПС; факультету комп'ютерних наук; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022 e-mail: ugryumov.mykhaylo52@gmail.com;* **https://orcid.org/0000-0003-0902-2735**

Шевченко Дмитро Олександрович — *студент Харківського національного університету імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022 e-mail: dimyich24@gmail.com;* **https://orcid.org/0000-0002-7897-250X**

Шматков Сергій Ігорович — *д.т.н., проф., завідувач кафедрою ТПС; факультету комп'ютерних наук; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022; e-mail: sershmat@gmail.com;* **https://orcid.org/0000-0002-0298-7174**.

Розроблене на сьогоднішній день інформаційне забезпечення не дозволяє з досить високим рівнем достовірності вирішувати завдання виявлення викидів у вибірках даних та часових рядах.

Тому, ця робота присвячена вибору метрик для оцінювання правильності виявлення викидів, найкращих математичних моделей та методів для вирішення проблеми виявлення викидів в пробних вибірках при управлінні процесами в системах за станом.

Були використані математичні моделі та методи виявлення викидів (аномальних значень) та програмні засоби на основі мови Python, такі як scikit-learn, Tensorflow, NumPy, Pandas і інші.

В ході виконання роботи було отримано: огляд метрик, які використовуються для оцінки ефективності математичних моделей та методів виявлення викидів; огляд традиційних методів та методів глибокого навчання для виявлення викидів; результати дослідження, щодо ефективності та якості математичних моделей і методів виявлення викидів, використовуючи 12 пробних вибірок; висновки про найкращу метрику та найкращі математичні моделі і методи для вирішення проблеми виявлення викидів в пробних вибірках при управлінні процесами в системах за станом.

Головним напрямком використання обраних методів є моніторинг рівня аномальних значень в різних вибірках при управлінні процесами в системах за станом, що робить ці методи універсальними для використання.

Ключові слова: *виявлення викидів, машинне навчання, управління процесів, метрики оцінки якості, глибоке навчання*

Anomaly detection methods in sample datasets when managing processes in systems by the state

Lykhach Oleh — *stydent V.N. Kaeazin National Universite, Svobody Sq 6, Kharkiv, Ukraine, 61022; e-mail: lykhach2018@gmail.com;* **https://orcid.org/0000-0002-4598-2912**

Ugryumov Mykhaylo *Doctor of Technical Sciences, Professor, Professor of the Department of Theoretical&Applied Systems; Faculty of Computer Science; VN Kharkiv National University Karazina, Maidan Svobody, 4, Kharkiv-22, Ukraine, 61022;*
e-mail: ugryumov.mykhaylo52@gmail.com;
https://orcid.org/0000-0003-0902-2735

Shevchenko Dmytro *stydent V.N. Karazin National Universite, Svobody Sq 6, Kharkiv, Ukraine, 61022;*
e-mail: dimyich24@gmail.com;
https://orcid.org/0000-0002-7897-250X
.

Shmatkov Sergei *Doctor of Technical Sciences, Professor, Head of the Department of Theoretical&Applied; Faculty of Computer Science; VN Kharkiv National University Karazina, Maidan Svobody, 4, Kharkiv-22, Ukraine, 61022;*
e-mail: sershmat@gmail.com;
https://orcid.org/0000-0002-0298-7174

The current information software does not allow solving the problems of detecting outliers in data samples and time series with a sufficiently high level of reliability.

Therefore, this work is devoted to the choice of metrics for assessing the correctness of detecting outliers, as well as the best mathematical models and methods for solving the problem of detecting outliers in test samples when managing processes in systems by state. Mathematical models and methods for detecting outliers (anomalous values) and Python-based software tools such as scikit-learn, Tensorflow, NumPy, Pandas and others have been used.

The results of our work are the overview of the metrics used to assess the effectiveness of mathematical models and methods for detecting outliers; the overview of traditional and deep learning techniques of detecting outliers; the results of researching the efficiency and quality of mathematical models and methods for detecting outliers using 12 datasets; the conclusions about the best metric and the best mathematical models and methods for solving the problem of detecting outliers in test samples when managing processes in systems by state.

The selected methods are mainly used for monitoring the level of anomalous values in various datasets when managing processes in systems by state, which makes these methods universal.

Keywords: *outlier detection, machine learning, process control, quality assessment metrics, deep learning*

1 Постановка проблеми та її актуальність

Будемо розглядати в якості об’єкту дослідження управління процесів в системах за їх станом, наприклад, економічних систем, заснованих на даних моніторингу контрольованих змінних стану. Результати моніторингу – це вибірки даних та часові ряди. Вибірка даних представляє собою набір значень за певній проміжок часу, які можуть суттєво змінюватись в залежності від ситуації в системі. В той час як часові ряди являють собою сукупність виміряних значень змінних, одержуваних на певних інтервалах часу, що нерозривно примикають один до одного та протягом яких значення змінних істотно не змінюються. Часові ряди, будучи дискретною моделлю контролю стану динамічних систем, зазвичай містять параметричну невизначеність, є нестаціонарними і зашумленими.

При розв’язанні задачі виявлення викидів у вибірках даних та часових рядах потрібно попередньо обробити вхідні дані та видалити пропущені значення для того, щоб можна було використати методи виявлення викидів. Після того, як вхідні дані були оброблені та пропущені значення були видалені потрібно знайти викиди (також відомі, як аномалії) і видалити їх з наборів даних та часових рядів. Це дозволить підвищити рівень достовірності інформації та покращити управління процесів в системі.

Задача виявлення викидів в результаті її декомпозиції повинна бути представлена, як послідовність вирішення взаємопов’язаних задач таких як:

- моніторинг стану системи (вибір та вимірювання значень контрольованих змінних стану системи через певні проміжки часу);
- попередня обробка даних, яка призводить до приведення даних моніторингу до виду, придатного для виявлення викидів;

- знаходження аномальних значень у часових рядах та вибірках даних моніторингу системи.

Розгляду завдань теорії та практики управління процесів у динамічних системах приділяється велика увага як науковцям в Україні, так і за її межами. На цей час опубліковано безліч робіт, присвячених опису математичних моделей та методів виявлення викидів у процесах управління систем за станом [1-15].

Розглядаючи класифікацію методів (моделей) виявлення викидів в системах можна визначити, що їх можна поділити на традиційні методи та методи глибокого навчання. Класифікація цих моделей та методів виявлення викидів представлена на рис. 1.1-1.2.

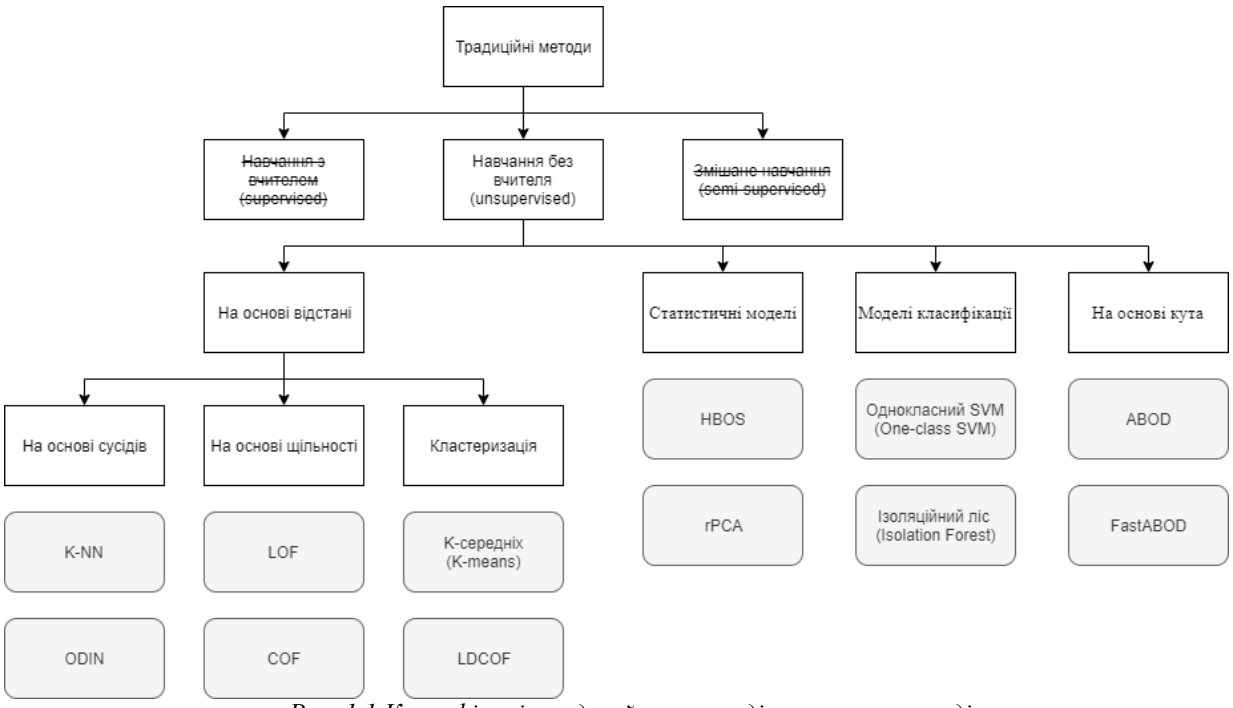

Рис. 1.1 Класифікація моделей та методів виявлення викидів

Можна виділити вищу ланку ієрархії моделей та методів виявлення викидів – це традиційні моделі та методи глибокого навчання (рис. 1.2). Серед них можна виділити чотири основних типи моделей: моделі на основі відстані, статистичні моделі, моделі класифікації та моделі на основі кута. Аналізуючи типи моделей для глибокого навчання можна виділити три типи навчання: глибоке навчання для вилучення функцій, навчання особливостей представлення нормальності та наскрізне навчання для визначення викидів.

До множини методів, моделей на основі відстані відносяться наступні: K-NN[16], ODIN [17], LOF [18] та K-means [19]. Серед статистичних моделей можна виділити: HBOS [20] та rPCA [21]. Прикладами моделей класифікації для виявлення викидів є: SVM [22] та Isolation Forest [23]. До методів, моделей побудованих на основі кута можна віднести ABOD [24] та FastABOD.

До множини методів, моделей на основі відстані відносяться наступні: K-NN[16], ODIN [17], LOF [18] та K-means [19]. Серед статистичних моделей можна виділити: HBOS [20] та rPCA [21]. Прикладами моделей класифікації для виявлення викидів є: SVM [22] та Isolation Forest [23]. До методів, моделей побудованих на основі кута можна віднести ABOD [24] та FastABOD.

Аналіз існуючих літературних джерел показує, при розробці математичних моделей та методів вирішення завдань виявлення викидів у вибірках даних та часових рядах виникає низка проблем:

- невизначеність вхідних даних (обмежений обсяг вибірок, наявність пропущених значень, корельованість змінних станів);
- велика розмірність множини змінних стану;
- невизначеність у виборі форми приведення вхідних даних до нормального вигляду, придатного для моделей виявлення викидів;
- невизначеність у виборі критеріїв якості математичних моделей та методів;
- невизначеність вибору правильних результатів рішень щодо аномальних значень.

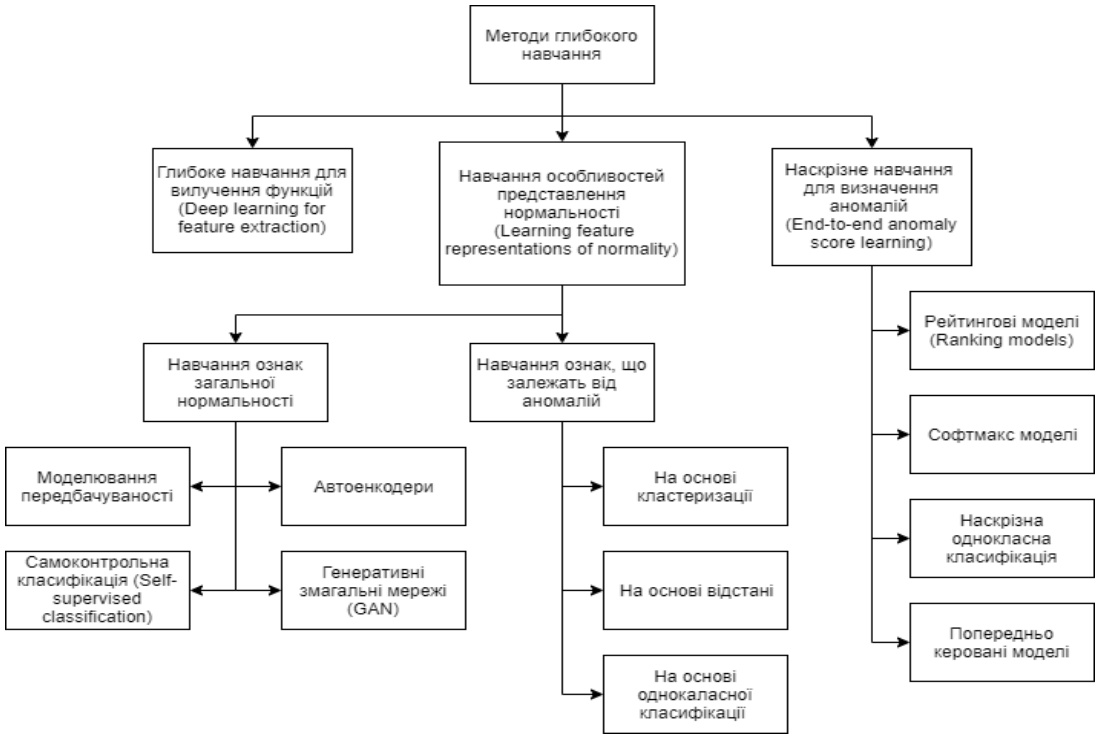

Рис. 1.2 Класифікація методів глибокого навчання для виявлення викидів

Слід зазначити, що у роботах, присвячених вирішенню завдань виявлення викидів, відсутнє визначення щодо найкращих серед існуючих математичної моделі або методу для виявлення аномалій та не враховується метрика для визначення точності цих моделей.

Розроблене на сьогоднішній день інформаційне забезпечення не дозволяє з досить високим рівнем достовірності вирішувати завдання виявлення викидів у вибірках даних та часових рядах.

Таким чином, виникає потреба у виборі метрик для оцінювання правильності виявлення викидів, найкращих математичних моделей, методів та засобів реалізації інформаційної технології при виявленні аномалій у часових рядах та вибірках даних при управлінні процесів в системах за їх станом.

Ця робота присвячена вибору метрик, найкращих математичних моделей та методів для вирішення проблеми виявлення викидів в пробних вибірках при управлінні процесами в системах за станом.

2 Постановка завдання виявлення викидів в пробних вибірках

Наділі будемо розглядати пробні вибірки в яких присутні аномалії (або викиди). Аномалії і викиди – це два терміни, які найчастіше використовуються в контексті виявлення аномалій; іноді взаємозамінні. Аномалії або викиді – це екземпляри даних, які виділяються і несхожі на інші

Необхідно отримати функціональну залежність, яка буде відображати зв’язок викидів, реальних даних та пробних вибірок. А також визначити якість цієї залежності за допомогою єдиної метрики.

Результатом вирішення задачі повинен бути математична модель (або метод) при використанні яких можливо отримати результати, які вказували би на приналежність певного спостереження до викидів, використовуючи лише дані із пробної вибірки.

3 Метрика для оцінювання якості виявлення викидів

Метрика в загальному сенсі машинного навчання – це еталон вимірювання якості математичного методу або моделі. Існує проблема у виборі єдиної метрики для визначення якості математичних методів, моделей виявлення викидів у пробних вибірках.

В даний час є велика кількість метрик, тому розглянемо популярні метрики, які підтвердили свою використовуваність за довгі роки у задачах виявлення викидів.

1. Точність (Accuracy) – це найпростіший та інтуїтивно зрозумілий показник продуктивності класифікатора. Він розраховується по формулі 3.1, як відношення кількості правильно передбаченого класу до загальної кількості передбачень:

$$\text{Точність} = \frac{TP+TN}{T}, \tag{3.1}$$

де TP – кількість значень правильної класифікації позитивного класу, TN – кількість значень правильної класифікації негативного класу, T – загальна кількість передбачень.

Коли ми стикаємося з проблемою дисбалансу класів, точність є неправильною метрикою для використання. Наприклад, нехай існує 2 класи: клас A становить 99% набору даних, а клас B – це решта 1%. Якщо прогнозувати клас A кожного випадку, буде досягнута точність 99%. За метрикою точності можна вважати, що модель працює чудово, але насправді модель працює дуже погано.

2. Точність (Precision) і відклик (Recall). Загалом, існує компроміс між відкликом (відсоток дійсно позитивних випадків, які були класифіковані як такі) і точністю (відсоток позитивних класифікацій, які дійсно позитивні). Ці метрики можна розрахувати за формулами 3.2-3.3:

$$\text{Відклик} = \frac{TP}{TP+FN}, \tag{3.2}$$

де TP – кількість значень правильної класифікації позитивного класу, FN – кількість значень неправильної класифікації негативного класу.

$$\text{Точність} = \frac{TP}{TP+FP}, \tag{3.3}$$

де TP – кількість значень правильної класифікації позитивного класу, FP – кількість значень неправильної класифікації позитивного класу.

У ситуаціях, коли потрібно виявити екземпляри класу меншості, зазвичай більш важливою метрикою є відклик, ніж точність. Але, як сказано раніше, повинен бути компроміс між цими метриками.

3. Метрика F. Ця метрика використовується в тих випадках, коли потрібно досягнути високого значення точності (Precision) і відклику (Recall) та отримати лише один показник якості. F1 можна розрахувати за формулою:

$$F1 = \frac{2*precision*recall}{precision+recall}, \tag{3.4}$$

де precision – точність, recall – відклик.

Метрику F1 не можна коригувати в залежності від потреб класифікації, тому виник окремий випадок метрики, відомої як F-Beta, яка має, корегувальний параметр β. Цей параметр надає змогу корегувати значення відклику та точності в такій залежності, що чим більше значення β тим більша залежність метрики від відклику і тим менша від точності. Дана метрика розраховується за формулою:

$$F_\beta = (1+\beta^2) * \frac{precision*recall}{\beta^2*precision+recall}, \tag{3.5}$$

де precision – точність, recall – відклик, β – параметер корегування.

4. Метрика Kappa. Kappa або Cohen's Kappa схожа на точність класифікації, за винятком того, що вона нормалізується на базовій лінії випадкових шансів у наборі даних. Метрика розраховується за формулою:

$$k = \frac{p_0-p_e}{1-p_e}, \tag{3.6}$$

де p_0 – це дотримана угода, p_e – це очікувана угода.

Ця метрика відображає наскільки краще працює класифікатор (p_0) порівняно з продуктивністю класифікатора, який просто вгадує випадковим чином відповідно до частоти кожного класу (p_e).

Стандартизованого способу інтерпретації його значень не існує. Ландіс і Кох [25] пропонують спосіб характеристики цінностей. Згідно з їхньою схемою, значення < 0 вказує на відсутність згоди, 0–0.20 як незначну, 0.21–0.40 як справедливу, 0.41–0.60 як помірну, 0.61–0.80 як суттєву та 0.81–1 як майже ідеальну згоду.

5. Метрика ROC-AUC. Крива ROC-AUC є вимірюванням продуктивності для проблем класифікації, виявлення викидів при різних порогових налаштуваннях. ROC – це крива ймовірності, а AUC – ступінь або міра відокремленості. Ця метрика відображує, наскільки математична модель здатна розрізняти класи. Чим вища міра AUC, тим краще модель правильно прогнозує класи.

Крива ROC зображена на рис. 3.1, де відображена залежність TPR (True Positive Rate) від FPR (False Positive Rate). По осі ординат відображена міра TPR, по осі абсцис — FPR. TPR – це частка правильних прогнозів у прогнозах позитивного класу. Ця міра розраховується за формулою:

$$TPR = \frac{TP}{TP+FN}, \tag{3.7}$$

де TP – кількість значень правильної класифікації позитивного класу, FN – кількість значень неправильної класифікації негативного класу.

FPR – це частка неправильних прогнозів у прогнозах позитивного класу. Ця міра розраховується за формулою :

$$FPR = \frac{FP}{FP+TN}, \tag{3.8}$$

де FP – кількість значень неправильної класифікації позитивного класу, TN – кількість значень правильної класифікації негативного класу.

Площа під кривою ROC визначає значення метрики ROC-AUC. Для відмінної моделі значення цієї метрики буде дорівнювати одиниці. Діагональна пряма, яка відображена на рис. 3.1, вказує на модель, яка не має змоги відрізнити позитивний клас від негативного, тобто не має можливості розділення класів.

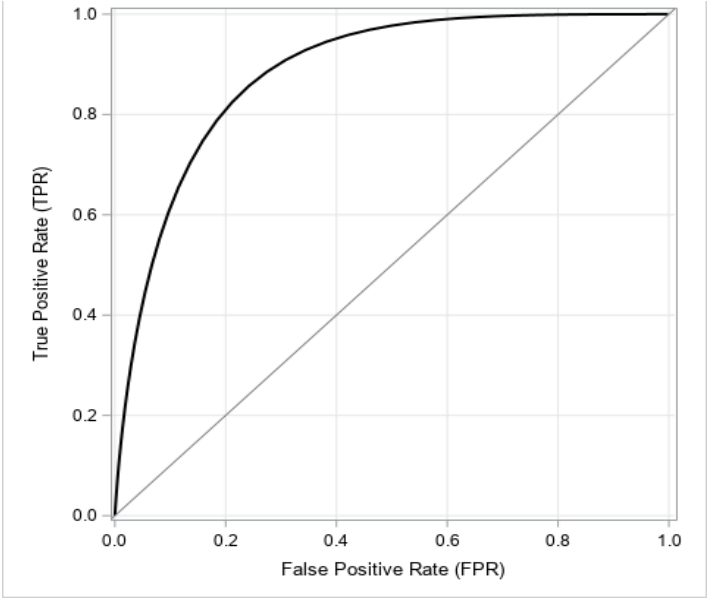

Рис. 3.1 Крива ROC

Цю метрику слід використовувати, коли модель повинна працювати однаково добре як на позитивному, так і на негативному класі.

6. Метрика PR-AUC. Ця метрика розраховується так само, як ROC-AUC. Відмінною рисою є те, що крива будується на основі точності (Precision) і відклику (Recall).

Цю метрику слід використовувати, коли модель повинна відмінно визначати або позитивні класи, або негативні. Тобто за допомогою PR-AUC можна сфокусуватися на правильному прогнозуванні одного із класів.

7. Метрика pAUC (Partial AUC). Часткова AUC була запропонована як альтернативний захід до стандартної AUC. При використанні часткової AUC враховується лише конкретна область простору ROC. Приклад кривої pAUC відображено на рис.3.2, де pAUC розраховується для затіненої області.

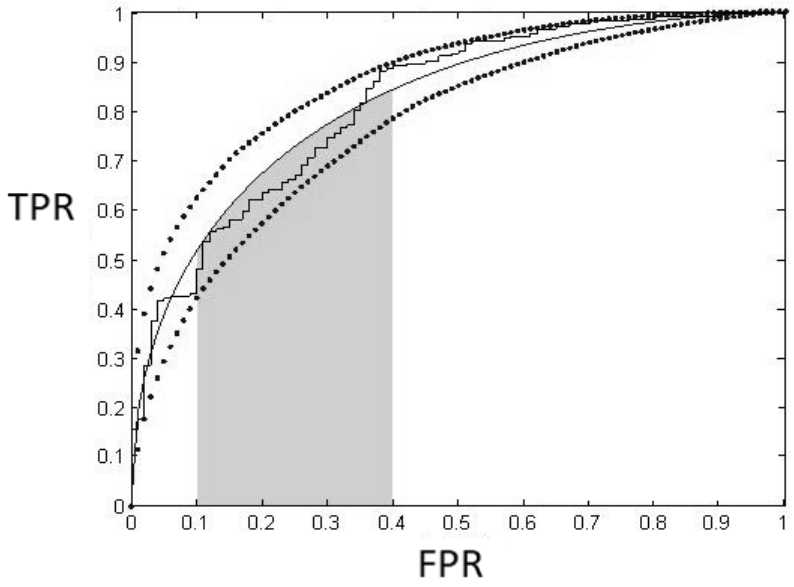

Рис. 3.2 Крива pAUC

8. Метрика двосторонній pAUC (Partial AUC). На відміну від pAUC, замість обмеження лише частоти помилкових позитивних результатів (FPR), двосторонній pAUC фокусується на частковій площі під кривою з обмеженнями як по горизонталі, так і по вертикалі.

Приклад кривої двостороннього pAUC відображено на рис.3.3. Двосторонній pAUC позначає площу заштрихованої області A. Ця заштрихована область безпосередньо визначається явною верхньою межею FPR (p0 = 0,5) і нижньою межею TPR (q0 = 0,65). На відміну від цього, pAUC позначає площу обох регіонів A і B.

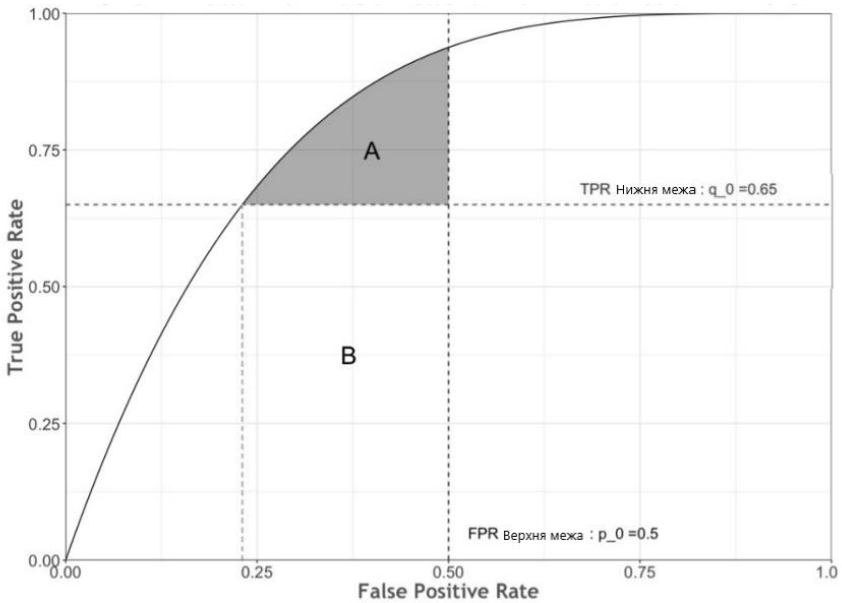

Рис. 3.3 Крива двостороннього pAUC

9. Метрика LogLoss. Вона є однією з найважливіших класифікаційних метрик на основі ймовірностей. Ця метрика розраховується за формулою 3.9. LogLoss вказує на те, наскільки ймовірність прогнозу близька до відповідного фактичного/істинного значення. Та чим більше прогнозована ймовірність відрізняється від фактичного значення, тим вище значення логарифмічних втрат.

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)], \quad (3.9)$$

де y_i – справжній клас, p_i – вірогідність того, що y_i вказує на позитивний клас, n – загальна кількість спостережень з вибірки.

До недоліків цієї метрики можна віднести те, що при незбалансованості класів, мажоритарний клас може домінувати над LogLoss.

Проаналізувавши усі метрики, які використовуються при оцінці якості методів (моделей) виявлення викидів, можна визначити, що найкращою є метрика PR-AUC. Ця метрика була обрана найкращою, через те що вона фокусується на малих позитивних класах, в нашому випадку викидах та дозволяє об'єктивно оцінити якість математичних моделей та методів.

4 Методи виявлення викидів (аномальних значень)

4.1 Традиційні методи для виявлення викидів

Серед традиційних методів можна виділити такі як:

1) Z-оцінка (стандартна оцінка спостереження) – це індикатор, що визначає положення вихідної оцінки з погляду її відстані від середнього значення при вимірі в одиницях стандартного відхилення, при умові гаусовського розподілу.

Метод дозволяє побачити, наскільки вище або нижче середнього знаходиться це значення на кривій розподілу (рис. 4.1).

Це робить z-оцінку параметричним методом. Інколи точки даних не описуються гауссовим розподілом. Ця проблема може бути вирішена шляхом застосування перетворень до даних, таких як масштабування.

Під час обчислення z-оцінки для кожної вибірки в наборі даних необхідно вказати поріг. Аналізуючи точки даних, які лежать за певним порогом, можна визначити чи відносяться вони до аномальних.

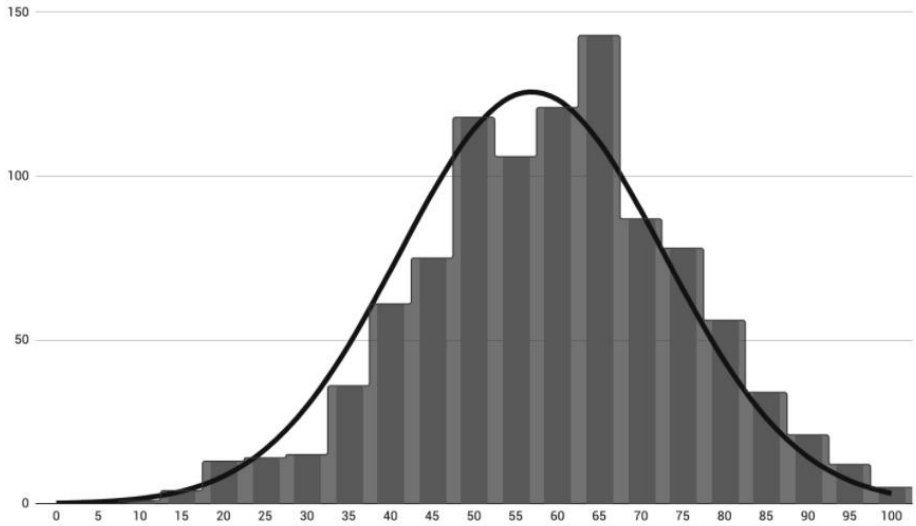

Рис. 4.1 Приклад роботи методу Z-оцінка [26]

Z-оцінку для будь-якої точки даних можливо розрахувати за формулою:

$$z = \frac{x-\mu}{\sigma}, \tag{4.1}$$

де x – вхідний показник, μ - середнє значення набору даних, σ – стандартне відхилення для набору даних.

Даний метод ефективний та простий для виявлення викидів в наборі даних з параметричними розподілами в малорозмірному просторі об'єктів.

Визначимо переваги та недоліки методу Z-оцінки. До переваг можна віднести наступні:

- ефективний метод, щоб описати значення у просторі ознак за допомогою розподілу Гауса.
- проста реалізація.

Серед недоліків можна виділити наступні:

- метод ефективний у просторі об'єктів низької розмірності;
- якщо розподіли не можна вважати параметричними, метод працює не точно.

2) DBSCAN — це метод кластеризації на основі щільності. Цей метод зосереджений на пошуку сусідів за щільністю на n-вимірній сфері з радіусом ε (рис.4.2.). Кластер можна визначити як максимальний набір точок, пов’язаних із щільністю в просторі ознак.

Це ефективний метод для роботи з наборами даних середнього розміру. DBSCAN самостійно оцінює кількість кластерів, немає потреби вказувати кількість бажаних кластерів, це модель машинного навчання без вчителя.

Dbscan визначає такі класи точок як:

- основна точка: A є основною точкою, якщо її околиця (визначена ε) містить принаймні стільки ж або більше точок, ніж параметр MinPts(мінімальна кількість спостережень у кластері);
- гранична точка: C — це гранична точка, яка лежить у кластері, і її околиці не містять більше точок, ніж MinPts, але вона все ще є досяжною для щільності іншими точками в кластері;
- викид: N — це точка викиду, яка не лежить у жодному кластері, і вона не пов’язана з будь-якою іншою точкою.

Таким чином, ця точка матиме власний кластер [27].

Визначимо переваги та недоліки даного методу. До переваг можна віднести наступні:

- точно працює, якщо простір ознак пошуку викидів є багатовимірним;
- легка візуалізація результатів.

Серед недоліків можна виділити наступні:

- значення у просторі функцій необхідно відповідно масштабувати;
- це неконтрольована модель, і її необхідно повторно калібрувати щоразу, коли аналізується новий пакет даних.

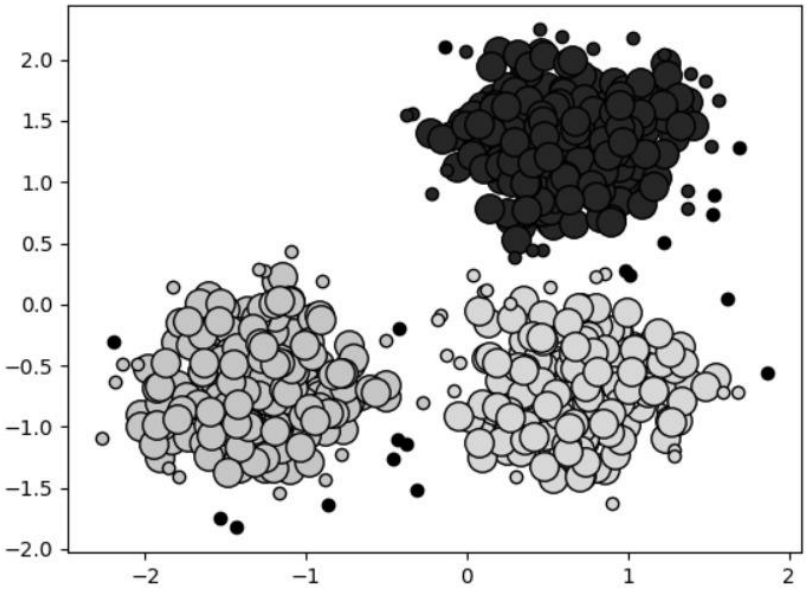

Рис. 4.2 Приклад роботи методу DBSCAN[28]

3) Ізоляційний ліс – ефективний метод для виявлення викидів (аномальних значень) в наборах даних. Цей метод заснований на бінарних деревах рішень. Основний принцип ізоляційного лісу полягає в тому, що викиди є нечисленними і далекими від решти спостережень.

Щоб побудувати дерево, алгоритм випадковим чином вибирає об’єкт із простору ознак і випадкове розділене значення в діапазоні між максимумами та мінімумами. Це робиться для всіх

спостережень у навчальному наборі. Для побудови лісу складається ансамбль дерев із усередненням усіх дерев у лісі.

Даний метод порівнює спостереження зі значенням розщеплення у вузлі, цей вузол матиме два дочірніх вузла, на яких буде зроблено ще одне випадкове порівняння. Кількість розщеплень, зроблених алгоритмом для екземпляра, називається довжиною шляху. Викиди матимуть коротшу довжину шляху, ніж решта спостережень.

Оцінку аномалії можна обчислити за формулою:

$$S(x,n) = 2^{-\frac{E(h(x))}{e(n)}}, \quad (4.2)$$

де E(h(x)) – середня довжина шляху вибірки, C(n) –невдалий пошук довжини, N – кількість зовнішніх вузлів.

Розглянемо переваги та недоліки методу. До переваг можна віднести відсутність необхідності масштабувати значення у просторі функцій.

Серед недоліків можна виділити наступні:

- складна візуалізація результатів;
- при неправильній оптимізації, для великого набору даних, довгий час навчання.

4) Local Outlier Factor (LOF) — метод машинного навчання без вчителя для виявлення аномалії, який обчислює локальне відхилення щільності точки щодо її сусідів. Даний метод вважає викидами зразки, які мають значно нижчу щільність, ніж їхні сусіди.

Кількість сусідів, що розглядаються зазвичай встановлюється:

- більшою, ніж мінімальна кількість вибірок, яку повинен містити кластер, так що інші вибірки можуть бути локальними викидами щодо цього кластера;
- менше, ніж максимальна кількість близьких сусідів за вибірками, які потенційно можуть бути локальними викидами [29].

LOF дає кращі результати, аніж глобальний підхід до пошуку викидів. Оскільки граничного значення LOF немає, вибір точки як викиду залежить від користувача.

Розглянемо переваги та недоліки методу. До переваг можна віднести, що точка буде вважатися викидом, якщо вона знаходиться на невеликій відстані від надзвичайно щільного скупчення, глобальний підхід може не розглядати цей момент як вибій. Але LOF може ефективно ідентифікувати локальні викиди.

Серед недоліків можна виділити наступні:

- немає певного порогового значення, вище якого точка визначається як викид;
- ідентифікація викиду залежить від проблеми та користувача.

4.2 Методи глибокого навчання для виявлення викидів (аномальних значень)

Серед методів глибокого навчання можна виділити наступні.

1) Автоенкодери - це нейронні мережі, призначені для вивчення низькорозмірного уявлення за деякими вхідними даними (рис. 4.3). Вони складаються з двох компонентів: кодувальника, який вчиться відображати вхідні дані в низькорозмірне уявлення (так зване вузьке місце), і декодер, який вчиться відображати це низькорозмірне уявлення назад у вихідні вхідні дані. Структуруючи проблему навчання таким чином, мережа кодувальника вивчає ефективну функцію стиснення, яка відображає вхідні дані в помітне уявлення нижчої розмірності, так що мережа декодера може успішно відновлювати вихідні вхідні дані.

Модель навчається шляхом мінімізації помилки відновлення, яка є різницею (середньоквадратичною помилкою) між вихідним введенням і відновленим висновком, створеним декодером [30].

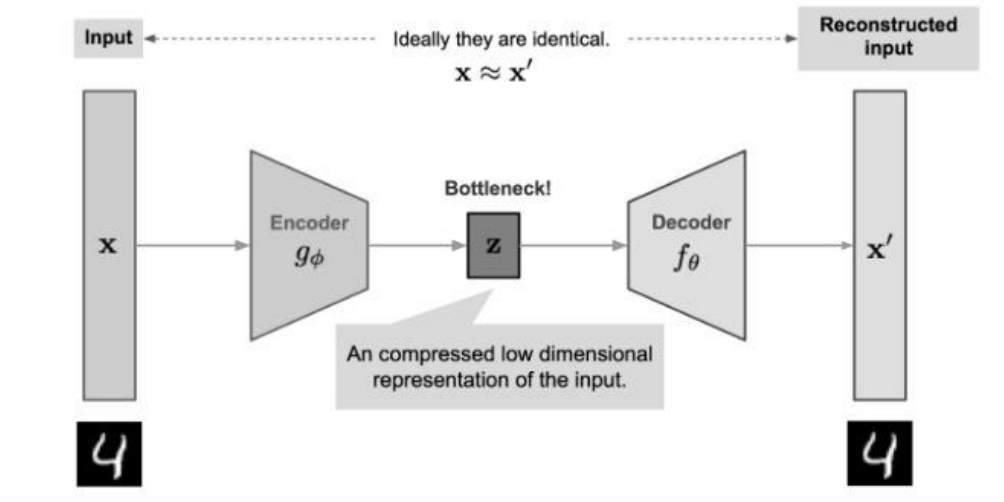

Рис. 4.3 Архітектура AutoEncoder[31]

Застосування автоенкодера для виявлення аномалій слідує загальному принципу: спочатку потрібно змоделювати нормальну поведінку, а потім генерувати оцінку аномалії кожної нової вибірки даних. Щоб змоделювати нормальну поведінку, слід використовувати напівконтрольований підхід, тобто коли навчання моделі проходить на нормальних вибірках даних. Таким чином модель вивчає функцію відображення, яка успішно відновлює нормальні вибірки даних з дуже невеликою помилкою відновлення. Така поведінка відтворюється під час тестування, коли помилка відновлення мала для нормальних вибірок даних та велика для аномальних вибірок даних.

Щоб ідентифікувати аномалії, використовується обробка помилки відновлення як оцінка аномалії, після цього можливо вибрати зразки з помилками відновлення, що перевищують заданий поріг.

Розглянемо переваги та недоліки методу. До переваг можна віднести наступні:

- компактність та швидкість кодування;
- зменшення розмірності даних, що прискорює навчання моделі.

Серед недоліків можна виділити такі як:

- складна візуалізація результатів;
- складне навчання.

2) Варіаційний автоенкодер (VAE) - це розширення автоенкодера (рис. 4.4). Подібно до автоенкодера, він складається з кодувальника та мережевого компонента декодера, але він включає важливі зміни в структурі завдання навчання, щоб пристосуватися до варіаційного висновку.

На відміну від навчання зіставлення вхідних даних з фіксованим вектором вузького місця (точкова оцінка), VAE вивчає зіставлення вхідних даних з розподілом і вчиться відновлювати вихідні дані шляхом вибірки цього розподілу з використанням прихованого коду.

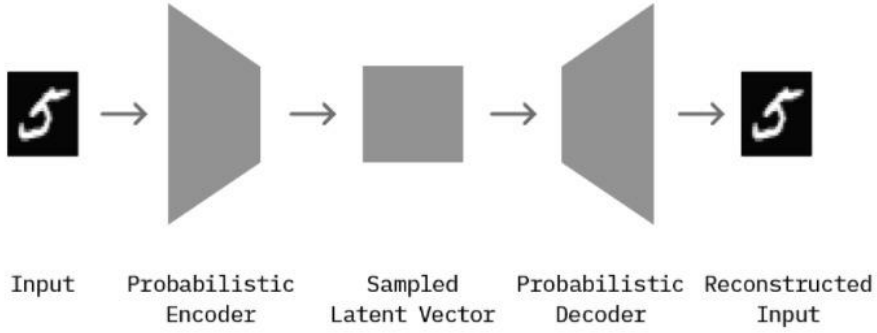

Рис. 4.4 Модель варіаційного автоенкодера [31]

Модель VAE навчається шляхом мінімізації різниці між розрахунковим розподілом, створеним моделлю, та реальним розподілом даних. Ця різниця оцінюється за допомогою дивергенції Кульбака-Лейблера, яка кількісно визначає відстань між двома розподілами, вимірюючи, скільки інформації втрачається, коли один розподіл використовується для подання іншого.

Використання VAE для виявлення аномалій: подібно автоенкодеру, навчання VAE починається на нормальних вибірках даних. Під час тестування можна визначити оцінку аномалії двома способами. По-перше, вилучити зразки прихованого коду з кодера з урахуванням наших вхідних даних, відстежити відновлені значення з декодера та обчислити середню помилку відновлення. Аномалії позначаються з урахуванням порогу помилки відновлення.

В якості альтернативи потрібно вивести середнє значення та параметр дисперсії з декодера та обчислити ймовірність того, що нова точка даних належить розподілу нормальних даних, на якому була навчена модель. Якщо точка даних знаходиться в області низької щільності (нижче деякого порога), це позначається як аномалія (рис. 4.5).

Розглянемо переваги та недоліки методу. До переваг можна віднести наступні:

- чіткий спосіб оцінки якості моделі (логарифмічна ймовірність);
- легка візуалізація.

Серед недоліків можна виділи такі як:

- неоптимальні фактори варіації;
- може мати градієнти високої дисперсії.

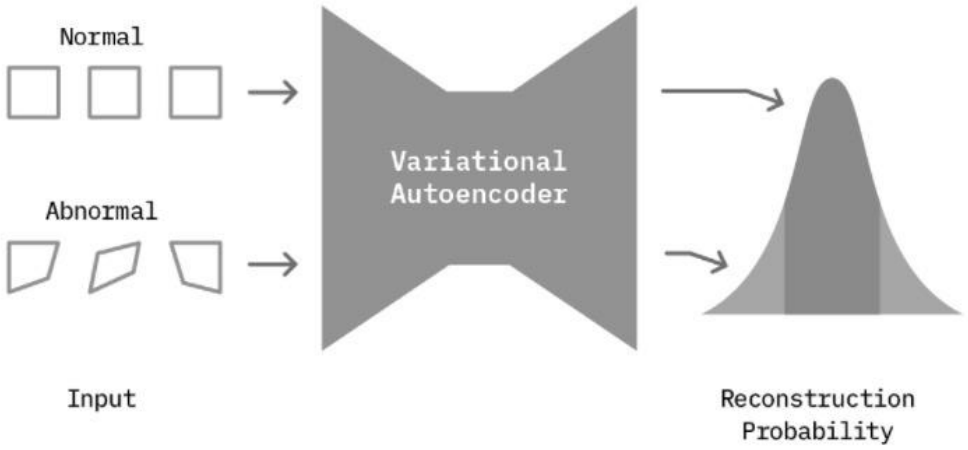

Рис.4.5 Оцінка аномалій за допомогою VAE. [31]

3) Генеративні змагальні мережі (GAN) – це нейронні мережі, розроблені для вивчення генеративної моделі розподілу вхідних даних (рис. 4.6). У класичному формулюванні вони складаються з пари (зазвичай з прямим зв'язком) нейронних мереж, а саме з генератора G і дискримінатора D. Обидві мережі навчаються спільно з кінцевою метою визначити розподіл вихідних даних X.

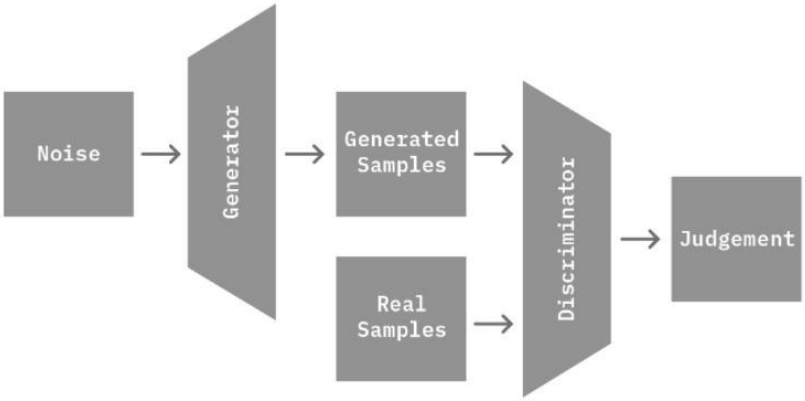

Рис. 4.6 Класична модель GAN [31]

Щоб ідентифікувати аномалії, використовується напівкерований підхід, тобто потрібно навчати модель від послідовності до послідовності на нормальних даних. Моделі послідовність-послідовність (sequence-to-sequence) - це клас нейронних мереж, в основному призначений для вивчення зіставлень між даними, які найкраще представлені у вигляді послідовностей. Потім під час тестування можна порівняти різницю (середньоквадратичну помилку) між вихідною послідовністю, згенерованою моделлю, та її вхідними даними. Це значення використовується як показник аномалій.

Розглянемо переваги та недоліки методу. Недоліком генеративних змагальних мереж є складність оцінки моделі. До переваг можна віднести наступні:

- GAN генеруює дані, схожі вхідні;
- GAN деталізує дані та може легко інтерпретуватися у різні версії.

5. Огляд результатів та ефективності методів

В якості наборів даних були обрані різні вибірки, які використовуються для оцінки якості методів виявлення викидів, наприклад ForestCover, Satellite та інші. Також було створено власний набір даних, який буде позначатися, як «Economical (наш)». Цей набір даних відображає соціально-економічний розвиток країн за 2012-2020 роки. Він представляє 115 альтернатив з 32 змінними стану. Набір даних відображено на рис. 5.1.

Country Name	EGI_2020	EPI_2020	OSI_2020	HCI_2020	TII_2020	EGI_2018	EPI_2018	OSI_2018	HCI_2018	TII_2018	...	ICT_16	ICT_15	ICT_13	ICT_12	SPI_20	SPI_18	SPI_17	SPI_
Albania	73.99	84.52	84.12	80.01	57.85	65.19	75.84	73.61	78.77	43.18	...	4.90	4.73	4.62	4.42	75.41	71.77	71.65	70.
Algeria	51.73	15.48	27.65	69.66	57.87	42.27	20.22	21.53	66.40	38.89	...	4.32	3.71	4.46	4.22	69.92	66.83	66.83	65.
Argentina	82.79	85.71	84.71	91.00	72.65	73.35	62.36	75.00	85.79	59.27	...	6.68	6.40	6.62	6.38	80.66	74.98	74.84	74.
Armenia	71.36	75.00	70.00	78.72	65.36	59.44	56.74	56.25	75.47	46.60	...	5.56	5.32	5.64	5.55	76.46	70.87	70.53	69.
Australia	94.32	96.43	94.71	100.00	88.25	90.53	98.31	97.22	100.00	74.36	...	8.08	8.29	8.23	8.14	91.29	88.32	88.23	87.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
United States of America	92.97	100.00	94.71	92.39	91.82	87.69	98.31	98.61	88.83	75.64	...	8.13	8.19	7.78	7.75	85.71	84.78	85.27	86.
Uruguay	85.00	85.71	84.12	85.14	85.74	78.58	91.57	88.89	77.19	69.67	...	6.75	6.70	7.05	6.80	82.99	79.40	NaN	N
Viet Nam	66.67	70.24	65.29	67.79	66.94	59.31	69.10	73.61	65.43	38.90	...	4.18	4.28	4.48	4.39	68.85	NaN	NaN	N
Zambia	42.42	30.95	25.88	67.45	33.94	41.11	39.89	47.92	56.89	18.53	...	2.19	2.04	2.68	2.63	55.34	NaN	NaN	N
Zimbabwe	50.19	45.24	52.35	61.35	36.88	36.92	27.53	32.64	56.68	21.44	...	2.85	2.90	3.12	2.99	52.26	45.26	43.76	42.

Рис. 5.1 Набір даних «Economical (наш)»

Було проведено дослідження щодо ефективності традиційних математичних моделей та методів для виявлення викидів. В якості показника ефективності було значення обраної метрики PR-AUC.

Для тестування та оцінки якості було розроблено програмне забезпечення, яке дозволяє:

- виконувати тренування математичних моделей та методів;
- кластеризувати вибірки на різних наборах даних;
- розраховувати показник ефективності.

Програмне забезпечення розроблено за допомогою мови Python. Також були використані наступні бібліотеки: scikit-learn, NumPy, Pandas і інші.

В якості наборів даних були використані різні вибірки, які були створені саме для оцінки якості математичних моделей та методів виявлення викидів (аномальних значень).

Дослідження якості проводилося таким чином, що набори даних подавалися до програмного забезпечення, яке в свою чергу виконувало усі необхідні дії з ними. В результаті роботи програми видавалися результати виявлення викидів та значення метрики PR-AUC.

Результати оцінки якості виявлення викидів для різних наборів даних приведені в таблиці 1.

Таблиця 1. Результати тестування традиційних методів для виявлення викидів

Набір даних	PR-AUC			
	Ізоляційний ліс	DBSCAN	LOF	Z-оцінка
Http (KDDCUP99)	**0.90**	0.60	0.45	0.50
ForestCover	**0.85**	0.83	0.72	0.60
Mulcross	**0.89**	0.33	0.37	0.41
Smtp (KDDCUP99)	**0.83**	0.80	0.65	0.60
Shuttle	**0.81**	0.60	0.55	0.67
Mammography	**0.86**	0.77	0.67	0.65
Satellite	**0.82**	0.68	0.72	0.66
Pima	**0.71**	0.65	0.52	0.60
Breastw	0.67	**0.71**	0.49	0.85
Arrhythmia	**0.87**	0.86	0.37	0.41
Ionosphere	**0.80**	0.78	0.73	0.63
Economical (наш)	**0.84**	0.73	0.65	0.54

Також було проведено дослідження щодо ефективності математичних моделей та методів глибокого навчання для виявлення викидів. Це дослідження є подібним до того, що проводилося для оцінки якості традиційних методів виявлення викидів. Відмінність полягає тільки у різних математичних моделях і методах, шо оцінюються. Тому, було проведена імплементація автоенкодера, варіаційного автоенкодера та генеративних змагальних мереж. Головною бібліотекою для їх реалізації була Tensorflow, яка є комплексною платформою з відкритим вхідним кодом для машинного навчання.

Результати дослідження оцінки якості виявлення викидів (аномальних значень) для визначених наборів даних приведені в таблиці 2.

Таблиця 2. Результати тестування методів глибокого навчання для виявлення викидів

Набір даних	PR-AUC		
	Автоенкодер	Варіаційний автоенкодер	GAN
Http (KDDCUP99)	0.90	**0.97**	0.95
ForestCover	0.88	0.93	**0.97**
Mulcross	0.92	**0.95**	0.76
Smtp (KDDCUP99)	0.89	**0.91**	0.90
Shuttle	0.93	**0.95**	0.94
Mammography	0.87	0.89	**0.90**
Satellite	0.85	0.87	**0.88**
Pima	0.81	0.84	**0.86**
Breastw	0.75	0.77	**0.79**
Arrhythmia	0.89	0.90	**0.93**
Ionosphere	0.85	0.86	**0.90**
Economical (наш)	0.88	**0.90**	0.86

Вибираючи найкращі методи серед традиційних та глибокого навчання, не важко бачити, що кращими є ізоляційний ліс та метод, що базується на генеративних змагальних мережах (GAN).

Побудуємо графіки ящиків з вусами для кожної змінної стану із вхідного набору даних та результуючого набору, де викиди були видалені за допомогою найкращих методів. В якості вхідної вибірки буде набір даних «Economical (наш)». Графік ящиків з вусами для вхідної вибірки представлено на рис. 5.2.

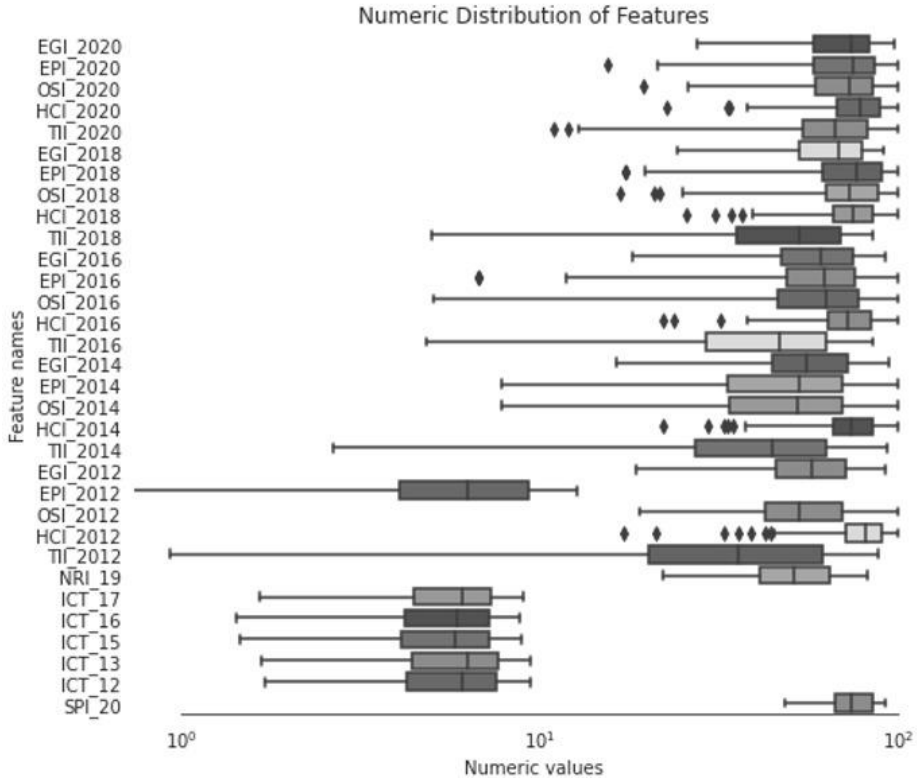

Рис. 5.2 Графік ящиків з вусами для змінних вхідної вибірки

На рис 5.2 можна побачити, що деякі змінні стану мають викиди, які відображені точками на графіку ящиків з вусами. Використаємо ізоляційний ліс для виявлення та видалення викидів з вхідної вибірки і побудуємо графік ящиків з вусами для набору даних, де викиди були видалені за допомогою ізоляційного лісу. Цей графік відображено на рис. 5.3.

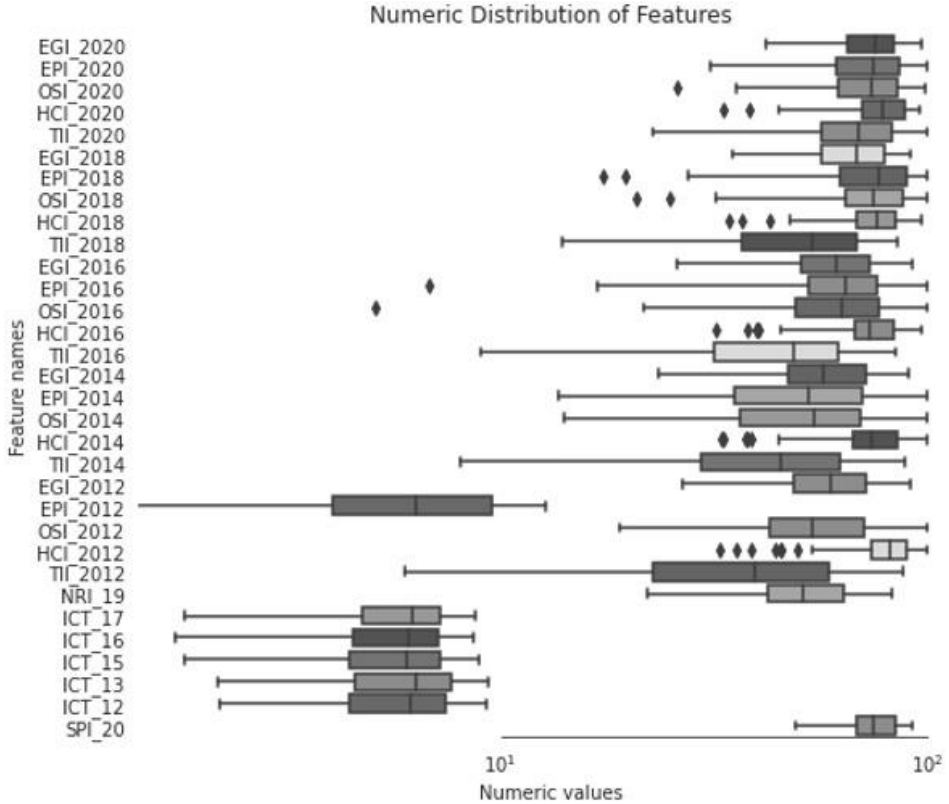

Рис. 5.3 Графік ящиків з вусами для набору даних, де викиди були видалені за допомогою ізоляційного лісу

Також використаємо генеративні змагальні мережі для виявлення та видалення викидів з вхідної вибірки і побудуємо графік ящиків з вусами для набору даних, де викиди були видалені за генеративних змагальних мереж. Цей графік відображено на рис. 5.4.

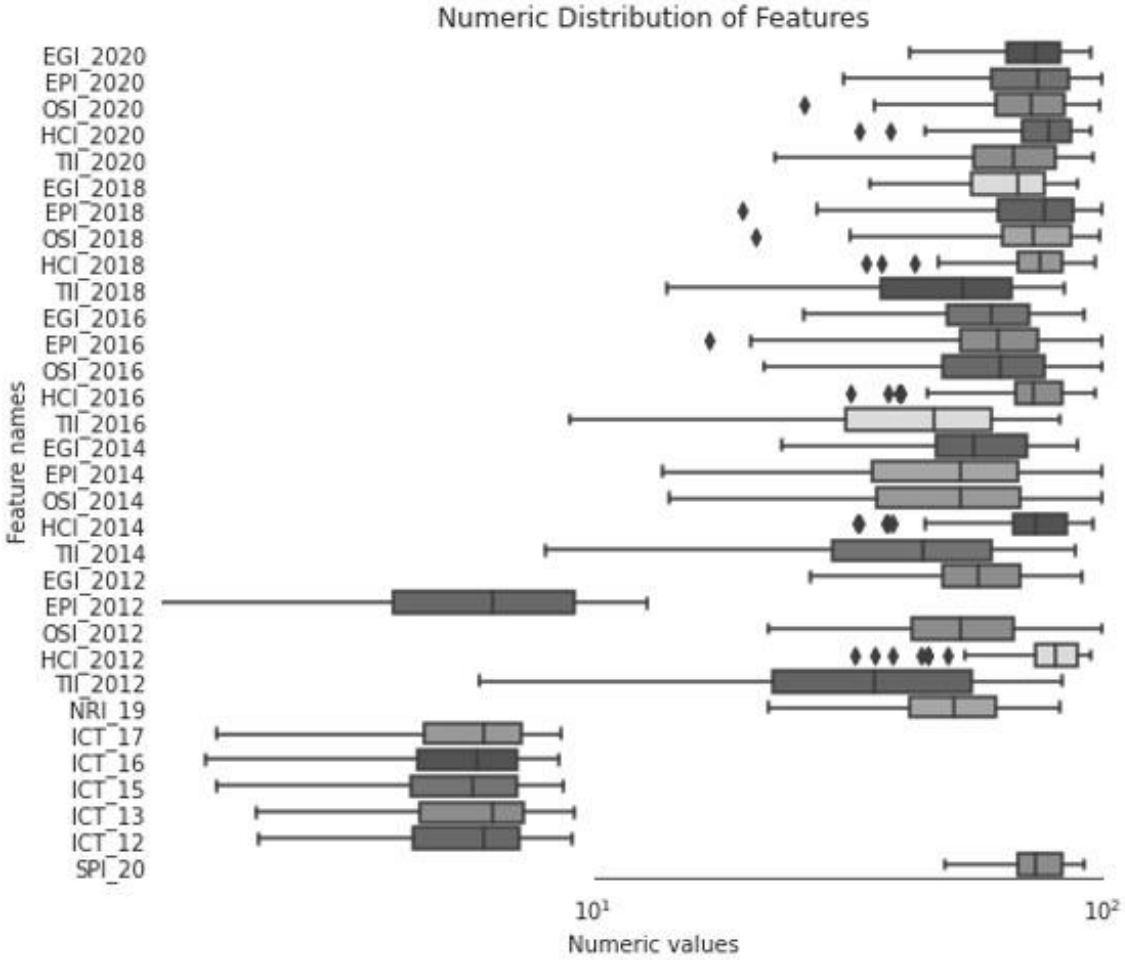

Рис. 5.4 Графік ящиків з вусами для набору даних, де викиди були видалені за допомогою генеративних змагальних мереж

6 Результати та висновки

Проведено аналіз метрик, які використовуються при оцінці якості математичних моделей та методів виявлення викидів. Визначено, що найкращою є метрика PR-AUC. Ця метрика була обрана найкращою, через те що вона фокусується на малих позитивних класах, в нашому випадку викидах, та дозволяє об'єктивно оцінити якість математичних моделей та методів.

Виконано класифікацію методів (моделей) виявлення викидів. Ці методи поділяються на традиційні та глибокого навчання. На основі аналізу результатів дослідження щодо ефективності традиційних математичних моделей та методів і аналізу їх переваг та недоліків, було визначено, що найбільш ефективним є метод «ізоляційний ліс». Цей метод здобув найкращі результати показників ефективності. Перевагою методу є відсутність необхідності масштабувати дані в просторі та відсутність великої кількості параметрів, що робить його надійним та простим в підборі параметрів для покращення роботи методу.

Серед математичних моделей та методів глибокого навчання для виявлення викидів було визначено, що найбільш ефективним є метод, що базується на генеративних змагальних мережах (GAN). Він досяг найкращих показників ефективності серед усіх математичних моделей та методів виявлення викидів, які були розглянуті в роботі, використовуючи різні набори даних. Перевагами даного методу є те, що метод деталізує дані та може легко інтерпретуватися у різні версії. Цей метод потребує великі обчислювальні можливості, тому його слід використовувати при їх наявності.

Обираючи між ізоляційним лісом та GAN слід підкреслити, що обидва методи мають досить високі показники ефективності виявлення викидів на пробних вибірках. Вибір між ними залежить від обчислювальних можливостей та поставлених цілей, щодо якості виявлення викидів.

Таким чином, в роботі було отримано:

- огляд метрик, які використовуються для оцінки ефективності математичних моделей та методів виявлення викидів;
- огляд традиційних методів та методів глибокого навчання для виявлення викидів;
- результати дослідження, щодо ефективності та якості математичних моделей і методів виявлення викидів, використовуючи 12 пробних вибірок;
- висновки про найкращу метрику та найкращі математичні моделі і методи для вирішення проблеми виявлення викидів в пробних вибірках при управлінні процесами в системах за станом.

ЛІТЕРАТУРА

1. В.П. Шкодирєв, К.І. Ягафаров, В.А. Баштовенко, Є.Е. Ільїна. Огляд методів виявлення аномалій в потоках даних. URL: http://ceur-ws.org/Vol-1864/paper_33.pdf (дата звернення: 10. 11. 2021).

2. М.В. Ломоносова. Виявлення аномалій у роботі механізмів методами машинного навчання. URL: http://ceur-ws.org/Vol-2022/paper59.pdf (дата звернення: 10. 11. 2021).

3. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. URL: https://arxiv.org/abs/1901.03407 (дата звернення: 10. 11. 2021)

4. Srikanth Thudumu, Philip Branch, Jiong Jin & Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data.

5. Deep Learning for Anomaly Detection: A Review: ACM Computing Surveys: Vol 54, No 2. URL: https://dl.acm.org/doi/10.1145/3439950 (дата звернення: 10. 11. 2021).

6. Muruti G., Rahim F., bin Ibrahim Z. A Survey on Anomalies Detection Techniques and Measurement Methods // 2018 IEEE Conference on Application, Information and Network Security (AINS). 2018.

7. Shikha Agrawal, JitendraAgrawal. Survey on Anomaly Detection using Data Mining Techniques.

8. Pang G. и др. Deep Learning for Anomaly Detection // ACM Computing Surveys. 2021. Т. 54. № 2. С. 1-38.

9. Nassif A. и др. Machine Learning for Anomaly Detection: A Systematic Review // IEEE Access. 2021. Т. 9. С. 78658-78700.

10. Izhak Golan, Ran El-Yaniv. Deep Anomaly Detection Using Geometric Transformations. URL: https://proceedings.neurips.cc/paper/2018/file/5e62d03aec0d17facfc5355dd90d441c-Paper.pdf (дата звернення: 10. 11. 2021).

11. Mohammad Braei, Sebastian Wagner. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. URL: https://www.semanticscholar.org/paper/Anomaly-Detection-in-Univariate-Time-series%3A-A-on-Braei-Wagner/cf45bce52cca1f6e450ddaa1d19fe6e30661dffb (дата звернення: 10. 11. 2021).

12. Atiq ur Rehman & Samir Brahim Belhaouari. Unsupervised outlier detection in multidimensional data

13. Victoria J. Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. URL: https://core.ac.uk/download/pdf/58585.pdf (дата звернення: 12. 11. 2021).

14. Karanjit Singh and Dr. Shuchita Upadhyaya. Outlier Detection: Applications And Techniques. URL:https://www.researchgate.net/publication/267964435_Outlier_Detection_Applications_And_Techniques (дата звернення: 12. 11. 2021).

15. Wang S. и др. Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discriminative Network. URL:

https://proceedings.neurips.cc/paper/2019/hash/6c4bb406b3e7cd5447f7a76fd7008806-Abstract.html (дата заернення: 12. 11. 2021).

16. Karanjit Singh and Dr. Shuchita Upadhyaya. Outlier Detection: Applications And Techniques. URL: https://www.researchgate.net/publication/228686398_k-Nearest_neighbour_classifiers (дата звернення: 14. 11. 2021).

17. Yen-Chang Hsu, Yilin Shen, Hongxia Jin, Zsolt Kira. Generalized ODIN: Detectiong Out-of-distribution Image without Learning from Out-jf-distribution Data. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Hsu_Generalized_ODIN_Detecting_Out-of-Distribution_Image_Without_Learning_From_Out-of-Distribution_Data_CVPR_2020_paper.pdf (дата звернення: 14. 11. 2021).

18. Breunig M. и др. LOF // Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. 2000. URL: https://www.researchgate.net/publication/221214719_LOF_Identifying_Density-Based_Local_Outliers (дата звернення: 15. 11. 2021).

19. Na S., Xumin L., Yong G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm // 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. 2010.

20. Markus Goldstein, Andreas Dengel. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm. URL: https://www.researchgate.net/publication/231614824_Histogram-based_Outlier_Score_HBOS_A_fast_Unsupervised_Anomaly_Detection_Algorithm (дата звернення: 15. 11. 2021).

21. Warp-core. URL: https://workday.github.io/warp-core/contents/anomaly_detection/ (дата звернення: 17. 11. 2021).

22. PDF) Support Vector Machines: Theory and Applications. URL:https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications (дата звернення: 17. 11. 2021).

23. Fei Tony Liu, Kai Ming TingGippsland School of Information TechnologyMonash University, Victoria, Australia. Isolation Forest. URL: https://www.researchgate.net/publication/224384174_Isolation_Forest (дата звернення: 18. 11. 2021).

24. Xuehui Wang, Yong Zhang, Hao Liu, Yang Wang, Lichun Wang , and Baocai Yin. An Improved Robust Principal Component Analysis Model for Anomalies Detection of Subway Passenger Flow.

25. JR L., GG K. The measurement of observer agreement for categorical data. URL: https://pubmed.ncbi.nlm.nih.gov/843571/ (дата звернення: 18. 11. 2021).

26. Study Finance. URL: https://studyfinance.com/static/media/z-score.png (дата звернення: 21. 11. 2021).

27. A Brief Overview of Outlier Detection Techniques .URL: https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561 (дата звернення: 21. 11. 2021).

28. Demo of DBSCAN clustering algorithm. URL: https://scikit-learn.org/stable/_images/sphx_glr_plot_dbscan_001.png (дата звернення: 22. 11. 2021).

29. Outlier detection with Local Outlier Factor (LOF). URL: https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html#:~:text=The%20Local%20Outlier%20Factor%20(LOF,lower%20density%20than%20their%20neighbors. (дата звернення: 22. 11. 2021).

30. Anomaly Detection using Autoencoders.URL: https://towardsdatascience.com/anomaly-detection-using-autoencoders-5b032178a1ea (дата звернення: 22. 11. 2021).

31. Cloudera Fast Forward. Deep Learning for Anomaly Detection. URL: https://ff12.fastforwardlabs.com/ (дата звернення: 24. 11. 2021)

REFERENCES

1. V.P. Shkodyrev, K.I. Yafagorov, B.A. Bashtovenko, Y.E. Ilyina. Review of methods for detecting anomalies in data streams.URL: http://ceur-ws.org/Vol-1864/paper_33.pdf (Last accessed: 10. 11. 2021). [in Russian]

2. M.V. Lomonosov. Detection of anomalies in the work of mechanisms by machine learning methods.URL: http://ceur-ws.org/Vol-2022/paper59.pdf (Last accessed: 10. 11. 2021). [in Russian]

3. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. URL: https://arxiv.org/abs/1901.03407 (Last accessed: 10. 11. 2021)

4. Srikanth Thudumu, Philip Branch, Jiong Jin & Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data.

5. Deep Learning for Anomaly Detection: A Review: ACM Computing Surveys: Vol 54, No 2. URL: https://dl.acm.org/doi/10.1145/3439950 (Last accessed: 10. 11. 2021).

6. Muruti G., Rahim F., bin Ibrahim Z. A Survey on Anomalies Detection Techniques and Measurement Methods // 2018 IEEE Conference on Application, Information and Network Security (AINS). 2018.

7. Shikha Agrawal, JitendraAgrawal. Survey on Anomaly Detection using Data Mining Techniques.

8. Pang G. Deep Learning for Anomaly Detection // ACM Computing Surveys. 2021. Т. 54. № 2. С. 1-38.

9. Nassif A. Machine Learning for Anomaly Detection: A Systematic Review // IEEE Access. 2021. Т. 9. С. 78658-78700.

10. Izhak Golan, Ran El-Yaniv. Deep Anomaly Detection Using Geometric Transformations. URL: https://proceedings.neurips.cc/paper/2018/file/5e62d03aec0d17facfc5355dd90d441c-Paper.pdf (Last accessed: 10. 11. 2021).

11. Mohammad Braei, Sebastian Wagner. Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art. URL: https://www.semanticscholar.org/paper/Anomaly-Detection-in-Univariate-Time-series%3A-A-on-Braei-Wagner/cf45bce52cca1f6e450ddaa1d19fe6e30661dffb (Last accessed: 10. 11. 2021).

12. Atiq ur Rehman & Samir Brahim Belhaouari. Unsupervised outlier detection in multidimensional data

13. Victoria J. Hodge and Jim Austin. A Survey of Outlier Detection Methodologies. URL: https://core.ac.uk/download/pdf/58585.pdf (Last accessed: 12. 11. 2021).

14. Karanjit Singh and Dr. Shuchita Upadhyaya. Outlier Detection: Applications And Techniques. URL:https://www.researchgate.net/publication/267964435_Outlier_Detection_Applications_And_Techniques (Last accessed: 12. 11. 2021).

15. Wang S. и др. Effective End-to-end Unsupervised Outlier Detection via Inlier Priority of Discriminative Network. URL: https://proceedings.neurips.cc/paper/2019/hash/6c4bb406b3e7cd5447f7a76fd7008806-Abstract.html (Last accessed: 12. 11. 2021).

16. Karanjit Singh and Dr. Shuchita Upadhyaya. Outlier Detection: Applications And Techniques. URL: https://www.researchgate.net/publication/228686398_k-Nearest_neighbour_classifiers (Last accessed: 14. 11. 2021).

17. Yen-Chang Hsu, Yilin Shen, Hongxia Jin, Zsolt Kira. Generalized ODIN: Detectiong Out-of-distribution Image without Learning from Out-jf-distribution Data. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Hsu_Generalized_ODIN_Detecting_Out-of-Distribution_Image_Without_Learning_From_Out-of-Distribution_Data_CVPR_2020_paper.pdf (Last accessed: 14. 11. 2021).

18. Breunig M. и др. LOF // Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. 2000. URL: https://www.researchgate.net/publication/221214719_LOF_Identifying_Density-Based_Local_Outliers (Last accessed: 15. 11. 2021).

19. Na S., Xumin L., Yong G. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm // 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. 2010.

20. Markus Goldstein, Andreas Dengel. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm. URL: https://www.researchgate.net/publication/231614824_Histogram-based_Outlier_Score_HBOS_A_fast_Unsupervised_Anomaly_Detection_Algorithm (Last accessed: 17. 11. 2021).

21. Warp-core. URL: https://workday.github.io/warp-core/contents/anomaly_detection/ (Last accessed: 17. 11. 2021).

22. PDF) Support Vector Machines: Theory and Applications. URL:https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications (Last accessed: 17. 11. 2021).

23. Fei Tony Liu, Kai Ming TingGippsland School of Information TechnologyMonash University, Victoria, Australia. Isolation Forest. URL: https://www.researchgate.net/publication/224384174_Isolation_Forest (Last accessed: 18. 11. 2021).

24. Xuehui Wang, Yong Zhang, Hao Liu, Yang Wang, Lichun Wang , and Baocai Yin. An Improved Robust Principal Component Analysis Model for Anomalies Detection of Subway Passenger Flow.

25. JR L., GG K. The measurement of observer agreement for categorical data. URL: https://pubmed.ncbi.nlm.nih.gov/843571/ (Last accessed: 18. 11. 2021).

26. Study Finance. URL: https://studyfinance.com/static/media/z-score.png (Last accessed: 21. 11. 2021).

27. A Brief Overview of Outlier Detection Techniques .URL: https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561 (Last accessed: 21. 11. 2021).

28. Demo of DBSCAN clustering algorithm. URL: https://scikit-learn.org/stable/_images/sphx_glr_plot_dbscan_001.png (Last accessed: 22. 11. 2021).

29. Outlier detection with Local Outlier Factor (LOF). URL: https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html#:~:text=The%20Local%20Outlier%20Factor%20(LOF,lower%20density%20than%20their%20neighbors. (Last accessed: 22. 11. 2021).

30. Anomaly Detection using Autoencoders.URL: https://towardsdatascience.com/anomaly-detection-using-autoencoders-5b032178a1ea (date of application: 22. 11. 2021).

31. Cloudera Fast Forward. Deep Learning for Anomaly Detection. URL: https://ff12.fastforwardlabs.com/ (Last accessed: 24. 11. 2021)

Надійшла у першій редакції 06.11.2021, в останній - 11.02.2022.