Міністерство освіти та науки України
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова
праця на правах рукопису

АЩЕПКОВ ВАЛЕРІЙ ОЛЕГОВИЧ

УДК 004.8:006.91

ДИСЕРТАЦІЯ
ОБРОБКА РЕЗУЛЬТАТІВ ВИМІРЮВАНЬ ВИТРАТИ РІДИНИ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

Спеціальність: 152 Метрологія та інформаційно-вимірювальна техніка
Галузь знань: 15 Автоматизація та приладобудування

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____ В.О. Ащепков

Науковий керівник:
Скляров Володимир Васильович
доктор технічних наук, с.н.с

Харків - 2024

АНОТАЦІЯ

Ащепков В.О. Обробка результатів вимірювань витрати рідини з використанням машинного навчання. - Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук (доктора філософії) за спеціальністю 152 “Метрологія та інформаційно-вимірювальна техніка” (15 - Автоматизація та приладобудування). - Харківський національний університет радіоелектроніки, Харків, 2024.

Дисертаційна робота присвячена розробці підходів до обробки результатів вимірювань витрати рідини з використанням методів машинного навчання для виявлення викидів і підвищення якості метрологічних даних. Основний акцент зроблено на вдосконаленні процесів ідентифікації та виключення викидів із даних вимірювань, що дозволяє підвищити точність і стабільність результатів.

Метою дисертаційної роботи є підвищення стабільності вимірювань витрати рідини та достовірності метрологічних даних шляхом розробки та адаптації алгоритму на основі моделі ізольованого лісу. Виключення надмірних похибок дозволяє зменшити стандартну невизначеність за типом А, що сприяє забезпеченню надійності отриманих результатів.

Об'єкт дослідження — це процес вимірювання витрати рідини та обробки метрологічних даних.

Предмет дослідження — є виявлення викидів у метрологічних даних і їх вплив на стабільність і достовірність результатів вимірювань.

У *вступі* дисертаційної роботи детально обґрунтовано актуальність дослідження, що полягає у застосуванні методів машинного навчання для вдосконалення обробки результатів вимірювань витрати рідини, для підвищення точності й надійності метрологічних процесів. Зважаючи на зростаючу складність вимірювальних систем та необхідність забезпечення високої точності, використання алгоритмів машинного навчання стає перспективним напрямом для сучасної метрології. У роботі визначено мету, завдання, об’єкт і предмет

дослідження, що стосуються впровадження нових методів для виявлення аномалій у вимірювальних даних, що дозволяє значно знизити стандартну невизначеність вимірювань за типом А.

У вступі дисертаційної роботи детально обґрунтовано актуальність дослідження, що полягає у вдосконаленні обробки результатів вимірювань витрати рідини з використанням методів машинного навчання для підвищення стабільності, точності та достовірності метрологічних даних. Зважаючи на зростаючу складність вимірювальних систем та необхідність забезпечення високих стандартів точності, використання алгоритмів машинного навчання, визнано перспективним напрямом для сучасної метрології. Розкрито наукову новизну та практичне значення отриманих результатів, а також особистий внесок автора в розробку й реалізацію нових методик. Представлено етапи проведених досліджень, надано відомості про апробацію та впровадження отриманих результатів у практику, а також наведено кількість публікацій за темою роботи, що підкреслює важливість та актуальність дослідження.

Перший розділ дисертації присвячено глибокому аналізу існуючих методів машинного навчання та статистичних методів обробки даних, які використовуються для виявлення викидів у метрологічних даних. Наведено класифікацію та розкрито принципи роботи основних моделей машинного навчання, таких як лінійна та логістична регресія, метод найближчих сусідів, дерева рішень, метод опорних векторів та нейронні мережі. Для кожного методу проаналізовано особливості, переваги та обмеження при застосуванні до завдань виявлення викидів у вимірювальних даних. Окрім цього, у розділі представлено огляд основних статистичних методів для ідентифікації викидів, таких як параметричні та робастні методи, а також порівняння цих підходів із методами машинного навчання з точки зору їх точності, надійності та гнучкості. Цей порівняльний аналіз дозволив виділити переваги машинного навчання, його здатність адаптуватися до різноманітних типів даних і забезпечувати більшу точність у порівнянні з класичними статистичними методами.

Другий розділ роботи зосереджений на дослідженні метрологічних характеристик «Державного первинного еталона одиниці об’ємної і масової витрати рідини, об’єму і маси рідини, що протікає по трубопроводу» включаючи стабільність витрати та оцінку невизначеності вимірювань коріолісових витратомірів. Представлено аналіз стабільності вимірювань та особливостей вимірювального процесу, що дає змогу оцінити точність і надійність вимірювального обладнання, особливо в умовах міжнародних звірень. Проведені дослідження заклали основи для подальшого застосування методів машинного навчання для підвищення точності вимірювань.

Третій розділ дисертації розглядає процес вибору моделі машинного навчання для виявлення викидів. У цьому розділі детально обґрунтовано вибір моделі ізольованого лісу як оптимального інструмента для ідентифікації викидів у вимірювальних даних. Описано основні параметри та методологію налаштування моделі, такі як «кількість дерев», «кількість припущених аномалій» та «критерій зупинки», які відіграють ключову роль у досягненні точних результатів. Крім того, у розділі представлено результати експериментальних досліджень роботи моделі на реальних даних, включаючи аналіз отриманих результатів. Порівняння з робастними статистичними методами підкреслило надійність отриманих результатів та підтвердило, що модель ізольованого лісу є ефективним інструментом, який можна успішно використовувати для практичних метрологічних завдань нарівні з існуючими статистичними методами. Це забезпечує можливість її інтеграції у сучасні системи обробки метрологічних даних для підвищення точності та стабільності результатів.

Четвертий розділ роботи зосереджено на вдосконаленні алгоритму ізольованого лісу для метрологічних задач. У роботі запропоновано підхід до стабілізації результатів шляхом усереднення шкал аномальності при різних значеннях параметра "кількість припущених аномалій". Цей підхід дозволяє зменшити залежність результатів від випадкових факторів у вибірках та покращити стабільність шкали аномальності.

Практичне застосування алгоритму продемонстровано на вимірювальних даних, де вдосконалений алгоритм дозволив ідентифікувати та виключити надмірні похибки, що забезпечило зниження стандартної невизначеності за типом А та підвищення точності результатів. Запропонований підхід довів свою ефективність у задачах підвищення стабільності метрологічних даних, підвищуючи надійність та точність результатів.

У *висновках* дисертаційної роботи узагальнено основні результати дослідження, підсумовано значення наукових і практичних досягнень. Показано, що запропоновані методи виявлення викидів у вимірювальних даних сприяють зниженню стандартної невизначеності вимірювань за типом А, що забезпечує підвищення стабільності та достовірності результатів вимірювань.

Отримані результати підтверджують ефективність використання алгоритму на основі моделі ізольованого лісу для вирішення метрологічних задач та демонструють його перспективність як інструмента для вдосконалення процесів обробки даних.

Результати роботи формують основу для подальших досліджень, а саме аналізу причин нестабільності на окремих діапазонах витрат, вдосконалення алгоритмів виявлення викидів та розширення сфер їх застосування в метрології.

Наукова новизна результатів дисертаційного дослідження:

- Вперше проведено дослідження технічних факторів, що впливають на стабільність витрати рідини в умовах державного еталона, із розробкою заходів для зменшення варіативності даних;
- Вперше запропоновано використання моделі ізольованого лісу для виявлення викидів у метрологічних даних із малими вибірками та невідомим розподілом, що дозволило підвищити достовірність результатів вимірювань;
- Порівняння результатів роботи алгоритму ізольованого лісу з робастними статистичними методами підтвердило його переваги для метрологічних задач, що забезпечило нові підходи до виявлення викидів у вимірювальних даних.

- Розроблено та адаптовано алгоритм на основі ізольованого лісу для специфіки метрологічних даних, що дозволило знизити стандартну невизначеність за типом А та підвищити стабільність і достовірність результатів вимірювань.

Практичне значення отриманих результатів дисертації полягає у впровадженні алгоритму на основі моделі ізольованого лісу в процеси метрологічної діяльності, які здійснюються з використанням «Державному первинному еталоні одиниці об'ємної та масової витрати рідини, об’єму та маси рідини, що протікає по трубопроводу». Розроблений алгоритм дозволяє автоматично ідентифікувати та виключати аномальні значення у вимірювальних даних, що особливо важливо для забезпечення високої точності і стабільності під час метрологічних досліджень. Це, у свою чергу, сприяє зниженню похибок і підвищенню надійності результатів, що особливо важливо для забезпечення достовірності під час проведення міжнародних звірень.

Розроблений у дисертації підхід до обробки результатів вимірювань витрати рідини під час міжнародних звірень впроваджено як актуальний метод аналізу вимірювальних даних у дисципліні «Державні еталони України» в навчальному процесі Харківського національного університету радіоелектроніки на кафедрі Інформаційно-вимірювальних технологій. Він сприяє набуттю студентами практичних навичок роботи з сучасними вимірювальними технологіями, доповнюючи навчальний процес і підвищуючи обізнаність щодо сучасних методів метрологічного контролю та аналізу даних

Отримані результати використані в ПрАТ «Енергооблік» для покращення якості продукції та оптимізації виробничих процесів. Це стосується, зокрема:

- використання результатів дослідження для аналізу експериментальних даних під час розробки нових моделей витратомірів-лічильників, що дозволить ідентифікувати потенційні джерела похибок та покращити конструкцію пристроїв на етапі розробки;
- забезпечення стабільності роботи проливних установок завдяки виявленню аномалій, що сприятиме швидкому виявленню відхилень від норми,

знижуючи ризик виходу обладнання з ладу та зменшуючи кількість позапланових обслуговувань;

- можливості інтеграції алгоритму на основі моделі ізольованого лісу у систему контролю якості на виробництві, що дозволить автоматично виявляти аномалії у вимірювальних даних, підвищуючи надійність і точність результатів.

Ключові слова: похибка, невизначеність вимірювань, еталон, масова витрата, вимірювання, метод, обробка результатів вимірювань, випробуванння, викиди, машинне навчання, класифікація, ізольований ліс, дерево рішень, виявлення аномалій.

Список публікацій здобувача

Наукові праці, в яких опубліковано основні наукові результати:

1. V. Aschepkov, "Methods of machine learning in modern metrology," *Measuring Equipment and Metrology*, vol. 85, no. 1, pp. 57–60, 2024. doi: 10.23939/istcmtm2024.01.

2. В.О. Ащепков, "Використання моделі ISOLATION FOREST для виявлення аномалій у даних вимірювань," *Сучасний стан наукових досліджень та технологій в промисловості*, no. 1 (27), pp. 98–113, 2024. doi: 10.30837/ITSSI.2023.26.

3. В.О. Ащепков, "Дослідження метрологічних характеристик державного первинного еталона одиниці об’ємної та масової витрати рідини при підготовці до участі у міжнародних звіреннях," *Український метрологічний журнал*, no. 1 (77), pp. 31–37, 2024. doi: 10.24027/2306-7039.1.2024.300937.

4. V. Aschepkov, "Methods for outlier detection in metrological studies," *Measuring Equipment and Metrology*, vol. 85, no. 3, pp. 25–29, 2024. doi: 10.23939/istcmtm2024.03.025.

5. В.О. Ащепков, Д.Ю. Бяллович, В.В. Скляров, "Вплив порогових значень на стандартну невизначеність типу А при вимірюваннях масової витрати рідини," *Український метрологічний журнал*, no. 3 (30), 2024. doi: 10.24027/2306-7039.3.2024.312469.

Результати, які засвідчують апробацію матеріалів дисертації:

6. V. Aschepkov, "Improving the efficiency of processing of measurement results using the machine learning method," in *Theses of Reports 1st European Competition of Young Best Metrologists in Ukraine*, Ivano-Frankivsk, Ukraine, June 24–28, 2024, pp. 6–8.

ABSTRACT

Aschepkov V.O. Processing of Liquid Flow Measurement Results Using Machine Learning. - Qualification scientific work as a manuscript.

Dissertation for the degree of Candidate of Technical Sciences (Doctor of Philosophy) in specialty 152 “Metrology and Information-Measuring Technology” (15 - Automation and Instrumentation). - Kharkiv National University of Radio Electronics, Kharkiv, 2024.

The dissertation focuses on the development of approaches for processing liquid flow measurement results using machine learning methods to detect outliers and improve the quality of metrological data. The main emphasis is on improving the processes of identifying and excluding outliers from measurement data, which allows increasing the accuracy and stability of results.

The purpose of the dissertation is to enhance the stability of liquid flow measurements and the reliability of metrological data by developing and adapting an algorithm based on the isolation forest model. Eliminating excessive errors reduces standard uncertainty of type A, contributing to the reliability of obtained results.

Object of the research: the process of measuring liquid flow and processing metrological data.

Subject of the research: the detection of outliers in metrological data and their impact on the stability and reliability of measurement results.

The **introduction** substantiates the relevance of the research, which lies in the application of machine learning methods to improve the processing of liquid flow measurement results to enhance the accuracy and reliability of metrological processes. Considering the growing complexity of measurement systems and the need to ensure high precision, the use of machine learning algorithms is recognized as a promising direction for modern metrology. The purpose, objectives, object, and subject of the research are defined, focusing on the implementation of new methods for detecting anomalies in measurement data, which significantly reduces the standard uncertainty of type A.

The dissertation emphasizes the improvement of data processing methods for liquid flow measurements using machine learning to increase stability, accuracy, and reliability. The scientific novelty and practical significance of the obtained results are revealed, as well as the author's personal contribution to the development and implementation of new methodologies. The stages of the research are presented, along with information on the approbation and implementation of the results in practice, and the number of publications on the topic, highlighting the importance and relevance of the research.

The first of the dissertation is devoted to a thorough analysis of existing machine learning and statistical methods for processing data used to detect outliers in metrological measurements. It provides a classification and detailed description of the principles behind the main machine learning models, including linear and logistic regression, k-nearest neighbors, decision trees, support vector machines, and neural networks. Each method is analyzed for its features, advantages, and limitations when applied to the task of detecting outliers in measurement data. Additionally, an overview of the main statistical methods for identifying outliers, such as parametric and robust methods, is presented. A comparative analysis of these approaches with machine learning methods highlights the advantages of machine learning, including its ability to adapt to various data types and achieve greater accuracy compared to classical statistical methods.

The second section focuses on studying the metrological characteristics of the "State Primary Standard of Unit of Volume and Mass Flow of Liquids, Volume and Mass of Liquid Flowing Through a Pipeline," including the stability of flow and the evaluation of measurement uncertainty for Coriolis flowmeters. It presents an analysis of flow stability and the specific features of the measurement process, enabling the evaluation of the accuracy and reliability of measurement equipment, especially in the context of international comparisons. The research lays the foundation for further application of machine learning methods to improve measurement accuracy.

The third section of the examines the process of selecting a machine learning model for detecting outliers. This chapter provides a detailed justification for choos-

ing the isolation forest model as an optimal tool for identifying anomalies in measurement data. Key parameters and setup methodologies, such as the number of trees, the number of presumed anomalies, and stopping criteria, are described. The chapter also includes experimental results from applying the model to real data and a thorough analysis of the outcomes. A comparison with robust statistical methods emphasizes the reliability and effectiveness of the isolation forest model as a practical tool for metrological tasks

The fourth section of the centers on improving the isolation forest algorithm for metrological applications. A new approach is proposed to stabilize results by averaging anomaly scales under different parameter settings, thereby reducing dependence on random factors in the datasets and improving the stability of the anomaly scale. Practical application of the algorithm demonstrated its ability to identify and exclude outliers, reducing standard uncertainty of type A and improving result accuracy. This approach proved effective in enhancing the stability of metrological data, thereby improving reliability and accuracy in practical tasks.*The conclusions* of the dissertation summarize the main research results and the scientific and practical achievements obtained during the work. The importance of the research for the further development of technologies in metrology, particularly in anomaly detection in measurement data, is highlighted. Prospects for further research are outlined, including the possibility of algorithm improvement and expanding its application to ensure even higher quality data processing in metrology.

The conclusions of the dissertation summarize the main findings of the research and assess the significance of its scientific and practical achievements. It has been shown that the proposed methods for detecting outliers in measurement data contribute to reducing the standard uncertainty of type A, which ensures improved stability and reliability of measurement results.

The results of this dissertation form a foundation for further studies, specifically the analysis of causes of instability in specific flow rate ranges, improvement of outlier detection algorithms, and the expansion of their application areas in metrology.

Scientific novelty of the dissertation research:

- For the first time, technical factors influencing the stability of liquid flow under the conditions of the state standard were studied, resulting in the development of measures to reduce data variability;
- For the first time, the isolation forest model was proposed for detecting outliers in metrological data with small samples and unknown distributions, improving the reliability of measurement results;
- A comparison of the isolation forest algorithm's performance with robust statistical methods confirmed its advantages for metrological tasks, introducing new approaches to outlier detection in measurement data.; An algorithm based on the isolation forest model was developed and adapted for the specifics of metrological data, reducing standard uncertainty of type A and enhancing the stability and reliability of measurement results.

The practical significance of the results lies in the implementation of the isolation forest-based algorithm in metrological processes conducted using the "State Primary Standard of Unit of Volume and Mass Flow of Liquids, Volume and Mass of Liquid Flowing Through a Pipeline." The developed algorithm enables automatic identification and exclusion of outliers in measurement data, which is particularly important for ensuring high accuracy and stability during metrological research. This, in turn, reduces errors and improves reliability, which is especially critical for ensuring accuracy during international comparisons.

The approach to processing liquid flow measurement results developed in this dissertation was introduced as a relevant method of data analysis within the discipline "State Standards of Ukraine" in the educational process at Kharkiv National University of Radio Electronics, at the Department of Information and Measurement Technologies. It helps students gain practical skills in working with modern measurement technologies, enriching the educational process and raising awareness of modern methods of metrological control and data analysis.

The obtained results were utilized at PrJSC "Energooblik" to improve product quality and optimize production processes, specifically:

- Using the research results to analyze experimental data during the development of new flowmeter models, enabling the identification of potential error sources and improving device design during the development stage;
- Ensuring the stability of flow rigs by detecting anomalies, facilitating the rapid identification of deviations from norms, reducing the risk of equipment failure, and minimizing unscheduled maintenance;
- Integrating the isolation forest-based algorithm into the quality control system at the production level, enabling automatic detection of anomalies in measurement data, thereby enhancing the reliability and accuracy of results.

Keywords: error, measurement uncertainty, standard, mass flow, measurement, method, measurement result processing, testing, outliers, machine learning, classification, isolation forest, decision tree, anomaly detection.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАК, ОДИНИЦЬ І ТЕРМІНІВ

ML — машинне навчання

ІЛ — ізольований ліс

ЛР — лінійна регресія

LR — логістична регресія

k-NN — метод найближчих сусідів

ДР — дерева рішень

SVM — метод опорних векторів

НМ — нейронні мережі

IQR — інтерквартильний розмах

MAD — медіанне абсолютне відхилення

ДЕТУ 03-04-04 — Державний первинний еталон одиниці об’ємної і масової витрати рідини та об’єму і маси рідини, що протікає по трубопроводу

ЗВТ — засоби вимірювальної техніки

ВД — вимірювальна ділянка трубопроводу

One-Class SVM —Однокласова машина опорних векторів

ВСТУП

Актуальність теми дослідження.

Швидкий розвиток технологій та зростаючі вимоги до точності та надійності вимірювань ставлять перед науковцями та фахівцями галузі метрології завдання пошуку надійних і точних методів обробки результатів вимірювань та виявлення викидів.

На сьогоднішній день методи машинного навчання (далі — ML) здобули широкого застосування у різних галузях науки та техніки. Вони дозволяють якісно вирішувати складні завдання аналізу даних та прогнозування на основі навчання на великих обсягах інформації. У зв'язку з цим використання методів ML для обробки результатів вимірювань та виявлення викидів є перспективним напрямом досліджень.

У роботі проведений критичний аналіз існуючих методів обробки даних та виявлення викидів з метою визначення їхніх переваг та недоліків. Впровадження методів ML, зокрема методу ізольованого лісу (далі — ІЛ), у цю сферу відкриває нові можливості для підвищення точності та надійності вимірювань, що має велике значення для сучасних метрологічних процесів.

Обгрунтування вибору теми дослідження. На сьогоднішній день особливо важливою є потреба в надійних та точних методах виявлення викидів у вимірювальних даних. Виявлення викидів у процесі обробки даних вимірювань на еталонах механічних вимірювань є актуальним завданням, яке вирішується за допомогою різних методів та підходів, серед яких методи ML та статистичного аналізу даних є особливо перспективними.

Саме тому для вирішення задачі виявлення викидів у вимірювальних даних у рамках підготовки "Державного первинного еталона об'ємної та масової витрати рідини, об'єму та маси рідини, що протікає по трубопроводу" (далі — ДЕТУ 03-04-04) до міжнародних звірень у напрямку масової витрати рідини було обрано метод ІЛ. Його надійність та точність у виявленні викидів було

перевірено та підтверджено за допомогою статистичних методів, що дозволило забезпечити високу якість обробки даних вимірювань.

Мета та завдання дослідження. Метою дисертаційної роботи є підвищення стабільності вимірювань витрати рідини та достовірності метрологічних даних шляхом розробки та адаптації алгоритму на основі моделі ІЛ. Виключення надмірних похибок дозволяє зменшити стандартну невизначеність за типом А, що сприяє забезпеченню надійності отриманих результатів

Для досягнення цієї мети поставлені наступні завдання:

- Провести аналіз статистичних методів і методів ML для виявлення викидів, визначивши їх переваги, обмеження та придатність до задач метрологічних досліджень із малими вибірками та невідомим розподілом.
- Дослідити технічні фактори, що впливають на стабільність витрати рідини в умовах державного еталона, та розробити заходи для зменшення варіативності результатів вимірювань.
- Використати метод ІЛ для виявлення викидів у метрологічних даних, перевіривши його надійність через порівняння з робастними статистичними методами.
- Розробити алгоритм на основі моделі ІЛ, адаптований до специфіки метрологічних даних, та оцінити його ефективність у зниженні стандартної невизначеності за типом А і покращенні достовірності вимірювань.

Об'єкт дослідження — це процес вимірювання витрати рідини та обробки метрологічних даних.

Предмет дослідження — є виявлення викидів у метрологічних даних і їх вплив на стабільність і достовірність результатів вимірювань.

Методи дослідження. У роботі застосовано метод ІЛ, метод міжквартильного розмаху (далі — IQR) та метод медіанного абсолютного відхилення (далі — MAD).

Наукова новизна результатів дисертаційного дослідження:

- Вперше проведено дослідження технічних факторів, що впливають на стабільність витрати рідини в умовах державного еталона, із розробкою заходів для зменшення варіативності даних;
- Вперше запропоновано використання моделі ІЛ для виявлення викидів у метрологічних даних із малими вибірками та невідомим розподілом, що дозволило підвищити достовірність результатів вимірювань;
- Порівняння результатів роботи алгоритму ІЛ з робастними статистичними методами підтвердило його переваги для метрологічних задач, що забезпечило нові підходи до виявлення викидів у вимірювальних даних;
- Розроблено та адаптовано алгоритм на основі ІЛ для специфіки метрологічних даних, що дозволило знизити стандартну невизначеність за типом А та підвищити стабільність і достовірність результатів вимірювань.

Особистий внесок здобувача. Всі етапи дослідження, від обґрунтування вибору теми до аналізу отриманих результатів були отримані здобувачем. Він був відповідальним за планування та виконання експериментальних досліджень, обробку та аналіз вимірювальних даних, а також за розробку та налаштування алгоритму на основі методу ІЛ для виявлення викидів. При цьому, всі ідеї, аналіз та підготовка матеріалів для дисертації є результатом власного особистого внеску здобувача, хоча деякі з аспектів дослідження були обговорені та затверджені спільно з науковим керівником та колегами. У додаток до цього, було опубліковано п’ять наукових робіт, з яких чотири виконані без співавторів.

У науковій праці, виконаній у співавторстві, особистим внеском здобувача є:

- розрахунок викидів методом ІЛ;
- розрахунок викидів статистичними методами;
- розрахунок стандартної невизначеності вимірювань за типом А;
- аналіз отриманих результатів та формулювання висновків.

Список наукових праць здобувача представлений у Додатку А.

Апробація результатів дисертації. Результати роботи були представлені та обговорені на 1st European Competition of Young Best Metrologists in Ukraine, Ivano-Frankivsk, Ukraine, June 24, 2024 – June 28, 2024.

Публікації. Основні положення та результати дисертаційної роботи досить повно відображені у 5 друкованих працях наукових журналів, включених до «Переліку наукових фахових видань України», з них 2 – категорія А, що входить до наукометричної бази Web of Science; 3 – категорії Б.

Структура та обсяг дисертації. Дисертаційна робота складається з вступу, 4 розділів, висновків, списку використаних джерел і додатків. Повний обсяг дисертації становить 198 сторінках, включаючи 70 рисунків за текстом; 8 таблиць за текстом; 4 додатки на 14 сторінках; 108 найменувань використаних джерел інформації на 14 сторінках.

Практичне значення отриманих результатів дисертації полягає у впровадженні алгоритму на основі моделі ІЛ в процеси метрологічної діяльності, які здійснюються з використанням «Державного первинного еталона одиниці об'ємної та масової витрати рідини, об’єму та маси рідини, що протікає по трубопроводу». Розроблений алгоритм дозволяє автоматично ідентифікувати та виключати викиди у вимірювальних даних, що особливо важливо для забезпечення високої точності і стабільності під час метрологічних досліджень. Це, у свою чергу, сприяє зниженню похибок і підвищенню надійності результатів, що особливо важливо для забезпечення достовірності під час проведення міжнародних звірень.

Розроблений у дисертації підхід до обробки результатів вимірювань витрати рідини під час міжнародних звірень впроваджено як актуальний метод аналізу вимірювальних даних у дисципліні «Державні еталони України» в навчальному процесі Харківського національного університету радіоелектроніки на кафедрі Інформаційно-вимірювальних технологій. Він сприяє набуттю студентами практичних навичок роботи з сучасними вимірювальними технологіями, доповнюючи навчальний процес і підвищуючи обізнаність щодо сучасних методів метрологічного контролю та аналізу даних

Отримані результати використані в ПрАТ «Енергооблік» для покращення якості продукції та оптимізації виробничих процесів. Це стосується, зокрема:

- використання результатів дослідження для аналізу експериментальних даних під час розробки нових моделей витратомірів-лічильників, що дозволить ідентифікувати потенційні джерела похибок та покращити конструкцію пристроїв на етапі розробки;
- забезпечення стабільності роботи проливних установок завдяки виявленню викидів, що сприятиме швидкому виявленню відхилень від норми, знижуючи ризик виходу обладнання з ладу та зменшуючи кількість позапланових обслуговувань;
- можливості інтеграції алгоритму на основі моделі ІЛ у систему контролю якості на виробництві, що дозволить автоматично виявляти викиди у вимірювальних даних, підвищуючи надійність і точність результатів.

Документи, що підтверджують впровадження результатів дисертації представлені в Додатку Б.

1. АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА МЕТОДІВ СТАТИСТИЧНОЇ ОБРОБКИ ДАНИХ

1.1 Аналіз методів машинного навчання

1.1.1 Введення в машинне навчання та його застосування в метрології

ML є підгалуззю штучного інтелекту, що зосереджена на розробці алгоритмів і моделей, здатних навчатися з даних без необхідності програмувати кожен крок процесу. Традиційно в програмуванні кожен етап виконання задачі чітко визначається людиною, однак ML дозволяє системам адаптуватися до нових умов, аналізуючи великі обсяги даних і виявляючи в них закономірності без прямої участі людини. Розвиток обчислювальних технологій і збільшення доступних даних (Big Data) сприяли швидкому розвитку ML, що забезпечило широке застосування цих методів у багатьох сферах — від медицини та фінансів до транспорту і енергетики. У результаті таких досягнень з’явилася можливість вирішувати складні завдання та значно покращити ефективність аналізу даних, підвищуючи точність та швидкість прийняття рішень.

Зокрема, в медицині алгоритми ML використовуються для обробки медичних зображень, аналізу генетичних даних, діагностики захворювань і прогнозування результатів лікування, що дозволяє покращити якість медичних послуг і знизити ймовірність помилок. У фінансовому секторі ML допомагає виявляти аномалії в транзакціях, аналізувати ринкові тренди та прогнозувати економічні зміни, що робить фінансові операції безпечнішими і точнішими. У транспорті ці технології дозволяють автоматизувати управління транспортними засобами, оптимізувати маршрути і підвищувати безпеку перевезень. У промисловості ML використовується для прогнозування технічного стану обладнання, що дозволяє значно знизити витрати на обслуговування та уникати позапланових зупинок виробництва.

Метрологія, як наука про вимірювання, є однією з тих сфер, де ML має значний потенціал для підвищення ефективності та точності вимірювальних процесів. Впровадження методів ML у метрології дозволяє значно полегшити обробку великих обсягів вимірювальних даних і виявляти складні закономірності, що можуть бути непомітними для традиційних методів. Використання ML в цій галузі відкриває нові можливості для автоматизації аналізу та підвищення надійності вимірювань, скорочуючи людський фактор і збільшуючи об'єктивність процесів.

Одним з основних напрямків застосування ML в метрології є автоматичне виявлення викидів у вимірювальних даних. Алгоритми ML дозволяють швидко виявляти відхилення від норми, що може сприяти зниженню ймовірності помилок, пов’язаних з людським фактором. Це дає змогу значно підвищити точність вимірювань і забезпечити більшу надійність результатів. Зокрема, виявлення викидів допомагає своєчасно коригувати вимірювальні процеси, якщо наявні відхилення можуть впливати на результати досліджень чи виробничі показники [1].

ML також знаходить своє застосування в автоматизації калібрування вимірювальних приладів, що дозволяє створювати адаптивні системи, здатні самостійно налаштовувати пристрої в залежності від змін умов експлуатації. Це не тільки підвищує точність вимірювань, але й зменшує час, необхідний для проведення налаштувань, що особливо важливо в умовах високої швидкості технологічних процесів. Крім того, ML активно використовується для поліпшення методів прогнозування та утримання еталонів, що є критично важливим для підтримання високої точності вимірювань в галузях, де точність є вирішальним фактором, наприклад, у енергетиці, аерокосмічній промисловості чи в хімічному виробництві.

Іншою важливою сферою застосування ML є автоматизація процесів метрологічної верифікації та сертифікації продукції. Завдяки цим методам, можна забезпечити більш ефективний контроль за відповідністю продукції встановле-

ним стандартам і вимогам, що сприяє зниженню ризиків помилок у процесі перевірки та покращенню ефективності сертифікаційних процедур [2].

ML також застосовується для моніторингу та діагностики технічного стану обладнання. Алгоритми, здатні працювати з даними в реальному часі, допомагають виявляти відхилення від норми, що дозволяє своєчасно проводити технічне обслуговування та уникати аварійних ситуацій. Такі системи можуть прогнозувати потенційні проблеми ще до їх виникнення, що знижує ймовірність виникнення несправностей і знижує витрати на ремонт [3].

Загалом, використання ML в метрології відкриває нові можливості для автоматизації, підвищення точності та надійності вимірювальних процесів. Це дозволяє значно підвищити ефективність роботи вимірювальних приладів, зменшити вплив людського фактора і створити більш гнучкі системи, здатні адаптуватися до змін у технологічних умовах без необхідності постійного втручання людини. Можливість обробляти великі обсяги даних і виявляти навіть незначні аномалії дозволяє підвищити точність вимірювань, що має важливе значення для наукових досліджень, промислових процесів та розвитку нових технологій.

1.1.2 Класифікація моделей машинного навчання

Одним з ключових аспектів ML є методи навчання. Ці методи включають в себе різноманітні алгоритми та підходи, які використовуються для навчання моделей на основі даних. Основні класи методів ML включають у себе: навчання з учителем; навчання без учителя; навчання з підкріпленням.

Навчання з учителем передбачає навчання моделі на основі пар дані-відповідь. Ці пари даних використовуються для навчання моделі таким чином, щоб вона могла навчитися прогнозувати відповіді на нових даних. Такі методи включають у себе класифікацію, де модель намагається прогнозувати класи або категорії, регресію, де модель прогнозує числові значення, а також інші підходи, які передбачають наявність правильних відповідей для навчання.

Навчання без учителя, з іншого боку, не вимагає наявності правильних відповідей у наборі даних. Моделі навчаються здійснювати припущення та знаходити внутрішню структуру даних, такі як кластери або залежності. Такі методи використовуються для завдань, де важко або неможливо навчити модель на основі правильних відповідей, таких як аналіз текстів або виявлення аномалій.

Навчання з підкріпленням передбачає взаємодію моделі з довколишнім середовищем шляхом спроб і помилок. Модель отримує винагороду або штраф за свої дії і навчається оптимальним стратегіям. Ці методи часто використовуються для задач, де потрібно приймати послідовні рішення, таких як управління роботами чи вирішення задач з великим простором дій [4].

Усі ці методи ML мають свої переваги та недоліки, і їх вибір залежить від конкретного завдання та властивостей даних. Тому існує класифікації моделей ML за типами , наприклад навчання з учителем, навчання без учителя та навчання з підкріпленням можна класифікувати за типом методу навчання. Ось декілька основних типів класифікації моделей ML:

- **За типом вирішення задачі:**

1. **Класифікація:** Ця задача полягає в призначенні кожному вхідному прикладу одного з певного набору класів або міток. Наприклад, розпізнавання об'єктів на зображеннях, визначення категорії електронних листів як спам або не спам, визначення того, чи пацієнт має певне захворювання на основі його медичних даних;

2. **Регресія:** У цій задачі модель намагається передбачити числове значення для вихідної змінної на основі вхідних даних. Наприклад, прогнозування ціни акцій на основі ринкових даних, прогнозування рівня забруднення повітря на основі погодних умов та інших факторів;

3. **Кластеризація:** У цьому типі задач модель намагається групувати схожі елементи даних у класи або кластери, не маючи заздалегідь позначених категорій. Наприклад, сегментація клієнтів для рекламних кампаній на основі їхніх покупок та поведінки, або групування статей за схожими темами [5-6];

4. **Рекомендації:** У цьому типі задач модель намагається рекомендувати користувачам продукти, послуги або інші елементи на основі їхніх попередніх взаємодій або відповідності з іншими користувачами. Наприклад, системи рекомендацій для інтернет-магазинів або стрімінгових платформ;

5. **Виявлення аномалій:** У цій задачі модель намагається виявити аномалії або відхилення від звичайного шаблону в даних. Наприклад, виявлення фінансових шахрайств на основі аномальних транзакцій, або виявлення несправностей у виробництві на основі даних моніторингу обладнання [7].

- **За типом даних:**

– **Структуровані дані**: Моделі для структурованих даних використовуються, коли дані організовані у вигляді таблиць, де кожен стовпець відповідає певній ознаці, а кожен рядок представляє окремий елемент або спостереження. Це типово для даних баз даних та в ексель-подібних форматів;

– **Неструктуровані дані**: Моделі для неструктурованих даних використовуються для аналізу тексту, аудіо, відео та зображень. Ці дані не мають чіткої організації та вимагають спеціалізованих методів обробки.

- **За типом роботи моделі** :

- **Лінійні моделі**: Вони використовують лінійну функцію для моделювання залежності між вхідними та вихідними даними. Лінійні моделі, можуть бути ефективними для простих завдань, де взаємозв'язки між ознаками є лінійними;

- **Нелінійні моделі**: Ці моделі можуть моделювати більш складні залежності між вхідними та вихідними даними, які не можуть бути адекватно виражені лінійними функціями;

- **Ансамблеві моделі**: Ці моделі комбінують декілька базових моделей з метою покращення їх загальної ефективності. До них належать випадковий ліс, градієнтний бустінг, багатокласовий класифікатор AdaBoost тощо;

- **Байєсовські моделі**: Ці моделі використовують теорему Байєса для класифікації об'єктів на основі ймовірностей входження в певний клас.

Найвідомішими прикладами є наївні байєсовські класифікатори, такі як наївний байєсівський класифікатор;

- **Порядкові моделі**: Ці моделі призначені для прогнозування порядку або рангу елементів у вибірці. Вони можуть бути використані для ранжування результатів пошуку, оцінки схожості тощо;
- **Кластерні моделі**: Ці моделі використовуються для групування схожих елементів у кластери або групи без заздалегідь визначених категорій. Вони дозволяють виявляти природні структури у даних та знаходити схожі об'єкти;
- **Порогові моделі**: Ці моделі розділяють дані на дві або більше категорії на основі порогового значення. Вони часто використовуються для бінарної класифікації, де об'єкти класифікуються як належні до одного з двох класів в залежності від значення певної ознаки.

1.1.3 Основні методи машинного навчання

В сфері ML існує величезна кількість методів, але серед них можна виділити базові, основні методи, які лежать в основі більшості алгоритмів ML. Ці базові методи визначають загальні принципи роботи і допомагають у формуванні основ для розробки більш складних та спеціалізованих методів.

Лінійна регресія (далі — ЛР) є одним з найпростіших і широко використовуваних методів у ML. Вона відіграє важливу роль в статистиці та аналізі даних, дозволяючи моделювати зв'язки між неперервними змінними та прогнозувати значення однієї змінної на основі інших. Основна ідея полягає в тому, щоб знайти лінійну залежність між предикторами (незалежними змінними) та цільовою змінною (залежною змінною). Ця залежність виражається у вигляді лінійної функції, яка наближає розподіл даних.

В контексті ЛР передбачається, що зв'язок між предикторами та цільовою змінною може бути апроксимований за допомогою простої лінійної функції. Ми припускаємо, що існує лінійний зв'язок між предикторами та цільовою

змінною, і що помилка між передбаченими та спостережуваними значеннями є випадковою.

Математично ЛР виражається наступним чином [8]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots \ + \beta_n X_n + \varepsilon \ , \tag{1.1}$$

де Y - цільова змінна (залежна змінна),

X_1, X_2, ..., X_n - предиктори (незалежні змінні),

β_0, β_1, ..., β_n - коефіцієнти регресії (параметри моделі),

ε - випадкова помилка.

Мета ЛР полягає в оцінці коефіцієнтів β_0, β_1, ..., β_n, які найкращим чином апроксимують залежність між предикторами та цільовою змінною. Це зазвичай робиться шляхом мінімізації суми квадратів різниць між спостережуваними та передбаченими значеннями. Цей метод відомий як метод найменших квадратів (MLK).

Оцінка параметрів моделі здійснюється на основі навчального набору даних, який складається з пар значень предикторів та відповідних значень цільової змінної. Потім, використовуючи оцінені параметри, модель може бути використана для прогнозування значень цільової змінної для нових спостережень.

ЛР має декілька важливих властивостей та припущень. Одне з ключових припущень полягає в тому, що відношення між предикторами та цільовою змінною дійсно лінійне. Це означає, що зміна предиктора призводить до постійної зміни цільової змінної. Крім того, важливо перевірити припущення про нормальний розподіл залишків та рівність дисперсій.

ЛР є надзвичайно гнучким методом, який може бути адаптований для вирішення різних задач. Вона може бути застосована як для простих моделей з одним предиктором, так і для складних моделей з численними предикторами та їх взаємодіями. Крім того, ЛР має безліч розширень, які дозволяють враховувати особливості даних та вирішувати різні завдання.

Логістична регресія (“Logistic regression”, далі – LR) є потужним статистичним методом, який широко використовується для розв'язання задач класи-

фікації. Вона дозволяє моделювати ймовірність віднесення об'єкта до певного класу залежно від значень його ознак.

LR часто застосовується в медичній діагностиці, фінансовому аналізі, маркетингових дослідженнях та інших галузях, де важливо прогнозувати ймовірність настання події на основі наявних даних. Вона також може бути використана для оцінки впливу різних факторів на ймовірність настання події та для прийняття рішень на основі цих оцінок.

Ключовою особливістю LR є можливість виводу ймовірностей належності до класу, що дозволяє приймати більш обґрунтовані рішення та оцінювати ступінь невизначеності прогнозів. Вона також дозволяє оцінювати важливість різних ознак та їх вплив на прогнози моделі.

LR також є лінійним методом, що означає, що вона старається знайти лінійну гіперплощину, яка розділяє класи. Важливо зауважити, що LR не передбачає самі класи, а лише ймовірності належності до кожного класу. Потім ці ймовірності можуть бути використані для прийняття рішення щодо класифікації, наприклад, шляхом вибору класу з найбільшою передбаченою ймовірністю.

Одним з переваг LR є її інтерпретованість. Коефіцієнти, отримані під час навчання моделі, можна інтерпретувати як ваги ознак, тобто як внесок кожної ознаки у передбачення класу. Це робить LR дуже корисною для виявлення важливих ознак в даних.

Ще однією перевагою є відносна простота реалізації та навчання моделі. LR має помітно менше гіперпараметрів порівняно з іншими методами ML, що спрощує її використання, особливо в завданнях з невеликими обсягами даних.

Однак, як і у будь-якого методу, у LR є свої обмеження. Вона працює краще у випадку, якщо дані добре розділяються лінійно, що означає, що вона може бути неефективною у випадку складних не лінійних залежностей в даних. Крім того, вона може бути чутливою до мультиколінеарності ознак, коли ознаки сильно корелюють між собою.

Метод найближчих сусідів (k-Nearest Neighbors, далі - k-NN) є одним з найпростіших і широко використовуваних методів ML. Він застосовується для

вирішення завдань класифікації та регресії. Основна ідея методу полягає в передбаченні класу або значення цільової змінної для нового спостереження на основі його найближчих сусідів у просторі ознак.

Процес класифікації з використанням методу k-NN починається з визначення числа сусідів (k), які будуть використані для прийняття рішення. Потім для кожного нового спостереження визначаються k найближчих сусідів за допомогою певної метрики відстані, зазвичай використовується евклідова відстань.

Для задачі класифікації кожен з найближчих сусідів "голосує" за належність нового спостереження до певного класу. Наприклад, якщо більшість з k найближчих сусідів належать до класу "1", то нове спостереження також буде віднесено до класу "1".

У випадку задачі регресії значення цільової змінної для найближчих сусідів усереднюються, і отримане середнє значення використовується як прогноз для нового спостереження.

Метод k-NN не має явної структурної моделі, що робить його досить гнучким для різних типів даних і завдань. Однак він вимагає значних обчислювальних ресурсів при роботі з великими обсягами даних, оскільки для кожного нового спостереження необхідно обчислювати відстані до всіх існуючих об'єктів у вибірці. Крім того, вибір оптимального значення параметра k є важливим аспектом в роботі методу і може вплинути на його продуктивність.

Перевагами методу k-NN є його простота в реалізації, відносна незалежність від припущень про дані та здатність обробляти як числові, так і категоріальні ознаки. Однак його недоліками є неефективність при роботі з великими обсягами даних та низька швидкість передбачення для нових спостережень.

Метод k-NN знаходить своє застосування в різних галузях, включаючи біоінформатику, фінанси, медицину, маркетинг та інші. Він може бути використаний для вирішення завдань прогнозування, класифікації, рекомендаційних систем та аналізу даних .

Дерева рішень (далі –ДР) і їх ансамблі - це потужний інструмент у сфері ML, широко використовуваний для вирішення задач класифікації та регресії. Вони базуються на ідеї побудови деревоподібної структури, яка розділяє простір ознак на рекурсивні розбиття, що дозволяють робити прогнози для нових спостережень.

ДР починається з кореневого вузла, який містить усі навчальні дані. Потім алгоритм обирає ознаку та порогові значення, які найкраще розділяють дані на дві підгрупи з метою максимізації критерію розділення, такого як критерій Джині або ентропія. Цей процес повторюється рекурсивно для кожного вузла, поки не буде виконаний критерій зупинки, такий як мінімальна кількість спостережень у вузлі або досягнення максимальної глибини дерева.

Одне ДР може бути схильне до перенавчання, тому часто використовуються методи ансамблювання, такі як випадковий ліс і градієнтний бустинг, щоб покращити узагальнювальну здатність моделі.

Випадковий ліс представляє собою ансамбль ДР, де кожне дерево будується на основі випадкової підвибірки навчальних даних та випадкової підмножини ознак. Потім прогнози всіх дерев усереднюються, щоб отримати остаточний прогноз моделі. Цей підхід дозволяє знизити перенавчання та покращити узагальнювальну здатність моделі.

Градієнтний бустинг, з іншого боку, будує ансамбль дерев послідовно, кожне нове дерево спрямоване на покращення прогнозів попередніх дерев. Цей процес дозволяє моделі поступово покращувати свою продуктивність та досягати високої точності прогнозування.

Такі ансамблі дозволяють покращити якість прогнозів та знизити ризик перенавчання, зробивши модель більш стабільною та надійною на нових даних. Однак вони потребують налаштування гіперпараметрів та можуть бути більш витратними за обчислювальними ресурсами, ніж окремі дерева.

ДР мають кілька переваг, які роблять їх привабливим вибором для моделювання даних. По-перше, вони легко інтерпретовані, що означає, що прийняття рішень моделлю може бути легко пояснено людині. Кожне розбиття в дереві

представляє собою просте правило, засноване на значеннях ознак, що робить його зрозумілим та доступним для аналізу. Також вони здатні автоматично обробляти категоріальні та числові дані без необхідності попередньої обробки, такої як кодування категоріальних змінних. Вони також можуть працювати з даними, що містять пропущені значення, що робить їх зручним інструментом для аналізу реальних даних.

Ще однією перевагою ДР є їх здатність моделювати нелінійні залежності в даних. У той час як лінійні моделі можуть обмежуватися лише лінійними відносинами між змінними, ДР можуть виявляти та моделювати більш складні взаємозв'язки, що робить їх корисним інструментом для аналізу даних з нелінійними структурами.

Метод опорних векторів (Support Vector Machine, далі — SVM) належить до сімейства алгоритмів ML, що ґрунтуються на принципі статистичного навчання, який спрямований на знаходження оптимального розділення між різними класами даних. Основна ідея полягає в пошуку такої гіперплощини у багатовимірному просторі ознак, яка максимально розділяє об'єкти різних класів. Цей процес досягається шляхом максимізації відстані між класами, тобто відстані від найближчих об'єктів кожного класу до розділювальної гіперплощини.

Важливим аспектом SVM є його здатність працювати з нелінійними даними. У випадку, коли дані не можуть бути лінійно розділені, SVM використовує ядерні функції, що дозволяють відображати дані в простір більшої вимірності, де вони стають лінійно роздільними. Це дозволяє SVM ефективно обробляти складні та нелінійні залежності в даних.

Однією з ключових переваг SVM є його здатність до регуляризації. Це дозволяє контролювати складність моделі та запобігати перенавчанню шляхом покарання за великі ваги. Такий механізм регуляризації сприяє підвищенню узагальнюючої здатності моделі та поліпшенню її продуктивності на нових даних.

Крім того, SVM має хорошу масштабованість та може ефективно працювати з даними великих обсягів і високої вимірності. Це робить його привабли-

вим вибором для розв'язання різних завдань ML, включаючи класифікацію, регресію, виявлення аномалій та багато інших.

Штучні нейронні мережі (далі — НМ) є потужним класом алгоритмів ML, натхненних біологічною нейронною системою людини. Вони складаються з множини взаємозв'язаних нейронів, організованих у шари, кожен з яких виконує певні обчислювальні операції.

Ключовим елементом НМ є нейрон, який є обчислювальною одиницею, що обробляє вхідні дані та передає результати іншим нейронам. Кожен нейрон отримує вхідні сигнали, множить їх на відповідні ваги та застосовує активаційну функцію для генерації вихідного сигналу. Таким чином, нейрони можуть виявляти складні залежності в даних та виконувати різноманітні обчислювальні завдання.

НМ можуть мати різні архітектури, включаючи прості одношарові мережі та більш складні багатошарові мережі. У багатошарових НМ нейрони організовані у послідовні шари: вхідний шар, приховані шари та вихідний шар. Кожен шар має свої ваги, які визначають внесок кожного нейрона у кінцевий результат.

Останнім часом зростає популярність глибоких НМ, які представляють собою багатошарові архітектури з великою кількістю нейронів. Глибокі НМ можуть навчатися на великих обсягах даних та досягати високої точності в різних завданнях, таких як розпізнавання зображень, обробка природної мови, генерація тексту та багато іншого.

Навчання НМ відбувається шляхом налаштування ваг нейронів за допомогою алгоритму зворотного поширення помилки. Під час навчання мережа порівнює передбачені значення з фактичними та коригує ваги таким чином, щоб мінімізувати помилку. Цей процес повторюється багато разів, поки мережа не досягне достатньої точності на навчальному наборі даних.

НМ мають кілька ключових особливостей:

- **Гнучкість:** НМ можуть навчатися на широкому спектрі даних та розв'язувати різноманітні задачі, включаючи класифікацію, регресію, кластеризацію та інші;

- **Автоматичне витягування ознак.** Вони автоматично витягують значущі ознаки з вхідних даних, що дозволяє їм виявляти складні залежності та розв'язувати складні задачі;

- **Здатність до узагальнення**. Добре навчені НР здатні узагальнювати свої знання на нові дані та досягати високої точності на тестових наборах даних;

- **Масштабованість.** Є можливість масштабованості для роботи з великими обсягами даних та великою вимірності.

НМ продовжують активно розвиватися, і їх потенціал у ML залишається великим. З розвитком обчислювальних технологій та методів оптимізації НМ стають все потужнішими та універсальнішими інструментами для розв'язання різноманітних задач у сфері штучного інтелекту та аналізу даних.

1.1.4 Знаходження аномалій в даних

Кожна з основних моделей може бути використана для знаходження аномалій.

ЛР може бути використана для аналізу залишкової моделі. Залишки представляють собою різницю між фактичними та передбаченими значеннями і можуть використовуватися для виявлення аномальних точок.

Аномалії можуть проявитися у вигляді великих значень залишків, що може свідчити про незвичайні або непредставницькі спостереження. Однак лінійні моделі можуть бути нечутливими до складних аномалій або викидів, що вимагає уважного аналізу результатів.

LR може використовуватися для визначення ймовірності аномальності кожної точки даних. Наприклад, якщо точка має високу ймовірність бути аномальною, це може бути ознакою аномалії. LR може бути ефективною в вияв-

ленні аномалій у випадку, коли вони проявляються у формі викидів або незвичайних подій.

k-NN може бути застосований шляхом аналізу відстаней між точками даних. Точки, які знаходяться далеко від своїх найближчих сусідів, можуть бути розглянуті як потенційні аномалії. Цей метод ґрунтується на припущенні про те, що нормальні точки утворюють компактні групи в просторі ознак, і точки, які знаходяться далеко від цих груп, можуть бути аномальними. Однак k-NN може бути чутливим до вибору метрики відстані та кількості сусідів, що вимагає ретельного підбору параметрів для досягнення оптимальної продуктивності.

ДР є графічною моделлю, яка розбиває дані на підгрупи на основі певних правил. Вони можуть бути ефективними в виявленні аномалій завдяки їх здатності адаптуватися до складних структур даних. ДР можуть виявляти аномалії шляхом ідентифікації незвичних шляхів або розгалужень у даних, які можуть свідчити про наявність аномалій. Однак вони можуть бути схильні до перенавчання та вимагають обережного налаштування параметрів.

SVM може бути адаптований для пошуку аномалій шляхом побудови границі рішення таким чином, щоб мінімізувати вплив аномальних точок на класифікацію. SVM може бути особливо ефективним для виявлення аномалій у випадку лінійно нероздільних даних.

НМ можуть бути ефективними завдяки їх здатності екстрагувати складні ієрархічні ознаки з даних. НМ можуть виявляти аномалії як точки, які значно відрізняються від нормального розподілу даних. Однак їх використання може потребувати великої кількості даних та обчислювальних ресурсів для навчання та оцінки. Інформація про використання різних моделей ML під конкретні задачі представлено в літературі [9-13].

1.2 Аналіз статистичних методів обробки даних

1.2.1 Основні поняття

Статистичний аналіз даних являє собою комплексний підхід до вивчення та інтерпретації зібраних даних. Цей аналіз варіюється від базової описової статистики, яка допомагає зрозуміти основні тенденції та характеристики в даних, до складних багатофакторних аналізів і моделей прогнозування, призначених для глибокого розуміння та прогнозування майбутніх подій. Статистичний аналіз застосовний до широкого спектра завдань: від визначення характеристик вибірок і порівняння груп до виявлення взаємозв'язків, причинно-наслідкових зв'язків та інтерпретації викидів у даних [14].

Викиди можуть виникати з різних причин, включаючи помилки вимірювань, незвичайні події або зміни в процесі збору даних. Ці аномалі або нетипові значення можуть суттєво вплинути на результати аналізу, спотворити висновки та призвести до неправильних рішень. Тому коректне виявлення та обробка викидів є критично важливими етапами аналітичного процесу, що дозволяють покращити точність та надійність результатів.

Критерії визначення викидів становлять фундамент статистичних методів та алгоритмів для ідентифікації аномальних даних. Основна мета цих критеріїв — виділення даних, які виходять за межі звичайного розподілу і можуть спотворити загальний аналіз. Це досягається шляхом порівняння кожного спостереження з загальним розподілом і визначення тих, що значно відрізняються від решти даних.

Гіпотеза представляє собою припущення про характеристики популяції або процесу, яке потрібно перевірити на основі зібраних даних. У контексті статистичної оцінки викидів зазвичай використовуються два типи гіпотез: нульова гіпотеза, яка припускає відсутність викидів або значних відмінностей між групами, і альтернативна гіпотеза, яка вказує на наявність істотних відмінностей або аномалій.

Ефективність та обґрунтованість будь-якого критерію викидів залежать від відповідності даних певному розподілу. Це означає, що перед застосуванням конкретного критерію необхідно мати уявлення про розподіл даних, щоб правильно ідентифікувати викиди та забезпечити достовірність результатів аналізу. Розуміння цих фундаментальних понять є ключем до глибокого аналізу даних і досягнення вірних висновків на основі статистичного аналізу.

1.2.2 Типи статистичних методів для виявлення викидів

Типи статистичних методів для виявлення викидів можна поділити на два основних підходи: **параметричні та робастні методи**. Параметричні методи передбачають, що дані відповідають певній математичній моделі або розподілу, тоді як робастні методи є більш універсальними та стійкими до різних форм даних, включаючи викиди та порушення передбачень про розподіл [15].

Параметричні методи є основним інструментом у статистичному аналізі, оскільки вони передбачають, що дані підпорядковуються певному розподілу. Наприклад, t-тести та ЛР передбачають нормальний розподіл даних. Параметричні методи використовують параметри цієї моделі для проведення статистичних тестів або оцінки параметрів. Однак вони можуть бути чутливими до викидів або порушень передбачень про дані. Наприклад, у разі викидів або відхилень від нормального розподілу параметричні методи можуть давати неправильні або ненадійні результати.

У відміну від параметричних методів, **робастні методи** не передбачають жорсткої відповідності даним певного розподілу або моделі. Вони розроблені таким чином, щоб бути більш стійкими до порушень передбачень та викидів в даних. Наприклад, замість використання середнього та стандартного відхилення робастні методи можуть використовувати медіану та квартилі. Це дозволяє надійніше оцінювати параметри або проводити статистичні тести у випадку наявності викидів або відхилень від передбаченого розподілу.

1.2.3 Статистичні тести

Статистичні тести — це методи аналізу даних, що дозволяють робити висновки про характеристики популяцій на основі вибіркових даних. Вони знаходять широке застосування для перевірки гіпотез, оцінки різниць між групами та ідентифікації взаємозв'язків між змінними. Ґрунтуючись на статистичних моделях, ці тести передбачають певні умови або розподіли даних, зокрема нормальне розподіл.

Основна концепція статистичних тестів включає перевірку нульової гіпотези (H0) — припущення про відсутність різниць або ефектів — проти альтернативної гіпотези (H1), яка стверджує наявність таких. Прийняття або відхилення нульової гіпотези здійснюється на основі p-значення, отриманого в результаті тесту. В контексті метрології та вимірювань, статистичні тести відіграють ключову роль, допомагаючи визначити, чи відповідають зібрані дані очікуваним розподілам, існують чи статистично значущі різниці між групами даних, і чи є спостережувані ефекти результатом випадкових варіацій.

Статистичні тести для даних, що слідують нормальному розподілу, базуються на припущенні, що більшість вимірюваних значень, включаючи систематичні та випадкові помилки, підпорядковуються нормальному розподілу. Це припущення ґрунтується на центральній граничній теоремі, яка стверджує, що середні значення великої кількості незалежних та однаково розподілених випадкових змінних будуть наближатися до нормального розподілу, незалежно від форми вихідного розподілу даних. Це робить нормальний розподіл особливо значущим для метрології при оцінці та інтерпретації результатів вимірювань.

Для аналізу даних, що передбачають нормальний розподіл, використовуються різні статистичні тести. Основними серед них є тест Шапіро-Уілка для перевірки нормальності розподілу даних та тест Граббса для ідентифікації потенційних викидів у нормально розподілених даних.

Тест Граббса призначених для виявлення викидів серед вимірювальних даних, які слідують нормальному розподілу. Цей тест широко використовуєть-

ся в аналітичній хімії, метрології та інших галузях науки для перевірки одиничних екстремальних значень, які можуть вказувати на помилки вимірювання або інші аномалії.

Тест Граббса зосереджений на максимальному відхиленні окремого спостереження від середнього значення, нормалізованому стандартним відхиленням усієї вибірки. Це відхилення порівнюється з критичним значенням з розподілу Стьюдента, що дозволяє оцінити, чи є дане спостереження статистично значущим викидом.

Тест Граббса визначається наступною формулою:

$$G = \frac{\left| X_{\max} - \bar{X} \right|}{s}, \tag{1.2}$$

де X_{max} — спостереження, що має найбільше абсолютне відхилення від середнього значення;

$\bar{X}$ — середнє значення всієї вибірки;

s — стандартне відхилення вибірки.

Викид вважається значущим, якщо розрахункове значення G перевищує критичне значення $G_{critical}$, яке можна визначити з таблиць розподілу Стьюдента або за допомогою статистичного програмного забезпечення. Критичне значення залежить від розміру вибірки *n* і заданого рівня значущості α.

Важливо зазначити, що при використанні тесту Граббса слід уважно ставитися до визначення рівня значущості α, який впливає на чутливість тесту до викидів. Занадто низьке значення α може призвести до помилкових відкидань коректних даних, тоді як занадто високе — до недооцінки кількості викидів. Збалансований підхід до вибору α забезпечує оптимальне використання тесту для конкретного дослідження.

Тест Шапіро-Уілка вимірює, наскільки добре вибірка даних відповідає нормальному розподілу. В основі методу лежить порівняння очікуваного порядку статистик з фактичними значеннями вибірки.

Статистика тесту Шапіро-Уілка визначається наступним чином:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}, \tag{1.3}$$

де $x_{(i)}$ — значення з вибірки, впорядковані від найменших до найбільших;

a_i — коефіцієнти, які визначаються зі значень порядкових статистик нормального розподілу, залежно від кількості спостережень у вибірці;

$\overline{x}$ — середнє арифметичне усієї вибірки.

Тест Шапіро-Уілка часто використовується у статистичних дослідженнях, де важливо переконатися, що дані не відхиляються від нормального розподілу, оскільки багато статистичних методів та тестів (наприклад, т-тест) передбачають нормальність даних. В аналітичній хімії, соціальних науках, медицині та інших галузях, де аналіз даних є критичним компонентом дослідження, цей тест допомагає забезпечити валідність статистичних висновків. Якщо значення W є маленьким, і відповідне p-значення нижче заданого рівня значущості (зазвичай 0.05), нульова гіпотеза про те, що дані мають нормальний розподіл, відхиляється. Це вказує на те, що дані мають аномальний розподіл, що може вимагати іншого підходу до їх аналізу або інтерпретації.

Т-критерій Стьюдента, також відомий як t-тест, застосовується для порівняння середніх значень двох груп. Цей тест використовується в різноманітних галузях науки та інженерії для визначення статистичної значущості різниці між двома середніми. Т-тест може бути застосований в двох основних варіантах: як одновибірковий тест або як двовибірковий тест (залежні або незалежні вибірки).

Т-тест базується на припущенні, що дані мають нормальний розподіл у кожній з груп. Він оцінює, чи є відмінності між середніми значеннями двох груп значущими, чи вони можуть бути випадковими через природну варіативність даних.

Формула для розрахунку t-статистики у найпростішому випадку, для одновибіркового t-тесту, виглядає так [15]:

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}, \tag{1.4}$$

де $\overline{x}$ — середнє значення вибірки;

μ — теоретичне середнє значення, з яким порівнюється вибірка;

s — стандартне відхилення вибірки;

n — розмір вибірки.

Для двовибіркового t-тесту з незалежними вибірками формула стає складнішою, враховуючи середні та стандартні відхилення обох груп, а також розміри кожної вибірки

Т-тест широко застосовується у клінічних дослідженнях для порівняння ефективності лікування між двома групами пацієнтів, у соціальних науках для аналізу результатів експериментів, а також у бізнесі для порівняння продуктивності двох команд або ефективності двох маркетингових стратегій.

Статистичне значення t-тесту визначається через p-значення, яке вказує на ймовірність отримати спостережені результати, якщо нульова гіпотеза (гіпотеза про відсутність різниці) є правдивою. Якщо p-значення менше заданого порогу (наприклад, 0.05), то нульова гіпотеза відхиляється, що свідчить про статистичну значущість різниці між групами.

1.2.4 Критерій вибору тесту

Оскільки в статистиці та метрології існує велика різноманітність розподілів, вибір відповідного тесту є критично важливим для правильної інтерпретації даних.

Крім нормального розподілу, який часто зустрічається у природі та застосуваннях, існують інші типи розподілів, такі як експоненціальний, Вейбулла, лог-нормальний та біноміальний розподіли, серед інших. Кожен розподіл має унікальні характеристики, які можуть значно вплинути на поведінку та аналіз даних.

При виборі статистичного тесту для аналізу даних важливо враховувати наступні критерії:

1. **Тип розподілу даних.** Тест повинен бути адаптований до розподілу, який найкраще відображає природу даних. Наприклад, для даних, що слідують експоненціальному розподілу, слід застосовувати тест, який спеціально розроблений для цього типу розподілу;
2. **Розмір вибірки.** Деякі тести краще працюють з великими вибірками, тоді як інші можуть бути більш чутливими до малих вибірок. Вибір тесту часто залежить від кількості доступних даних;
3. **Чутливість до викидів.** Важливо враховувати чутливість тесту до аномалій або викидів у даних, оскільки вони можуть істотно вплинути на результати тесту;
4. **Мета аналізу.** Залежно від того, чи аналіз спрямований на порівняння груп даних, перевірку асоціацій між змінними, чи оцінку відповідності даних до теоретичного розподілу, вибір статистичного тесту може змінюватися

Вибір правильного статистичного тесту є не лише технічним завданням, але й ключовим аспектом для забезпечення валідності висновків, отриманих з даних. Наприклад, для перевірки нормальності розподілу можна застосувати тест Шапіро-Уілка або Андерсона-Дарлінга, залежно від особливостей даних і вимог до аналізу.

1.2.5 Метод 3-х стандартних відхилень

Метод трьох стандартних відхилень — це відомий статистичний інструмент, який використовується для ідентифікації потенційних викидів у наборах даних. Цей метод базується на припущенні про нормальне розподілення значень у наборі даних. Згідно з правилом трьох сигм, приблизно 99.7% усіх значень нормально розподіленої величини знаходяться в межах трьох стандартних

відхилень від їхнього середнього. Отже, значення, які лежать за цими межами, можуть вважатися аномаліями або викидами.

Ось кроки для застосування цього методу:

1. Обчислення середнього значення (μ):

Середнє значення визначається як сума усіх спостережень, поділена на їх кількість:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \tag{1.5}$$

де x_i — окремі значення в наборі даних,

n — загальна кількість спостережень.

2. Обчислення стандартного відхилення (σ):

Стандартне відхилення вимірює розкидання даних відносно середнього [15]:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \,. \tag{1.6}$$

Ця формула показує середньоквадратичне відхилення значень від середнього.

3. Встановлення граничних меж для визначення викидів:

- Верхня межа визначається як μ + 3σ;
- Нижня межа визначається як μ - 3σ.

Значення, що випадають за ці межі, можна розглядати як потенційні викиди. Цей метод особливо корисний для ідентифікації екстремальних значень у даних, де припущення про нормальність є прийнятним. Втім, важливо пам'ятати, що реальні дані можуть не завжди дотримуватися нормального розподілу, тому рекомендується проводити додатковий аналіз розподілу даних перед використанням цього методу. Спостереження, значення яких виходять за межі визначені верхньою та нижньою межами, вважаються викидами. Цей метод дозволяє швидко ідентифікувати значення, які значно відрізняються від більшості даних, що може вказувати на помилки у зборі даних, особливості вимірювального процесу або інші аномалії.

Метод трьох стандартних відхилень широко застосовується в статистиці, дослідженні даних, фінансах та інженерії для оцінки стабільності процесів, контролю якості та аналізу ризиків. Важливо зазначити, що метод найефективніший при роботі з даними, які близькі до нормального розподілу, і його застосування до ненормально розподілених даних може потребувати додаткового аналізу або використання інших методів.

1.2.6 Непараметричні тести

Непараметричні тести - це клас статистичних методів, які використовуються для аналізу даних, коли дані не відповідають передбачуваному параметричному розподілу або коли параметричні припущення не можуть бути виконані через обмеження вибірки або природи даних. Ці тести не вимагають строгого припущення про форму розподілу даних і використовуються для перевірки гіпотез про різницю між групами або для оцінки зв'язків між змінними. Вони особливо корисні в ситуаціях, коли дані категоріальні, асиметричні або містять викиди, що робить їх менш придатними для застосування параметричних методів.

Основні переваги непараметричних тестів включають їх універсальність і неприхильність до розподілу даних. Таким чином, їх можна застосовувати до даних будь-якого типу розподілу, включаючи розподіли, що відрізняються від нормального, такі як розподіли з тяжкими хвостами або асиметричні розподіли. Це особливо корисно в реальних дослідженнях, де дані можуть бути складними і не відповідати стандартним припущенням про розподіл.

Приклади непараметричних тестів включають наступні тести:

- **Тест Манна-Уїтні** (U-тест) є одним з базових непараметричних методів для аналізу двох незалежних вибірок. Цей тест використовує ранги спостережень замість їх фактичних значень. Загальна ідея полягає в ранжуванні усіх спостережень від найменшого до найбільшого, не зважаючи на те, до якої вибірки вони належать. Потім сумуються ранги для кожної

групи окремо. Статистика U обчислюється на основі цих рангових сум, допомагаючи визначити, чи існують статистично значущі відмінності між групами. Тест вимірює розміщення спостережень однієї групи відносно спостережень іншої групи, індикуючи вищі або нижчі середні значення [16];

- **Тест Уілкоксона** для двох залежних вибірок використовується в аналізі повторних вимірювань на одних і тих же суб'єктах. Основна мета цього тесту — порівняти медіани двох зв'язаних груп, щоб визначити, чи є статистично значущі різниці в їх розподілах до і після досліджуваного впливу. Для цього кожній парі значень (до і після) присвоюється різниця, а потім визначається ранг цих різниць незалежно від їх знаків. Статистика тесту обчислюється на основі суми рангів для позитивних і негативних різниць, що дозволяє оцінити напрямок змін [17];
- **Тест Крускала-Уолліса** є розширенням U-тесту Манна-Уїтні для випадків, коли аналізується більше двох груп. Цей тест схожий за своєю суттю на однофакторний дисперсійний аналіз (ANOVA), але він не вимагає припущення про нормальний розподіл даних. Всі спостереження ранжуються разом, після чого для кожної групи обчислюється сума їх рангів. Цей тест вимірює відмінності в середніх рангах між групами, надаючи змогу визначити, чи існують статистично значущі відмінності між трьома або більше групами [18];
- **Тест Спірмена** використовується для оцінки ступеня кореляції між двома змінними. Цей тест аналізує, наскільки добре відносини між рангами двох змінних можна описати за допомогою монотонної функції. Коефіцієнт кореляції Спірмена визначається шляхом порівняння рангів кожної змінної, що робить його стійким до викидів та ефективним у випадках, коли змінні не розподілені нормально. Він особливо корисний для вимірювання взаємозв'язків в умовах, коли звичайні параметричні методи кореляції можуть бути неефективними [19].

Хоча непараметричні тести мають безліч переваг, вони також мають свої обмеження. Зокрема, вони можуть бути менш потужними, ніж параметричні тести, особливо коли дані дійсно відповідають передбачуваному параметричному розподілу. Крім того, непараметричні тести можуть вимагати більшого обсягу вибірки для досягнення статистичної значущості. Тим не менш, у більшості випадків непараметричні тести є цінним інструментом для аналізу даних, особливо коли припущення про розподіл не можуть бути виконані.

1.2.7 Метод інтерквартильного размаху

IQR займає ключове місце у статистиці як міра розсіювання, яка відображає варіабельність середніх 50% даних у наборі. IQR ефективно виключає вплив екстремальних значень або викидів, що робить його особливо цінним для аналізу реальних даних, часто сповнених аномаліями.

Для визначення IQR дані спочатку упорядковуються від найменшого до найбільшого значення. Розділ відсортованого масиву даних на чотири рівні частини дозволяє ідентифікувати квартилі: перший квартиль (Q_1) розділяє нижні 25% даних, тоді як третій квартиль (Q_3) відокремлює верхні 25%. Медіана, що діє як другий квартиль, вказує на середину набору даних. Обчислення IQR шляхом віднімання Q_1 з Q_3 ($IQR = Q_3 - Q_1$) виявляє діапазон, в якому знаходиться основна маса даних, виключаючи крайні квартилі, що містять викиди.

IQR має широке застосування у статистиці для ідентифікації викидів, оцінки розсіювання та порівняння розподілів між наборами даних. Викиди, як правило, визначаються як спостереження, що виходять за межі 1.5×IQR від квартилів Q_1 та Q_3. Це дозволяє аналітикам виявляти аномалії, які можуть впливати на загальні висновки з даних, та адекватно їх адресувати - виправляти, ігнорувати чи детальніше аналізувати (1.6).

Діаграма **"Ящик з вусами"** є інструментом візуального аналізу даних, що базується на IQR. Ця діаграма забезпечує зручне та інтуїтивно зрозуміле

графічне представлення розподілу значень у наборі даних, відображаючи основну масу даних, викиди, а також центральну тенденцію.

"Ящик" у середині діаграми представляє діапазон між першим і третім квартилями (Q_1 та Q_3), тобто IQR, і відображає основну масу даних. Лінія всередині ящика вказує на медіану, що ділить набір даних на дві рівні частини. Це дозволяє швидко оцінити середнє розташування значень у наборі даних.

"Вуса" діаграми розтягуються від країв ящика до крайніх точок даних, які не вважаються викидами, зазвичай до значень, що знаходяться в межах 1.5× IQR від першого і третього квартилів. Це показує загальний розмах даних, виходячи за рамки основної концентрації значень.

Викиди, розташовані за межами "вусів", позначаються на діаграмі окремими точками або символами. Це дозволяє визначити та відокремити аномальні значення, які можуть свідчити про помилки вимірювань, особливості розподілу даних або інші фактори, що впливають на аналіз. Викиди вимагають додаткового розгляду, оскільки вони можуть мати значний вплив на висновки, отримані від аналізу даних.

Застосування діаграми "Ящик з вусами" є особливо корисним при порівнянні розподілів між різними групами або категоріями. Вона дозволяє наочно оцінити та порівняти медіани, розмахи та розподіли між групами, виявляючи можливі відмінності або подібності. Такий порівняльний аналіз може бути важливим для визначення груп з значно відмінними характеристиками або для підтвердження однорідності даних у різних группах [20].

1.2.8 Медіанна абсолютного відхилення

MAD є ще однією робастною мірою розсіювання, яка визначає стабільність та варіативність набору даних, мінімізуючи вплив викидів. На відміну від стандартного відхилення, що чутливе до екстремальних значень, MAD надає більш надійне уявлення про розсіювання даних, особливо в умовах, коли присутні аномалії.

Для розрахунку MAD спочатку знаходиться медіана всього набору даних, яка слугує точкою відліку. Далі обчислюється абсолютне відхилення кожного спостереження від медіани, а потім медіана цих абсолютних відхилень визначається як MAD. Таким чином, MAD відображає типове відхилення спостережень від середнього рівня набору даних, ігноруючи потенційно дестабілізаційний вплив викидів.

MAD може бути особливо корисним в статистичному аналізі для ідентифікації даних, які відрізняються від більшості спостережень. Викиди можуть бути визначені як спостереження, що лежать за межами певного кратного MAD від медіани, наприклад, більше ніж 3×MAD.

Застосування MAD надає статистичним аналітикам надійний інструмент для оцінки розсіювання даних, що є критично важливим у ситуаціях, де нормальність розподілу не може бути гарантована або де присутній високий ризик викидів. Ця міра розсіювання використовується у широкому спектрі застосувань, від фінансового аналізу до обробки сигналів, і є цінним доповненням до інших статистичних інструментів для глибшого розуміння структури даних [21].

1.3 Висновок до першого розділу

Аналіз існуючих методів виявлення викидів показує, що кожен із них має свої переваги, обмеження та застосовується для специфічних типів даних і умов. Вибір підходу залежить від розподілу даних, розміру вибірки та необхідного рівня чутливості до викидів.

Методи статистичного аналізу пропонують різні підходи до роботи з даними залежно від їхніх характеристик. IQR і MAD демонструють високу стійкість до впливу викидів та незалежність від припущень про розподіл, що робить їх універсальними для роботи з малими вибірками чи даними з невідомим розподілом. Однак у випадках сильно асиметричних розподілів їхня точність може знижуватися. Параметричні методи, такі як метод трьох стандартних від-

хилень або критерій Граббса, ефективні для нормально розподілених даних і великих вибірок, але втрачають точність у разі складних або асиметричних розподілів.

Методи ML значно розширюють можливості виявлення викидів, особливо в багатовимірних і складних наборах даних. Регресійні підходи, такі як ЛР та LR, дозволяють оцінювати залишки або ймовірності викиди, проте мають обмеження в роботі з багатовимірними даними. Методи, такі як k-NN, ДР та НМ, забезпечують високу гнучкість і здатність аналізувати складні залежності, хоча й вимагають значних обчислювальних ресурсів та налаштування параметрів. SVM добре адаптується до задач із багатовимірними даними, формуючи граничні рішення для мінімізації впливу аномалій.

У метрологічних досліджень, орієнтованих на виявлення викидів у малих вибірках із невідомим розподілом, найбільш ефективними є робастні статистичні методи, такі як IQR та MAD, а також методи ML без учителя. Ці підходи дозволяють врахувати складность та специфіку даних, що дозволяює адаптуватися до невизначеності розподілу та гарантують стабільність і надійність результатів.

2. ДОСЛІДЖЕННЯ МЕТРОЛОГІЧНИХ ХАРАКТЕРИСТИК ЕТАЛОНА

2.1 Міжнародні звірення

У сфері метрології, міжнародне співробітництво та узгодження стандартів вимірювань становлять необхідну складову розвитку наукових та промислових стандартів. Центральною ланкою у цьому процесі є проведення міжнародних звірень, що забезпечують взаємне визнання метрологічних засобів та методик. Ці заходи спрямовані на забезпечення консистенції та валідації національних еталонів з міжнародно прийнятими стандартами, що, в свою чергу, сприяє точності та надійності вимірювань на міжнародному рівні.

Участь у міжнародних звіреннях має важливе значення для підвищення рівня метрологічного забезпечення, вдосконалення та розробки новітніх технологій вимірювання. Через систематичний аналіз отриманих даних та виявлення потенційних джерел помилок, вдається підвищити рівень достовірності та відтворюваності вимірювань [22].

Регіональні та глобальні метрологічні інститути сприяють встановленню єдиної бази знань, публікуючи результати звірень у доступних звітах та наукових публікаціях. Це дозволяє науковцям та фахівцям у галузі вимірювальної техніки використовувати актуальну інформацію для розвитку власних проєктів та досліджень.

Результати міжнародних порівнянь за період з 2002 по 2024 рік, зокрема у сфері вимірювання витрати та об'єму рідин, демонструють зростаючу тенденцію до розширення обсягу та глибини досліджень. Ці дослідження, охоплюючи широкий діапазон від 50 кг/г до 36000 кг/г, підтверджують важливість міжнародної співпраці та обміну досвідом для вдосконалення методів вимірювання.

Реалізація міжнародних проєктів та звірень вимагає консолідованих зусиль наукових та метрологічних спільнот, що спрямовані на підвищення

точності, надійності та уніфікації вимірювань, а також на підтримку інноваційного розвитку в галузі. Результати цієї роботи, зібрані та систематизовані у наукових публікаціях і звітах, слугують основою для подальших досліджень та розробок у вимірювальній техніці. Звіти міжнародних організацій представлені в літературі [23-40].

2.2 Особливості роботи ДЕТУ 03-04-04

ДЕТУ 03-04-04 призначений для відтворення, зберігання і передавання одиниці масової витрати рідини в діапазоні від $2{,}8\cdot10^{-1}$ до 28 кг/с, об'ємної витрати рідини в діапазоні від $2{,}8\cdot10^{-4}$ до $2{,}8\cdot10^{-2}$ $м^3$/с, маси рідини в діапазоні від 100,0 до 3000 кг, об'єму рідини в діапазоні від 0,1 до 3,0 $м^3$, та передає розмір цих одиниць вторинним еталонам, робочим еталонам 1-го розряду, робочим засобам вимірювальної техніки (далі — ЗВТ) безпосереднім звіренням відповідно до ДСТУ 4403:2005 "Метрологія. Державна повірочна схема для засобів вимірювань об'ємної і масової витрати рідини, об'єму і маси рідини, що протікає по трубопроводу" [41].

ДЕТУ 03-04-04 було створено на базі еталонної витратомірної установки ВЗУ-180 в 2004 році, запозичивши від неї всю гідравлічну і витрато-вимірювальну системи. Саму ж ВЗУ-180 було введено в метрологічну практику України у січні 1995 року.

Основними складовими частинами еталона є: прецизійні ваги виробництва фірми Mettler Toledo, з діапазоном виміру від 0 до 150 кг та від 0 до 3000 кг; перекидальний пристрій перенаправлення потоку, що перенаправляє потік рідини, по черзі – у бак для зберігання рідини або у вимірювальний бак для її зважування; та електромагнітні витратоміри для встановлення та підтримки необхідної витрати рідини.

Схематично, робота ДЕТУ 03-04-04 добре відображена на його мнемосхемі яка нанесена на пульті керування еталона (рис 2,1).

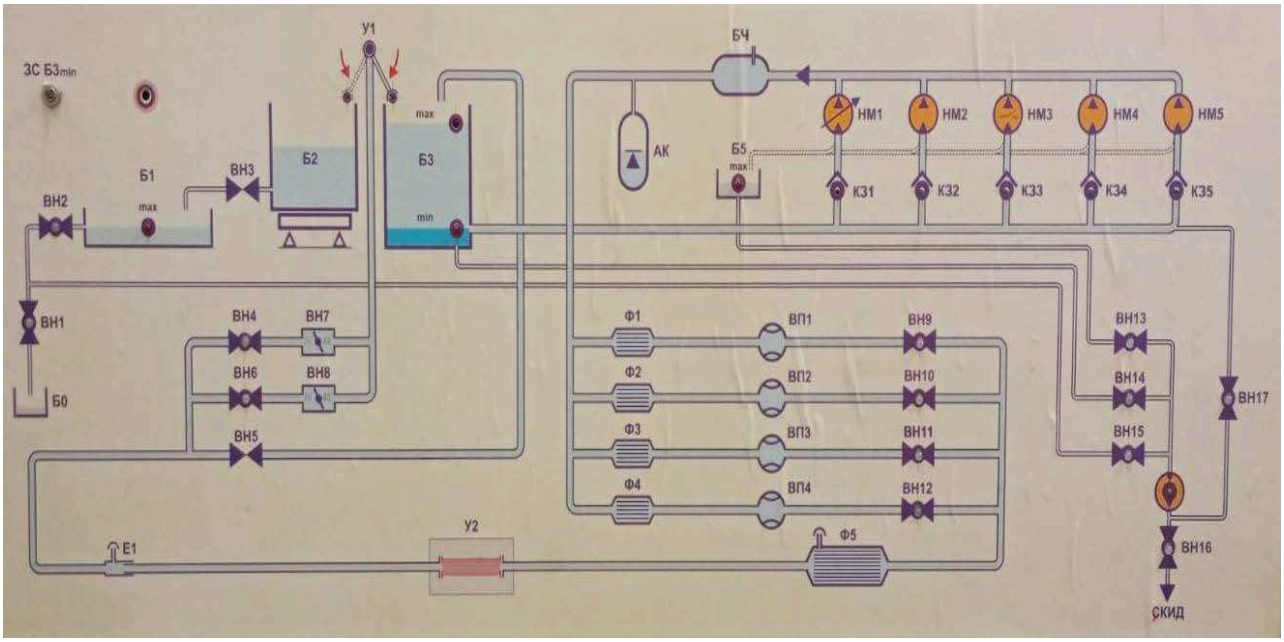

Рисунок 2.1- Мнемосхема еталона ДЕТУ 03-04-04

На рис. 2.1 введені наступні позначення: ВН1...ВН17 – вентилі (ВН7 і ВН8 – керуючі з пульту керування еталона кульовий кран DN50 та шиберна заслінка DN80, які призначені для регулювання витрати рідини у його вимірювальній ділянці $Е_1$, $У_2$, $Ф_5$); $У_1$ – пристрій перенаправлення потоку; У2 – ЗВТ, якому передається одиниця, що відтворюється еталоном; АК – двохступеневий гідродинамічний фільтр; Б0, Б1,...,Б5 – ємкості (Б0 і Б5 – відстійники, Б1 – бак для зливу рідини з баку Б2, що зважується, Б3 – бак із запасом оборотної води, Б4 – повітровідділювач); ЗС $Б3_{min}$ – звукова сигналізація мінімально допустимого рівня води в Б3 і максимального в Б5; НМ1,...,НМ5 – насоси (НМ1 – насос постійного струму з оборотами, що варіюються (відтворюваною витратою)); КЗ1...КЗ5 – зворотні клапани; Ф1...Ф5 – хонікомби; ВП1...ВП4 – еталонні витратоміри ЕМВВ-2; Е1 – компенсатор – пристрій, який дозволяє варіювати протяжність вимірювальної ділянки еталона.

ДЕТУ 03-04-04 працює на масовому методі вимірювання. Рідина з накопичувального бака Б3 всмоктується насосними агрегатами НМ1,...,НМ5 проходить через двохступеневі гідродинамічні фільтри, баки для скидання повітря, витратоміри, струменевипрямляч потоку, та інші складові частини еталону. Рідина, потрапляючи на вимірювальну ділянку трубопроводу (далі —

ВД), де протікає через встановлений ЗВТ до перекидного пристрою У1, який перенаправляє потік рідини або до зважувального бака Б2, який стоїть на прецензійних підлогових вагах, або назад до накопичувального бака Б3. Після встановлення обраного значення витрати рідини, і початку вимірювання, одночасно перекидній пристрій перенаправляє потік рідини до бака на зважування, та починається відлік часу вимірювання, і фіксується показання ЗВТ. По закінченню вимірювання, потік рідини перенаправляється назад до баку накопичування, фіксується час та показання ЗВТ. Потім вимірювальні показання ЗВТ і еталону записуються до протоколу вимірювань.

2.3 Дослідження стабільності витрати рідини

На ДЕТУ 03-04-04 проводилося дослідження стабільності витрати рідини, шляхом розрахунку стандартної невизначеності вимірювання за типом А масової витрати рідини на ВД еталона. Ці дослідження були виконані у рамках підготовки ДЕТУ 03-04-04 до міжнародних звірень.

Вимірювання відбувалися наступним чином: спочатку за допомогою вентилю встановлювалось обране значення витрати рідини Q_{em}.; вимірювався час t наповнення рідиною вимірювального баку, та маса m рідини у вимірювальному баку. Цю процедуру було повторено по 5 разів на 3 значеннях витрати рідини на ВД діаметром 50 мм, 15 мм, 6 мм. Після проведення вимірювань виявлено слабку стабільність потоку, що призвело до рішення замінити насосні агрегати еталона на нові, та провести повторні дослідження.

Додатково, для підвищення стабільності потоку витрати під час вимірювань та зменшення флуктуацій, був встановлений частотний перетворювач для всіх насосних агрегатів. Частотний перетворювач, завдяки регулюванню частоти обертання двигуна насосного агрегату, може доволі точно встановлювати необхідне значення витрати рідини та має плавний запуск, що зменшує флуктуації рідини при виході на робочий режим вимірювання на еталоні. Отримані результати до модернізації та після наведені в таблиці 2.1.

Таблиця 2.1 - Результати дослідження стабільності витрати рідини

До модернізації							Після модернізації				
№	DN	m (т)	t (с)	T, ^{0}C	Q ет. (т/ч)	Ua, %	Ua, %	Q ет. (т/ч	T, ^{0}C	t (с)	m (т)
1	50	0,35642	28,495	16,73	45,030	8,570 E-02	5,773E-02	45,483	16,49	28,452	0,35946
	50	0,35392	27,981	16,74	45,535			45,466	16,50	28,023	0,35392
	50	0,35583	28,399	16,76	45,107			45,207	16,52	28,338	0,35585
	50	0,35965	28,601	16,77	45,270			45,246	16,53	28,615	0,35964
	50	0,35566	28,216	16,78	45,377			45,340	16,54	28,241	0,35568
2	50	0,27454	40,082	16,28	24,658	2,140 E-01	8,530E-02	25,575	16,45	40,060	0,28460
	50	0,26196	37,865	16,29	24,906			25,282	16,46	35,900	0,25212
	50	0,25764	36,854	16,31	25,167			25,844	16,48	37,242	0,26735
	50	0,26061	37,143	16,32	25,259			25,676	16,49	35,129	0,25055
	50	0,25971	36,245	16,33	25,795			25,984	16,50	36,225	0,26147
3	50	0,21194	151,321	16,68	5,042	6,672 E-02	3,525E-02	5,141	16,78	140,481	0,20061
	50	0,21959	150,328	16,69	5,259			5,159	16,79	142,700	0,20451
	50	0,22615	151,120	16,71	5,387			5,206	16,81	139,464	0,20167
	50	0,21466	150,073	16,72	5,149			5,051	16,82	143,200	0,20094
	50	0,21578	151,194	16,73	5,138			5,261	16,83	141,225	0,20638
4	15	0,15097	106,674	16,29	5,095	1,094 E-01	7,403E-02	5,271	16,72	103,909	0,15213
	15	0,15955	107,933	16,30	5,322			5,159	16,73	105,513	0,15120
	15	0,15632	103,426	16,32	5,441			5,317	16,75	103,441	0,15278
	15	0,15916	104,486	16,33	5,484			5,015	16,76	108,142	0,15064
	15	0,15197	110,806	16,34	4,937			4,967	16,77	109,648	0,15129
5	15	0,11308	143,866	16,46	2,830	1,010 E-01	2,767E-02	2,532	16,23	144,201	0,10141
	15	0,10980	159,831	16,47	2,473			2,598	16,24	141,198	0,10191
	15	0,10050	149,024	16,49	2,428			2,651	16,26	141,307	0,10407
	15	0,10425	156,790	16,50	2,394			2,601	16,27	145,167	0,10489
	15	0,11402	160,063	16,51	2,565			2,506	16,28	144,460	0,10057
6	15	0,14156	387,850	16,23	1,314	1,097 E-01	2,556E-03	1,037	16,75	351,209	0,10120
	15	0,12957	401,182	16,24	1,163			1,049	16,76	353,593	0,10301
	15	0,11717	408,055	16,26	1,034			1,033	16,78	349,364	0,10020
	15	0,10055	439,651	16,27	0,823			1,046	16,79	357,472	0,10391
	15	0,10709	461,712	16,28	0,835			1,040	16,80	352,217	0,10178
7	6	0,12428	384,850	16,55	1,163	6,630 E-02	3,287E-03	0,988	16,77	368,427	0,10106
	6	0,11696	411,129	16,56	1,024			0,972	16,78	367,072	0,09906
	6	0,11398	467,754	16,58	0,877			0,986	16,80	369,541	0,10126
	6	0,10043	415,156	16,59	0,871			0,992	16,81	370,405	0,10202
	6	0,11782	403,856	16,60	1,050			1,002	16,82	374,095	0,10412
8	6	0,08399	408,546	16,42	0,740	9,865 E-02	2,406E-03	0,448	16,80	653,032	0,08125
	6	0,08085	700,448	16,43	0,416			0,459	16,81	644,401	0,08225
	6	0,08810	995,200	16,45	0,319			0,455	16,83	651,228	0,08227

Продовження таблиці 2.1

№	DN	m (т)	t (с)	T, ^{0}C	Q ет. (т/ч)	Ua, %	Ua, %	Q ет. (т/ч	T, ^{0}C	t (с)	m (т)
8	6	0,08369	405,522	16,46	0,743			0,458	16,84	653,917	0,08328
	6	0,08576	748,500	16,47	0,412			0,449	16,85	656,923	0,08202
9	6	0,04814	850,882	16,58	0,204	2,372E-02	1,997E-03	0,114	16,50	1324,250	0,04180
	6	0,04593	1091,995	16,11	0,151			0,119	16,51	1326,231	0,04376
	6	0,04096	1550,765	16,27	0,095			0,116	16,53	1329,038	0,04282
	6	0,04613	852,479	16,77	0,195			0,112	16,54	1322,286	0,04112
	6	0,04536	1350,937	16,36	0,121			0,123	16,55	1328,472	0,04535

В таблиці 2.1 використовуються наступні позначення:

DN — діаметр трубопроводу на ВД, мм;

m — маса рідини, т;

t — час вимірювання , с;

Q ет. — масова витрата рідини, т/г;

Ua — стандартна невизначеність вимірювань за типом А,%.

Результат вимірювання масової витрати рідини (Q_{em} 1;2;…;5) визначався за наступним рівнянням [42]:

$$\mathrm{Q}e\mathrm{T} = \frac{m * 3\ \ 6\ (}{t}. \tag{2.1}$$

Стандартну невизначеність вимірювання за типом А витрати рідини U_a у відсотках було визначено за формулою [43]:

$$u_A(\overline{x}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{\mathrm{i}} - \overline{x})^2}\ , \tag{2.2}$$

де $\overline{x}$ – середнє арифметичне значення витрати рідини на точці витрати, було визначено за формулою [43]:

$$\overline{x} = \frac{\sum_{i=1}^{n}\mathrm{Q}e\mathrm{T}_{\mathrm{i}}}{n}. \tag{2.3}$$

Значення стандартної невизначеності вимірювань за типом А масової витрати рідини до та після модернізації зображено на рис 2.2

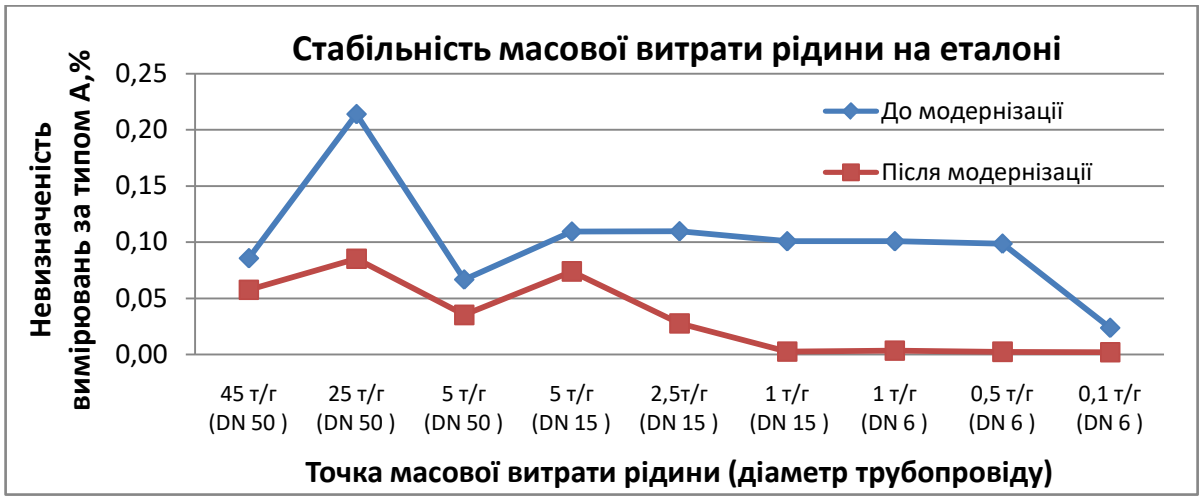

Рисунок 2.2 - Дослідження стабільності витрати рідини

Після проведення розрахунків та модернізації насосної складової частини еталону, спостерігається значне зниження стандартної невизначеності типу А масової витрати рідини. Це свідчить про успішність заходів, спрямованих на підвищення стабільності витрати. Однак суттєве зниження невизначеності в нижньому діапазоні може бути індикатором неефективної роботи старих насосних агрегатів малої потужності. Водночас, значення витрати в верхньому діапазоні значно вищі, що вказує на необхідність подальших досліджень у цьому напрямку. Це може потребувати впровадження нових стратегій або модифікацій системи з метою досягнення бажаних показників стабільності та точності вимірювань у всьому діапазоні витрати рідини.

Значення стандартної невизначеності за типом Б залишилися незмінними. Це може бути пояснено тим, що насосні агрегати, так само як і частотний перетворювач, впливають на процес проведення вимірювань, а не на обладнання, яким здійснюється вимірювання.

2.4 Дослідження розширеної невизначеності вимірювань коріолісових витратомірів

У міжнародних звіреннях за напрямком масової витрати рідини як еталон переносник використовується масовий коріолісовий витратомір. Тому, в рамках підготовки ДЕТУ 03-04-04 до міжнародних звірень, було проведено

вимірювання трьох масових витратомірів за методикою, описаною у фінальних звітах EUROMET [23-27].

2.4.1 Коріолісовий витратомір

Коріолісові масові витратоміри - це пристрої для вимірювання масового потоку рідини або газу, які використовують принцип коріолісового ефекту. Цей ефект виникає, коли об'єкт що рухається, наприклад, рідина або газ, змінює свій напрямок руху відносно точки обертання. У витратомірах коріоліса, трубки, через які проходить рідина або газ, піддаються вібраціям або коливанням. Коливання створюються за допомогою вбудованих датчиків. Коли рідина або газ протікає через ці трубки, вони відхиляються від свого звичайного руху, що призводить до зміни їх власної частоти коливань. Ця зміна частоти коливань пропорційна масовому потоку рідини або газу, який проходить через витратомір. Датчики вимірюють цю зміну, і за допомогою спеціального алгоритму обчислюють масовий потік.

Коріолісові витратоміри відрізняються високою точністю та стабільністю вимірювань, навіть при зміні властивостей рідини або газу, таких як температура, тиск чи в'язкість. Вони також можуть працювати в широкому діапазоні витрат і діаметрів трубок, що робить їх універсальними і ефективними для застосування у різних промислових галузях.

Загальну інформацію про витратоміри які використовувались при дослідженні наведено у таблиці 2.2, та рис. 2.3-2.5.

Таблиця 2.2 - Характеристики витратомірів.

Найменування	**Витратомір №1**	**Витратомір №2**	**Витратомір №3**
Назва витратоміра	Micro Motion модель CMF 025 (серія Elite).	Micro Motion модель CMF 050 (серія Elite).	Micro Motion модель CMF 200 (серія Elite).
Виробник	“Emerson Process Management Flow BV ”, Мексика.	“Emerson Process Management Flow BV ”, Мексика.	“Emerson Process Management Flow BV ”, Мексика.
Діапазон масової витрати	від 0,1 до 1,0 т/год.	Від 1 до 5,0 т/год.	від 5 до 45 т/год.
Номінальний діаметр	DN 6.	DN 15.	DN 50.

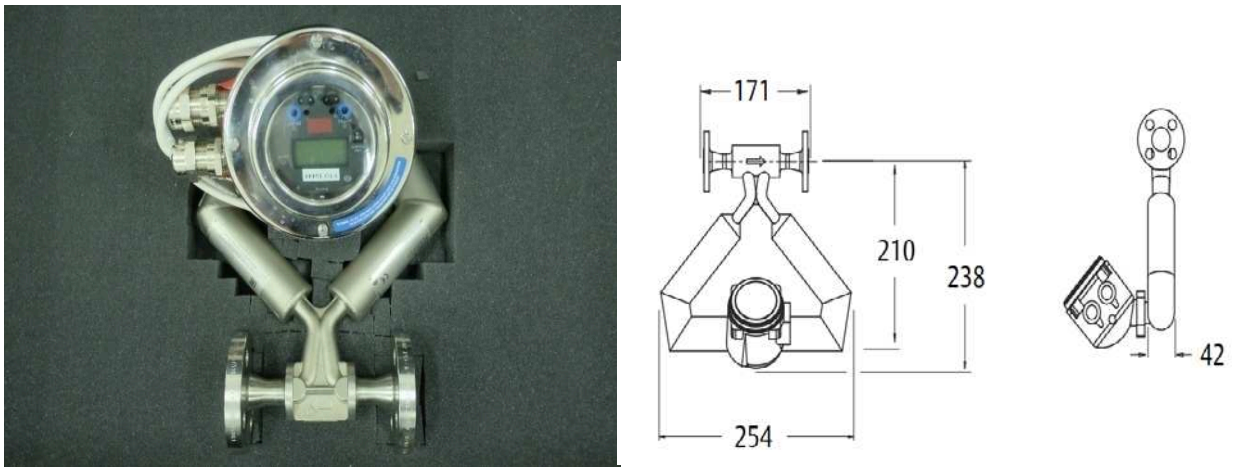

Рисунок 2.3 - Загальний вигляд та габаритні розміри витратоміра № 1.

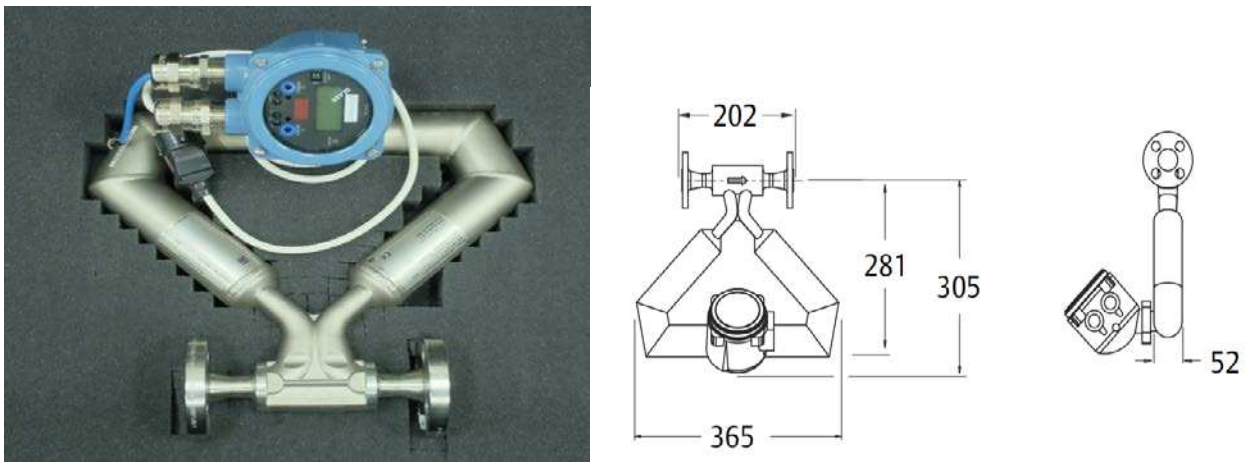

Рисунок 2.4 - Загальний вигляд та габаритні розміри витратоміра № 2.

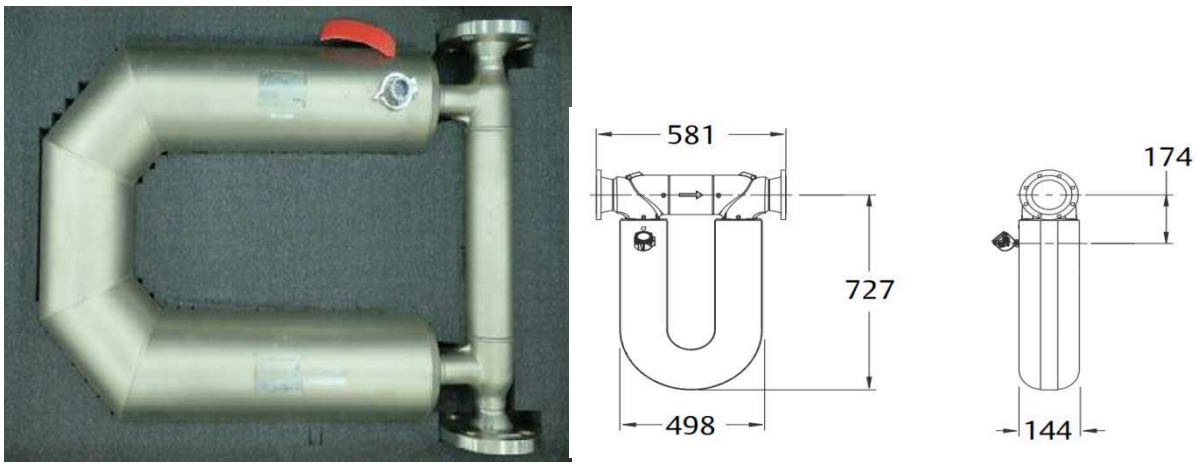

Рисунок 2.5 - Загальний вигляд та габаритні розміри витратоміра № 3.

2.4.2 Проведення вимірювань

Вимірювання масової витрати рідини відбувалось шляхом зважування рідини на прецензійних підлогових вагах за певний проміжок часу та порівняння результатів зі значеннями витратоміра.

Програмне забезпечення витратомірів ProLink III дає можливість фіксації маси рідини яка протікає по трубопроводу через витратомір, фіксації часу вимірювання, та кількість імпульсів за час вимірювання, які необхідні для подальших розрахунків.

Умови під час проведення вимірювань відповідали наступним вимогам:

– робоча рідина вода;

– температура робочої рідини: 20 ± 5 °С;

– температура навколишнього середовища: 20 ± 5 °С;

– вологість навколишнього повітря: від 30 до 80 %;

– атмосферний тиск: від 86 до 106 кПа;

– відсутність у вимірювальній лінії еталона вільного повітря.

Вимірювання проводилось в таких точках витрати:

– для витратоміра № 1 вимірювання проводиться в точках: 5,0 т/год, 25,0 т/год, 45,0 т/год.

– для витратоміра №2 вимірювання проводиться в точках: 1,0 т/год, 2,5 т/год, 5,0 т/год.

– для витратоміра №3 вимірювання проводиться в точках: 0,1 т/год, 0,5 т/год, 1,0 т/год.

Точки витрати обираються від найвищого до найнижчого. В кожній точці витрати проводиться 10 вимірювань.

Коефіцієнт перетворення (k-фактор) для кожному вимірювання, імп./кг, визначається за формулою [44]:

$$k_{ij} = \frac{N_{ij}}{M_{Refij}}, \tag{2.4}$$

де N – кількість імпульсів, отриманих від витратоміра під час вимірювання, імп.;

M_{Ref} – маса рідини, виміряна еталоном під час вимірювання, кг;τ_{Ref}

j, i – номер точки витрати та номер вимірювання

Розширена невизначеність вимірювання (для k = 2), оцінюється відповідно до рекомендацій WGFF [46]: U_{CMC}

$$U_{CMC} = \sqrt{U_{base}^2 + 2 \cdot u_{repeat}^2}\,, \tag{2.5}$$

де U_{base} – розширена невизначеність еталона %;

u_{repeat} – значення повторюваності k (розраховується як стандандартна невизначеність за типом А, формула 2.2), %.

Результати вимірювання масової витрати рідини та розрахунки коефіцієнта перетворення та розширеної невизначеності наведені в таблицях 2.3-2.5.

У таблицях 2.3-2.5 використовуються такі скорочення:

Q_{nom} – задана витрата, кг/год;

$T_{i,j}$ – температура рідини, ° С;

j,I – номер точки потоку та номер вимірювання відповідно;

Pij – тиск рідини в гідротракті під час вимірювання, МПа;

Nij – кількість імпульсів, отриманих від витратоміра під час вимірювання, імп .;

M_{TSij} – значення маси за показаннями витратоміра під час вимірювання, кг;

$\tau_{Ref\,ij}$ – час вимірювання, с;

$M_{Ref\,ij}$ – маса рідини, виміряна еталоном за час вимірювання, кг; $\tau_{Ref\,ij}$,

k_{ij} – коефіцієнт калібрування (k-фактор) для кожного вимірювання, імп./кг;

k_j – середньоарифметичне значення коефіцієнта калібрування (к-фактор) у точці потоку, імп./кг;

U_{CMC} – розширена невизначеність вимірювання (k = 2),%.

Таблиця 2.3 - Результати вимірювання витратоміром №1 на еталоні.

№ п/п	Q_{nom}	T_{ij}	P_{ij}	N_{ij}	M_{TSij}	$\tau_{Ref\ ij}$	$M_{Ref\ ij}$	k_{ij}	k_j	U_{CMC}
	[кг/ч]	[°C]	[МПа]	[имп.]	[кг]	[с]	[кг]	[имп./кг]	[имп./кг]	[%]
1	44,9109	19,04	0,105	136 757	0,45586	36,481	0,45511	300 494,35	300836,78	0,0362
	44,9079	19,20	0,103	136 766	0,45589	36,461	0,45483	300 699,13		
	44,8850	19,33	0,1035	136 439	0,4548	36,397	0,4538	300 661,05		
	44,8800	19,47	0,1035	136 286	0,45429	36,369	0,4534	300 588,85		
	44,6626	19,62	0,103	145 064	0,48355	38,890	0,48248	300 665,28		
	44,7313	19,73	0,103	137 432	0,45811	36,749	0,45662	300 978,90		
	44,9010	19,77	0,103	137 138	0,45713	36,538	0,45572	300 928,17		
	44,9008	19,88	0,103	136 970	0,45657	36,454	0,45467	301 253,63		
	44,6544	19,93	0,103	138 044	0,46015	36,905	0,45777	301 559,70		
	44,8862	20,03	0,102	137 234	0,45745	36,623	0,45663	300 538,70		
2	24,9413	18,70	0,35	105 356	0,35119	50,549	0,35021	300 839,46	301148,75	0,0998
	24,9243	18,71	0,34	105 974	0,35325	50,946	0,35272	300 450,75		
	24,8897	18,74	0,35	105 686	0,35229	50,752	0,35089	301 196,92		
	24,9098	18,82	0,34	111 080	0,37027	53,100	0,36742	302 327,01		
	24,9049	18,83	0,34	106 805	0,35602	50,909	0,35219	303 262,41		
	24,9050	18,82	0,34	106 292	0,35431	51,026	0,353	301 113,28		
	24,8938	18,83	0,35	107 216	0,35739	51,662	0,35724	300 125,93		
	24,8900	18,85	0,34	106 967	0,35656	51,463	0,35581	300 632,33		
	24,8881	18,87	0,34	106 922	0,35641	51,386	0,35525	300 979,56		
	24,8745	18,72	0,34	106 295	0,35432	51,184	0,35366	300 559,83		
3	5,0370	19,24	0,13	30 354	0,10118	72,071	0,10084	301 011,47	300902,60	0,0213
	5,0400	19,30	0,13	30 366	0,10122	72,143	0,101	300 653,43		
	5,0403	19,37	0,11	30 471	0,10157	72,239	0,10114	301 275,43		
	5,0395	19,34	0,11	30 261	0,10087	71,864	0,1006	300 805,14		
	5,0449	19,37	0,11	30 612	0,10204	72,594	0,10173	300 914,15		
	5,0508	19,39	0,12	30 342	0,10114	71,882	0,10085	300 862,64		
	5,0483	19,39	0,11	30 363	0,10121	71,889	0,10081	301 190,33		
	5,0390	19,39	0,12	30 456	0,10152	72,322	0,10123	300 859,40		
	5,0410	19,40	0,12	30 393	0,10131	72,171	0,10106	300 742,10		
	5,0479	19,40	0,12	30 411	0,10137	72,123	0,10113	300 711,92		

Таблиця 2.4 - Результати вимірювання витратоміром №2 на еталоні.

№ п/п	Q_{nom}	T_{ij}	P_{ij}	N_{ij}	M_{TSij}	$\tau_{Ref\ ij}$	$M_{Ref\ ij}$	k_{ij}	k_j	U_{CMC}
	[кг/ч]	[°C]	[МПа]	[имп.]	[кг]	[с]	[кг]	[имп./кг]	[имп./кг]	[%]
1	18,20	0,118	364 664,35	101	72,800	101,23	3 602,33	18,20	3605,73	0,0234
	18,30	0,118	368 618,84	102	73,459	102,28	3 604,02	18,30		
	18,22	0,118	371 104,21	103	73,965	102,96	3 604,35	18,22		
	18,22	0,118	366 533,94	101	73,136	101,64	3 606,20	18,22		
	18,25	0,118	366 374,53	101	72,982	101,58	3 606,76	18,25		
	18,32	0,118	368 633,35	102	73,401	102,19	3 607,33	18,32		
	18,27	0,118	367 856,15	102	73,261	101,96	3 607,85	18,27		
	18,28	0,118	366 115,69	101	73,165	101,56	3 604,92	18,28		
	18,30	0,118	364 569,24	101	72,689	101,20	3 602,46	18,30		
	18,35	0,118	368 148,68	102	73,195	101,95	3 611,07	18,35		
2	18,56	0,38	273 111,93	75	109,330	75,70	3 607,82	18,56	3606,96	0,0328
	18,55	0,38	271 845,16	75	108,968	75,44	3 603,46	18,55		
	18,55	0,38	272 695,77	75	109,385	75,71	3 601,85	18,55		
	18,57	0,38	273 209,42	75	109,445	75,64	3 611,97	18,57		
	18,53	0,38	272 334,26	75	109,035	75,51	3 606,60	18,53		
	18,52	0,38	272 810,07	75	109,296	75,66	3 605,74	18,52		
	18,54	0,38	273 475,39	75	109,463	75,80	3 607,85	18,54		
	18,55	0,38	273 130,58	75	109,475	75,82	3 602,36	18,55		
	18,56	0,38	273 544,11	75	109,374	75,72	3 612,57	18,56		
	18,55	0,38	273 389,02	75	109,498	75,74	3 609,57	18,55		
3	18,26	0,18	145 884,12	40	144,740	40,35	3 615,47	18,26	3614,13	0,0301
	18,22	0,18	145 756,72	40	144,224	40,33	3 614,10	18,22		
	18,22	0,18	145 491,97	40	143,861	40,32	3 608,43	18,22		
	18,23	0,18	145 949,17	40	144,113	40,38	3 614,39	18,23		
	18,27	0,18	145 898,56	40	144,164	40,38	3 613,14	18,27		
	18,26	0,18	145 750,92	40	143,875	40,32	3 614,85	18,26		
	18,28	0,18	145 605,48	40	143,524	40,27	3 615,73	18,28		
	18,29	0,18	145 518,58	40	143,544	40,23	3 617,17	18,29		
	18,30	0,18	145 534,27	40	143,904	40,33	3 608,59	18,30		
	18,31	0,18	146 117,76	40	143,882	40,37	3 619,46	18,31		

Таблиця 2.5 - Результати вимірювання витратоміром №3 на еталоні.

№ п/п	Q_{nom}	T_{ij}	P_{ij}	N_{ij}	M_{TSij}	$\tau_{Ref\ ij}$	$M_{Ref\ ij}$	k_{ij}	k_j	U_{CMC}
	[кг/ч]	[°C]	[МПа]	[имп.]	[кг]	[с]	[кг]	[имп./кг]	[имп./кг]	[%]
1	44,9109	19,04	105,00	136 757	0,45586	36,481	0,45511	300 494,35	300836,78	0,036274
	44,9079	19,20	103,00	136 766	0,45589	36,461	0,45483	300 699,13		
	44,8850	19,33	103,50	136 439	0,4548	36,397	0,4538	300 661,05		
	44,8800	19,47	103,50	136 286	0,45429	36,369	0,4534	300 588,85		
	44,6626	19,62	103,00	145 064	0,48355	38,890	0,48248	300 665,28		
	44,7313	19,73	103,00	137 432	0,45811	36,749	0,45662	300 978,90		
	44,9010	19,77	103,00	137 138	0,45713	36,538	0,45572	300 928,17		
	44,9008	19,88	103,00	136 970	0,45657	36,454	0,45467	301 253,63		
	44,6544	19,93	103,00	138 044	0,46015	36,905	0,45777	301 559,70		
	44,8862	20,03	102,00	137 234	0,45745	36,623	0,45663	300 538,70		
2	24,9413	18,70	0,35	105 356	0,35119	50,549	0,35021	300 839,46	301148,75	0,099806
	24,9243	18,71	34,00	105 974	0,35325	50,946	0,35272	300 450,75		
	24,8897	18,74	35,00	105 686	0,35229	50,752	0,35089	301 196,92		
	24,9098	18,82	34,00	111 080	0,37027	53,100	0,36742	302 327,01		
	24,9049	18,83	34,00	106 805	0,35602	50,909	0,35219	303 262,41		
	24,9050	18,82	34,00	106 292	0,35431	51,026	0,353	301 113,28		
	24,8938	18,83	35,00	107 216	0,35739	51,662	0,35724	300 125,93		
	24,8900	18,85	34,50	106 967	0,35656	51,463	0,35581	300 632,33		
	24,8881	18,87	34,50	106 922	0,35641	51,386	0,35525	300 979,56		
	24,8745	18,72	34,50	106 295	0,35432	51,184	0,35366	300 559,83		
3	5,0370	19,24	13,00	30 354	0,10118	72,071	0,10084	301 011,47	300902,60	0,021336
	5,0400	19,30	13,00	30 366	0,10122	72,143	0,101	300 653,43		
	5,0403	19,37	11,00	30 471	0,10157	72,239	0,10114	301 275,43		
	5,0395	19,34	11,00	30 261	0,10087	71,864	0,1006	300 805,14		
	5,0449	19,37	11,00	30 612	0,10204	72,594	0,10173	300 914,15		
	5,0508	19,39	12,00	30 342	0,10114	71,882	0,10085	300 862,64		
	5,0483	19,39	11,50	30 363	0,10121	71,889	0,10081	301 190,33		
	5,0390	19,39	12,00	30 456	0,10152	72,322	0,10123	300 859,40		
	5,0410	19,40	12,00	30 393	0,10131	72,171	0,10106	300 742,10		
	5,0479	19,40	12,00	30 411	0,10137	72,123	0,10113	300 711,92		

Результатом дослідження є розрахунок розширеної невизначеності вимірювань витратомірів згідно з процедурою міжнародних звірень. За результати було встановлено що розширеної невизначеності відповідає міжнародним стандартам. Але для відповідності у повному обсязі необхідно продовжити дослідження стабільності на надійності результатів вимірювань.

2.4.3 Методика обробки результатів міжнародних звірень

У міжнародних звіреннях за напрямом масової витрати рідини як еталон-переносник використовують масовий коріолісовий витратомір. Лабораторія-учасник проводить вимірювання витратоміра за своєю методикою, після чого розраховується похибка вимірювань із дотриманям вимог технічного протоколу звірень, який було узгоджено всіма учасниками ще до проведення звірень.

У міжнародній практиці проведення міжнародних звірень виявляється певна тенденція у проведенні обробки результатів звірень. Згідно з фінальними звітами міжнародних організацій [23-40], методика обробки результатів звірень полягає в розрахунку еталонного значення ключового порівняння, проходженні тесту "x_i – квадрат", обчисленні ступеня еквівалентності для оцінки успішності лабораторії-учасника та призначенні СМС-рядків. Еталоне значення ключового порівняння (CRV) визначається окремо для кожної витрати згідно з процедурою А з [45]. Визначення еталонного значення ключового порівняння включає перевірку на відповідність згідно з [46].

Еталонне значення ключового порівняння y розраховується відповідно до наступного рівняння [23]:

$$y = \frac{\frac{\varepsilon_{lab1}}{u^2_{\varepsilon lab1}} + \frac{\varepsilon_{lab2}}{u^2_{\varepsilon lab2}} + \ldots + \frac{\varepsilon_{labi}}{u^2_{\varepsilon labi}}}{\frac{1}{u^2_{\varepsilon lab1}} + \frac{1}{u^2_{\varepsilon lab2}} + \ldots + \frac{1}{u^2_{\varepsilon labi}}}, \tag{2.6}$$

де ε_{lab1} ε_{lab2}…, ε_{labi} – значення, виміряне в різних незалежних лабораторіях 1, 2, …, i, це можуть бути середні коефіцієнти перетворення, значення похибки вимірювань витратоміра або похибка витратоміра з урахуванням різних факторів впливу на вимірювання. У кожних окремих звіреннях це значення узгоджується технічним протоколом звірень.

$u_{\varepsilon lab1}$ $u_{\varepsilon lab2}$…, $u_{\varepsilon labi}$ – це стандартна невизначеність вимірювань значень ε_{lab1} ε_{lab2}…, ε_{labi}, які були виміряні в різних незалежних лабораторіях 1, 2, ..., *i*. Вона розраховується на основі наступного рівняння [23]:

$$\frac{1}{u_y^2} = \frac{1}{u_{\varepsilon lab1}^2} + \frac{1}{u_{\varepsilon lab2}^2} + \ldots + \frac{1}{u_{\varepsilon labi}^2}. \quad (2.7)$$

Розширена невизначеність еталонного значення U_y становить [47]:

$$U_y = 2 \cdot u_y. \quad (2.8)$$

Далі проводиться тест "x_i – квадрат", щоб перевірити, чи можна очікувати, що певні похибки та супутні їм невизначеності матимуть гаусівський розподіл. Якщо це так, то "*y*" буде прийнято. Тест "x_i – квадрат" виконується таким чином [23]:

$$\chi_{obs}^2 = \sum_{n}^{n-1} \left(\frac{\varepsilon_{lab\ i} - \varepsilon_{RV}}{u\left(\varepsilon_{lab\ i}\right)} \right)^2, \quad (2.9)$$

де X^2_{obs} – спостереження значень x_i – квадрат; ε_{RV} – значення похибки еталона-переносника.

Ступені свободи *v*, в цьому випадку, було встановлено згідно з рівнянням [25]:

$$v = n_i - 1. \quad (2.10)$$

Щоб реалізувати вихідне значення, лабораторії-учасники повинні відповідати таким умовам [23]:

$$Pr\{X_v^2 > X^2_{obs}\} < 0{,}05. \quad (2.11)$$

Якщо ця умова не виконується, еталонне значення ключового порівняння розраховується без даних лабораторії, яка має найвище значення "*X*". Ця процедура повторюється доти, поки умову не буде виконано. Після цього отримане таким чином значення *y* приймається як еталонне значення ε_{CRV}, а значення *U(y)* визнається як розширена невизначеність $U(\varepsilon_{CRV})$.

При підготовці до міжнародних звірень важливо враховувати значення проходження тесту "x_i – квадрат", який обчислюється на основі похибки вимірювань еталона-переносника та стандартної невизначеності. Тому важливо,

щоб стандартна невизначеність вимірювань за типом А еталона була якомога меншою. Крім того, при складанні технічних протоколів звірень необхідно визначати умови проведення вимірювань, такі як температура навколишнього середовища, вологість повітря, температура рідини, атмосферний тиск тощо. Серед таких умов часто враховуються й умови, що стосуються самої лабораторії, такі як герметичність гідротакту, відсутність повітря в системі, встановлення нуля витратоміра, допустимі зміни витрати при проведенні вимірювань на одній точці витрати.

2.5 Результати досліджень

Дослідження стабільності масової витрати рідини показали, що існує необхідність у виявленні можливих несправностей у обладнанні. Це може призводити до появи систематичних або випадкових помилок та похибок, які впливають на метрологічні характеристики еталону.

Дослідження розширеної невизначеності вимірювань витратомірів виконувалось аналогічно процедурі проведення міжнародних звірень, у результаті якої розраховується розширена невизначеність вимірювань. Ці значення допомагають оцінити межі можливих відхилень результатів вимірювань від середнього значення з певною ймовірністю. Вона враховує як систематичні, так і випадкові помилки вимірювань, а також ураховує додаткові фактори, що можуть впливати на точність результатів. Тому необхідно вчасно виявляти ці помилки для їх усунення під час проведення вимірювань.

Одним з підходів до виявлення випадкових помилок та похибок є методи ML. У якості вхідних даних для моделі ML необхідно використовувати відносну похибку вимірювань. Це пов’язано зі специфікою передачі одиниці еталоном ДЕТУ 03-04-04. При проведенні вимірювань немає необхідності у фіксованому значенні тривалості одного виміру, тобто кожне вимірювання часу або маси рідини відрізняється від інших значень у рамках однієї точки витрати, при

цьому значення витрати залишається незмінним. Це свідчить про те, що використання цих значень як вхідних даних є некоректним. Передача одиниці відбувається шляхом порівняння значення показань ЗВТ зі значенням показань еталону, тобто похибкою вимірювання, вираженою у відсотках, значення якої використовується для калібрування ЗВТ.

Відносна похибка вимірювань маси рідини ε, у відсотках, розраховується за наступною формулою [47]:

$$\varepsilon = \frac{(m_v - m_{ref})}{m_{ref}} \cdot 100\%, \tag{2.12}$$

де m_{ref} – значення маси рідини еталона,

m_v – значення маси рідини коріолісового витратоміра.

При дослідженні розширеної невизначеності вимірювань витратомірів на одній точці, необхідно виконувати по 11 вимірювань. Однак для дослідження виявлення викидів були проведені додаткові вимірювання в кожній вибірці. Це дозволило додатково перевірити працездатність моделі на вибірках різних розмірів, що варіювалися в діапазоні від 13 до 33 значень. Загалом на трьох витратомірах було виконано 169 вимірювань.Значення відносної похибки вимірювань будуть використані у якості вхідних даних для моделі ML, та представлені в таблиці 2.6

Таблиця 2.6 - Похибка вимірювання витратомірів

№	Точка масової витрати								
	Витратомір №1			Витратомір №2			Витратомір №3		
	45 т/г	25 т/г	5 т/г	5 т/г	2,5 т/г	1 т/г	1 т/г	0,5 т/г	0,1 т/г
1	0,16	0,28	0,34	0,06	0,22	0,43	0,42	0,26	2,90
2	0,23	0,15	0,22	0,11	0,10	0,39	0,65	0,37	2,86
3	0,22	0,40	0,43	0,12	0,05	0,23	0,46	0,29	2,65
4	0,20	0,78	0,27	0,17	0,33	0,40	0,42	0,30	2,68
5	0,22	1,09	0,30	0,19	0,18	0,36	0,47	0,33	2,30
6	0,33	0,37	0,29	0,20	0,16	0,41	0,46	0,34	2,82
7	0,31	0,04	0,40	0,22	0,22	0,44	0,42	0,48	2,58

Продовження таблиці 2.6

№	45 т/г	25 т/г	5 т/г	5 т/г	2,5 т/г	1 т/г	1 т/г	0,5 т/г	0,1 т/г
8	0,42	0,21	0,29	0,14	0,07	0,48	0,69	0,45	2,60
9	0,52	0,33	0,25	0,07	0,35	0,24	0,44	0,29	2,58
10	0,18	0,19	0,24	0,31	0,27	0,37	0,32	0,41	3,04
11	0,25	0,28	0,24	0,13	0,09	0,30	0,42	0,46	2,81
12	0,29	0,22	0,33	0,22	0,13	0,37	0,38	0,12	2,71
13	0,41	0,46	0,34	0,40	0,35	0,31	0,44	0,42	2,43
14	0,47	0,17	0,12	0,26	0,15	0,43	0,29	0,18	2,70
15	0,54	0,17	0,31	0,25	0,16	0,40	0,41	0,37	2,52
16	0,16		0,27	0,15	0,29	0,41	0,45	0,39	2,10
17	0,18		0,36	-0,13		0,33	0,47	0,44	
18				0,23		0,39	0,44	0,25	
19						0,38		0,28	
20						0,37		0,42	
21						0,42		0,42	
22						0,38		0,26	
23						0,35		0,43	
24						0,43		0,29	
25						0,41		0,26	
26						0,41		0,26	
27						0,42		0,23	
28						0,22		0,22	
29						0,31		0,38	
30						0,38		0,48	
31						0,41		0,37	
32						0,38		0,22	
33						0,39		0,46	

2.6 Висновок до другого розділу

Міжнародні звірення є важливим етапом у процесі забезпечення точності вимірювань. Вони дозволяють встановити відповідність між вимірювальними приладами та еталонами, що необхідно для забезпечення метрологічної надійності.

Стабільність витрати рідини була покращена завдяки проведеним дослідженням. У процесі дослідження було виявлено слабку стабільність витрати

рідини, яка виникала через погіршення працездатності насосної частини еталона, після чого було проведено зміну насосних агрегатів на нові. Також було помічено неможливість точного встановлення певного рівня витрати рідини, та причину погіршення характеристик насосних агрегатів, а саме вузол управління встановлення рівня витрати рідини. Регулювання рівнем рідини проводиться вентилями та електромеханічними кранами. Тому було встановлено частотний перетворювач який завдяки регулюванню частоти обертання двигуна насосного агрегату, може доволі точно встановлювати необхідний рівень витрати рідини та має плавний запуск, що зменшує флуктуації рідини при виході на робочий режим вимірювання на еталоні.

У результаті модернізації вдалося досягти суттєвого підвищення стабільності витрати рідини, що підтверджується зниженням стандартної невизначеності за типом А. Це зниження відображає покращену можливість відтворення витрати рідини та зменшення коливань витрати під час одного вимірювання.

На нижньому діапазоні витрати рідини від 0,5 до 2,5 т/г стандартна невизначеність знизилася з 0,1% до 0,03%, що стало результатом заміни насосного агрегату Н1 (рис. 2.1). На точці витрати 25 т/г невизначеність була зменшена з 0,21% до 0,085% завдяки модернізації насосного агрегату Н5 (рис. 2.1). Додатково, впровадження нового вузла управління рівнем витрати рідини дозволило знизити невизначеність вимірювань за типом А на всьому діапазоні витрати, забезпечуючи стабільні результати вимірювань.

Дослідження розширеної невизначеності вимірювань витратомірів проводилось згідно з процедурою міжнародних звірень, що забезпечує отримання результатів, які відповідають міжнародним стандартам. Проведений аналіз методики обробки результатів вимірювань показав, що ключовим фактором, який впливає на розрахунок розширеної невизначеності, є стандартна невизначеність за типом А. Тому подальші дослідження будуть зосереджені на аналізі викидів у даних та розробці підходів до їх виявлення та виключення з метою підвищення достовірності та надійності метрологічних результатів.

3. ВИКОРИСТАННЯ МОДЕЛІ МАШИННОГО НАВЧАННЯ ДЛЯ ЗНАХОДЖЕННЯ ВИКИДІВ

3.1 Вибір моделі машинного навчання

Виявлення викидів в даних є важливим завданням у сфері ML та аналізу даних. Ця задача може бути виконана за допомогою методів навчання з учителем, без учителя або комбінацією обох підходів. Методи навчання з учителем ґрунтуються на використанні розмічених даних, де кожному об'єкту присвоєна мітка класу - 'нормальний' або 'аномальний'. Модель навчається на таких даних для того щоб навчитися розділяти нормальні та аномальні об'єкти. Дані для методу з учителем мають бути попередньо оброблені та очищені від аномалій, щоб модель могла приймати ці значення як нормальні та виявляти все, що відрізняється від нормальних значень.

Використання даних дослідження розширеної невизначеності вимірювань витратомірів не передбачає можливості виділити дані для навчання моделі. Тому необхідно використовувати моделі без учителя, оскільки вони не потребують розмічених даних. Замість цього вони намагаються виявити структуру даних та виділяти об'єкти, які значно відрізняються від інших.

Важливо зауважити, що процес виявлення викидів може бути складним і вимагати ретельного налаштування параметрів моделі. Крім того, моделі виявлення викидів можуть стикатися з проблемою незбалансованих класів, коли кількість аномальних об'єктів значно менше, ніж нормальних.

Серед моделей ML без учителя виділяються декілька підходів:

1. **ІЛ (Isolation Forest).** Цей метод вирізняється здатністю ефективно виявляти аномалії, навіть у великих наборах даних. Він не потребує попереднього визначення кількості кластерів і ефективно виділяє аномалії завдяки принципу ізоляції аномальних точок в ДР. Цей метод особливо корисний у випадках, коли аномалії не мають типового розподілу або мають різноманітні форми [49];

2. **Локальний вибір (LOF - Local Outlier Factor).** Він надає міру локальної віддаленості або ізольованості об'єкта від його сусідів. Це дозволяє виявляти аномалії, які можуть бути невидимими на глобальному рівні. Локальний вибір особливо корисний у випадках, коли дані мають групи з різною щільністю, дозволяючи ідентифікувати аномалії в межах локальних груп [50];
3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**. Цей метод ідеально підходить для виявлення аномалій у наборах даних, де аномальні точки представлені як ізольовані точки, віддалені від основних кластерів. DBSCAN здатний виявляти кластери різних форм, що робить його гнучким інструментом для роботи з різноманітними типами даних [51];
4. **Автоенкодери (Autoencoders).** Ці НМ навчаються відтворювати вхідні дані, мінімізуючи реконструкційну помилку. Автоенкодери можуть ефективно виявляти аномалії, оскільки аномалії часто призводять до високої реконструкційної помилки, як автоенкодери навчаються відтворювати лише "нормальні" дані [52];
5. **Однокласова машина опорних векторів (далі — One-Class SVM).** Цей метод призначений для визначення границі нормальних даних, навчаючись лише на "нормальних" даних. Він дозволяє ідентифікувати дані, які не вписуються в цю границю, як аномалії. One-Class SVM є ефективним у ситуаціях, коли доступна обмежена кількість "нормальних" даних [53].

Ці моделі доповнюють одна одну, пропонуючи різноманітні підходи до виявлення викидів в залежності від характеристик даних та поставленої задачі. Вибір конкретної моделі або їх комбінації залежить від типу даних, наявності міток та специфічних цілей дослідження.

Враховуючи аналіз представлених моделей без учителя для виявлення викидів, ІЛ більш за все підходить для дослідження даних вимірювань на еталоні. Основні причини цього вибору обумовлені унікальними властивостями та

перевагами ІЛ, що робить цю модель особливо відповідною для дослідницького контексту:

- **універсальність та ефективність:** ІЛ демонструє вражаючу здатність ефективно обробляти великі обсяги даних незалежно від їх розмірності або розподілу. Він ідеально підходить для ситуацій, де викиди не мають чітко визначеного розподілу або мають різноманітні форми, що робить його оптимальним вибором для аналізу метрологічних даних;
- **мінімальні вимоги до попередньої обробки:** на відміну від інших моделей, таких як One-Class SVM або автоенкодери, ІЛ не потребує складної попередньої обробки даних або великої кількості "нормальних" даних для навчання. Це значно спрощує процес підготовки та аналізу даних, знижуючи ресурсні витрати та час, необхідний для підготовки моделі до використання;
- **швидкість та продуктивність**: ІЛ відзначається високою продуктивністю навіть при обробці великих наборів даних. Це дозволяє швидко ідентифікувати потенційні викидів, що є критично важливим для задач метрології, де швидкість аналізу може впливати на точність та надійність вимірювань.
- **гнучкість та адаптивність**: ІЛ демонструє виняткову гнучкість та адаптивність до різних типів даних та сценаріїв їх використання. Модель може бути ефективно налаштована та адаптована для конкретних потреб дослідження, що робить її відмінним вибором для різноманітних застосувань у метрології.

Враховуючи вищезазначені фактори, ІЛ виявляється оптимальним вибором для використання в сучасних дослідженнях в галузі метрології. Його висока ефективність, універсальність обробки даних, мінімальні вимоги до попередньої обробки та здатність швидко адаптуватися до складних умов дослідження роблять його ідеальним інструментом для аналізу метрологічних даних, де точність, швидкість та надійність аналізу мають вирішальне значення.

3.2 Модель ізольованого лісу

Модель ІЛ була вперше представлена Fei Tony Liu, Kai Ming Ting та Zhi-Hua Zhou 2008 році на міжнародній конференції IEEE International Conference on Data Mining [54]. Основна ідея методу полягає у використанні ізоляції як способу виявлення аномалій. Метод базується на припущенні, що аномалії, як правило, ізолюються швидше, ніж типові точки, завдяки своїй віддаленості або унікальності у багатовимірному просторі даних.

Метод був інтегрований у популярну бібліотеку ML — “Scikit-learn”, що зробило його доступним для широкого кола користувачів. Його застосування охоплює різні сфери, включаючи виявлення шахрайства у фінансах, аналіз мережевого трафіку в кібербезпеці, моніторинг стану обладнання в промисловості та аналіз медичних даних. Завдяки своїм унікальним властивостям і масштабованості ІЛ став стандартним інструментом для задач виявлення аномалій [55-63].

Аномалія — це точка даних, яка суттєво відрізняється від інших точок у вибірці. Аномалії можуть виникати внаслідок похибок вимірювання, шуму, рідкісних або виняткових подій у системі, або ж вказувати на проблеми, які потребують додаткового аналізу. У контексті ІЛ аномалія визначається як точка, яку можна ізолювати за допомогою меншої кількості поділів у порівнянні з типовими точками, що знаходяться у щільних кластерах. Це обумовлено тим, що аномалії зазвичай мають віддалене або унікальне положення в багатовимірному просторі ознак.

Механізм роботи моделі базується на побудові ансамблю ізоляційних дерев, кожне з яких створюється шляхом рекурсивного поділу вибірки. На кожному кроці вибирається випадкова ознака та поріг розділення. Поділи тривають до тих пір, поки кожна точка даних не буде ізольована в окремому вузлі або не буде досягнуто максимальної глибини дерева, яка зазвичай визначається як $log_2(n)$, де n — кількість точок у вибірці.

Ключовим параметром у моделі є довжина ізоляційного шляху *h(x)*, яка визначається як кількість поділів, необхідних для ізоляції точки *x* у дереві. Для нормалізації довжини ізоляційного шляху використовується функція *c(n)*, що враховує розмір вибірки:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}. \tag{3.1}$$

де *H(i)* — гармонійне число:

$$H(i) = \sum_{k=1}^{i} \frac{1}{k}, \text{ або } H(i) \approx \ln(i) + \gamma. \tag{3.2}$$

k — це змінна, яка приймає цілі значення від 1 до *i* під час обчислення суми;

γ — стала Ейлера-Маскероні ($\gamma \approx 0.577$).

Ця функція масштабує довжину ізоляційного шляху *h(x)*, забезпечуючи порівнянність результатів для різних наборів даних.

Нормалізований «коефіцієнт аномальності» або «ступінь аномальності» *s(x)* визначається за формулою:

$$s(x) = 2^{-\frac{h(x)}{c(n)}}. \tag{3.3}$$

Якщо *h(x)* мала, тобто точку легко ізолювати, *s(x)* наближається до 1, що вказує на високу ймовірність аномалії. Якщо *h(x)* велика, *s(x)* близьке до 0, що відповідає типовій точці.

У бібліотеці Scikit-learn реалізація ІЛ модифікована для використання шкали [−1,1], яка дозволяє спростити інтерпретацію результатів. Ця модифікація досягається шляхом перетворення шкали [0,1] у шкалу [−1,1] за допомогою формули:

$$s`(x) = 2s(x) - 1. \tag{3.4}$$

У цьому випадку значення $s'(x) < 0$ відповідають аномальним точкам, тоді як $s'(x) \geq 0$ — типовим точкам.

3.3 Налаштування моделі ізольованого лісу

Технічно модель ІЛ реалізована у вигляді програмного коду на мові Python, який наведено у Додатку В.

3.3.1 Програмний код

Програмний код включає в себе використання декількох бібліотек, зокрема:

- **scikit-learn**, яка надає готові інструменти для ML та аналізу даних;
- **pandas**, яка забезпечує високорівневі структури даних та функції для роботи з різними типами даних, такими як таблиці, часові ряди та інше;
- **matplotlib**, яка надає широкі можливості для створення різних типів графіків, включаючи лінійні графіки, стовпчасті діаграми, кругові діаграми, діаграми розсіювання та багато іншого;
- **numpy**, яка працює зі структурами даних, дозволяє виконувати математичні операції, лінійну алгебру, випадкові числа, перетворення даних та інше.

У наступних програмних кодах, також будуть використовуватись наступні бібліотеки:

- **os**, дозволяє взаємодіяти з операційною системою для виконання таких завдань, як створення шляхів до файлів, перелік файлів у директорії та інші операції, пов'язані з файловою системою;
- **re**: виконує складні завдання пошуку та заміни в текстових даних, включаючи передобробку текстових даних для аналізу;
- **scipy.stats,** частина бібліотеки SciPy, що надає багатий набір статистичних функцій, яка може використовуватися для глибокого аналізу даних, включаючи тестування гіпотез, розрахунок різних статистичних показників та моделювання.

Після імпорту необхідних бібліотек, відбувається читання даних з Excel-файлу. Для цього використовується функція `pd.read_excel("1.xlsx")`, яка зчитує дані з файлу "1.xlsx" і завантажує їх у форматі DataFrame. DataFrame - це основна структура даних в бібліотеці pandas, яка представляє дані у вигляді табличної форми. Ці дані будуть використовуватися для навчання моделі та аналізу аномалій.

У контексті моделі ІЛ, "навчанням моделі" є процес побудови ДР, які випадковим чином розділяють дані на менші групи. Ці дерева розбивають дані таким чином, щоб аномальні спостереження потрапляли в окремі гілки дерева, в той час як нормальні спостереження будуть мати менший шлях у дереві.

Після завантаження даних з файлу Excel йде налаштування параметрів моделі. Параметри моделі визначаються перед початком навчання. Після визначення параметрів моделі викликається метод `.fit()`, який навчає модель на навчальних даних. Тобто, модель адаптується до характеристик даних, враховуючи їх особливості і залежності.

Після проведення обчислень результати записуються в окремий файл Excel, у якому зберігається ступінь аномальності для кожного значення вхідних даних, а також визначається ознака аномальності для подальшого аналізу.

Ступінь аномальності обчислюється шляхом усереднення результатів серії запусків моделі з різними випадковими параметрами. Цей показник вказує на те, наскільки значення вхідних даних відхиляються від звичайних, нормальних значень. Чим менше значення ступеня аномальності, тим більша ймовірність, що це значення є аномалією. Визначається ознака аномальності, яка допомагає ідентифікувати, чи є кожне значення аномальним (1) або нормальним (-1) у порівнянні з рештою даних.

Після цього дані візуалізуються у вигляді графіків, де один графік відображає оригінальні дані, а інший - ступінь аномальності для кожного значення.

3.3.2 Параметри моделі.

Для досягнення оптимальних результатів необхідно належним чином налаштувати параметри моделі. Основні параметри налаштування включають: кількість дерев, максимальну глибину дерева, кількість прикладів у підвибірці, кількість припущених аномалій, вибірковість та випадковість. Для більш конкретних задач можуть бути використані додаткові параметри, такі як мінімальна кількість об'єктів у листках дерева, максимальне співвідношення аномалій та нормальних об'єктів, чи рівень впевненості в класифікації. Ці параметри допомагають налаштувати модель з урахуванням конкретних вимог задачі та властивостей даних.

Кількість дерев (n_estimators) є одним з основних параметрів у моделі ІЛ. Цей параметр визначає, скільки дерев буде створено в ансамблі. Кожне дерево в лісі використовує власну підвибірку даних та різні випадкові вибори ознак для побудови моделі. Зображення ДР при невеликій кількості ознак наведено на рисунку 3.1.

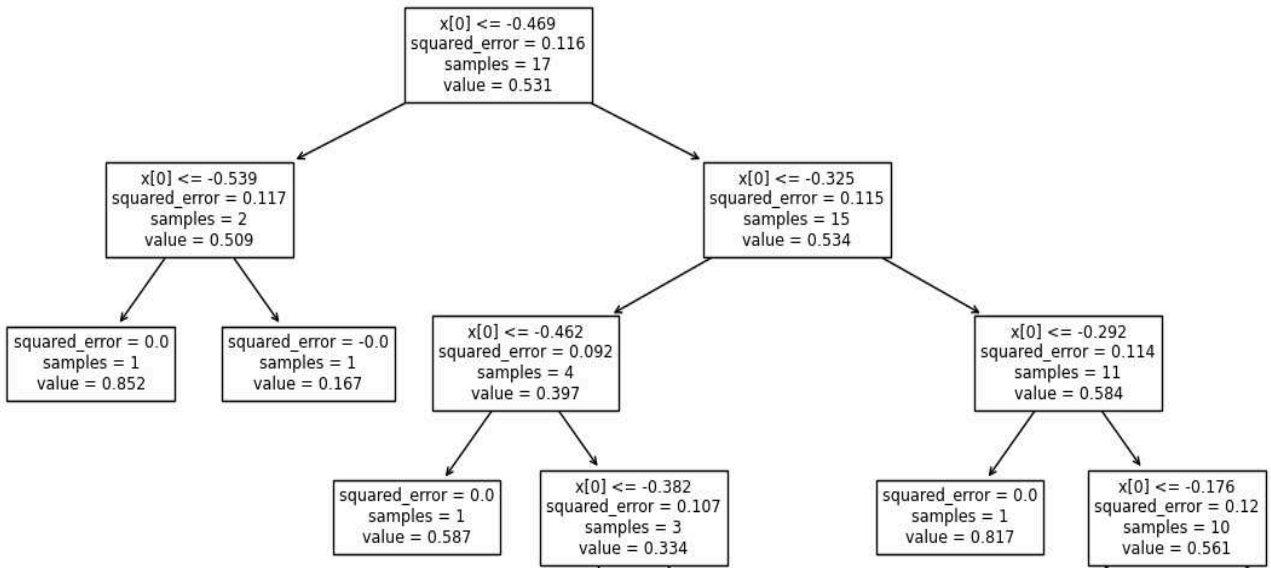

Рисунок 3.1 - ДР у моделі ІЛ.

ДР утворюються шляхом випадкового вибору підмножини ознак та поділу даних на дві частини на кожному рівні побудови. Кожне ДР починається з кореневого вузла, який включає в себе всі доступні ознаки. За допомогою

випадкового вибору підмножини ознак обирається певна кількість ознак для розгляду на даному рівні. Потім дані розділяються на дві групи відповідно до значень обраних ознак та певного порогу. Об'єкти, що мають значення ознаки менше порогу, потрапляють в одну групу, тоді як ті, що більше або дорівнюють порогу, - в іншу.

Після розділення даних кожна з отриманих груп стає новим піддеревом, і процес побудови повторюється для кожної з цих груп рекурсивно. Кожен новий вузол являє собою випадкову ознаку та поріг, за яким дані розділяються на дві частини. Цей процес триває до досягнення критерію зупинки, наприклад, досягнення максимальної глибини дерева або досягнення мінімальної кількості об'єктів у вузлі.

Кількість дерев у лісі має бути якомога більшим для максимальної продуктивності та точності роботи моделі. Виходячи з обчислювальних можливостей, цей параметр був збільшений до 1000 дерев.

При кожному повторному запуску моделі виявлялись незначні відмінності в значеннях ступеня аномальності, які розраховує модель. Для усунення цієї проблеми у програмному коді було виконано по 10 запусків моделі та усереднено 10 значень результатів, отриманих від кожного запуску. Такий підхід допоміг зменшити вплив незначних відхилень у значеннях, що можуть виникати при кожному новому запуску моделі. Після усереднення результатів роботи моделі було отримано стабільні та незмінні результати.

У даній моделі у якості критерію зупинки був обраний параметр **"max_samples"**, який контролює розмір випадкової підвибірки даних, що використовується для побудови кожного дерева у лісі. Використання цього параметра допомагає уникнути перенавчання моделі, зменшити обчислювальні витрати та забезпечити більшу різноманітність у навчанні кожного дерева.

Параметр "max_samples" був встановлений на значення "auto", що дозволяє автоматично розраховувати кількість об'єктів як **квадратний корінь від загальної кількості об'єктів у навчальній вибірці**. Цей підхід забезпечує розумне співвідношення між розміром підвибірки та загальною кількістю об'єктів,

що допомагає уникнути перенавчання та забезпечити стабільні результати моделі.

Параметр **"кількість припущених аномалій"** у моделі ІЛ визначає, скільки аномальних об'єктів очікується у навчальній вибірці. Цей параметр впливає на те, як модель розпізнає аномалії і може бути корисним, якщо заздалегідь відомо або припущено, що аномалій у навчальних даних немає великої кількості. Чим вище цей параметр, тим більш чутливою до аномалій буде модель.

Оскільки вхідними даними є значення вимірювань на державному первинному еталоні, то доцільно встановити параметр "кількість припущених аномалій" на максимальну чутливість, тобто 50%. Це дозволить моделі бути більш чутливою до аномалій і виявляти навіть дрібні відхилення в даних. Однак експериментально було встановлено, що при встановленні цього параметра на максимальне значення виникає несумісність декількох вибірок між собою через різний діапазон значень ступеня аномальності, а також відображення некоректно великої кількості аномалій, включаючи ті точки, які візуально можна розцінити як нормальні. Тому для даного типу витратомірів цей параметр було встановлено на значення 25%.

Для зменшення випадковості результатів моделі встановлюється **початкове значення генератора випадкових чисел**, яке використовується для ініціалізації внутрішніх випадкових процесів в алгоритмі. Наприклад, якщо встановлене значення "1000", це генерує випадкове число від 0 до 999 включно кожного разу, коли виконується алгоритм.

Важливо зазначити, що якщо цей параметр не встановлено, кожен новий запуск моделі буде генерувати нове випадкове число, що може призвести до відрізнення результатів. Суттєво, значення цього числа не впливає на результат моделі, проте встановлення початкового значення генератора випадкових чисел дозволяє забезпечити повторюваність результатів при повторних запусках моделі. Такий підхід корисний для порівняння результатів різних моделей або експериментів.

3.3.3 Ознаки в моделі ізольований ліс

У ML ознаки служать основою для аналізу та обробки даних. У моделі ІЛ кожна ознака являє собою індивідуальний атрибут або характеристику об'єкта, який аналізується. Ці ознаки можуть бути кількісними (наприклад, температура, тиск) або якісними (наприклад, колір, категорія), і кожна ознака повинна по-своєму відображати властивості об'єктів у наборі даних.

Розділення даних — це центральний елемент алгоритму ІЛ. Процес розпочинається з вибору однієї ознаки з набору доступних ознак. Потім алгоритм випадковим чином вибирає значення в межах діапазону цієї ознаки. На основі цього значення дані розділяються на дві групи: одна містить об'єкти зі значеннями ознаки нижче вибраного порогу, а інша — зі значеннями вище цього порогу. Цей процес розділення повторюється, поки не будуть досягнуті певні умови зупинки, такі як ізоляція усіх точок даних або досягнення максимальної глибини дерева.

Використання багатовимірного аналізу в ІЛ збільшує ймовірність успішного виявлення аномалій. Аномалії часто не проявляються ізольовано в одній ознаці, а радше у поєднанні змін за кількома взаємопов'язаними ознаками. Наприклад, при вимірюванні параметрів навколишнього середовища у приміщенні, таких як температура повітря, вологість або атмосферний тиск, і раптовій зміні вимірюваних показників, наприклад при виникненні пожежі, усі показники можуть одночасно змінюватися в тій чи іншій мірі. Якщо модель навчена розпізнавати такі комбіновані зміни, це значно підвищує її здатність визначати аномальні події.

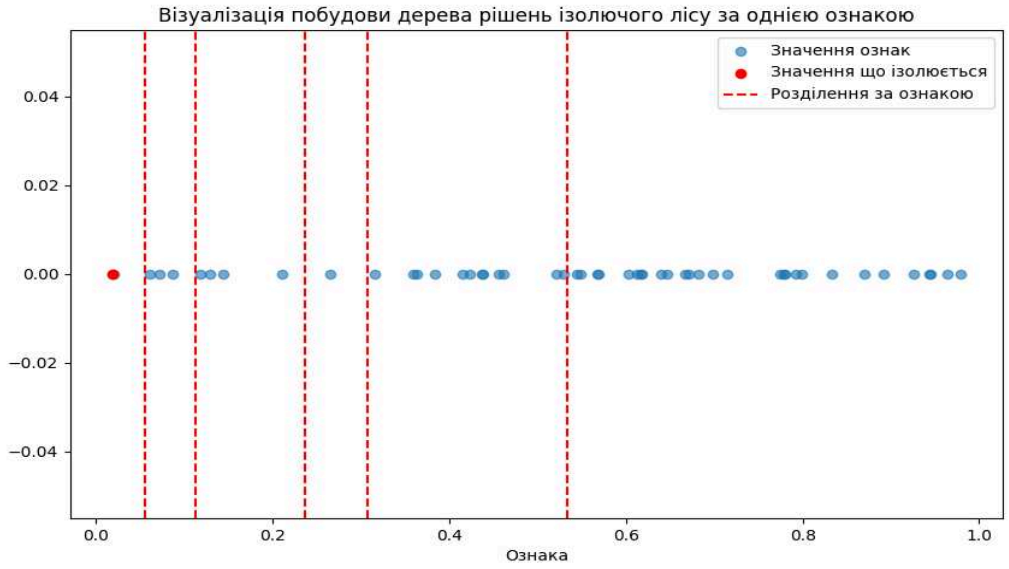

Рисунок 3.2 - Візуалізація процесу розділення за однією ознакою.

На рис 3.2 показано початок процесу розділення даних у рамках одного виміру (одновимірний простір) з використанням методу ІЛ. Вертикальні пунктирні лінії ілюструють, як у кожний момент часу вибірка розділяється за значенням однієї ознаки, вибраної випадковим чином з діапазону доступних значень цієї ознаки. У даному випадку ознака одна, і на графіку вона представлена по горизонтальній осі, а кожна точка відображає унікальне значення ознаки. Червона точка позначає об'єкт, який передбачається ізолювати. Рис 3.2 є ще однією з інтерпретацій побудови ДР моделі ІЛ (рис 3.1).

У контексті одновимірного простору виявлення аномалій легко візуалізується, оскільки всі розділення відбуваються уздовж однієї осі та представлені у вигляді ряду порогових значень. У двовимірному або багатовимірному просторі подібний процес також передбачає розділення, але вже у різних вимірах, що робить візуалізацію складнішою, проте основна концепція залишається незмінною, як показано на рис 3.3.

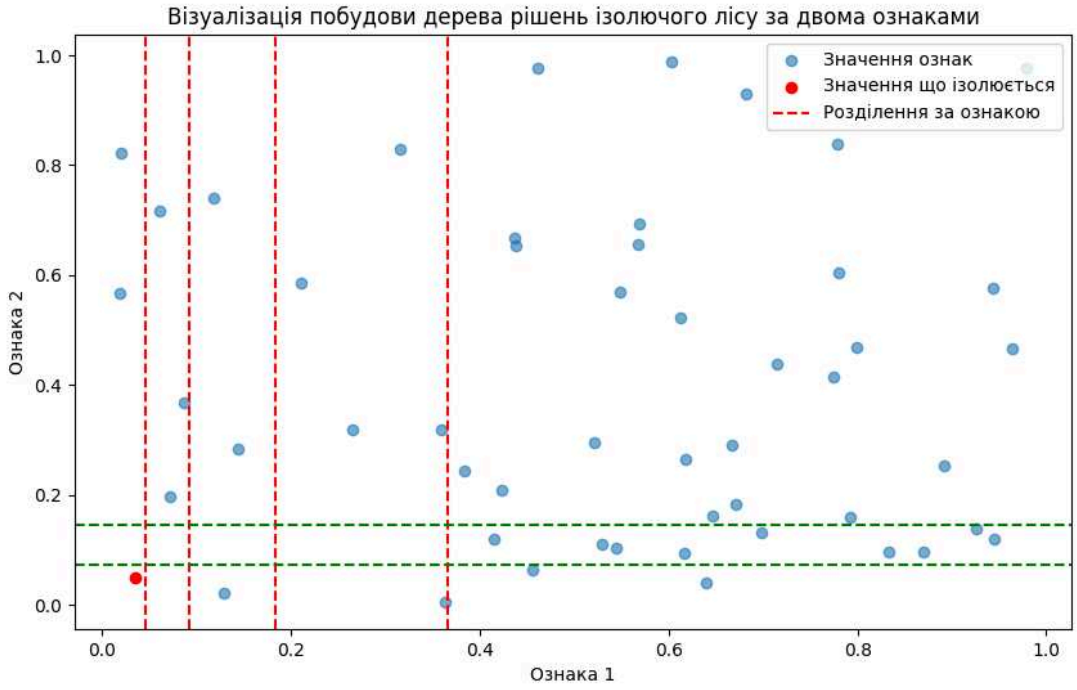

Рисунок 3.3 - Візуалізація процесу розділення за двома ознаками.

Рис 3.3 також ілюструє процес розділення, але у двовимірному просторі, де використовуються дві ознаки. Кожна лінія на графіку символізує розділення даних залежно від значення однієї з ознак. Ефективність ІЛ у багатовимірному просторі визначається здатністю ознак відображати аномальні дані. Чим більша релевантність ознак та їх комбінацій для виявлення аномалій, тим вища ймовірність успішного їх виявлення моделлю.

Тому доцільно використовувати більшу кількість ознак при роботі моделі, однак через специфіку досліджувальних даних ми будемо використовувати одну ознаку, а саме — відносну похибку вимірювань коріолісових витратомірів (табл. 2.6).

Під час дослідження стабільності витрати рідини, спрямованого на аналіз повторюваності вимірювань та стабільності потоку рідини, шляхом розрахунку стандартної невизначеності вимірювань за типом А на еталоні, були виміряні такі параметри, як маса рідини, час вимірювання та температура рідини. Ці величини могли бути використані як ознаки, оскільки вимірювання були націлені на виявлення повторюваності значень в рамках однієї точки витрати.

Під час дослідження розширеної невизначеності вимірювань коріолісових витратомірів цей аспект не брався до уваги, тому всі вимірювані значення мають великий і хаотичний діапазон відхилень одне від одного. Така ситуація виникає через те, що відносна похибка вимірювань визначається як різниця між показниками маси рідини, виміряної еталоном та витратоміром, і виражається у відсотках. Різке зміщення цієї різниці вказує на аномалію, яку потрібно виявляти, в той час як діапазон значень вимірювань може змінюватися в рамках допустимого діапазону відхилення витрати рідини (в рамках точки витрати рідини) і не впливати на показник похибки. Під час калібрування та перевірки ЗВТ на еталоні розраховується саме показник відносної похибки вимірювань, на основі якого приймається остаточне рішення.

3.4 Експериментальне дослідження роботи моделі ізольованого лісу

Експериментальне дослідження роботи моделі ІЛ на реальних вимірювальних даних дозволило отримати графіки ступеня аномальності, які наведені на рисунках 3.4-3.21. Після застосування моделі до вхідних даних, представлених на лівих графіках (рисунок 3.4-3.21), ми отримали графіки значень ступеня аномальності для кожного вимірювання, зображені на правих графіках (рисунок 3.4-3.21.). Цей ступінь аномальності вказує, наскільки модель вважає відповідні вхідні значення вимірювання аномальними [63].

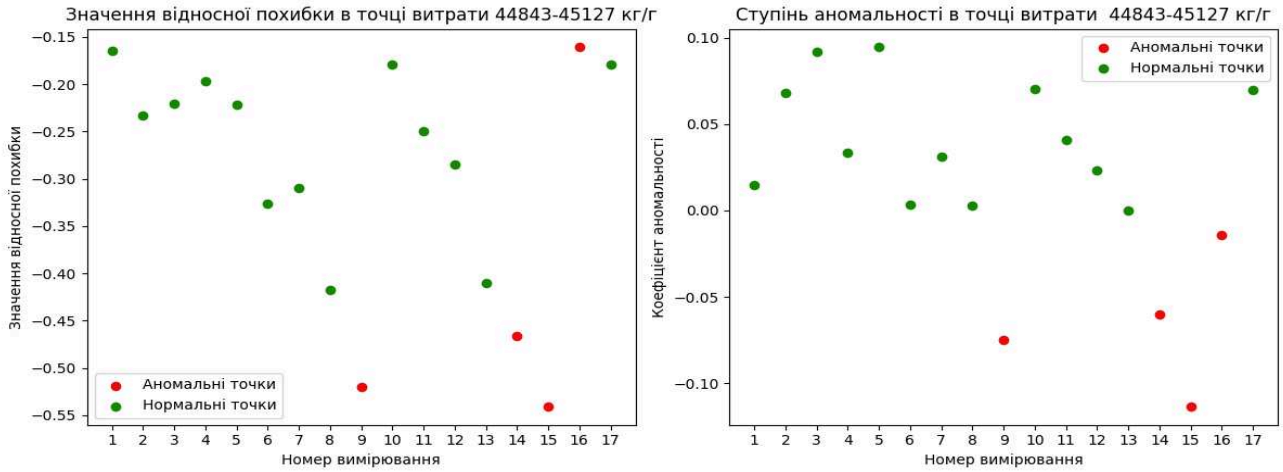

Рисунок 3.4 - Значення ступеня аномальності в точці витрати 45 т/г, витратоміра №1

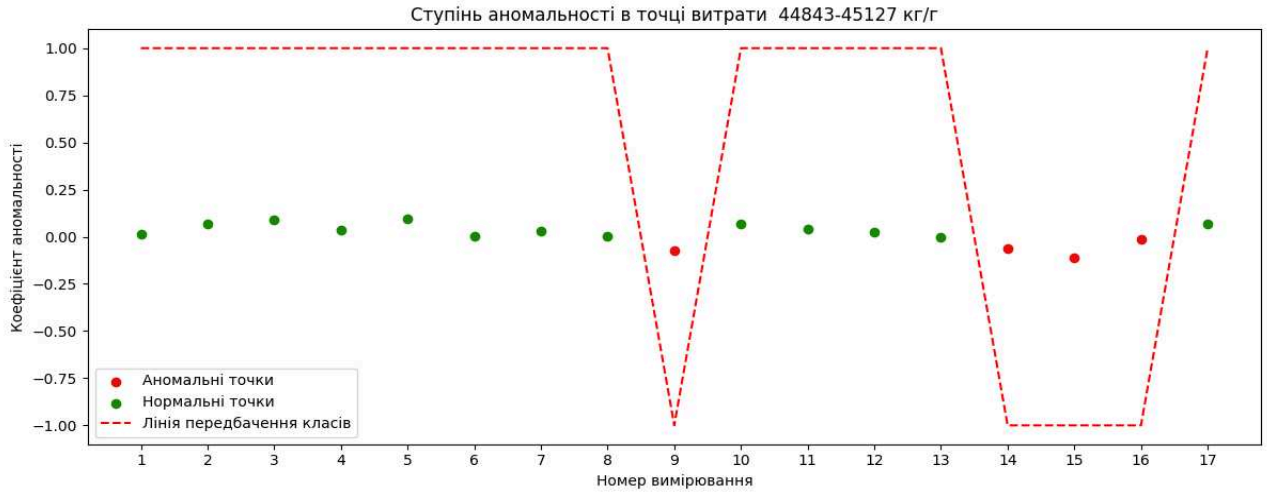

Рисунок 3.5 - Лінія передбачення класів в точці витрати 45 т/г, витратоміра №1

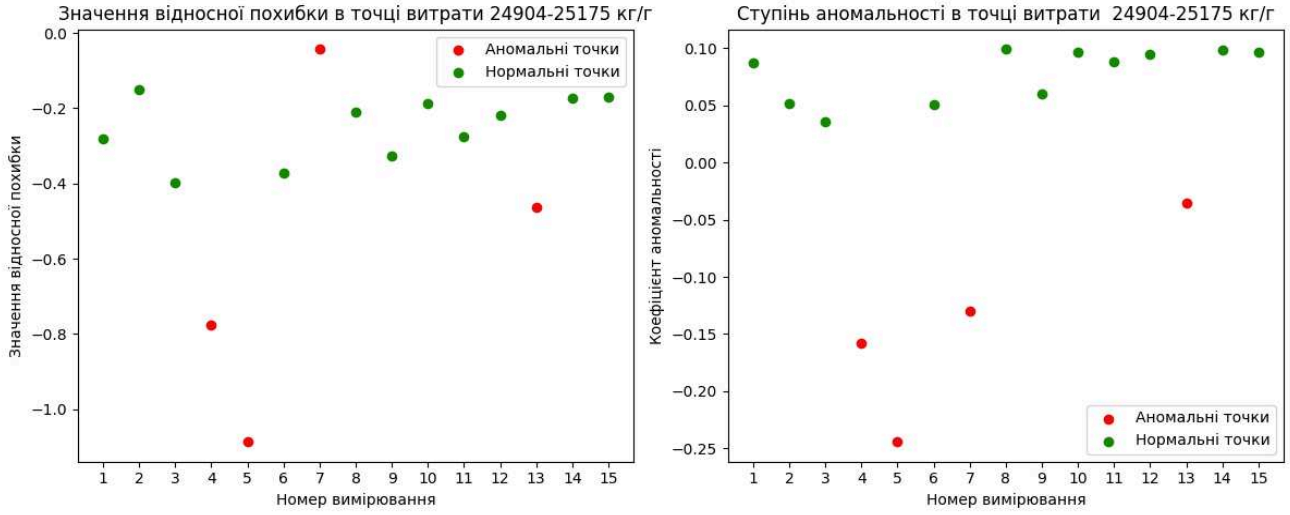

Рисунок 3.6 - Значення ступеня аномальності в точці витрати 25 т/г, витратоміра №1

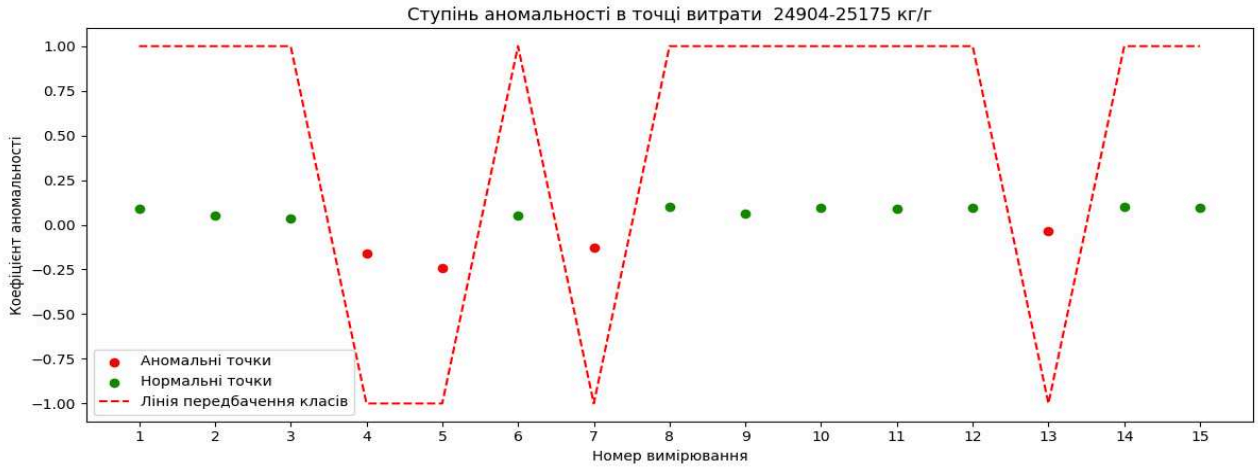

Рисунок 3.7 - Лінія передбачення класів в точці витрати 25 т/г, витратоміра №1

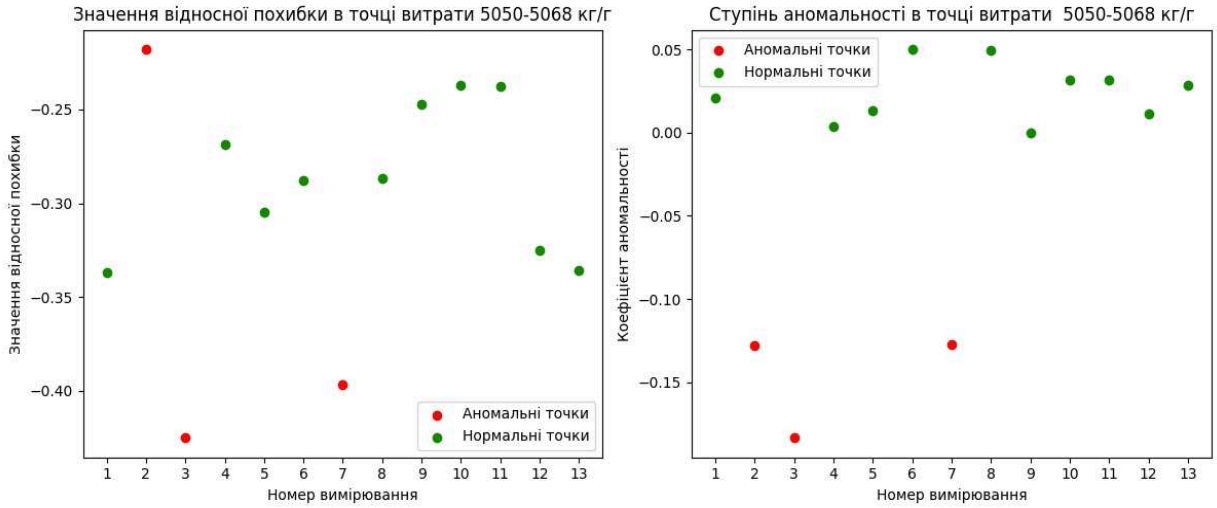

Рисунок 3.8 - Значення ступеня аномальності в точці витрати 5 т/г, витратоміра №1

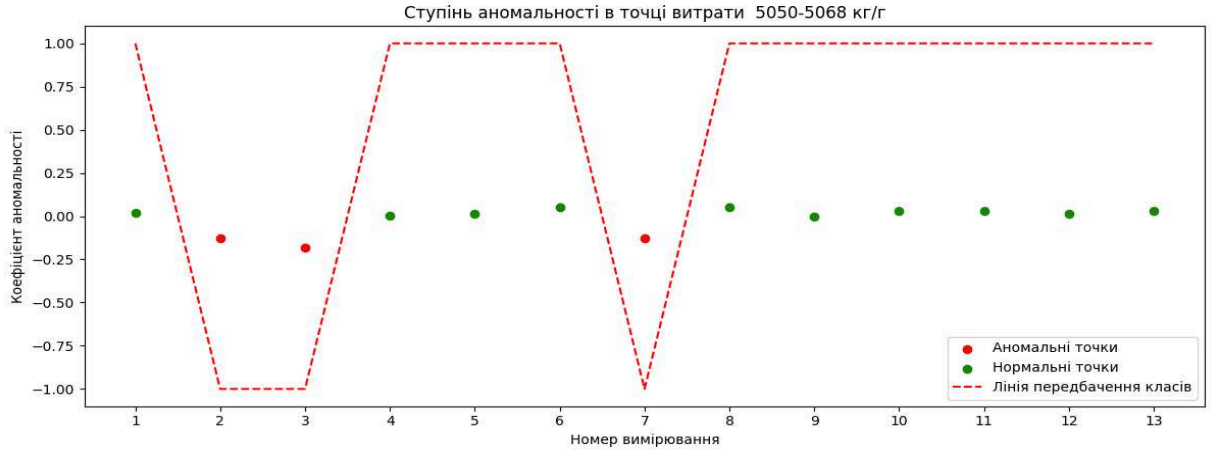

Рисунок 3.9 - Лінія передбачення класів в точці витрати 5 т/г, витратоміра №1

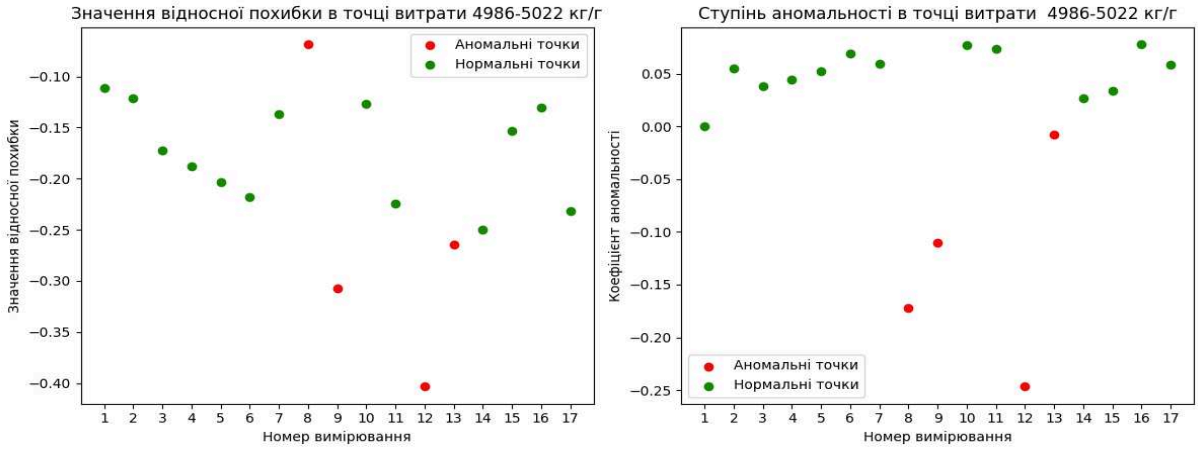

Рисунок 3.10 - Значення ступеня аномальності в точці витрати 5 т/г, витратоміра №2

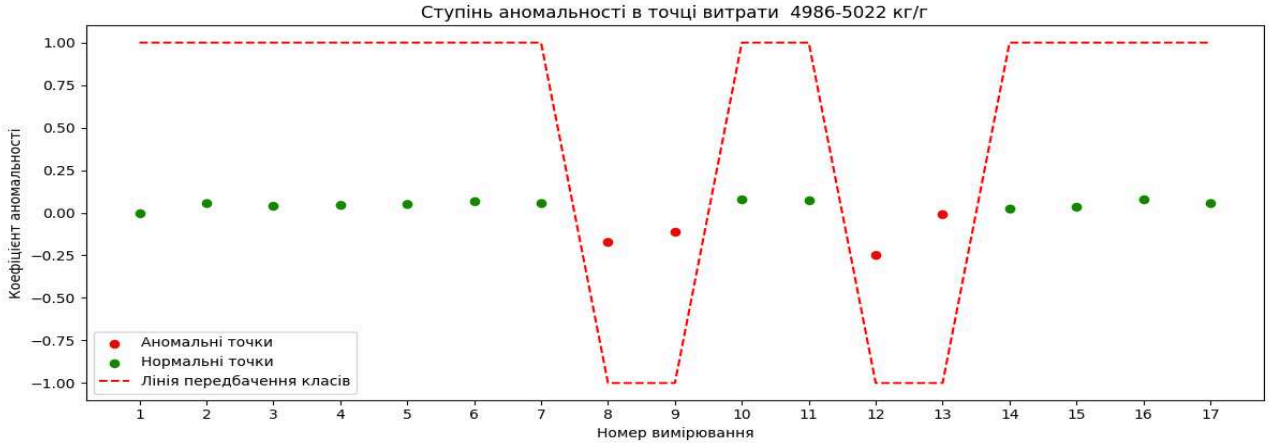

Рисунок 3.11 - Лінія передбачення класів в точці витрати 5 т/г, витратоміра №2

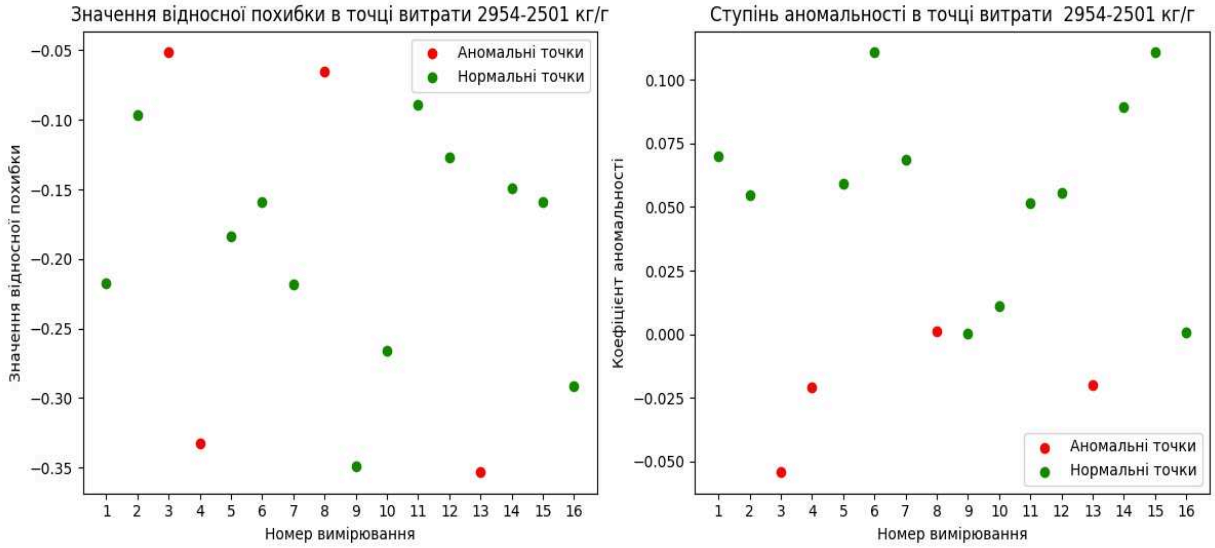

Рисунок 3.12 - Значення ступеня аномальності в точці витрати 2,5 т/г, витратоміра №2

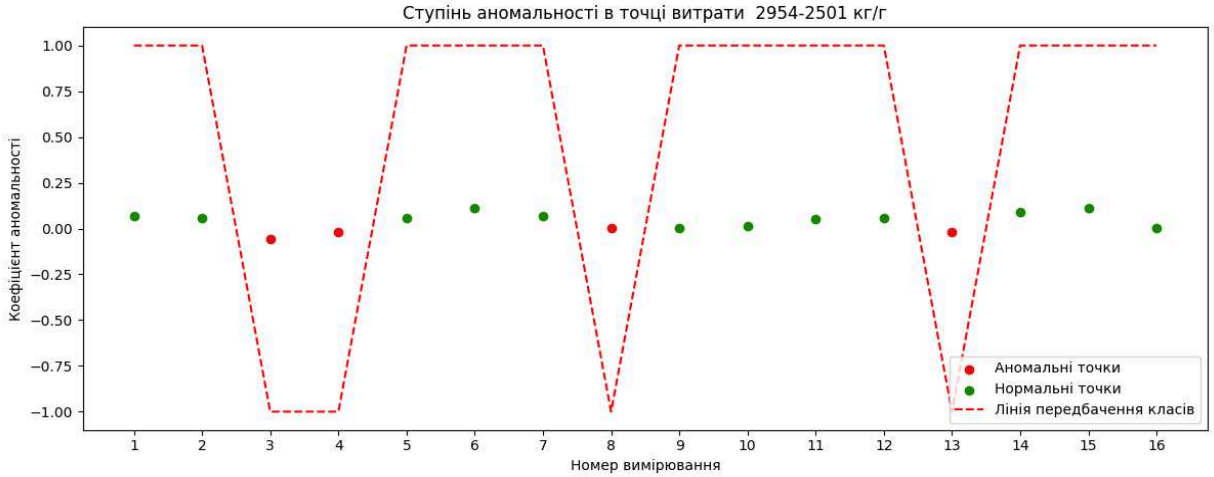

Рисунок 3.13 - Лінія передбачення класів в точці витрати 2,5 т/г, витратоміра №2

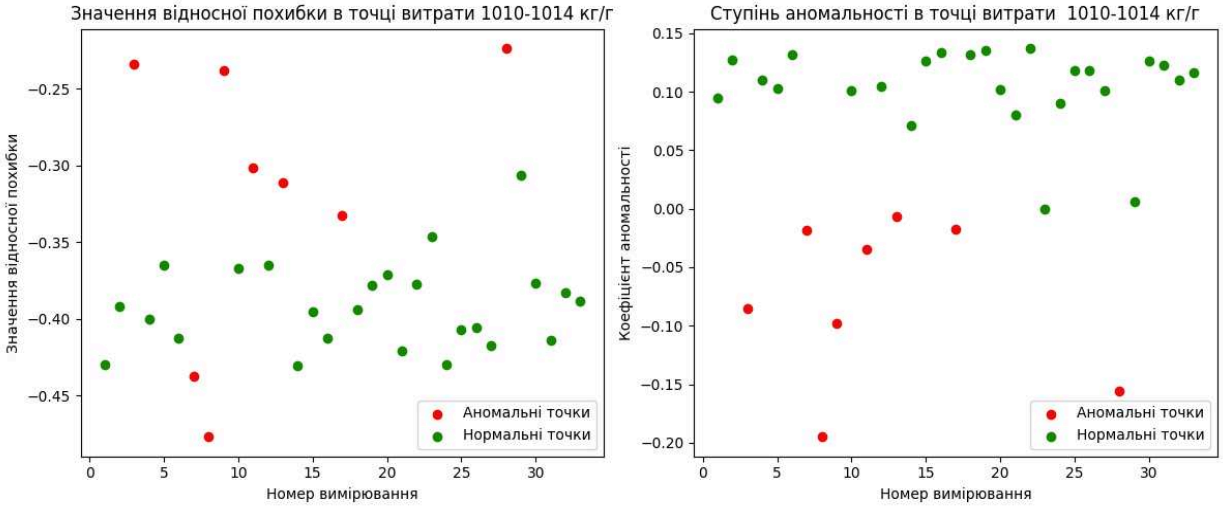

Рисунок 3.14 - Значення ступеня аномальності в точці витрати 1 т/г, витратоміра №2

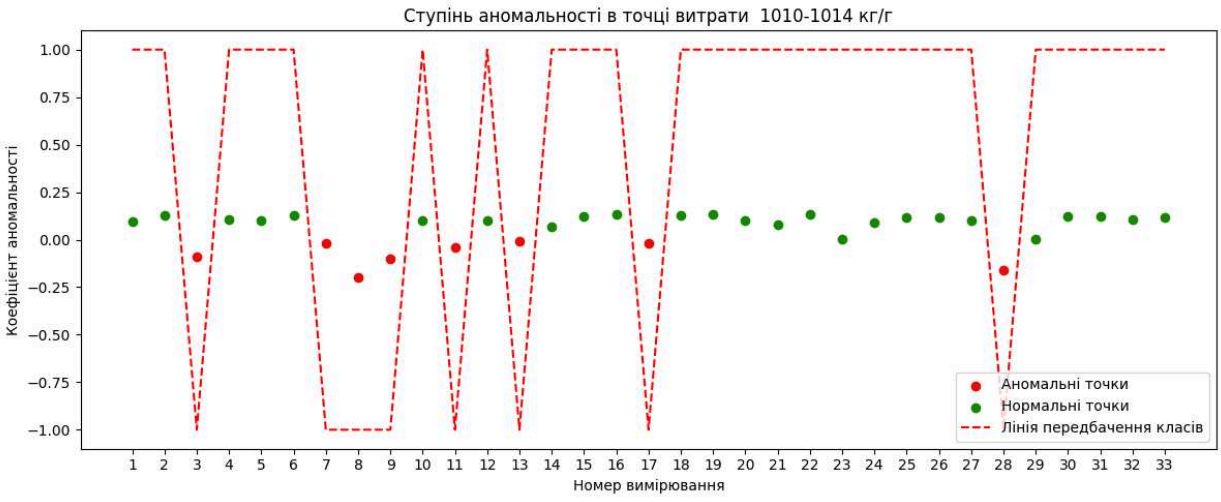

Рисунок 3.15 - Лінія передбачення класів в точці витрати 1 т/г, витратоміра №2

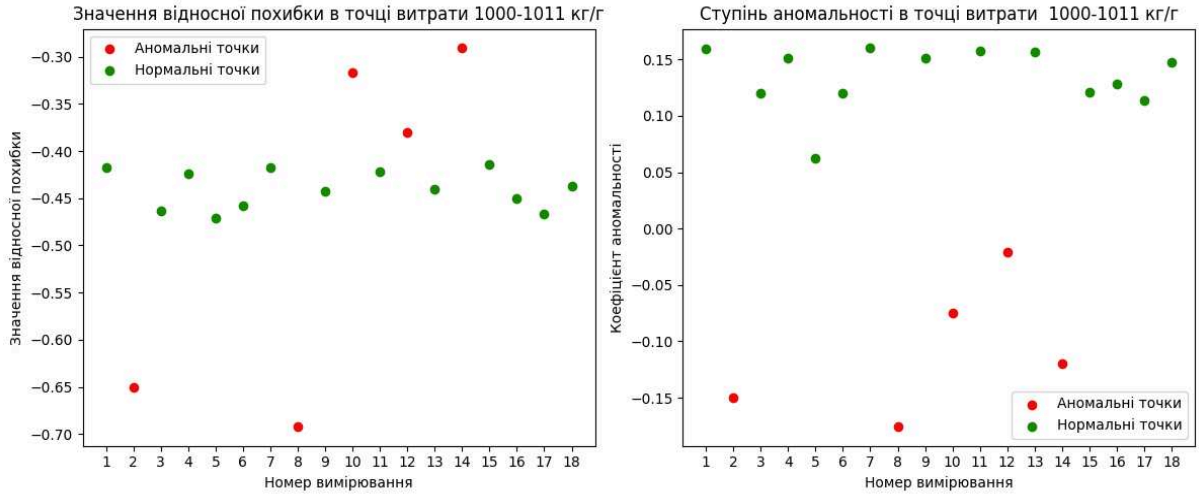

Рисунок 3.16 - Значення ступеня аномальності в точці витрати 1 т/г, витратоміра №3

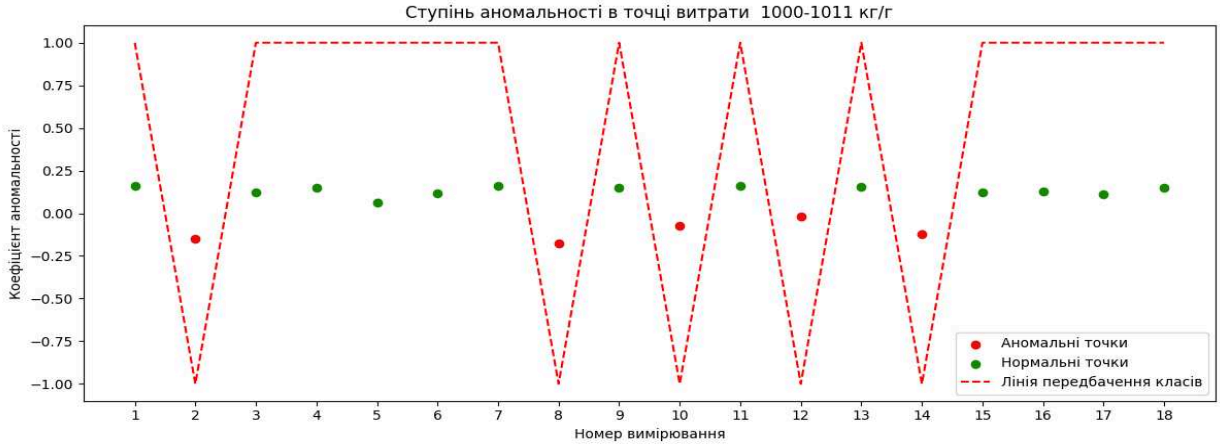

Рисунок 3.17 - Лінія передбачення класів в точці витрати 1 т/г, витратоміра №3

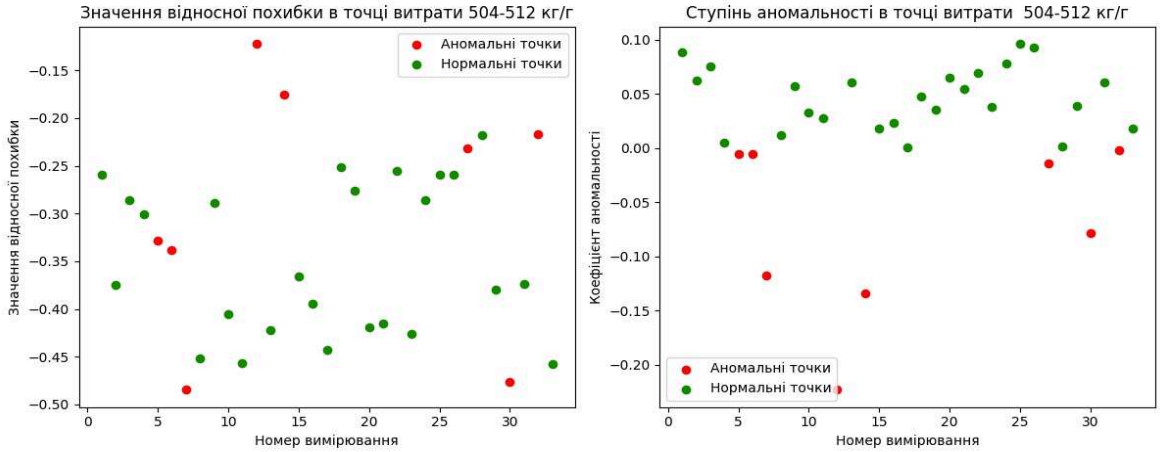

Рисунок 3.18 - Значення ступеня аномальності в точці витрати 0,5 т/г, витратоміра №3

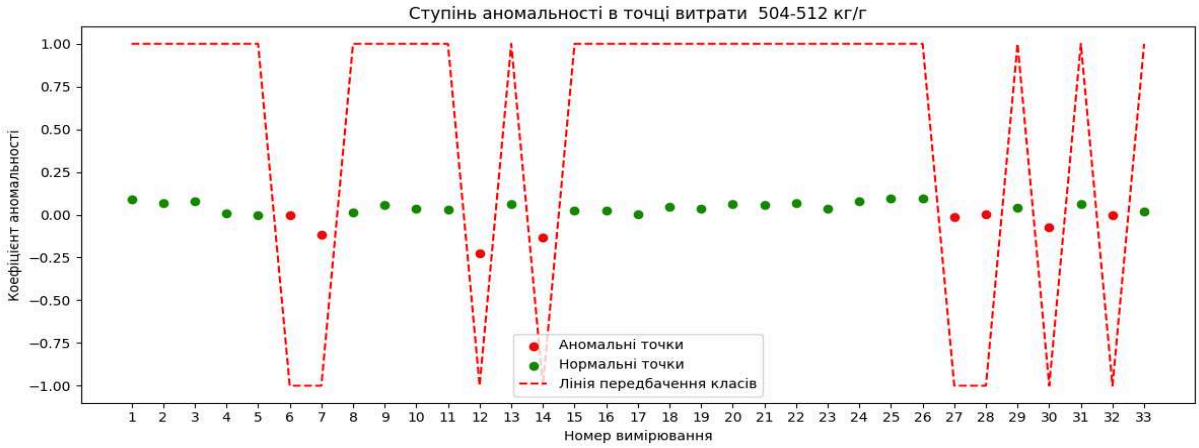

Рисунок 3.19 - Лінія передбачення класів в точці витрати 0,5 т/г, витратоміра №3

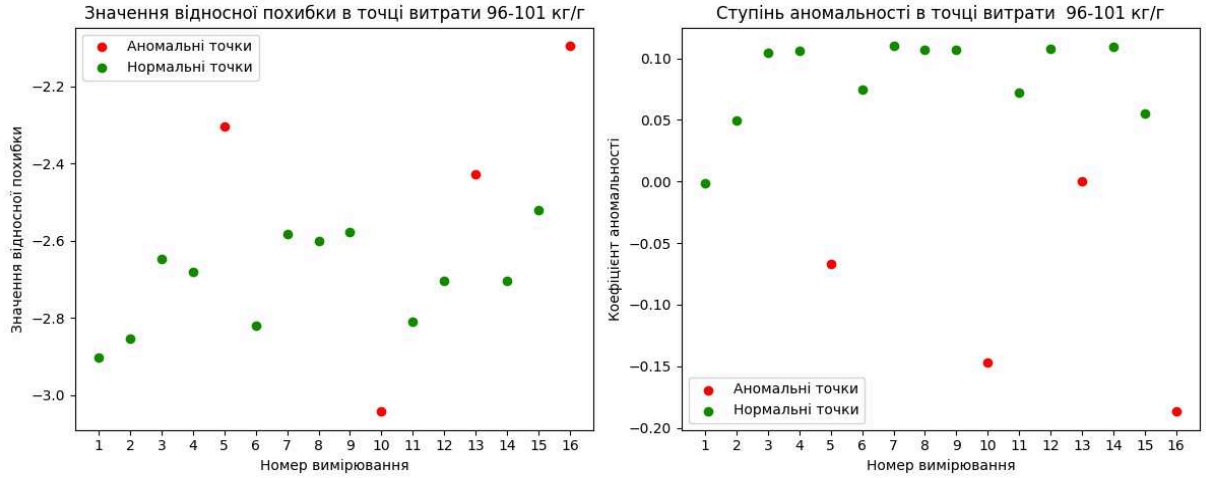

Рисунок 3.20 - Значення ступеня аномальності в точці витрати 0,1 т/г, витратоміра №3

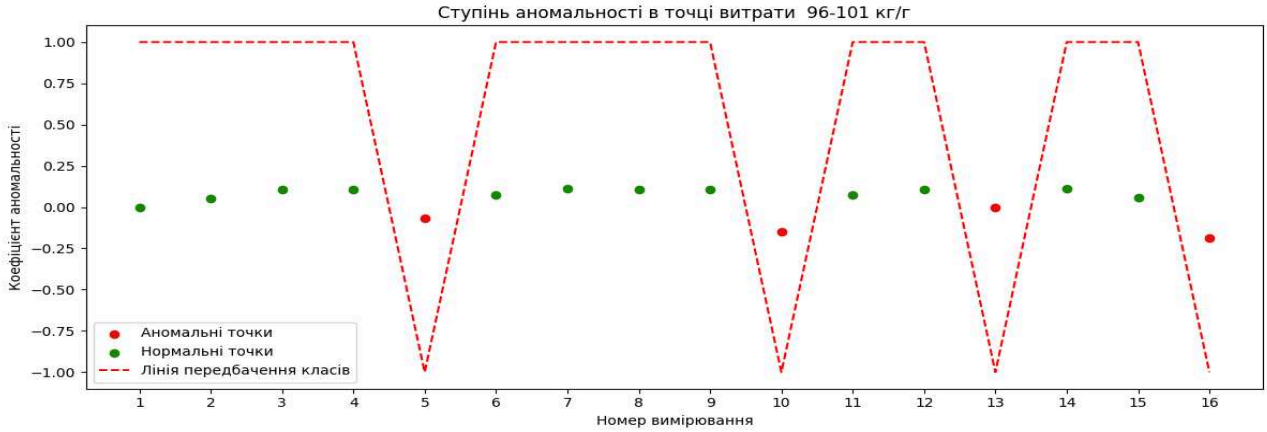

Рисунок 3.21 - Лінія передбачення класів в точці витрати 0,1 т/г, витратоміра №3.

3.5 Оцінка роботоздатності моделі ізольованого лісу

Оцінка роботоздатності моделі ІЛ для виявлення викидів полягає у порівнянні результатів цієї моделі зі статистичними методами виявлення викидів. Вибір методу для виявлення аномалій залежить від кількох факторів, таких як чутливість методу до викидів, тип розподілу даних та розмір вибірки. Одним із ключових завдань є забезпечення того, щоб обраний метод міг ефективно виявляти як великі, так і менш виражені викиди, не спотворюючи загальну картину даних [64].

У випадку малих вибірок, як у нашому дослідженні, проведення тестів для визначення розподілу даних є некоректним. Тому, як зазначалось у 1 розділі, найбільш ефективними виявляються робастні методи, такі як IQR та MAD. Ці методи дозволяють виявляти викиди без залежності від припущень щодо розподілу даних, що робить їх універсальними для різних типів вибірок.

Важливо зазначити, що мета цього розділу полягає не стільки в тому, щоб показати переваги моделі ІЛ над іншими методами, скільки у перевірці її роботоздатності та точності при порівнянні з робастними підходами. Важливо визначити, наскільки модель здатна коректно знаходити аномалії в даних порівняно з такими методами, як IQR і MAD, які не залежать від типу розподілу даних і можуть бути більш ефективними у випадках, коли вибірка мала або нестандартна.

Як зазначалося раніше (у підрозділах 1.2.4.2 — 1.2.4.3), IQR вимірює варіабельність даних на основі різниці між першим (Q_1) і третім (Q_3) квартилями [66]:

$$IQR = Q_3 - Q_1 \tag{3.1}$$

Викиди визначаються як спостереження, що виходять за межі інтервалу $[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$. Якщо значення спостереження виходить за межі цього інтервалу, його вважають викидом.

Формально, спостереження x є викидом, якщо виконується одна з умов:

$$x < Q_1 - 1.5 \times IQR, або \ \ x < Q_3 + 1.5 \times IQR. \tag{3.2}$$

Розрахунок MAD ґрунтується на медіані абсолютних відхилень кожного спостереження від медіани вибірки [67]:

$$MAD = med(|x_i - med(x)|), \tag{3.3}$$

де x_i — це окремі спостереження,

med(x) — медіана вибірки.

Для виявлення викидів використовується кратне MAD, тобто:

$$x \in викидом, якщо |x_i - med(x)| > 3 \times MAD. \tag{3.4}$$

У формулах (3.2) і (3.4) використовуються стандартні загальноприйняті порогові значення для виявлення викидів: для методу IQR — 1.5, а для методу MAD — 3. Ці значення зазвичай застосовуються в класичних статистичних дослідженнях для виявлення аномальних спостережень.

Однак, при аналізі даних вимірювань на ДЕТУ 03-04-04, який має високе значення стабільності витрати, такі стандартні порогові значення можуть бути недостатніми для виявлення викидів. Як було досліджено в розділі 2, ДЕТУ 03-04-04 характеризується високою стабільністю вимірювань, що робить викиди менш вираженими і складнішими для виявлення. Це означає, що стандартні значення порогів можуть бути занадто високими, щоб ефективно знаходити незначні аномалії, які мають слабко виражений характер [68].

Тому, для виявлення викидів у таких даних, порогові значення для методу IQR було знижено до 0.4, а для методу MAD — 1.5.

Результати розрахунку двох статистичних методів представлені на графіках (рис. 3.22 — 3.30), де також наведено результат розрахунку ІЛ для візуального порівняння.

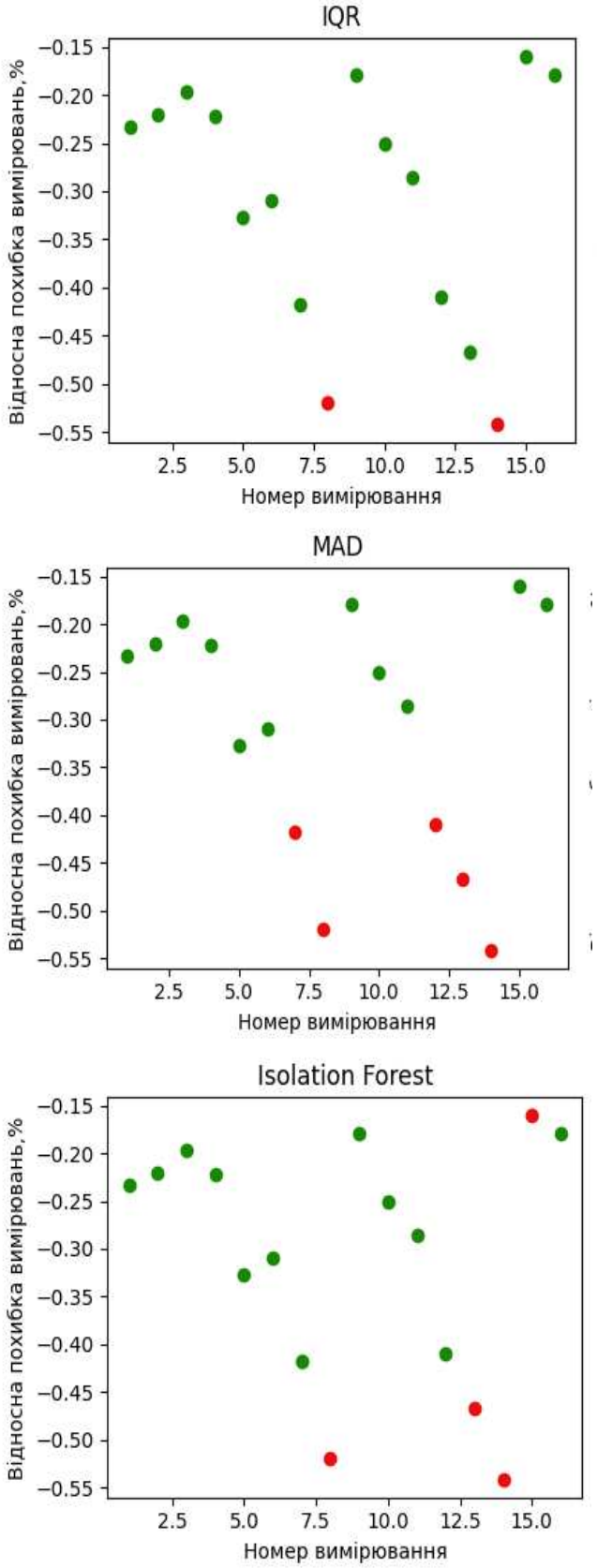

Рисунок 3.22 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 45 т/г витратоміра №1.

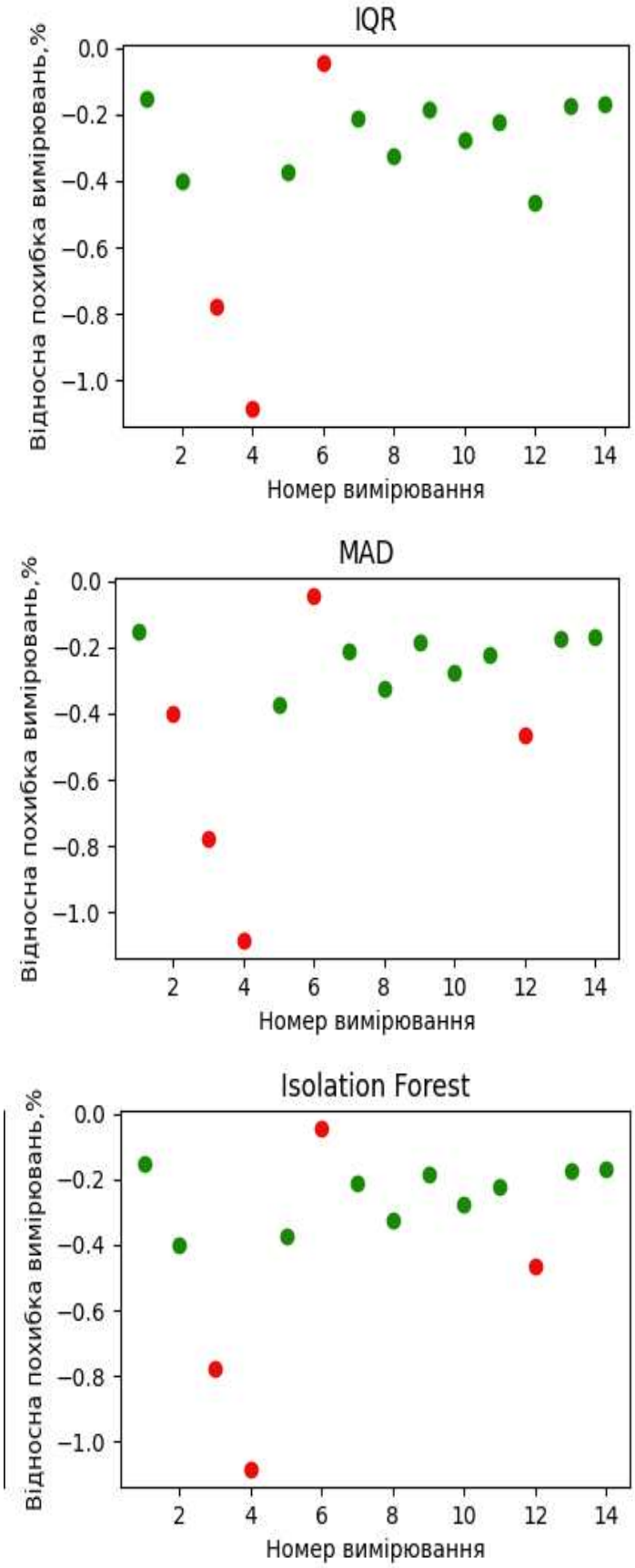

Рисунок 3.23 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 25 т/г витратоміра №1.

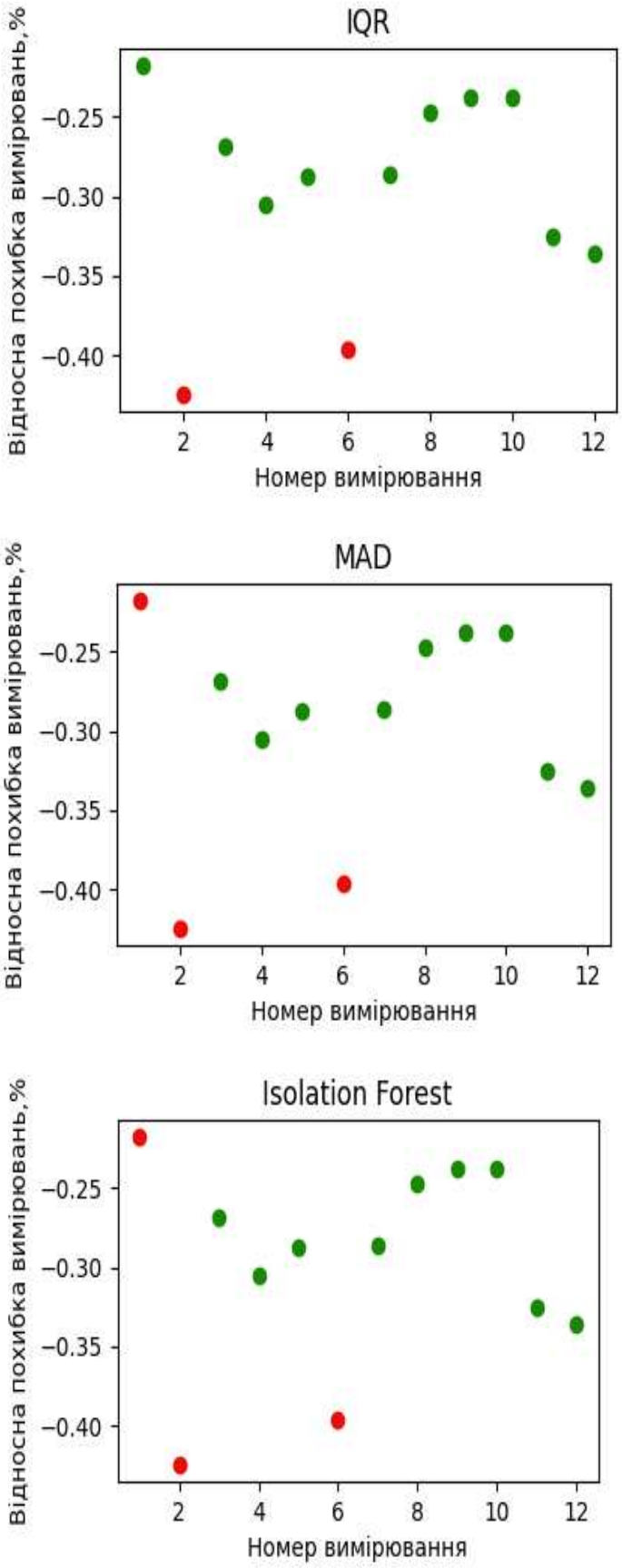

Рисунок 3.24 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 5 т/г витратоміра №1.

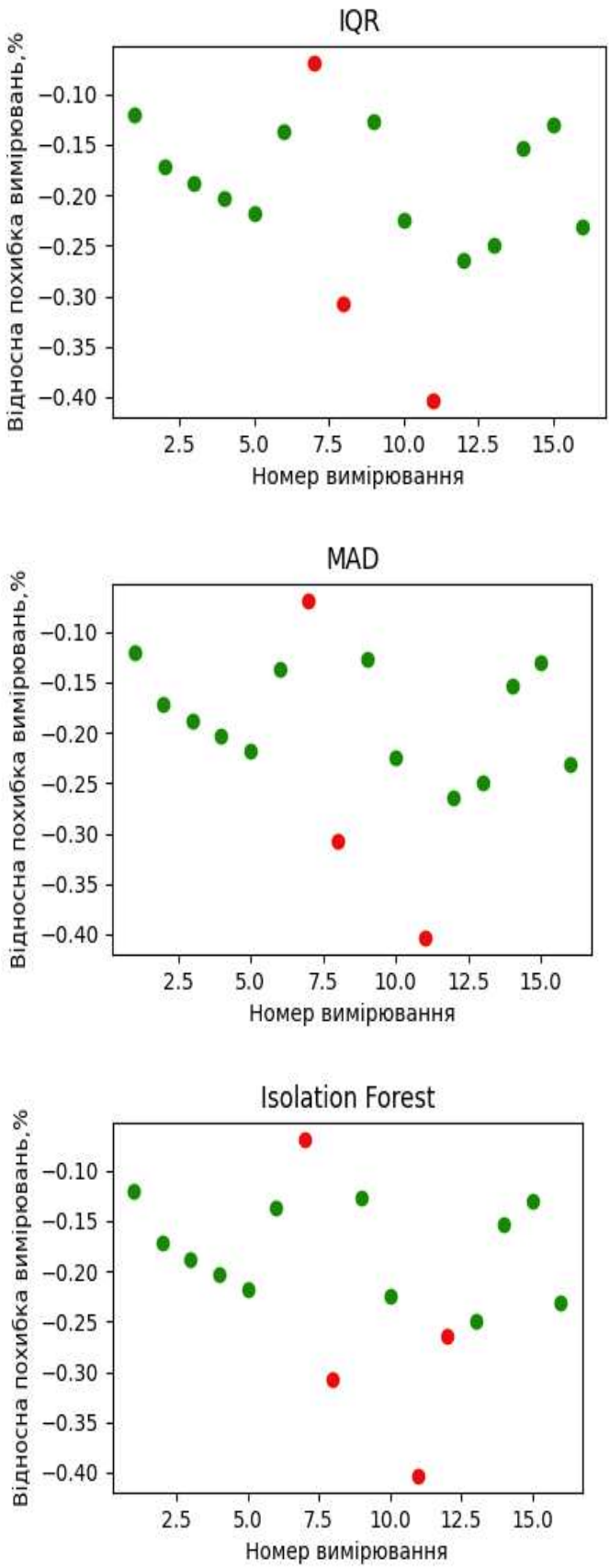

Рисунок 3.25 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 5 т/г витратоміра №2.

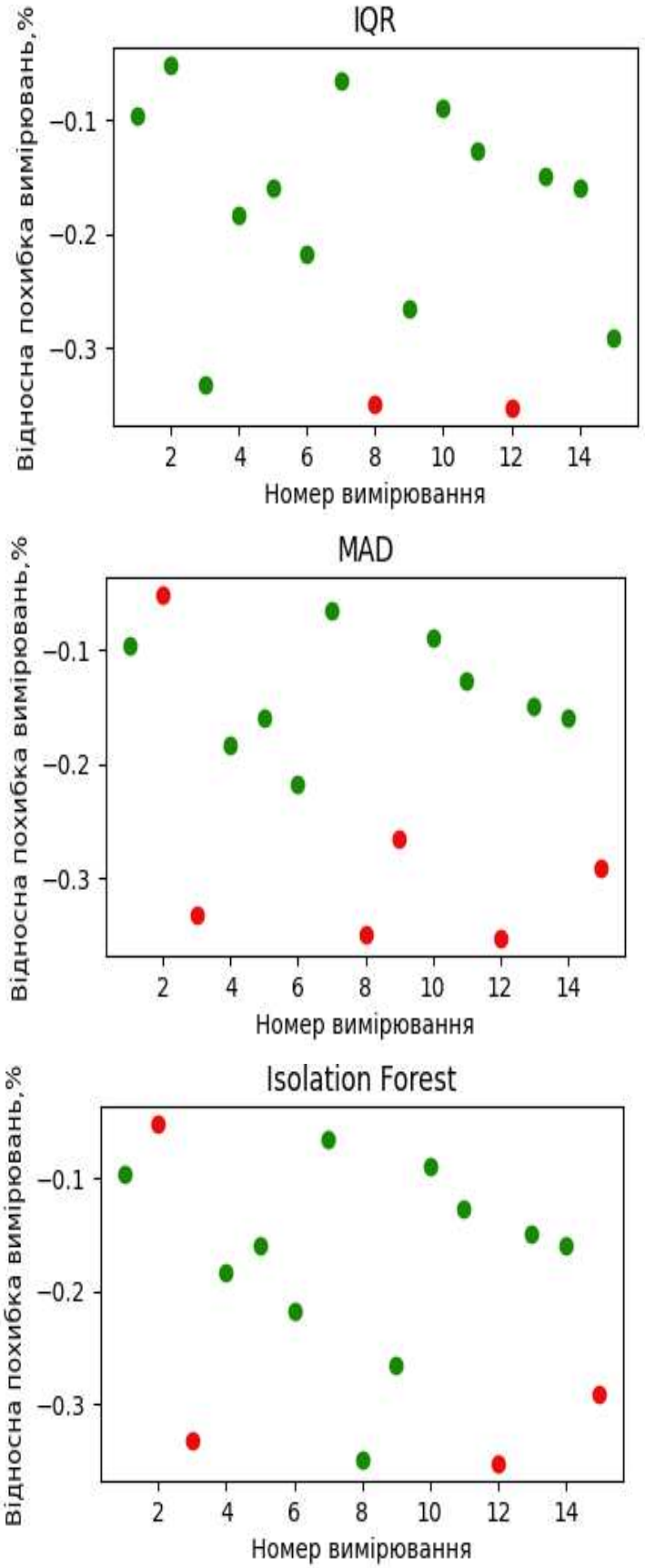

Рисунок 3.26 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 2.5 т/г витратоміра №2.

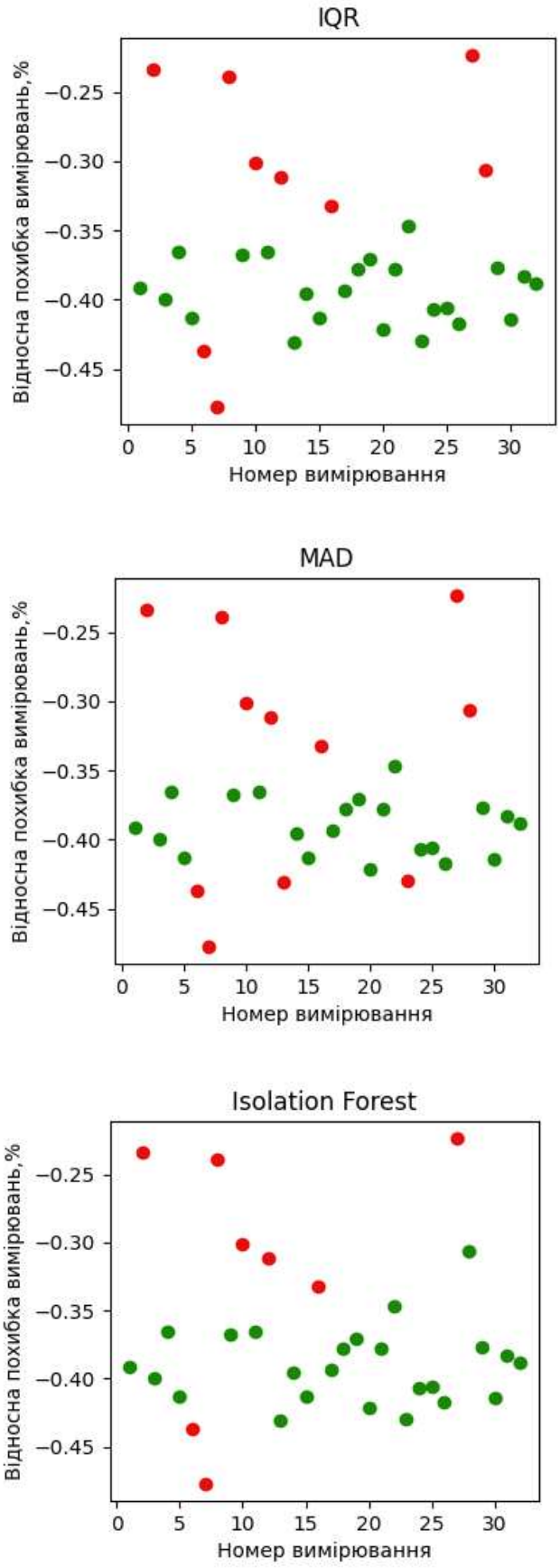

Рисунок 3.27 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 1 т/г витратоміра №2.

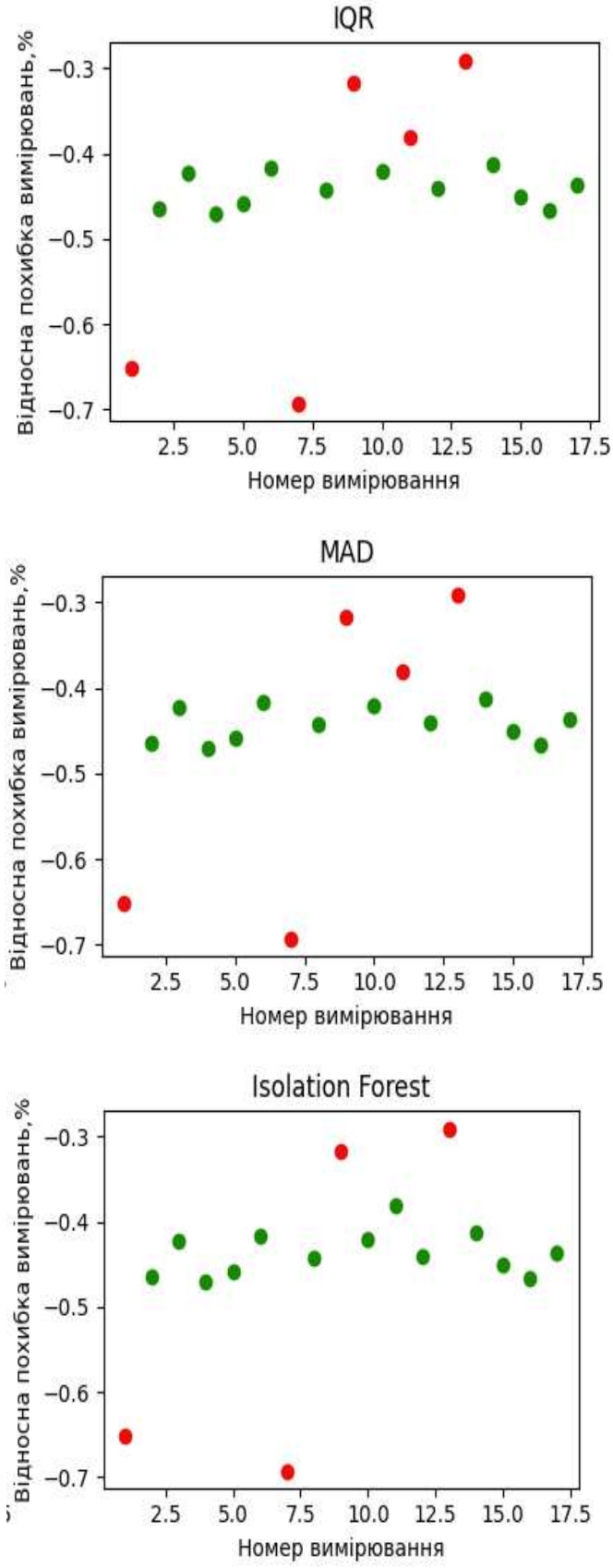

Рисунок 3.28 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 1 т/г витратоміра №3.

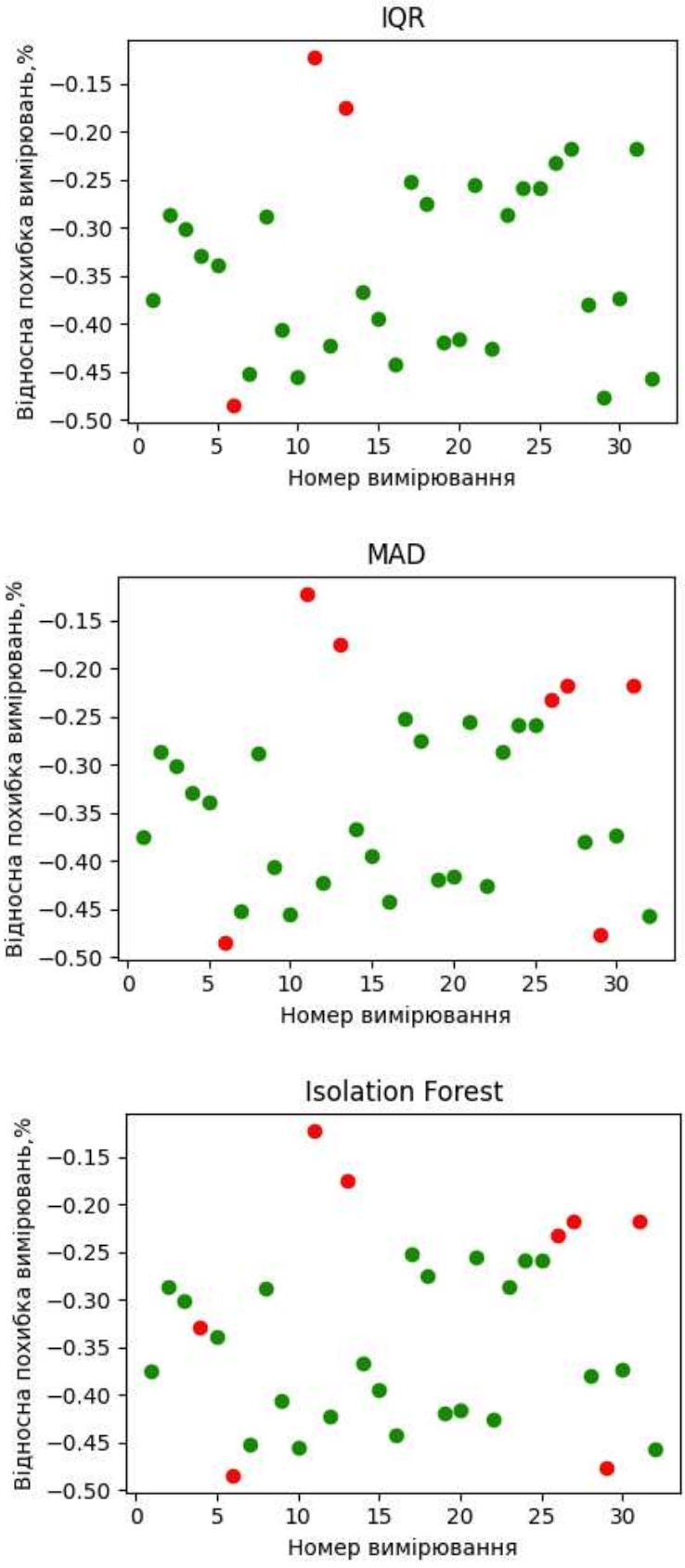

Рисунок 3.29 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 0.5 т/г витратоміра №3.

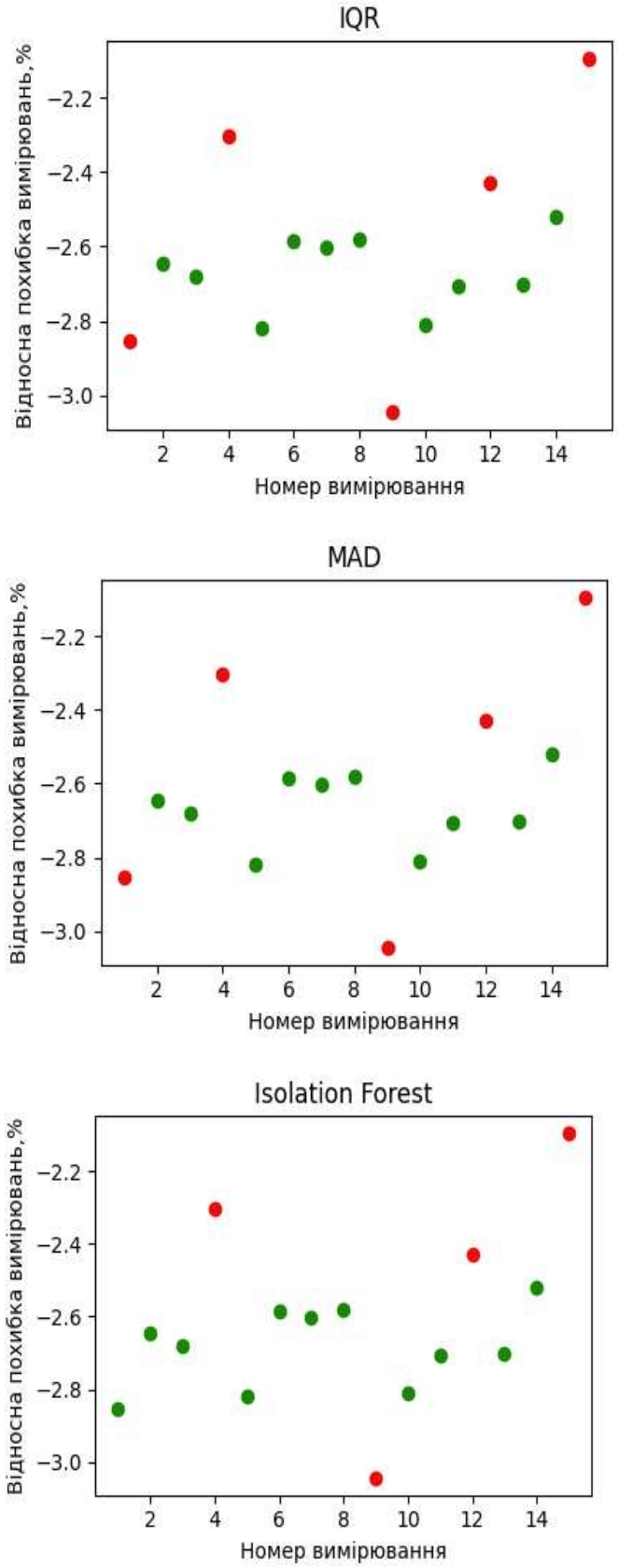

Рисунок 3.30 – Результат знаходження викидів методом IQR, MAD та ІЛ в точці витрати 0.1 т/г витратоміра №3.

За отриманими результатами можна стверджувати, що модель ІЛ успішно виявила викиди подібно до методів IQR та MAD. У вибірках, де викиди суттєво відрізнялися від основної маси даних, результати виявлення викидів усіма трьома методами були ідентичними, що підтверджує ефективність і роботоздатність моделі ІЛ.

Водночас у деяких вибірках спостерігалися незначні розбіжності між результатами різних методів. Ці відмінності можна пояснити різницею у порогових значеннях. Для методів IQR та MAD було обрано порогові значення 0.4 та 1.5 відповідно, що дозволило досягти максимальної схожості з чутливістю моделі ІЛ, налаштованої на виявлення 35% припустимих аномалій. Такий вибір порогових значень виявився оптимальним для забезпечення аналогічної чутливості всіх методів, з урахуванням специфіки даних та обмежень кожного підходу.

Якщо проаналізувати результати методів детальніше, можна побачити, що відмінності складають лише 1-2 точки, які при підвищеній чутливості також могли б бути класифіковані як викиди. Це додатково підтверджує різницю у порогових значеннях між методами.

Результати знаходження викидів методами IQR, MAD та ІЛ також представлені в таблиці 3.1, де зазначено відсоткове співвідношення викидів у відповідній вибірці.

Таблиця 3.1 – Результат знаходження викидів методом IQR, MAD та ІЛ.

Точка масової витрати	Розмір вибірки	Кількість викидів (метод IQR)	Кількість викидів (метод MAD)	Кількість викидів (метод ІЛ)
45 т/г, (витр. №1)	17	2 (11.76%)	5 (29.41%)	4 (23.52%)
25 т/г, (витр. №1)	15	3 (20%)	5 (33,33%)	4 (26.66%)
5 т/г , (витр. №1)	18	2 (11.11%)	3 (16.66%)	3 (16.66%)
5 т/г , (витр. №2)	17	3 (17.64%)	3 (17.64%)	4 (23.52%)
2.5 т/г, (витр. №2)	16	2 (12.5%)	6 (37,5%)	4 (25%)
1 т/г , (витр. №2)	33	9 (27.27%)	11 (33.33%)	8 (24,24%)
1 т/г , (витр. №3)	17	5 (29.41%)	5 (29.41%)	4 (23.52%)
0.5 т/г , (витр. №3)	33	3 (9.09%)	7 (21.21%)	8 (24.24%)
0.1 т/г , (витр. №3)	16	5(31,25%)	5(31,25%)	4 (25%)

3.6 Ступінь аномальності

Ступінь аномальності – це показник, що кількісно визначає відхилення кожної конкретної точки даних від "нормальних" значень у вибірці (розраховується за формулою 3.3). Він використовується для оцінки нетиповості або аномальності точки у контексті загальної структури даних.

Цей показник є основою для аналізу результатів роботи моделі. На відміну від традиційних методів, які бінарно класифікують дані як нормальні або аномальні, ІЛ використовує безперервну шкалу для оцінки інтенсивності аномалій. Такий підхід дає змогу гнучко налаштовувати модель під різні завдання та вимоги до точності виявлення викидів. Аналіз результатів моделі показав, що при використанні різних порогових значень можна регулювати чутливість до викидів: вищі пороги дозволяють виявляти тільки значні викиди, тоді як нижчі пороги збільшують чутливість до дрібніших відхилень.

Одним із ключових аспектів цього підходу є перехід від традиційних одиниць вимірювання до узагальненого показника ступеня аномальності, що дозволяє порівнювати результати між собою. Оскільки всі значення переводяться в одні й ті самі одиниці, це дає змогу об'єднувати вимірювання, навіть якщо вони були проведені на різних типах витратомірів або за різних умов. Така стандартизація спрощує аналіз, оскільки всі дані перетворюються на єдину шкалу, і ми можемо оцінювати ступінь аномальності незалежно від вихідних параметрів вимірювань.

Цей підхід дозволяє об'єднати результати моделі для вибірок, отриманих в рамках одного витратоміра (рис.3.31 — 3.33), що забезпечує порівнянність усіх вимірювань незалежно від їх початкових одиниць. Також це дає змогу об'єднувати результати різних витратомірів одного типу (рис.3.35), що дозволяє аналізувати викиди в межах одного типу ЗВТ, включаючи як випадкові, так і систематичні помилки.

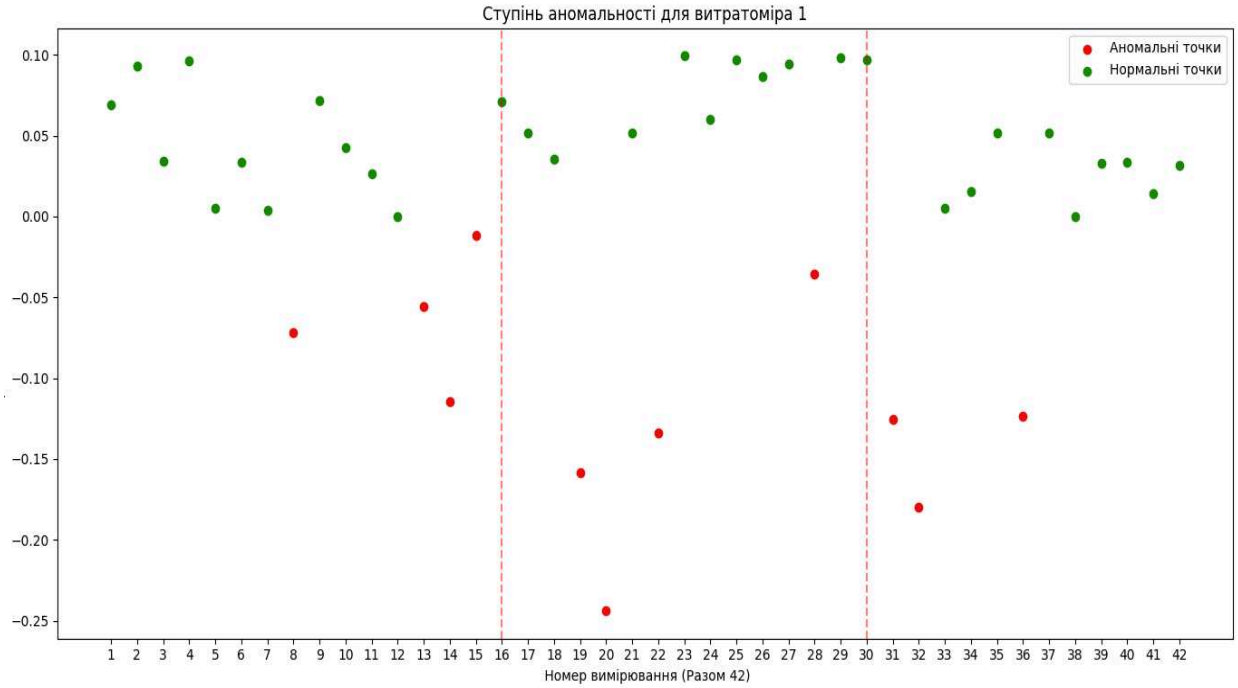

Рисунок 3.31 - Ступінь аномальності для витратоміра №1.

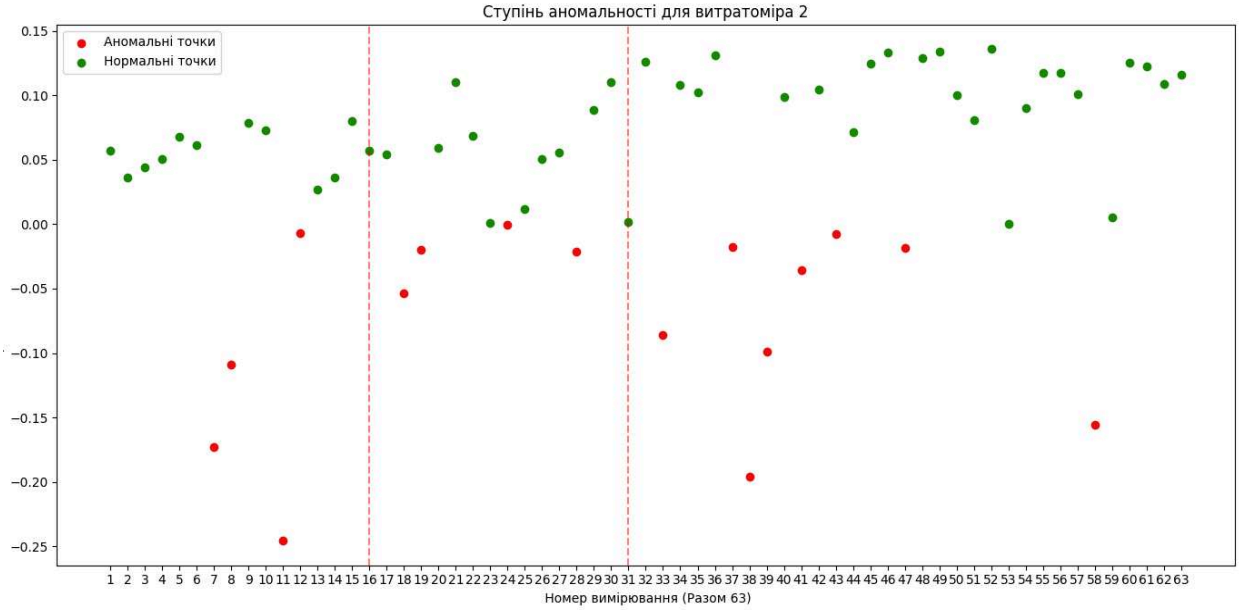

Рисунок 3.32 - Ступінь аномальності для витратоміра №2.

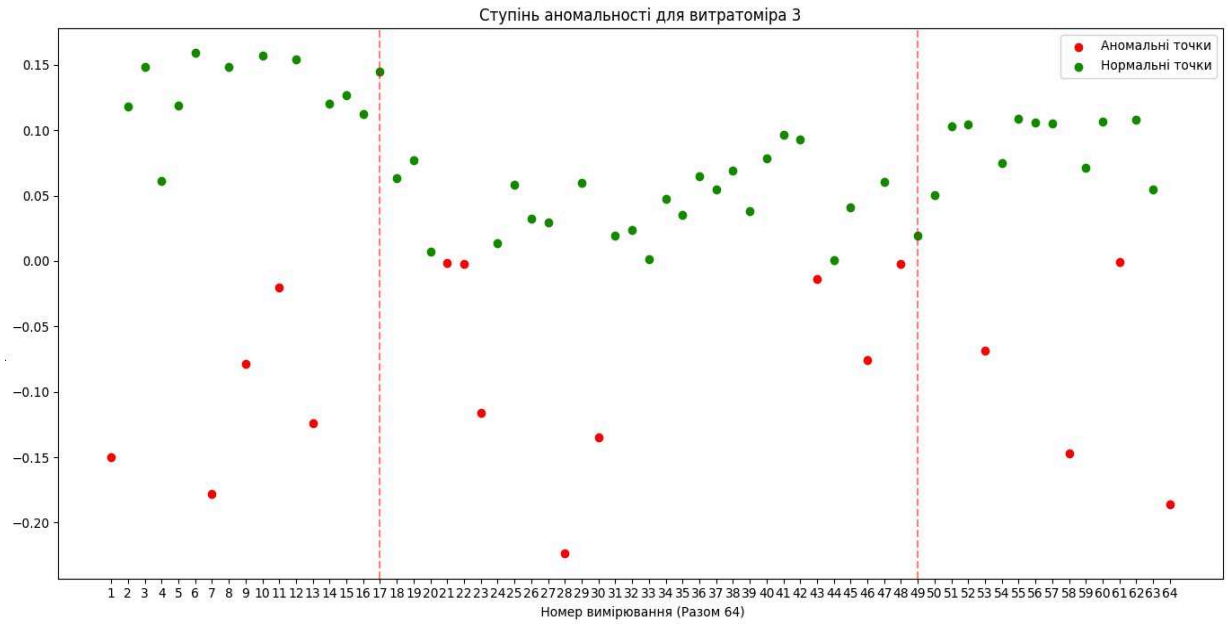

Рисунок 3.33 - Ступінь аномальності для витратоміра №3.

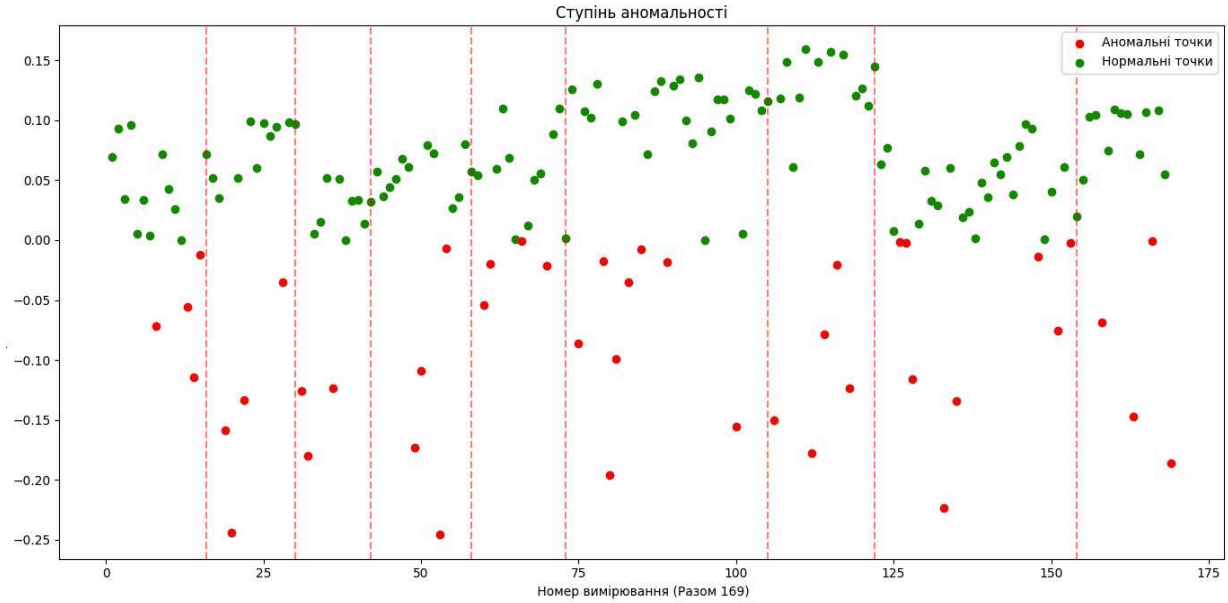

Рисунок 3.34 - Ступінь аномальності коріолісових витратомірів.

Така інтеграція результатів вимірювань на основі ступеня аномальності є важливою для аналізу надмірних похибок та їх причин, особливо у випадках, коли вимірювання проводяться на різних стандартних приладах або в різних умовах. Вона дозволяє уникнути роздроблення даних і дає можливість будувати більш комплексні моделі похибок і неточностей, для аналізу випадкових, систематичних та надмірних похибок.

3.7 Порівняльний аналіз методу ізольованого лісу з робастиними статистичними методами.

Для розуміння можливостей застосування кожного підходу важливо проаналізувати їхні переваги та недоліки. Такий аналіз дозволяє оцінити, наскільки ефективно ІЛ та робастні статистичні методи можуть вирішувати різні задачі виявлення викидів. Огляд сильних та слабких сторін кожного методу допомагає визначити оптимальні умови для їхнього використання, а також потенціал комбінованого підходу для досягнення більш точних та надійних результатів у різноманітних контекстах і структурах даних.

Переваги методу ІЛ:

1. **Адаптивність до різних розподілів даних.**

ІЛ не вимагає попередніх припущень щодо розподілу даних, що є суттєвою перевагою у випадках, коли структура даних є складною або не відповідає жодному з відомих розподілів (наприклад, нормальному, рівномірному тощо). Така адаптивність дозволяє алгоритму ефективно працювати з вибірками, які можуть містити зміщення, асиметрію або значний розкид значень. Це робить ІЛ універсальним інструментом для виявлення аномалій у реальних умовах, де дані часто мають нетиповий розподіл і можуть містити структурні зміни, що не можуть бути коректно оброблені стандартними статистичними методами. Відсутність обмежень на форму розподілу також робить ІЛ більш стійким до різноманітних типів даних, що забезпечує його надійність у широкому спектрі застосувань;

2. **Можливість обробки багатовимірних даних.**

Метод ІЛ відрізняється своєю здатністю обробляти багатовимірні дані, враховуючи взаємозв'язки між різними змінними одночасно. Це особливо важливо для задач, де прості статистичні методи обмежені одномірним аналізом і не можуть ефективно враховувати залежності між параметрами. ІЛ створює множинні ізоляційні дерева, які дозволяють враховувати всі параметри набору даних, виділяючи аномальні точки на основі їхнього положення в

багатовимірному просторі. Це забезпечує ІЛ перевагу у випадках, коли дані мають складну структуру і необхідно враховувати взаємодію між численними параметрами для коректного виявлення аномалій;

3. Стійкість до шуму.

ІЛ має вбудовану стійкість до шуму завдяки своїй основній концепції — ізоляції точок, що відрізняються від основної маси даних. На відміну від методів, які фокусуються на середніх значеннях або інших центральних тенденціях, ІЛ ізолює точки на основі їхньої віддаленості від більшості, що дозволяє мінімізувати вплив випадкових відхилень чи шуму. Це робить метод ефективним для обробки даних, які можуть містити коливання, що не є аномаліями, але можуть заважати іншим підходам коректно виявляти справжні аномальні точки. Таким чином, ІЛ забезпечує високу точність при виявленні аномалій навіть у випадках, коли дані містять значні випадкові коливання;

4. Гнучке налаштування параметра чутливості.

ІЛ дозволяє користувачеві налаштовувати параметр, який визначає очікувану кількість аномальних точок у вибірці. Це забезпечує можливість контролювати чутливість до аномалій, дозволяючи адаптувати метод під конкретні потреби та умови задачі. Наприклад, у ситуаціях, коли потрібно виявити лише значні аномалії, параметр можна зменшити, тоді як для виявлення слабко виражених відхилень його можна збільшити. Така гнучкість робить ІЛ привабливим інструментом для аналізу різних типів даних, оскільки він може бути налаштований на різні рівні чутливості залежно від специфіки конкретного завдання;

5. Не потребує попередньої обробки даних або виділення тренувального набору.

ІЛ може виявляти аномалії без попереднього розділення даних на тренувальні та тестові набори. Це значно спрощує процес аналізу, оскільки зменшує вимоги до підготовки даних. Відсутність потреби у виділенні окремого набору даних для навчання дозволяє економити ресурси і час, що є важливим для задач, де дані обмежені або доступ до них обмежений. Така властивість також забезпечує більшу гнучкість у роботі з даними, особливо в умовах, коли

необхідно проводити швидкий аналіз без додаткових кроків підготовки. Це робить ІЛ привабливим інструментом для швидких, інтуїтивно зрозумілих аналізів у випадках, коли обробка даних є ключовою ланкою;

6. Автоматичне виділення аномалій.

ІЛ автоматично ізолює аномальні точки в процесі побудови ізоляційних дерев. Алгоритм побудови дерева самостійно виділяє точки, які розташовані на "периферії" вибірки, як аномалії, не вимагаючи складних розрахунків або обчислення статистичних меж. Це дає можливість швидко ідентифікувати аномальні точки без необхідності додаткових обчислень чи порівняння з будь-якими іншими статистичними характеристиками даних. Такий автоматичний підхід робить метод ІЛ зручним і надійним для виявлення аномалій у різних наборах даних, особливо коли точний опис аномалії не визначений або коли аномалії є неявними.

Недоліки методу ІЛ:

1. **Складність інтерпретації результатів.**

Результати, отримані за допомогою ІЛ, можуть бути складними для інтерпретації через специфічний принцип роботи цього алгоритму. На відміну від традиційних статистичних методів, де аномалії можуть бути пояснені через відстань від медіани, квартилів чи інших центральних показників, ІЛ класифікує точки як аномальні на основі ізоляції у деревоподібній структурі. Це означає, що користувачі, особливо ті, хто не має досвіду роботи з алгоритмами ML, можуть не розуміти, чому певна точка визначена як аномальна. Такі результати часто вимагають додаткового аналізу або візуалізації для пояснення, що може ускладнити прийняття рішень на їх основі, особливо в критичних або регламентованих областях, де потрібна прозорість і обґрунтованість кожного висновку;

2. **Залежність від параметра "кількість припущених аномалій".**

ІЛ залежить від попереднього налаштування параметра, який визначає частку припущених аномалій у вибірці. Хоча цей параметр надає користувачеві певну гнучкість, його значення повинно бути точно підібрано для кожної

конкретної задачі, щоб уникнути помилкових виявлень. Наприклад, завищене значення може призвести до того, що алгоритм розпізнає як аномальні точки, які насправді є нормальними для конкретного набору даних. З іншого боку, занижене значення може спричинити пропуск важливих аномалій. Через це виникає потреба у трудомісткому процесі калібрування параметра для кожного конкретного набору даних, що не завжди є зручним і може вимагати попереднього дослідження структури даних;

3. **Складність налаштування для різнорідних даних.**

ІЛ може демонструвати обмежену ефективність при роботі з даними, які мають складну або змішану структуру. Якщо в наборі є кілька типів аномалій, наприклад, короткочасні "сплески" значень і довготривалі відхилення, ІЛ може некоректно ізолювати деякі типи аномалій або вимагати складного налаштування для їх розпізнавання. Також алгоритм може сприймати змішані дані як нормально розподілені, що призведе до неправильного розмежування нормальних і аномальних точок. Така складність налаштування особливо актуальна для задач, де дані мають неоднорідну структуру і можуть включати різні види аномалій, для яких алгоритм потребує специфічного калібрування;

4. **Можливість пропуску аномалій при високому рівні шуму.**

Незважаючи на вбудовану стійкість до шуму, ІЛ може пропускати деякі аномалії, якщо рівень шуму є занадто високим або якщо шум змінює структуру даних до такої міри, що алгоритм не може чітко відокремити аномальні точки. ІЛ не має вбудованих механізмів для чіткого розрізнення між шумом і справжніми аномаліями. У випадках, коли дані містять надмірний шум, частина аномалій може "загубитися" серед випадкових коливань, що знижує точність моделі. Це обмеження може стати критичним для задач, де важливо виявляти всі відхилення навіть за умов високого рівня шуму;

5. **Висока обчислювальна складність для малих наборів даних.**

ІЛ є обчислювально інтенсивним алгоритмом і може бути надмірним для невеликих вибірок, де його використання може виявитися невиправданим у порівнянні з більш простими робастними методами, такими як IQR або MAD.

Це обмеження може мати значення у випадках, коли доступні обмежені обчислювальні ресурси або коли точність не є критично важливою, а пріоритет надається економії ресурсів. Для невеликих наборів даних простіші методи можуть забезпечити аналогічну точність при значно меншому обсязі обчислень;

6. **Не підходить для динамічних даних у режимі реального часу без перекалібрування.**

ІЛ не призначений для обробки динамічних або потокових даних, де розподіл даних може змінюватися з часом. Алгоритм будує модель на основі "статичного" набору даних і не оновлюється автоматично, якщо структура даних змінюється. У разі застосування ІЛ для аналізу потокових даних необхідно проводити періодичну перекалібровку моделі, щоб забезпечити її актуальність. Без регулярного оновлення модель може давати невірні результати через зміни у вибірці або структури даних, що ускладнює її застосування в умовах, де дані змінюються в реальному часі.

Переваги робастних статистичних методів:

1. Стійкість до викидів.

Робастні статистичні методи, такі як IQR та MAD, характеризуються високою стійкістю до впливу викидів, що є однією з їхніх головних переваг. Ці методи зосереджуються на основних тенденціях, таких як медіана або квартилі, що знижує вплив аномальних значень на результати аналізу. У випадках, коли дані містять кілька окремих викидів, ці методи дозволяють уникнути спотворення основної тенденції, яку часто можуть викривити традиційні статистичні методи. Стійкість до викидів забезпечує надійність і точність аналізу, зберігаючи загальну картину навіть за наявності поодиноких аномальних точок;

2. Простота інтерпретації.

Робастні методи є зрозумілими і легко інтерпретуються, що робить їх ідеальними для прикладного використання. Вони дозволяють швидко визначити аномальні точки на основі чітко визначених відстаней від центральних значень, таких як медіана або межі квартилів. Наприклад, точки, що виходять за межі 1,5-кратного розмаху IQR, автоматично вважаються потенційними вики-

дами. Це спрощує процес аналізу і підвищує прозорість результатів, що є важливим для користувачів, які можуть не мати глибоких знань у галузі статистики або ML. Простота інтерпретації робить ці методи доступними для швидкого прийняття рішень;

3. Низька обчислювальна складність.

Робастні статистичні методи є обчислювально легкими та не потребують значних ресурсів, що робить їх оптимальними для роботи з невеликими вибірками та обмеженими обчислювальними потужностями. У порівнянні з більш складними алгоритмами, такими як методи ML, ці методи є швидшими, що особливо важливо у ситуаціях, коли результати потрібні терміново. Ця швидкість забезпечує ефективне використання обчислювальних ресурсів і дозволяє застосовувати робастні методи на малих вибірках або для задач, де обчислювальна ефективність є критичною;

4. Незалежність від розподілу даних.

Робастні методи, такі як IQR та MAD, не залежать від форми розподілу даних, що робить їх універсальними інструментами для аналізу вибірок з невизначеним або нетиповим розподілом. Ця властивість дозволяє ефективно застосовувати ці методи у випадках, коли дані можуть не відповідати нормальному або іншому теоретичному розподілу. Незалежність від розподілу означає, що ці методи можуть надавати надійні результати навіть у ситуаціях, коли структура даних є нестандартною або динамічною;

5. Простота застосування.

Робастні статистичні методи є простими у застосуванні, оскільки вони не потребують налаштування складних параметрів або попереднього підлаштування. Вони можуть бути впроваджені без значної підготовки даних, що робить їх ідеальними для використання в умовах, де немає можливості або потреби в детальному налаштуванні. Ця простота робить робастні методи доступними для широкого кола користувачів, незалежно від рівня їхньої статистичної підготовки, і сприяє їх популярності серед прикладних фахівців;

6. Підходять для малих вибірок.

Робастні методи ефективно працюють на малих вибірках, що робить їх особливо цінними в умовах обмеженості даних. На відміну від методів ML, які часто вимагають великої кількості даних для належної роботи, методи, як-от IQR та MAD, можуть надійно функціонувати навіть з невеликими обсягами даних. Це робить їх корисними для задач, де обмежена кількість спостережень або де збір даних є складним і дорогим процесом, наприклад, у медичних дослідженнях або при аналізі унікальних природних явищ;

7. Ефективність для одномірних даних.

Робастні статистичні методи особливо добре підходять для аналізу одномірних даних, оскільки дозволяють легко визначити відхилення від центральних показників. Це є їхньою основною перевагою у випадках, коли дані мають одномірну структуру, і немає необхідності аналізувати складні взаємозв'язки між змінними. Завдяки простоті і точності виявлення аномалій у одномірних даних, робастні методи є зручним інструментом для базового аналізу та швидкої діагностики виявлених відхилень;

8. Мінімальний вплив шуму на результати.

Робастні методи здатні ігнорувати незначні коливання або випадковий шум у даних, фокусуючись на основній тенденції. Це дозволяє уникнути хибних спрацьовувань та зберегти стабільність результатів навіть за наявності незначного шуму. Оскільки робастні методи не є чутливими до випадкових змін, вони забезпечують точніші результати для даних з мінімальним рівнем шуму, що є важливим для задач, де стабільність аналізу має критичне значення, наприклад, у фінансових дослідженнях або у виробничому контролі якості.

На основі проведеного аналізу можна зробити висновок, що метод ІЛ та робастні статистичні методи мають різні характеристики і підходять для вирішення специфічних задач виявлення аномалій. Кожен з цих підходів має свої сильні та слабкі сторони, що дозволяє ефективно їх комбінувати залежно від умов застосування. Метод ІЛ не вимагає припущень щодо розподілу даних, що робить його універсальним для аналізу вибірок із різноманітними розподілами та типами змінних [69].

Робастні статистичні методи є ефективним вибором для швидкого аналізу невеликих або одномірних вибірок, де важливо забезпечити простоту та швидкість обчислень. Вони характеризуються високою стійкістю до окремих аномалій і легко інтерпретуються, що є перевагою для задач, де важлива прозорість та інтуїтивність результатів. Однак, ці методи мають обмежену ефективність у багатовимірних просторах і можуть не виявити слабкі аномалії або специфічні типи відхилень у складних наборах даних.

У певних випадках комбіноване використання обох методів може забезпечити більш повне виявлення аномалій: метод ІЛ може бути застосований для первинного фільтрування та аналізу складних вибірок, тоді як робастні методи — для підтвердження та деталізації результатів.

3.8 Висновок до третього розділу

У третьому розділі дисертаційної роботи було обґрунтовано вибір моделі ІЛ серед методів ML без учителя для виявлення аномалій у метрологічних даних. Було налаштовано ключові параметри моделі, такі як «кількість дерев», «кількість припущених аномалій» та «критерій зупинки», які суттєво впливають на здатність алгоритму точно ідентифікувати аномалії та відхилення.

Модель була протестована на експериментальних даних із дослідження розширеної невизначеності вимірювань коріолісових витратомірів. У результаті було ідентифіковано певні точки вибірки як викиди. Для перевірки надійності результатів проведено порівняння з робастними статистичними методами (IQR і MAD). Аналіз показав, що кількість знайдених викидів змінюється залежно від порогових значень і налаштувань моделі. При порогових значеннях 0,4 і 1,5 та чутливості моделі ІЛ, налаштованої на виявлення 35% припустимих аномалій, результати обох підходів були схожими, з різницею у 1-3 викиди.

Ця відмінність підкреслює, що результати аналізу залежать від вибору порогових значень як у робастних методах, так і в моделі ІЛ. Робастні статис-

тичні методи вимагають універсального порогового значення, обґрунтування якого у випадку різнорідних або малих вибірок є складним завданням. Натомість числова оцінка ступеня аномальності, яка є особливістю ІЛ, дозволяє адаптувати порогові значення під специфіку конкретних вибірок, забезпечуючи більшу гнучкість.

Таким чином, у третьому розділі використано модель ІЛ для знаходження викидів, надійність результатів якої було підтверджено порівнянням із результатами робастних статистичних методів. Поряд із цим було виявлено проблему вибору порогового значення, яка є критичною для точності аналізу в обох підходах. Однак завдяки можливості адаптації моделі ІЛ її застосування є доцільним для аналізу метрологічних даних.

4. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ПАРАМЕТРІВ МОДЕЛІ ІЗОЛЬОВАНОГО ЛІСУ

4.1 Аналіз вдосконалення моделі ізольваного лісу

Модель ІЛ, яка входить до бібліотеки scikit-learn, є базовою версією алгоритму для виявлення аномалій. Її принцип роботи детально було описано у попередніх розділах. Однак, існують й удосконалені версії цієї моделі, які надають додаткові можливості та підходи для більш ефективного виявлення аномалій у даних.

Ось декілька вдосконалених версій моделі ІЛ:

1. Розширений ІЛ (Extended Isolation Forest - EIF).

Це розширення традиційної моделі ІЛ, що дозволяє робити розбиття даних під довільними кутами, замість виключно вздовж осей ознак. Це робить Extended Isolation Forest більш гнучким та часто ефективнішим у виявленні аномалій у даних із складними структурами або високим рівнем кореляції між ознаками. Це досягається шляхом використання більш складних математичних операторів для розбиття простору даних, що дозволяє Extended Isolation Forest ефективніше розділяти і виділяти аномальні дані. Такий підхід збільшує шанси знайти аномалії в наборах даних, де існує складна взаємодія або залежність між різними ознаками, наприклад, у випадках високої багатовимірної кореляції [79];

2. SCiForest (Isolation Forest з використанням підвибірок):

У цій версії для побудови дерев використовуються підвибірки даних, що може допомогти у більш точному налаштуванні моделі під конкретні завдання. У випадках наявності шуму чи невеликих аномалій у великих наборах даних, SCiForest може краще ідентифікувати ці відхилення, оскільки кожне дерево має унікальний вигляд даних, і тому вибірка, на якій воно навчається, менше схильна до впливу невеликих шумів або аномалій, які можуть бути присутніми в повній вибірці. Застосування підвибірок також сприяє зменшенню обчислювальної складності та часу навчання, оскільки кожне дерево обробляє лише фраг-

мент даних, що робить SCiForest привабливим варіантом для використання у сценаріях з обмеженими ресурсами. Це особливо актуально у випадках, коли потрібно швидко обробляти великі потоки даних або коли ресурси обмежені [79];

3. ForestASD (Anomaly Score Distribution).

Ця модифікація ІЛ включає механізми для адаптації порогів аномалій на основі розподілу оцінок аномальності, роблячи модель більш адаптивною до різних розподілів даних. Ця модель інкорпорує розширений механізм визначення аномальності, який базується на розподілі оцінок аномальності замість використання фіксованого порогу. Ця властивість робить iForestASD надзвичайно гнучкою та адаптивною до даних, які можуть мати різні рівні та типи аномалій. Такий підхід особливо ефективний у ситуаціях, де властивості аномалій можуть варіюватися від одного набору даних до іншого або навіть протягом одного й того ж набору. На практиці це означає, що iForestASD може автоматично калібрувати пороги для ідентифікації аномалій, засновуючись на внутрішній статистиці даних, що дозволяє більш чітко відділяти нормальні спостереження від аномальних. iForestASD забезпечує більш глибокий аналіз даних, розглядаючи не лише крайні випадки аномальності, але й тонші відмінності у поведінці даних, що можуть вказувати на потенційні аномалії. У такий спосіб модель відкриває нові перспективи для застосування в сферах, де необхідне динамічне або контекстно-адаптивне виявлення аномалій [80];

4. Feature Bagging Isolation Forest.

У цій версії моделі ІЛ для кожного дерева вибирається випадкова підмножина ознак, що дозволяє зменшити вплив шумових ознак та підвищити узагальнювальну здатність моделі. У цьому підході кожне дерево в ансамблі навчається на випадково вибраній підмножині ознак із загального набору даних, що має на меті зменшити варіабельність та вплив шумових ознак на процес класифікації. Вибірка підмножини ознак дозволяє моделі більш точно адаптуватися до різноманітності даних, адже кожне дерево отримує лише частину інформації та навчається ідентифікувати аномалії, базуючись на різних комбіна-

ціях ознак. Цей метод особливо ефективний у випадках, коли деякі ознаки можуть нести менше інформації про аномалії або бути шумовими, оскільки він дозволяє зменшити їх вагомість в кінцевому рішенні моделі. Feature Bagging Isolation Forest ефективно використовує принципи "мудрості натовпу", де кожне дерево вносить власний внесок у виявлення аномалій, а фінальне рішення є результатом агрегування думок усієї моделі. Це підход може значно підвищити загальну стійкість моделі до перенавчання та забезпечити більш стабільні прогнози у різноманітних даних. Крім того, робота з підмножинами ознак дозволяє виявляти аномалії, які можуть бути неявними при аналізі всього набору ознак одночасно. Це надає Feature Bagging Isolation Forest додаткову перевагу у виявленні складних патернів та неочікуваних відхилень у поведінці даних;

5. Isolation Forest з активним навчанням.

Поєднує в собі принципи ансамблевого виявлення аномалій з тактикою активного навчання. Ця модель розроблена для ефективної роботи у ситуаціях, коли вручну марковані дані є обмеженими або коли процес анотації даних є занадто трудомістким. Використання активного навчання в Isolation Forest дозволяє алгоритму вибірково запитувати мітки для об'єктів, що мають високий ступінь неоднозначності або підозру на аномалію, що оптимізує процес навчання моделі. У такому підході модель спершу визначає об'єкти, які з найбільшою ймовірністю можуть бути аномальними, і тоді запитує їхні мітки у експерта чи анотатора. Цей процес дозволяє відсіяти чіткі випадки нормальної поведінки та зосередитися на найбільш підозрілих об'єктах, ефективно використовуючи ресурси анотації;

6. Rotation Isolation Forest (RIF).

Ця модель вносить концепцію Rotation Forest, де дані трансформуються за допомогою PCA (аналіз головних компонентів) перед побудовою дерев. Цей підхід до трансформації даних значно підсилює можливості моделі у виявленні аномалій, особливо у складних багатовимірних просторах де дані часто характеризуються високим рівнем кореляції між ознаками. Застосування PCA дозволяє ротувати простір ознак таким чином, що основні варіативності у даних ви-

ражені через меншу кількість ознак. Це зменшує складність даних і водночас збільшує чутливість алгоритму до виявлення аномалій, оскільки дозволяє ІЛ ефективніше розділяти і виділяти аномальні зразки, які можуть бути приховані в оригінальному не коригованому дата сеті [81];

7. Isolation Mondrian Forest.

Розширення, яке інтегрує концепцію Mondrian Processes для оптимізації роботи з потоковими даними. Використання цього підходу дозволяє моделі динамічно адаптуватися до постійно змінюваних даних, що є критично важливим для систем, де нова інформація постійно надходить і потребує миттєвого аналізу та реакції. Основною ідеєю застосування Mondrian Processes є можливість "росту" моделі та її адаптації у відповідь на кожен новий елемент даних, що забезпечує швидку інтеграцію нової інформації та покращує ефективність обробки даних. Такий підхід дозволяє значно знизити обчислювальні витрати, адже модель оновлюється частинами і не потребує повної перебудови при кожному оновленні. Це робить Isolation Mondrian Forest ідеальним рішенням для застосувань, де необхідна висока швидкість обробки великих потоків даних та їхнє оперативне аналізування [82];

8. Kernel Isolation Forest.

Варіант ІЛ, який використовує ядерні методи для проекції даних у більш високий простір, де аномалії можуть бути легше ідентифіковані. Цей підхід дозволяє моделі ефективніше виявляти аномалії, особливо у складних наборах даних, де присутні нелінійні залежності між ознаками. Використання ядерних методів, зокрема таких як радіально-базисні функції (RBF) або поліноміальні ядра, дозволяє проектувати дані з початкового простору ознак у новий простір, де розділення між нормальними та аномальними спостереженнями стає виразнішим [83];

9. Liu and Ting’s Improved Isolation Forest.

Удосконалена версія, яка орієнтоване на покращення механізмів вибору розбиття в рамках побудови дерев. Відходячи від традиційної випадковості в процесі розділення, ця модифікована версія алгоритму вводить стратегію, яка

спеціально націлена на більш ефективне виявлення аномальних точок. Цей підхід передбачає використання алгоритмічних модифікацій для ідентифікації потенційних аномалій вже на ранніх етапах формування дерев, що дозволяє з фокусом ізолювати такі точки від більшості нормальних даних. Змінений алгоритм вибору точок розбиття включає аналіз властивостей даних, що сприяє визначенню сегментів, де аномалії мають найбільшу ймовірність виникнення. Це зменшує кількість випадкових рішень у процесі побудови дерев, забезпечуючи більш цілеспрямоване та обґрунтоване розділення. Такий метод збільшує ймовірність виявлення дійсно аномальних взірців і водночас знижує ризик включення до аномалій помилково класифікованих нормальних точок;

10. Subsampled Isolation Forest.

Реалізує ідею вибірки підвибірок фіксованого розміру для побудови кожного дерева в ансамблі. Цей метод дозволяє значно покращити продуктивність алгоритму, знижуючи час, необхідний для тренування моделі, а також забезпечуючи більш високу стабільність результатів при роботі з великими обсягами даних. Використання підвибірок допомагає уникнути перенавчання, оскільки кожне дерево базується на випадково вибраній, обмеженій кількості даних, що зменшує ймовірність того, що модель занадто точно "підганятиметься" під конкретні особливості тренувального набору даних. Це також забезпечує кращу загальну здатність моделі до узагальнення, оскільки різні дерева в ансамблі використовують різні аспекти даних, що дозволяє їм колективно охопити більш широкий спектр потенційних аномалій. Крім того, застосування методу підвибірок у Subsampled Isolation Forest знижує вимоги до обчислювальних ресурсів, що робить цю модель особливо придатною для застосувань у середовищах з обмеженими ресурсами, таких як мобільні пристрої або вбудовані системи. Такий підхід дозволяє ефективно обробляти великі набори даних, навіть на апаратному обладнанні з обмеженими можливостями, забезпечуючи при цьому достатньо високу точність та швидкість обробки даних.

Удосконалення моделі ІЛ зосереджене на різних аспектах її функціональності, що сприяє досягненню високої точності та адаптивності при вирішенні

різноманітних задач ідентифікації аномалій. Одним із ключових напрямів є точне налаштування параметрів моделі, включаючи калібрування кількості дерев, розміру підвибірок для тренування кожного дерева, та порогів для ідентифікації аномалій. Це дозволяє підвищити чутливість і специфічність моделі, адаптуючи її до особливостей конкретних даних.

Інший важливий аспект вдосконалення стосується ознак, які використовуються для навчання моделі. Застосування методів, таких як Feature Bagging та Kernel methods, зменшує вплив шумових ознак і покращує здатність моделі узагальнювати, дозволяючи ефективніше аналізувати комбінації ознак. Такі техніки також допомагають моделі краще виявляти складні аномалії, які не завжди очевидні при аналізі окремих ознак.

Останнім напрямком удосконалення є оптимізація процесу прийняття остаточних рішень про аномальність даних. Це включає розробку більш досконалих механізмів для встановлення порогів аномальності, таких як застосування адаптивних порогів, базованих на розподілі оцінок аномальності, та інтеграція методів активного навчання, що дозволяють моделі запитувати мітки для найбільш неоднозначних точок даних. Такі підходи забезпечують вищу точність ідентифікації аномалій і зменшують кількість помилкових сигналів.

4.2 Дослідження параметрів моделі ізольованого лісу

При аналізі вдосконалених моделей ІЛ особливу увагу приділяється точному налаштуванню параметрів моделі, що дозволяє значно підвищити її ефективність та адаптивність до різних завдань виявлення аномалій. В межах базової моделі акцентуються такі ключові параметри: “кількість припущених аномалій”, ”кількість дерев у лісі”, “критерій зупинки побудови дерева” та “початкове значення генератора випадкових чисел”. Кожен із цих параметрів був детально розглянутий у розділі 3.3, де були описані їхні функції та вплив на результати моделювання. Опираючись на попередній аналіз, можна зробити декілька ключових висновків:

- **«Кількість дерев у лісі»** безпосередньо впливає на стабільність та надійність результатів моделі. Збільшення кількості дерев сприяє підвищенню точності узагальнення за рахунок усереднення прогнозів по більшій кількості дерев. Це допомагає моделі краще справлятися з варіативністю даних та більш ефективно виявляти складні аномалії. Однак слід зазначити, що значне збільшення кількості дерев, наприклад до 1000 і більше, тягне за собою підвищення обчислювальних витрат та часу обробки даних. При цьому експериментальні дані показують, що подальше збільшення кількості дерев до 5000 не призводить до помітного поліпшення результатів, що робить недоцільним подальше нарощування цього параметра у зв'язку з обмеженнями обчислювальних ресурсів;
- **«Початкове випадкове число»**, що використовується для генерації випадкових підмножин даних та вибору атрибутів для розділення у деревах, відіграє ключову роль у забезпеченні відтворюваності результатів. Встановлення одного і того ж початкового числа гарантує консистентність результатів між різними запусками моделі. Однак при значенні параметра кількості дерев у лісі, рівному 1000, зміни у початковому випадковому числі не призводять до суттєвих змін у поведінці моделі, що вказує на його відносно менший вплив у даній конфігурації;
- **«Критерій зупинки побудови дерева»** та «**кількість припущених аномалій»** мають значний вплив на якість та точність моделі. Ці параметри критично важливі для визначення моменту, коли подальше розділення вузлів дерева стає невиправданим, або коли модель досягає оптимального балансу між виявленням аномалій та уникненням помилкових спрацьовувань. Оптимізація цих параметрів вимагає детального розуміння даних та специфіки завдання, для якої модель налаштовується.

4.2.1 Аналіз критерію зупинки

Критерій зупинки побудови дерева відіграє важливу роль у запобіганні перенавчанню моделі. Перенавчання виникає, коли модель надто точно адаптована до навчальних даних, що погіршує її здатність до узагальнення на нових даних, які не були раніше видимі. Для уникнення цієї проблеми використовується критерій зупинки, який визначає, коли подальше розділення вузлів дерева стає небажаним.

Критерій зупинки може бути реалізований кількома способами:

1. Максимальна глибина (Max Depth):

Цей параметр встановлює верхній ліміт кількості рівнів для кожного з дерев в ансамблі. Обмеження глибини дерева знижує ризик перенавчання, коли модель адаптується до нюансів тренувальних даних, ігноруючи загальні закономірності, що може призвести до поганої продуктивності на нових даних. Обмеження глибини також може прискорити навчання моделі, оскільки зменшується кількість варіантів розділення, що розглядаються;

2. Мінімальна кількість зразків для розділення (Min Samples Split):

Встановлює критерій мінімальної кількості зразків, які необхідно мати в вузлі, перш ніж він може бути розгалужений на підвузли. Це запобігає створенню надто дрібних правил, які можуть відображати випадкові особливості тренувальних даних, а не істинні закономірності, тим самим зменшуючи вірогідність перенавчання. Більш високе значення цього параметра сприяє побудові більш узагальнених моделей, які краще працюють на невідомих даних;

3. Мінімальна кількість зразків у листку (Min Samples Leaf):

Цей параметр забезпечує, що кожний листовий вузол (завершення дерева, де робиться передбачення) міститиме мінімально допустиму кількість зразків. Це підвищує надійність моделі, забезпечуючи, що передбачення робляться на основі достатньої кількості даних. Встановлення цього порогу важливе для за-

побігання створення листків, заснованих на малих і потенційно аномальних вибірках даних, що могло б призвести до нестабільних передбачень;

4. Максимальна кількість вибірок (Max Samples):

Встановлює обмеження на кількість даних, які використовуються для навчання кожного дерева в ансамблі. Це створює варіативність у навчанні різних дерев і сприяє "мудрості натовпу", де агрегація різноманітних моделей призводить до сильнішої та стійкішої до змін ансамблевої моделі. Крім того, використання підмножин даних для навчання різних дерев сприяє зниженню ризику перенавчання і допомагає моделі краще узагальнювати, роблячи її стійкішою до шуму і викидів у даних;

5. Максимальна кількість ознак (Max Features):

Цей параметр встановлює верхній ліміт кількості ознак, які розглядаються під час визначення найкращого способу розподілу даних у кожному вузлі рішення. Зменшення числа ознак, використовуваних для побудови кожного вузла, не тільки прискорює навчання за рахунок скорочення обчислювальних витрат, але й сприяє зниженню ризику перенавчання. Коли різні дерева використовують різні підмножини ознак, це підвищує різноманітність в ансамблі, що в кінцевому підсумку може призвести до покращення узагальнювальної спроможності моделі;

6. Максимальна кількість листків (Max Leaf Nodes):

Цей параметр визначає максимальну кількість листових вузлів, які можуть бути в дереві. Встановлення межі на кількість листків у дереві є ще одним способом контролювання його складності та уникнення перенавчання. Чим менше листків, тим простіша структура моделі, що зазвичай призводить до більш стійких і узагальнених передбачень. Таким чином, модель стає менш чутливою до невеликих коливань та шумів у даних, що важливо для підтримки гарної продуктивності моделі на нових даних.

Автоматичний критерій зупинки у моделі ІЛ дозволяє деревам рости до досягнення визначеної глибини або до моменту, коли кількість даних у вузлі стає занадто малою для подальшого розділення. Цей критерій спрощує

налаштування параметрів моделі, забезпечуючи водночас високу адаптивність до різних наборів даних. Така адаптивність критично важлива у ситуаціях, де характеристики даних можуть швидко змінюватися або коли вони не повністю відомі на етапі розробки моделі.

Застосування цього критерію сприяє збереженню високої узагальнювальної здатності моделі, оскільки воно запобігає перенавчанню на специфіці навчальної вибірки. В ІЛ, де основна ціль — виявлення аномалій, здатність моделі уникати надмірної уваги до аномальних або шумових даних безпосередньо впливає на якість і надійність її прогнозів.

В ІЛ випадковий вибір атрибутів для розділення дерев означає, що кожне дерево в ансамблі розвивається, враховуючи різні аспекти даних, що сприяє створенню робастної моделі, здатної ефективно функціонувати навіть при великих змінах у вхідних даних. Такий підхід знижує ризик переосмислення даних, що є звичайним явищем у більш традиційних моделях ДР, де вибір кращого поділу базується на конкретних статистичних критеріях, таких як приріст інформації або коефіцієнт Джині.

Цей метод випадкового відбору зменшує ймовірність того, що модель ІЛ стане чутливою до певних виразних аномалій у навчальних даних, замість того щоб розробляти загальні правила для виявлення аномалій, що ефективні на більш широкому спектрі сценаріїв.

Кожен обраний параметр, який може бути використаний у якості критерія зупинки, потребує детального практичного дослідження. Такий підхід забезпечить оптимальне налаштування моделі відповідно до специфіки даних, з якими вона працює. Водночас, це допоможе вирішити ряд проблем, пов'язаних з ефективністю моделі, її здатністю до узагальнення та відтворенням результатів у різних умовах.

4.2.2 Дослідження чутливості моделі ізольованого лісу

Чутливість моделі до аномалій значно залежить від параметра "кількість припущених аномалій", який є критично важливим для налаштування сприйнятливості моделі до класифікації даних як нормальних або аномальних. Цей параметр визначає, скільки відсотків даних модель вважатиме аномальними. При збільшенні значення цього параметра зростає ймовірність того, що модель класифікує більше точок даних як аномальні, навіть якщо вони є нормальними. Це може призвести до збільшення кількості помилково-позитивних результатів, тобто ситуацій, коли нормальні дані помилково ідентифікуються як аномалії. Такий підхід вимагає делікатного налаштування, щоб досягти балансу між виявленням справжніх аномалій та мінімізацією помилкових спрацьовувань.

Налаштування цього параметра вимагає детального аналізу, оскільки різні набори даних і різні прилади можуть потребувати різного рівня чутливості. Особливо це важливо у контексті метрологічних даних, де високий рівень точності є ключовим. У випадках, коли параметр "кількість припущених аномалій" встановлений занадто високо, це може призвести до включення нормальних вимірювань у категорію аномалій, що потенційно спотворить результати аналізу. З іншого боку, занадто низьке значення цього параметра може призвести до недооцінки справжніх аномалій, які можуть бути важливими для ідентифікації систематичних або випадкових помилок у роботі витратомірів.

Щоб дослідити вплив цього параметра на результати моделі, необхідно провести численні запускі моделі, змінюючи значення параметра "кількість припущених аномалій". Це дозволить отримати повну картину того, як зміна цього параметра впливає на результати класифікації даних. Такий підхід дасть можливість визначити оптимальне значення параметра для кожного конкретного витратоміра.

Подальший аналіз впливу параметра припущених аномалій буде ґрунтуватися на зведеному графіку, що об'єднує результати для трьох витратомірів, як це було зроблено у розділі 4.2. Такий графік дозволить візуалізувати, як різні

значення параметра впливають на кількість виявлених аномалій, а також на розподіл нормальних і аномальних точок для кожного з витратомірів. Це допоможе краще зрозуміти, як параметр "кількість припущених аномалій" впливає на загальну чутливість моделі та дасть змогу визначити найкращі налаштування для кожного типу вимірювань.

З кроком 5% у діапазоні від 5% до 50% було проведено налаштування параметра "кількість припущених аномалій", що дозволило отримати широкий спектр результатів для кожного значення параметра. Для кожного запуску моделі ми отримали результати, які були об'єднані у зведений графік. На рис. 4.1-4.10 представлені графіки для кожного з запусків, що відображають зміну кількості виявлених аномалій залежно від зміни параметра "кількість припущених аномалій". Кожен графік демонструє, як поступове збільшення цього параметра впливає на результативність моделі та змінює співвідношення між нормальними та аномальними точками.

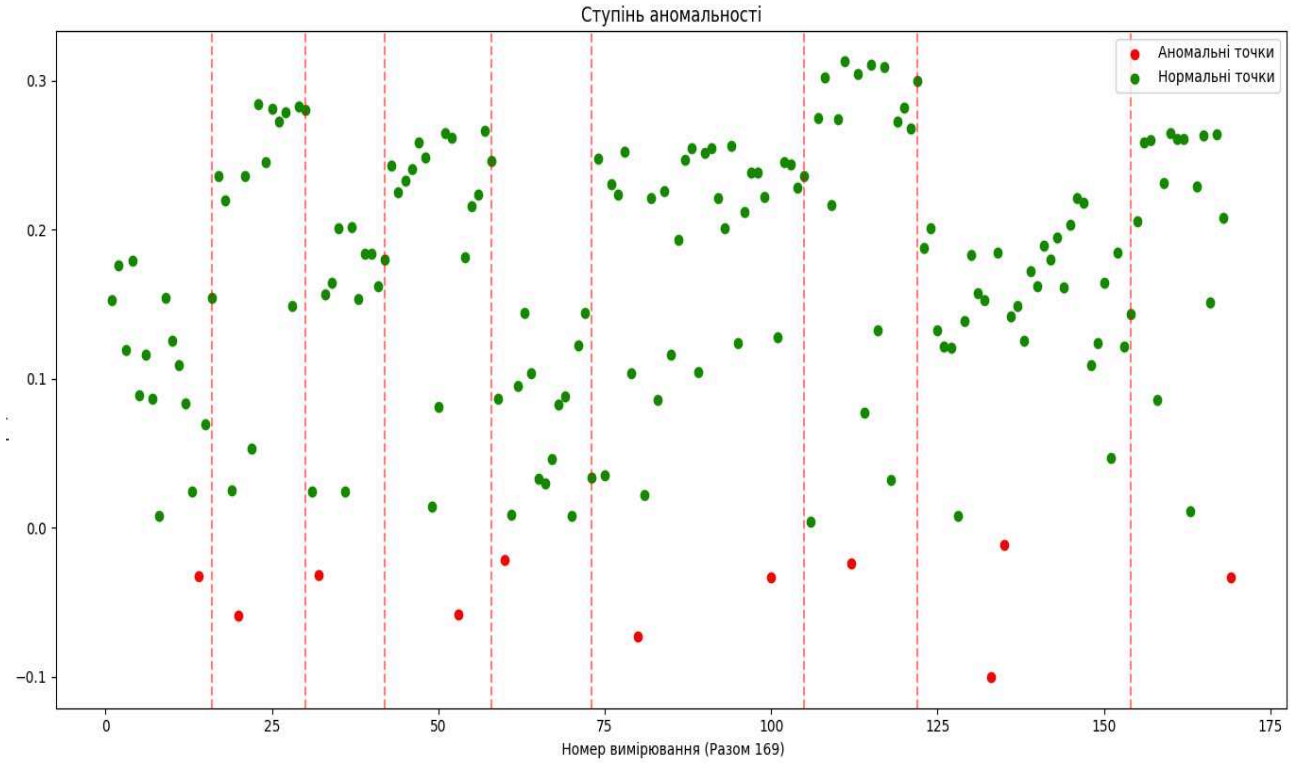

Рисунок 4.1 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 5%.

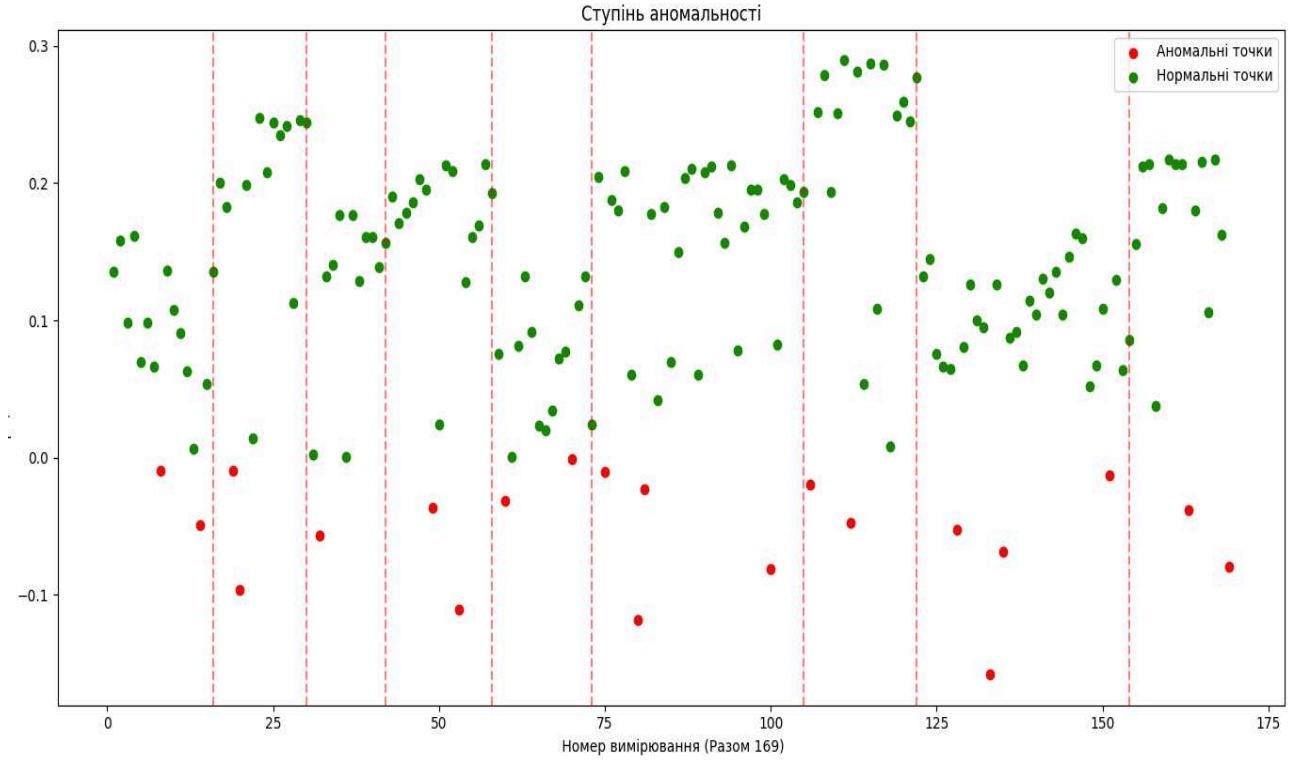

Рисунок 4.2 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 10%.

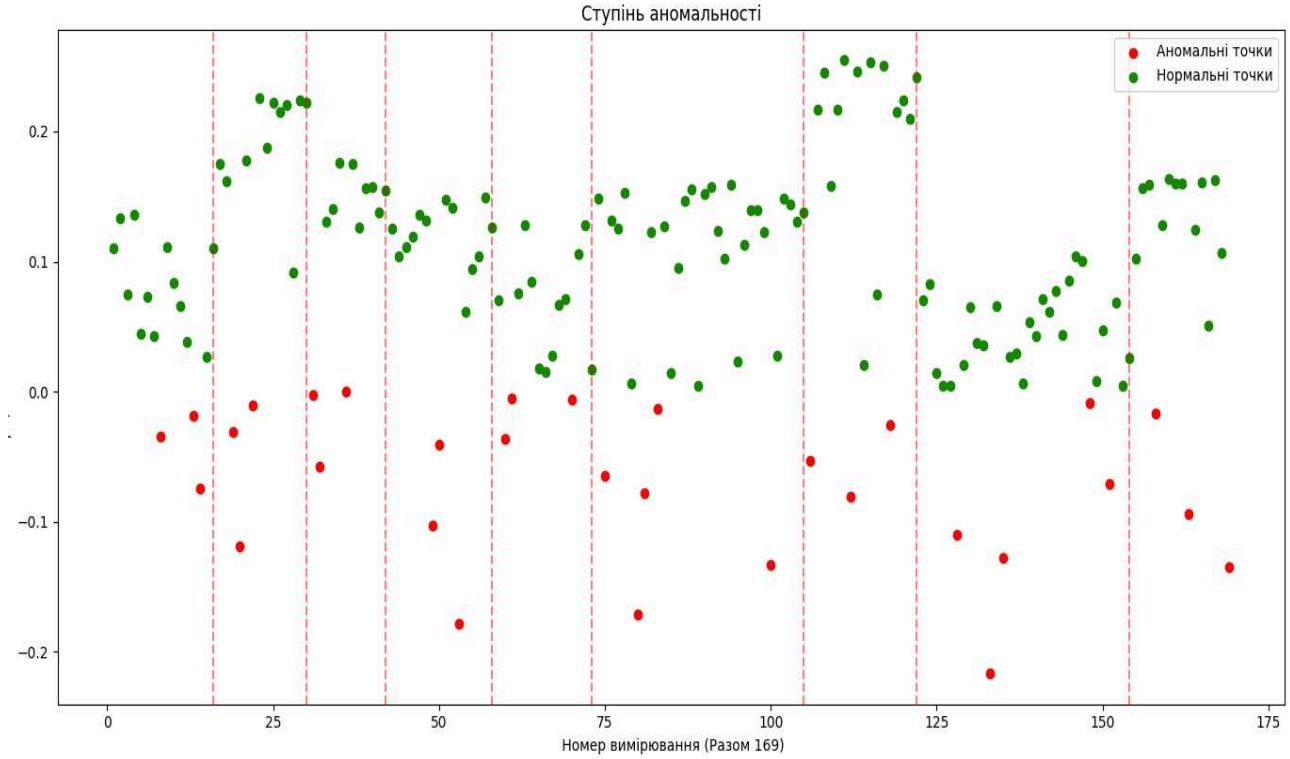

Рисунок 4.3 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 15%.

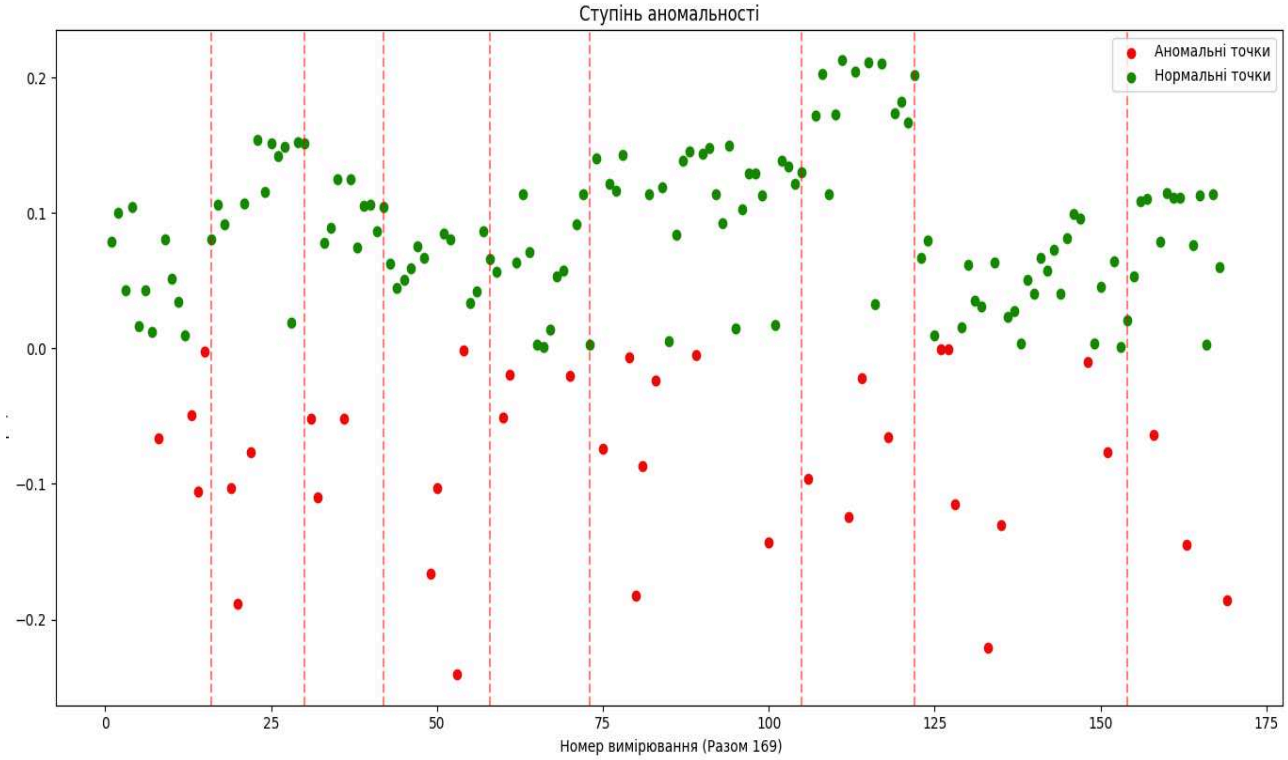

Рисунок 4.4 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 20%.

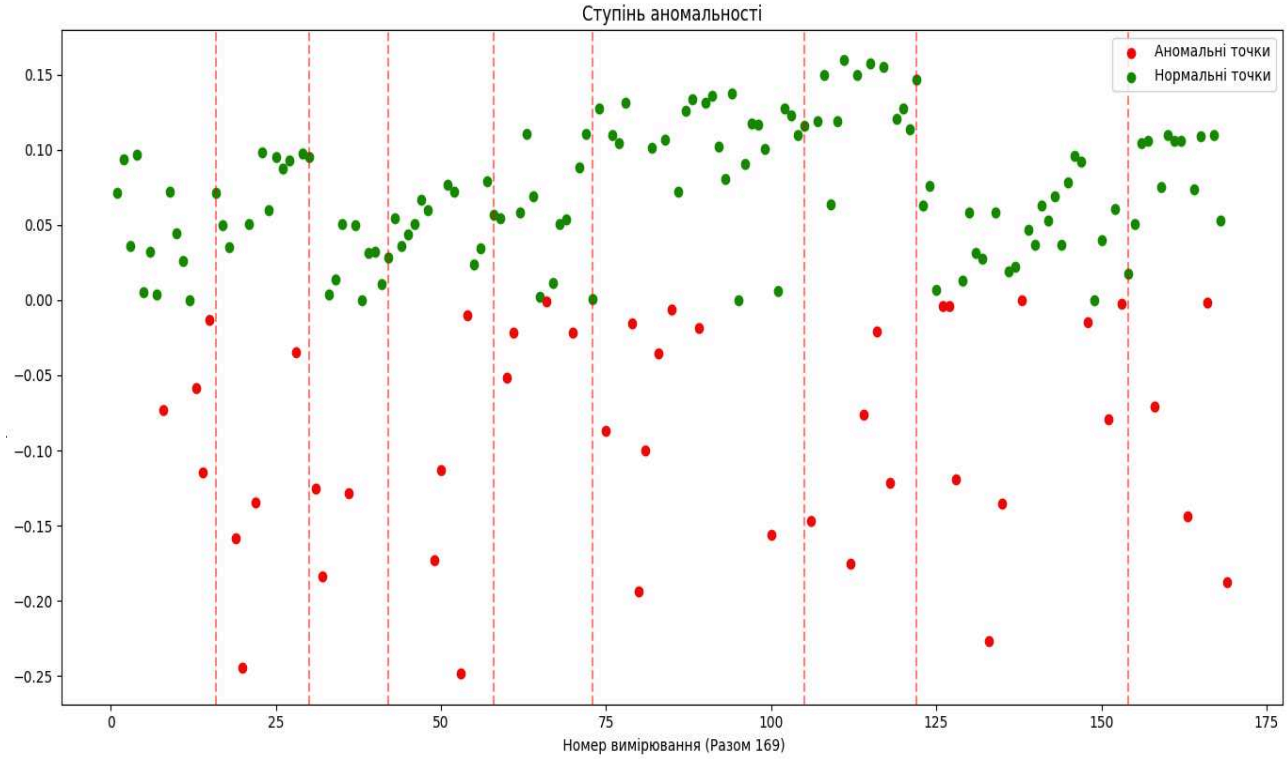

Рисунок 4.5 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на25%.

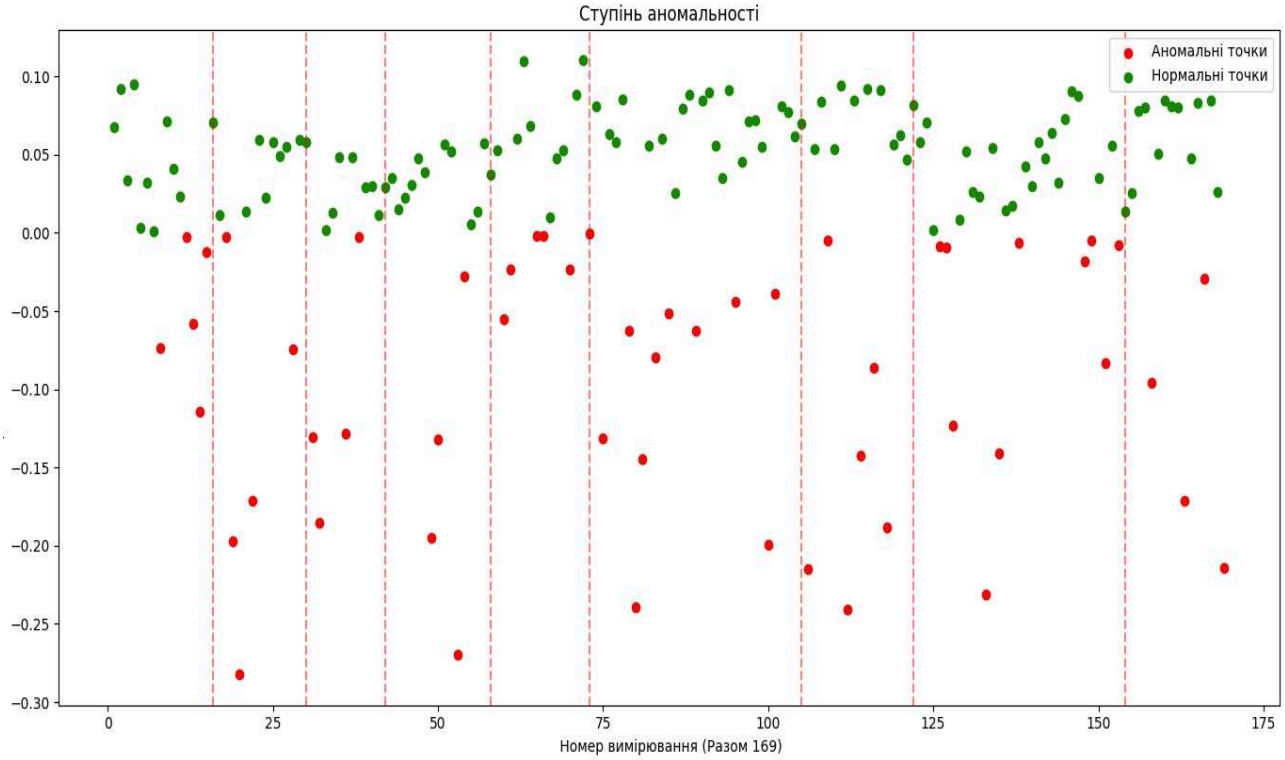

Рисунок 4.6 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 30%.

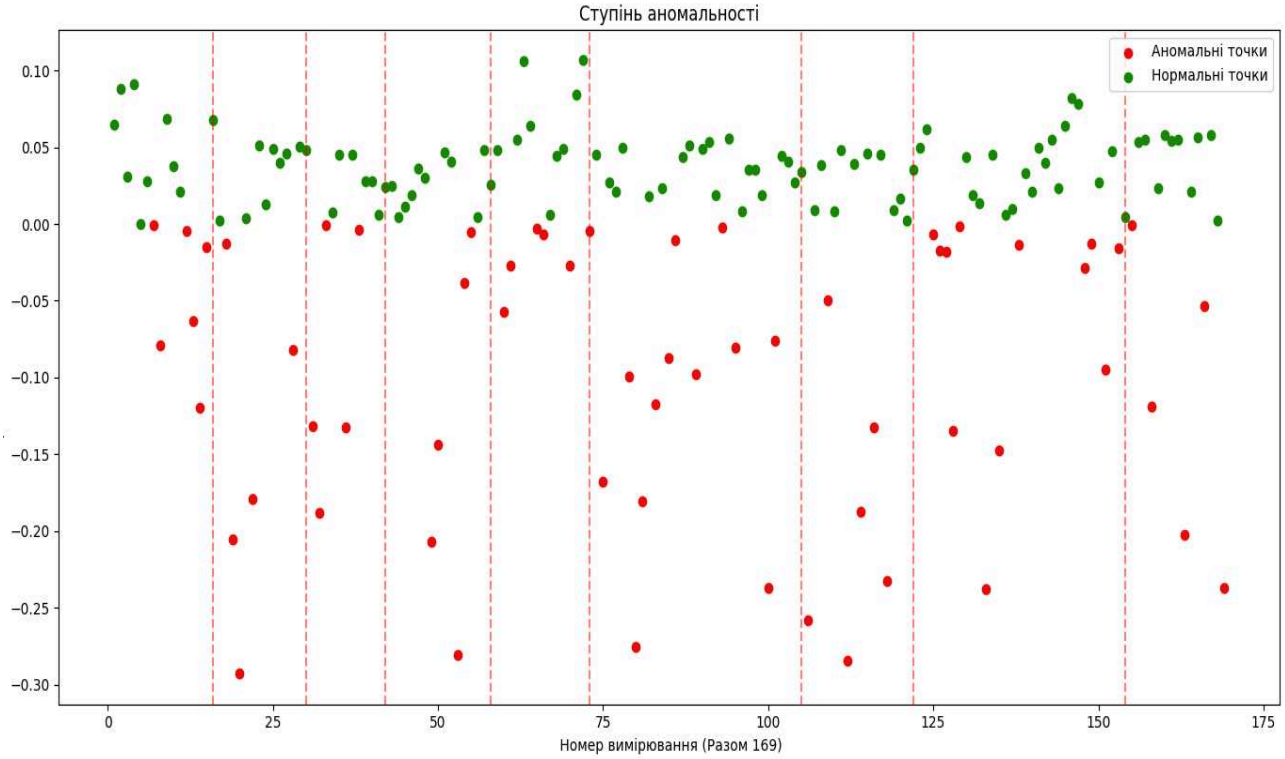

Рисунок 4.7 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 35%.

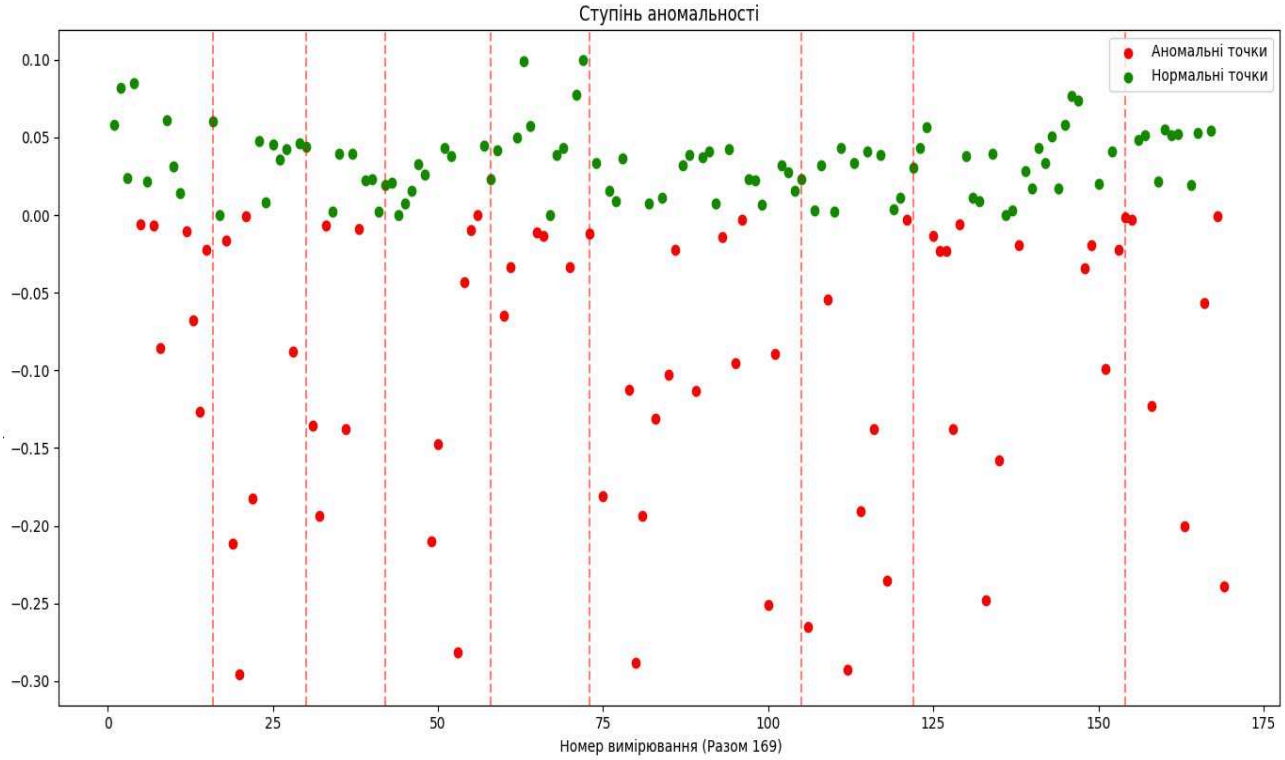

Рисунок 4.8 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 40%.

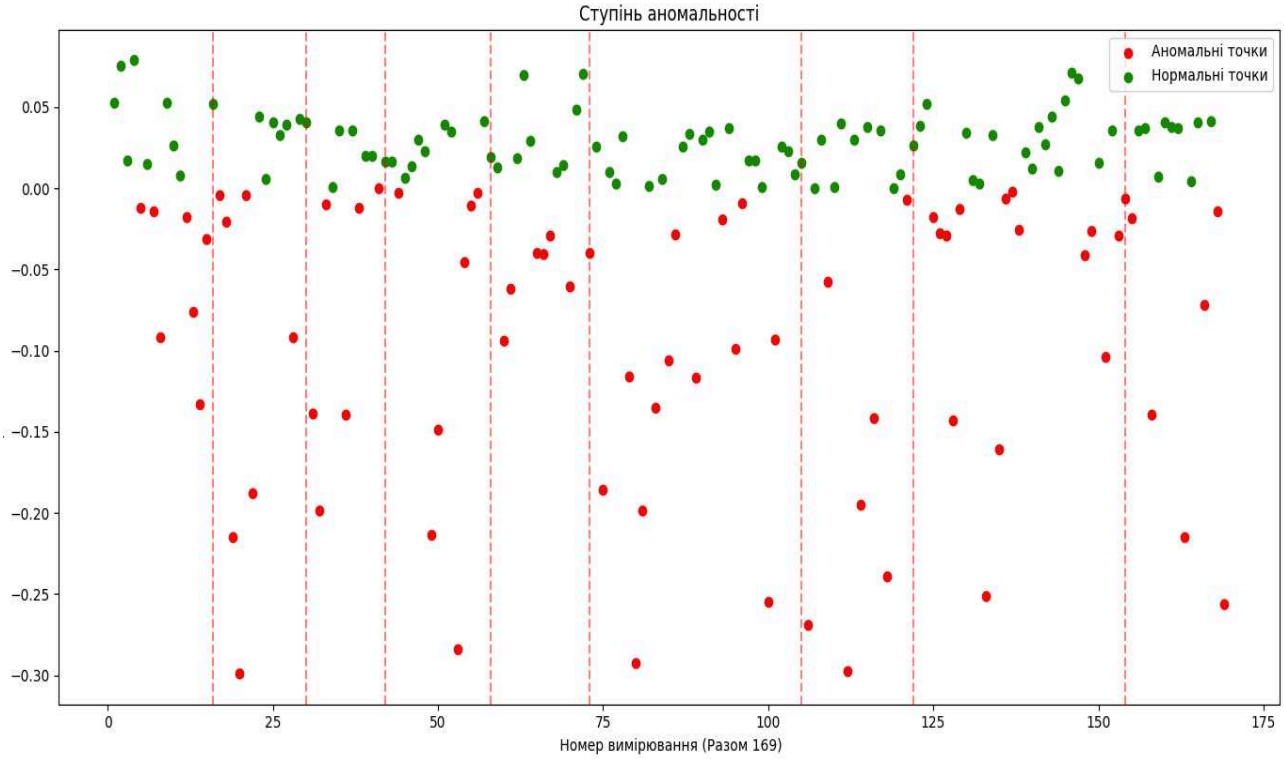

Рисунок 4.9 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 45%.

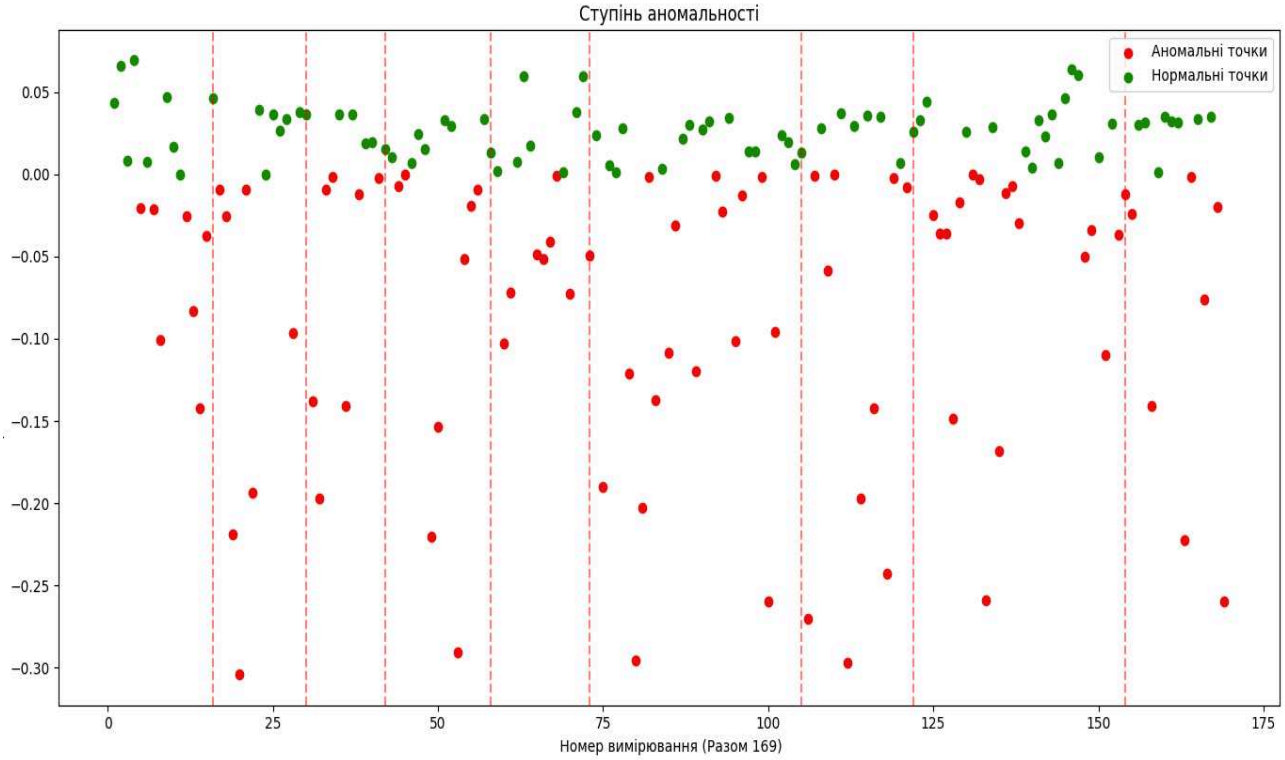

Рисунок 4.10 – Зведений графік результатів запуску моделі при налаштуванні параметра "кількість припущених аномалій" на 50%.

Аналіз результатів запусків моделі ІЛ демонструє декілька важливих аспектів, які варто враховувати при подальшій роботі з параметром "кількість припущених аномалій". Як було показано на графіках, збільшення цього параметра приводить до зростання кількості виявлених аномальних точок, що, з одного боку, дозволяє виявляти навіть найдрібніші відхилення, а з іншого боку, підвищує ризик класифікації нормальних точок як аномальних. Це відображає основну проблему налаштування порогу, оскільки кожна вибірка може вимагати індивідуального підходу.

На графіках видно, що при різних значеннях параметра "кількість припущених аномалій" формуються скупчення точок, які можна класифікувати як нормальні значення. Ці точки концентруються довкола певних діапазонів і не мають різких відхилень. У той час як викиди, які визначає модель, мають більш виражений характер і розміщуються на значно віддалених позиціях від основної маси даних. Це дозволяє зробити візуальне припущення про те, де

знаходяться нормальні значення, а де є викиди. Проте чітко видно, що поріг, на якому модель визначає аномалії, змінюється в залежності від налаштувань чутливості, що ускладнює встановлення єдиного оптимального порогу для всіх типів вибірок.

З цього можна зробити висновок, що встановити універсальний поріг для всіх наборів даних, у яких використовується модель ІЛ, неможливо. Кожна вибірка може мати свої унікальні характеристики, що потребують індивідуального налаштування параметра для виявлення аномалій. Одні вибірки можуть містити багато аномальних значень, тоді як інші можуть не мати взагалі жодних викидів. Однак, модель ІЛ працює таким чином, що вона завжди виявлятиме деякі точки як аномальні, навіть якщо у вибірці немає викидів. Це зумовлено самим принципом роботи моделі, яка базується на пошуку ізоляції точок, що відрізняються від інших.

Ще одним важливим аспектом є те, що якщо навіть видалити виявлені викиди та запустити модель ще раз на тій самій вибірці, вона знову знайде нові "аномальні" точки. Це пояснюється тим, що модель завжди намагається виділити ті точки, які найбільш ізольовані від основної маси даних. Навіть за відсутності явних викидів, модель буде класифікувати деякі точки як аномальні, оскільки шукає відхилення від структури даних. Це властивість моделі ІЛ є як її перевагою, так і недоліком: вона здатна знаходити навіть найменші відхилення, але при цьому завжди виявляє "аномалії", навіть там, де їх може не бути.

Для досягнення максимальної точності, необхідно враховувати характер даних у кожній вибірці та проводити індивідуальні налаштування моделі.

4.2.3 Класифікатор моделі ізольованого лісу

Як зазначалось раніше, результатом роботи моделі ІЛ є визначення ступеня аномальності кожної точки даних. Цей показник дозволяє оцінити відхилення точки від нормальних значень у вибірці. При зміні чутливості

моделі змінюється лише шкала ступеня аномальності, тобто числові значення показника, але просторове розташування точок у вибірці залишається незмінним. Це означає, що модель змінює тільки масштаби відхилень, не впливаючи на внутрішню структуру даних.

Для класифікації модель використовує поріг "0". Точки, значення яких нижчі за цей поріг, визначаються як аномальні, а ті, що мають значення вище, вважаються нормальними. Однак такий підхід є недостатньо точним, оскільки фіксований поріг не враховує специфіку та варіативність вибірок. Це призводить до того, що деякі нормальні точки можуть бути помилково класифіковані як аномальні лише через незначне відхилення їхніх значень від порогу.

Зміна чутливості моделі впливає на те, які точки будуть класифіковані як аномальні. Точки, що близькі до порогу "0", можуть легко перейти з категорії нормальних до аномальних через зміну чутливості, хоча їхнє фактичне положення відносно інших даних не змінюється. Ця проблема вказує на необхідність удосконалення підходу до класифікації, оскільки фіксований поріг не забезпечує достатньої гнучкості.

Замість використання жорсткого порогу для всіх вибірок, доцільно застосувати підхід, що базується на відстанях між точками та їхніх відхиленнях. Нормальні точки зазвичай мають схожі ступені аномальності і утворюють скупчення у певних діапазонах значень. Аномальні ж точки розташовані значно далі від основної групи і мають більші відхилення. Цей факт дає підстави для класифікації не лише через поріг, а через аналіз скупчень точок і відстаней між ними.

Нормальні значення можуть бути визначені як ті, що розташовані в межах певних діапазонів з мінімальними відхиленнями від одної групи до іншої. Викиди ж відрізняються від основної маси даних значними відхиленнями. Це дозволяє покращити класифікацію через детальний аналіз цих скупчень і створити більш точну систему для відокремлення аномальних даних від нормальних.

Замість фіксованого порогу "0", динамічний підхід передбачає визначення викидів на основі відстаней між точками. Відстані між нормальними точками є меншими, ніж між нормальними і аномальними. Це дозволяє точніше визначати відхилення і зменшити кількість помилкових спрацьовувань. Нормальні точки утворюють компактні скупчення, і ці скупчення можна легко відокремити від аномальних точок, які знаходяться на значній відстані від основної маси даних.

Такий підхід дозволяє гнучкіше адаптувати модель до різних типів вибірок, що особливо важливо, коли дані мають різну структуру. У випадках, коли вибірка містить велику кількість нормальних точок і лише кілька викидів, фіксований поріг може призвести до значної кількості помилкових класифікацій. Динамічна класифікація, яка базується на характеристиках вибірки, дозволяє точніше враховувати індивідуальні особливості даних.

Слід також враховувати проблему описану у попередньому розділі, що модель ІЛ за своєю природою завжди прагне визначити певний відсоток точок як аномальні, навіть якщо в реальності викиди можуть бути відсутні. Модель ізолює найменш схожі точки, і це призводить до класифікації деяких нормальних точок як аномалій, навіть якщо вибірка є однорідною. Це може призвести до ситуації, коли після виключення виявлених аномалій модель знову знайде нові викиди при повторному запуску. Для того щоб уникнути цієї проблеми, необхідно замінити фіксований класифікатор на підхід, що базується на аналізі відстаней між нормальними точками та їхніх відхилень. Це дозволить точніше ідентифікувати викиди і зменшити кількість помилково класифікованих даних.

4.3 Методи знаходження скупчень

Для вирішення задачі пошуку скупчень нормальних точок на основі ступеня аномальності, де відстані між точками не змінюються, можна використовувати методи, які ефективно виявляють групи точок, близькі одна до одної, без

зміни відносного положення в просторі. Одним із таких підходів є методи кластеризації, які дозволяють розбити дані на скупчення на основі їхньої просторової близькості та подібності.

Методи кластеризації:

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** – це алгоритм кластеризації, який ґрунтується на щільності точок. Він об'єднує точки в кластери на основі їхньої просторової близькості та щільності оточуючих точок. Для кожної точки визначається, чи належить вона до скупчення точок з високою щільністю. Якщо так, то вона стає частиною кластера, а якщо ні, то класифікується як ізольована або аномальна. DBSCAN можна ефективно використовувати для пошуку нормальних точок у задачі класифікації на основі ступеня аномальності, де відстані між точками залишаються незмінними. Метод дозволяє визначити групи точок, які мають схожі значення ступеня аномальності і формують щільні скупчення, що відповідають нормальним точкам.
- **k-means** є одним із найпоширеніших методів кластеризації, що базується на поділі даних на **k** кластерів, де кожен кластер представляється своїм центроїдом. Для даних з незмінними відстанями між точками ступеня аномальності, цей метод може бути корисним для групування нормальних точок в кластери, що мають схожі значення ступеня аномальності, та ідентифікації точок, що знаходяться далеко від центрів кластерів, як потенційних аномалій.У процесі роботи алгоритму кожна точка даних призначається до кластера на основі мінімальної відстані до центроїда. Центроїди оновлюються після кожної ітерації, поки алгоритм не досягне стабільного стану, при якому відстань між точками та центроїдами мінімальна. Для нашої задачі цей метод може використовуватися для пошуку скупчень нормальних точок, оскільки точки з близькими значеннями ступеня аномальності утворюватимуть компактні кластери.
- **Ієрархічна кластеризація** – це підхід, який будує ієрархію об'єднання точок даних у кластери на різних рівнях. Алгоритм послідовно об'єднує

найближчі точки в скупчення, поки всі точки не будуть об'єднані в одну групу. Потім, на основі дендрограми, можна виділити різні рівні кластеризації для аналізу структури даних. Для задачі пошуку нормальних точок на основі ступеня аномальності цей метод можна використовувати для визначення кластерів нормальних точок, що мають схожі значення ступеня аномальності, і відокремлення аномалій, які не входять до цих скупчень. Він також дозволяє гнучко налаштовувати рівень деталізації аналізу в залежності від потреб конкретної задачі [84-91].

Методи на основі щільності працюють за принципом оцінки локальної щільності точок. Для кожної точки визначається, скільки сусідів вона має на заданій відстані. Якщо кількість сусідів перевищує певний поріг, точка вважається частиною щільного скупчення, що складається з нормальних точок. Якщо ж точка має низьку щільність сусідів, її можна вважати аномальною.

Методи на основі щільності:

- **KDE (Kernel Density Estimation)–** це метод непараметричної оцінки щільності, що дозволяє оцінити ймовірність знаходження точки в конкретній області простору на основі її сусідів. У задачі класифікації нормальних точок за ступенем аномальності KDE може бути використаний для побудови карти щільності, що покаже, де знаходяться основні скупчення нормальних точок. Цей метод дозволяє оцінити, чи точка знаходиться в щільному скупченні (нормальні точки), чи в області з низькою щільністю (анормальні точки).
- **OPTICS (Ordering Points To Identify the Clustering Structure)** – це метод кластеризації, що також базується на щільності, але на відміну від DBSCAN, він будує впорядковану послідовність точок на основі їхньої щільності, без жорсткої вимоги до визначення радіусу пошуку сусідів. Це дозволяє методові виявляти скупчення точок різної щільності і гнучко працювати з нерівномірними вибірками. OPTICS може бути використаний для виявлення скупчень нормальних точок, незалежно від їхньої форми та щільності.

– **Метод локальної оцінки аномалій (LOF)** аналізує щільність точок у локальних областях і порівнює її з щільністю сусідів. Якщо щільність точки значно відрізняється від щільності її найближчих сусідів, то вона вважається аномальною. LOF дозволяє оцінювати локальні скупчення нормальних точок та ідентифікувати ті, що відхиляються. Цей метод підходить для задачі, оскільки враховує локальну структуру даних і дозволяє виявляти як щільні скупчення нормальних точок, так і окремі аномальні точки [92-107].

Метод **"перелому коліна" (Elbow Method)** зазвичай використовується для вибору оптимальної кількості кластерів у задачах кластеризації, проте його ідею можна адаптувати для задачі пошуку скупчень нормальних точок. Для цього аналізуються зміни у відстанях між точками або їх ступенем аномальності. Якщо на графіку залежності відстаней спостерігається різкий стрибок (перелом), це може свідчити про межу між нормальними точками та аномаліями. Використання цього методу дозволяє ідентифікувати моменти, коли точки перестають належати до нормальних скупчень і починають відхилятися від загальної структури.

4.4 Результати дослідження

Розглянуті варіанти пошуку скупчень нормальних точок, зокрема методи кластеризації та підходи, засновані на щільності, вимагають налаштування певних параметрів, що може бути суттєвим недоліком у задачах з різнорідними вибірками даних. Наприклад, для алгоритмів кластеризації, таких як **DBSCAN**, необхідно задавати **радіус пошуку (ε)** та **мінімальну кількість точок у кластері (minPts),** що впливає на форму та розмір кластерів. Якщо ці параметри підібрані невірно, то кластеризація може або об'єднати аномальні точки з нормальними, або, навпаки, не охопити всі нормальні точки у скупченні. В результаті, параметри, налаштовані для однієї вибірки, можуть не підходити

для іншої, що призводить до необхідності коригування порогових значень для кожного нового набору даних.

Схожа ситуація виникає з іншими методами, такими як метод **k-means**, де потрібно визначати кількість кластерів (**k**), що також є параметром, який безпосередньо впливає на результати кластеризації. Для кожної вибірки даних може знадобитися оптимальний параметр **k**, що не завжди можливо оцінити наперед, особливо для даних різного обсягу та структури. Крім того, алгоритми **ієрархічної кластеризації** вимагають вибору рівня деталізації у дендрограмі, що знову ж таки є своєрідним порогом для розподілу на кластери.

Методи, засновані на щільності, такі як **LOF (Local Outlier Factor)** та **KDE (Kernel Density Estimation),** також мають обмеження, пов'язані з налаштуванням параметрів. У випадку LOF, для оцінки локальної щільності використовується кількість найближчих сусідів, що потребує встановлення оптимального значення цього параметра. KDE, у свою чергу, вимагає вибору ширини ядра (bandwidth), від якої залежить рівень деталізації карти щільності. Оптимальні значення цих параметрів можуть суттєво змінюватися для різних вибірок, що потребує індивідуального налаштування для кожного випадку.

Метод "перелому коліна" (Elbow Method), який часто використовується для визначення оптимальної кількості кластерів, також вимагає ручного аналізу графіка змін відстаней, щоб ідентифікувати точку перелому. Це означає, що навіть цей метод опосередковано містить необхідність встановлення порогового значення, яке визначається на основі візуального аналізу та може відрізнятися для різних вибірок.

Усі ці підходи передбачають наявність певних параметрів, які потрібно підлаштовувати під кожну конкретну вибірку. Універсальність таких методів обмежується тим, що налаштування параметрів, оптимальних для одного набору даних, може бути неприйнятним для іншого. В результаті, хоча ці методи дозволяють визначати скупчення нормальних точок та відокремлювати аномалії, вони не забезпечують універсального підходу для різнорідних

вибірок, оскільки завжди містять елементи порогового налаштування, яке необхідно адаптувати до конкретних особливостей кожного набору даних.

Тому доцільно використовувати стандартну модель ІЛ без класифікатора та проводити візуальний аналіз результатів для встановлення необхідних умов її коректної роботи. Модель ІЛ дозволяє ефективно визначати аномалії, оскільки ґрунтується на принципах випадкової ізоляції точок, що забезпечує адаптивний підхід до виявлення відхилень. Однак, щоб модель працювала належним чином, необхідно забезпечити відповідні умови для аналізу результатів та встановлення порогів, які підходять для конкретної вибірки даних. Цей підхід дозволяє уникнути проблем, пов’язаних з налаштуванням жорстких параметрів, як у методах кластеризації або щільності, і дає змогу зосередитися на динамічній оцінці результатів на основі візуалізації та адаптивного аналізу.

При використанні моделі ІЛ виникають дві основні проблеми, які впливають на точність виявлення аномалій. Перша проблема полягає у зміні шкали ступеня аномальності між різними запусками моделі (розділ 4.2.2). Значення аномальності залежать від налаштування параметра “кількість припущених аномалій”, який визначає чутливість моделі до виявлення відхилень. Зміна цього параметра призводить до зміщення шкали ступеня аномальності, що ускладнює визначення порогу між нормальними та аномальними точками.

Друга проблема стосується можливості виявлення аномалій навіть у випадках, коли у вибірці відсутні явні відхилення (розділ 4.2.3). Модель ІЛ здатна ідентифікувати певні точки як аномальні, навіть якщо вибірка повністю складається з нормальних точок, оскільки вона завжди прагне виявити деякий відсоток точок, що відрізняються від основної маси, згідно із заданим параметром “кількість припущених аномалій”. Це може призводити до помилкового виявлення аномалій у випадках, коли дані є однорідними.

Для вирішення першої проблеми, пов’язаної зі зміною шкали ступеня аномальності, доцільно провидити серію запусків моделі ІЛ з різними значеннями параметра “кількість припущених аномалій”, у діапазоні від 0,05 до максимально можливого значення 0,5 з мінімальним кроком 0,01. Це дозволяє от-

римати набір результатів, які відображають ступінь аномальності для кожної точки при різних налаштуваннях чутливості моделі. Оскільки відстань між точками залишається незмінним, а змінюється лише масштаб ступеня аномальності, при отримані результатів розрахунку моделі ІЛ при різній чутливості необхідно усереднювати значення ступеня аномальності кожної точки для всіх запусків моделі. Таким чином, отримані результати відображають стабільні середні значення аномальності, де відстані між точками залишаються незмінними, і шкала ступеня аномальності стає менш залежною від вибору конкретного значення параметра “кількість припущених аномалій”, Такий алгоритм дозволяє вирішити проблему масштабування, зберігаючи відносне положення точок у просторі аномальності.

Для вирішення другої проблеми необхідно проаналізувати результати усереднення та виявити закономірності роботи моделі ІЛ.

Результати запусків алгоритму на основі моделі ІЛ, який усереднює результати запусків при різних значеннях параметра “кількість припущених аномалій” та встановлює порогове значення на основі заданих умов, наведено на рис. 4.11-4.19.

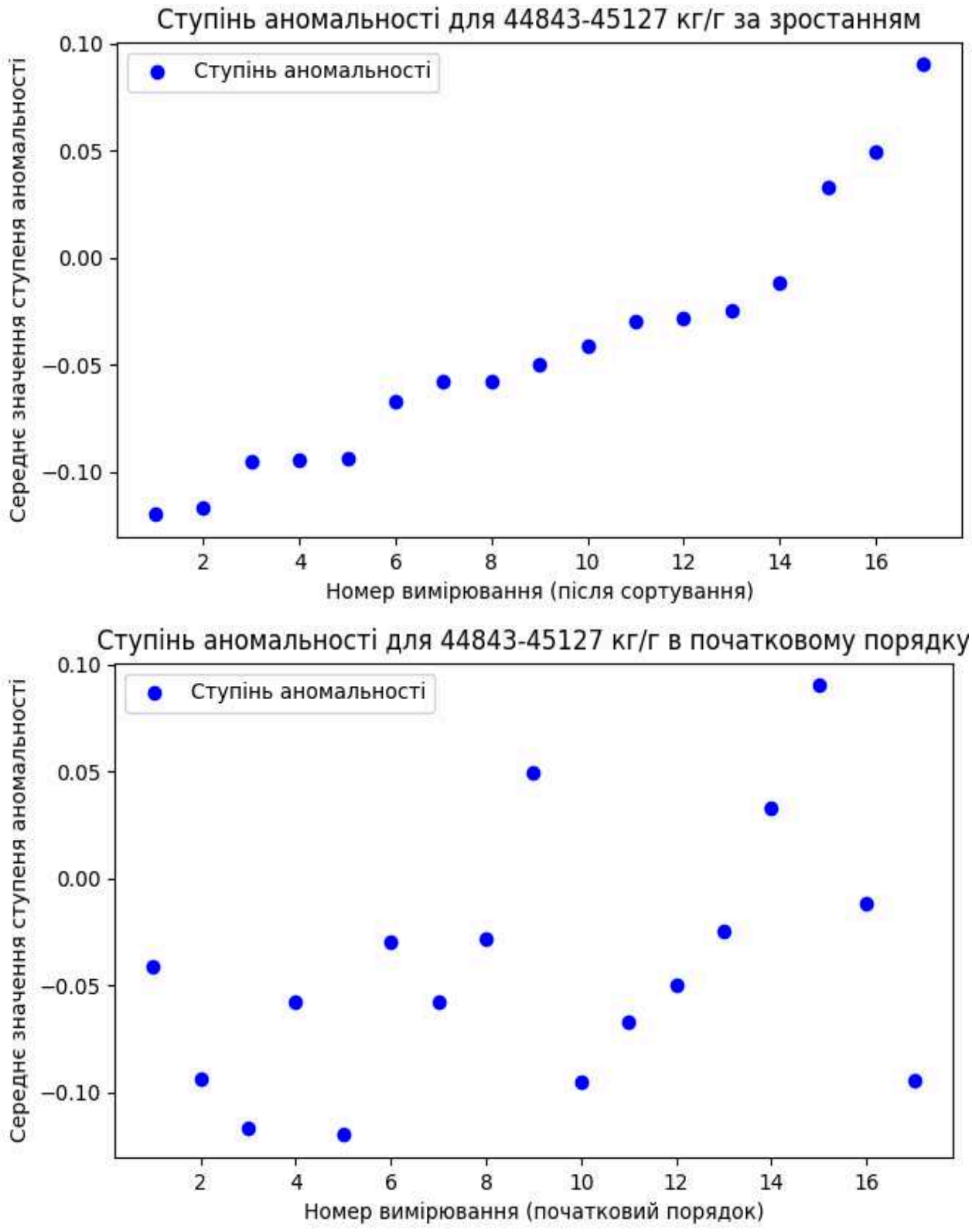

Рисунок 4.11 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 45 т/г витратоміра №1.

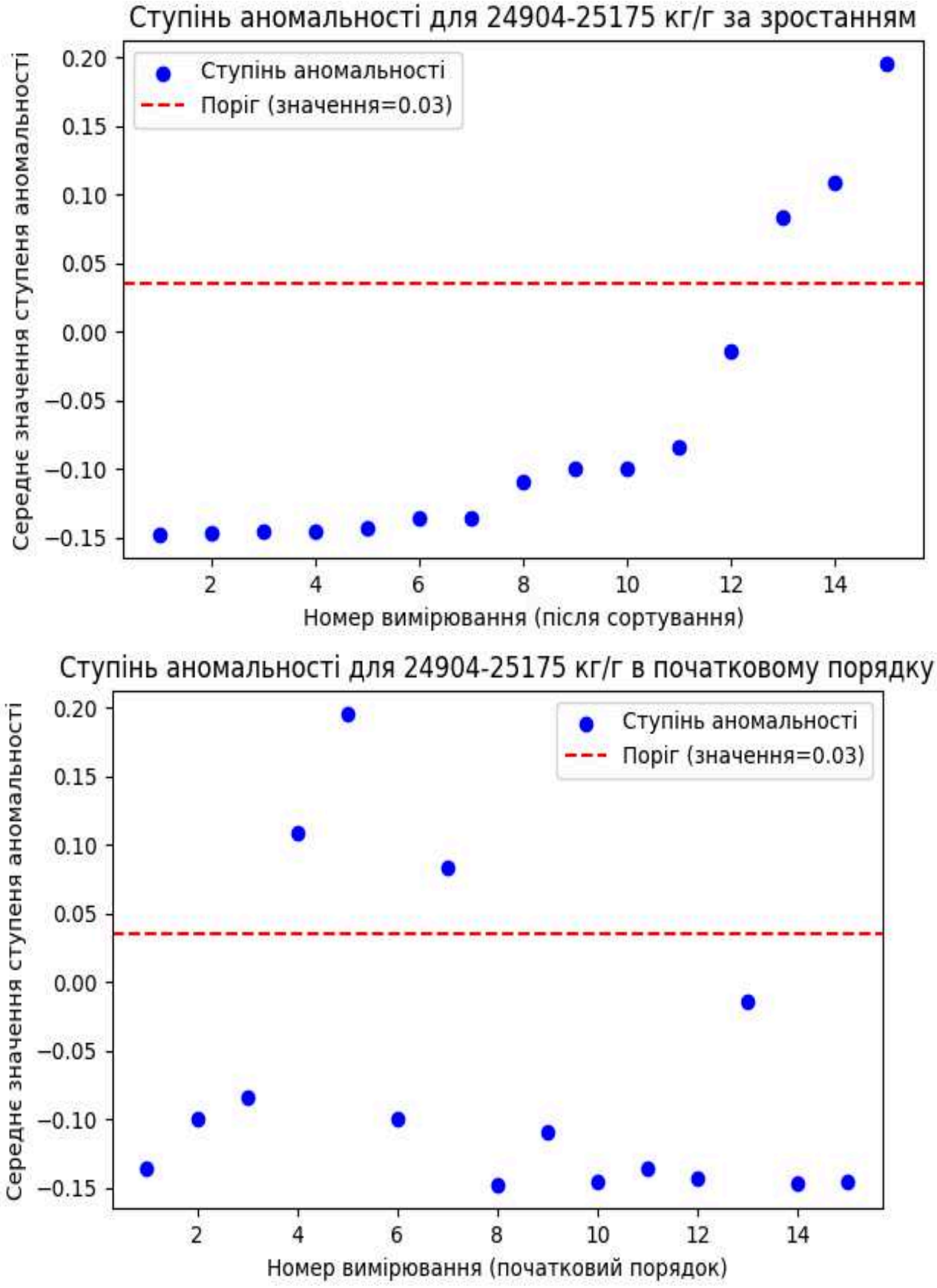

Рисунок 4.12 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г витратоміра №1.

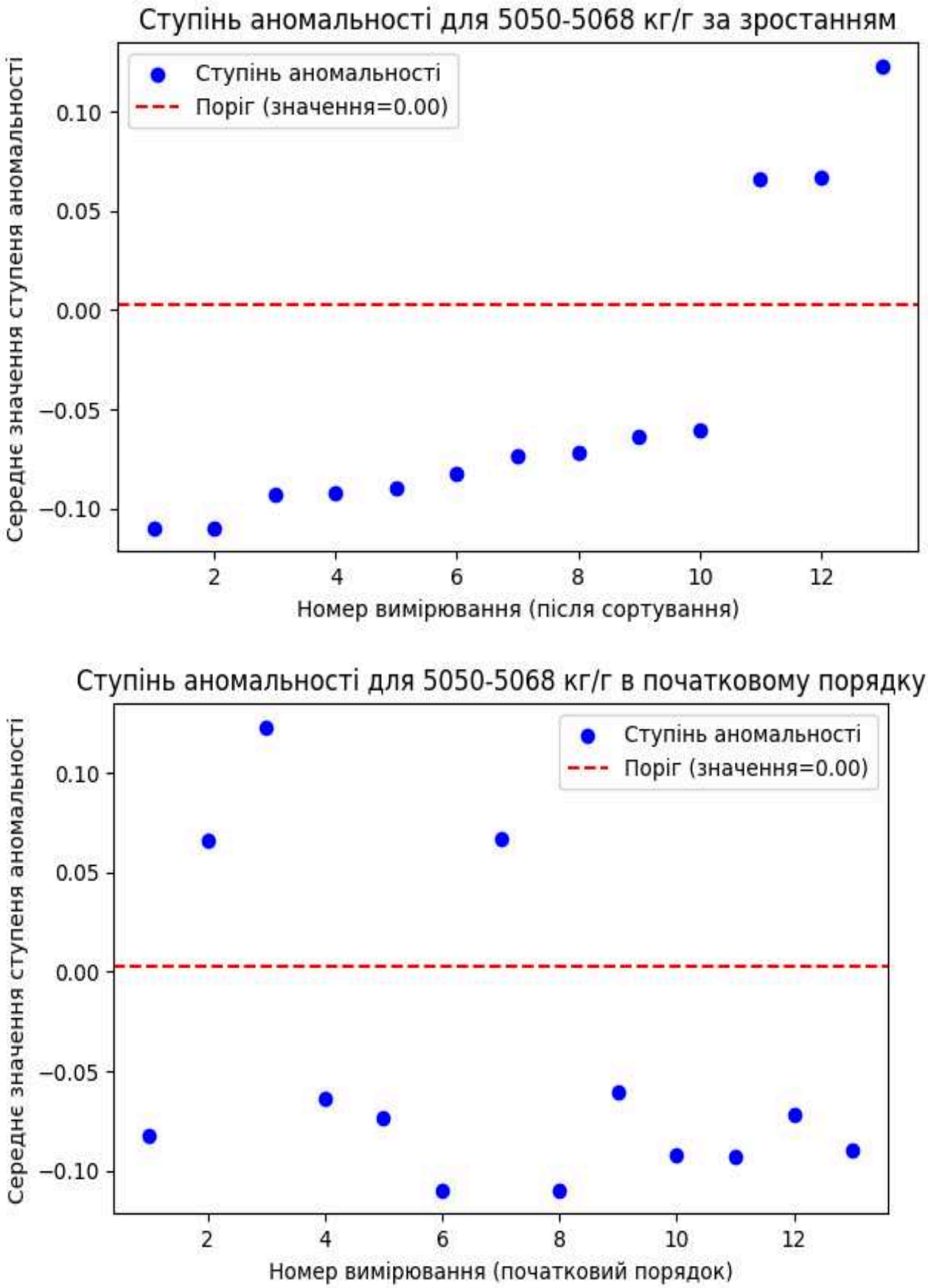

Рисунок 4.13 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 5 т/г витратоміра №1.

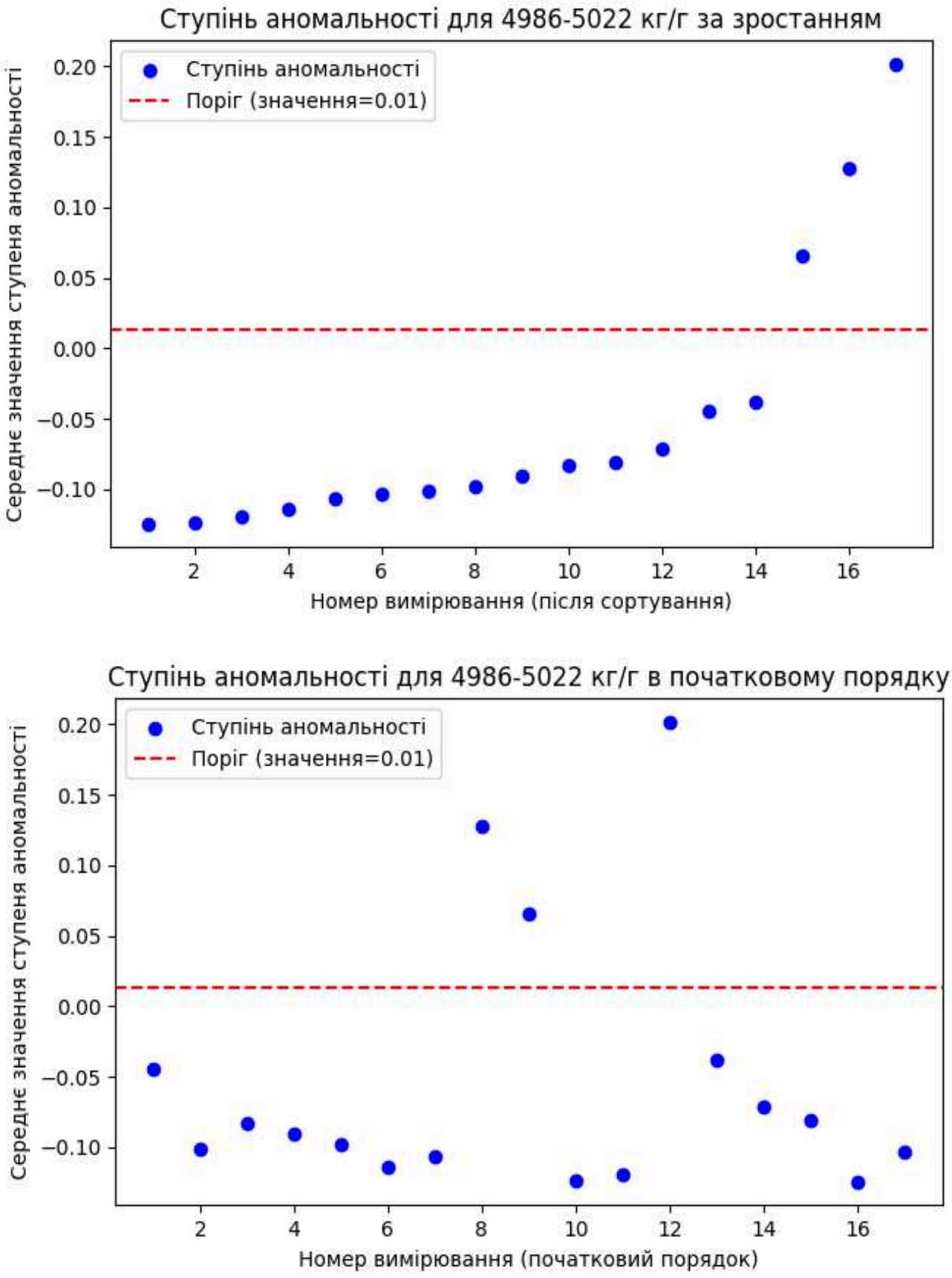

Рисунок 4.14 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 5 т/г витратоміра №2.

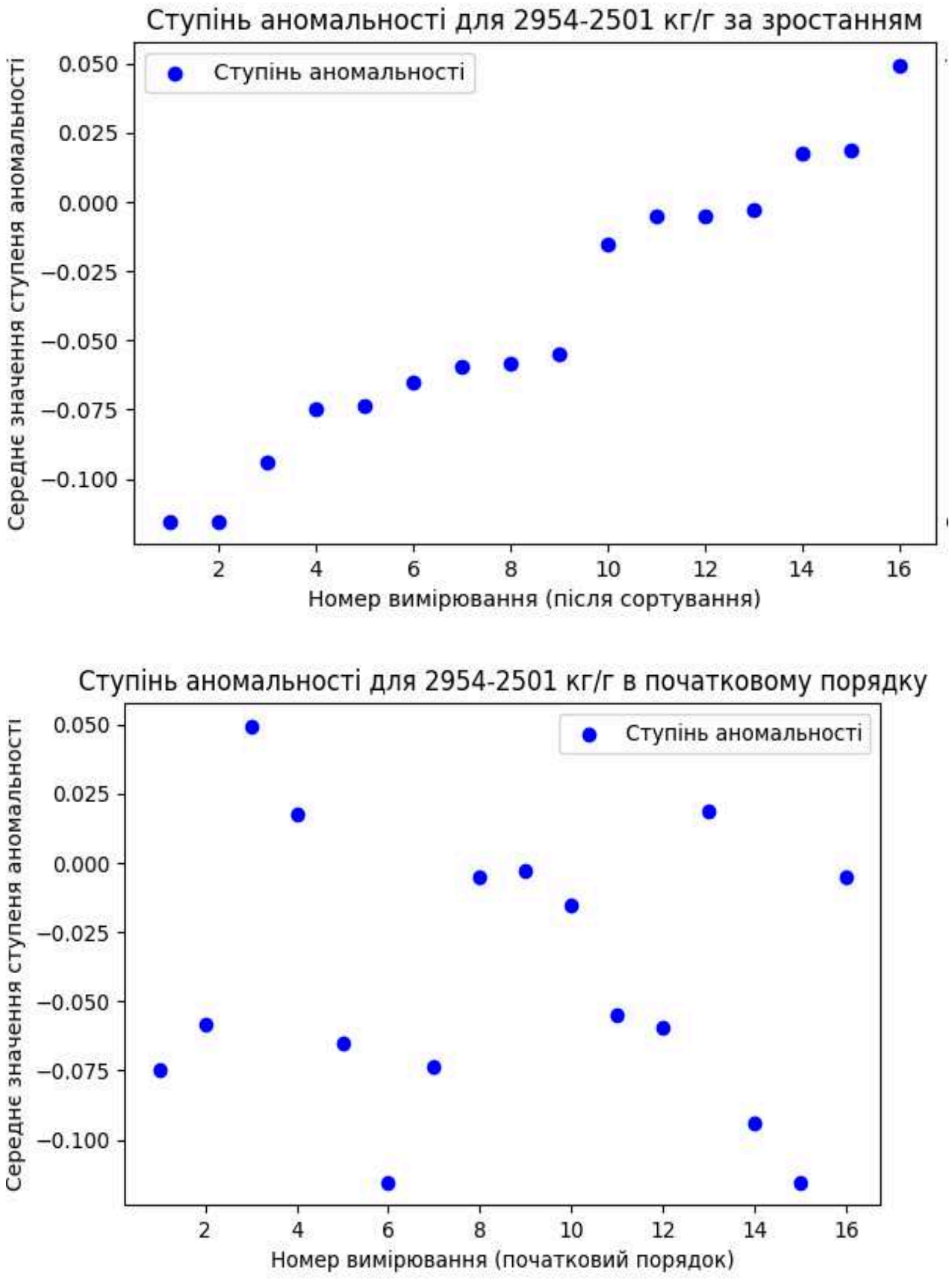

Рисунок 4.15 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 2,5 т/г витратоміра №2.

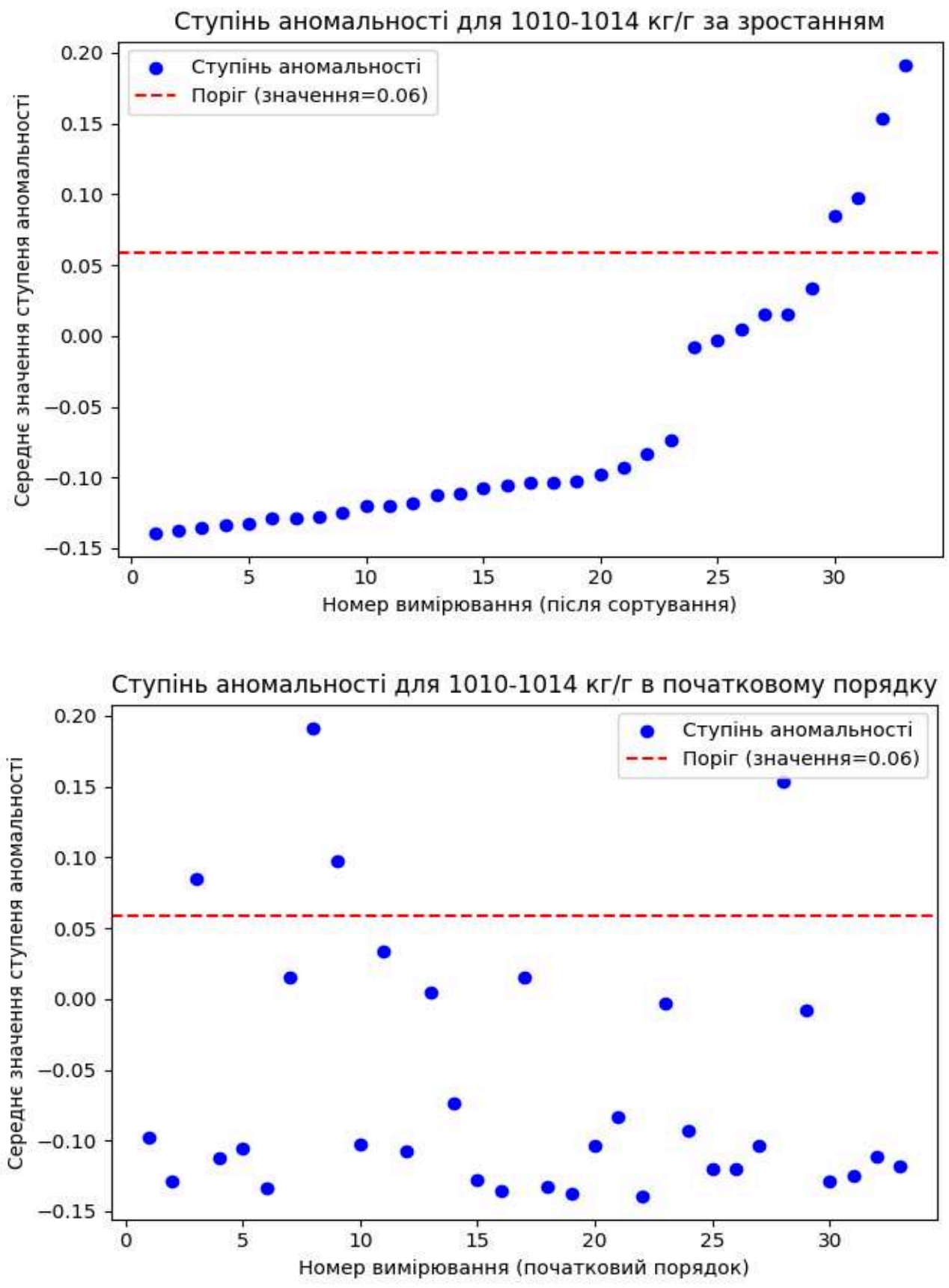

Рисунок 4.16 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 1 т/г витратоміра №2.

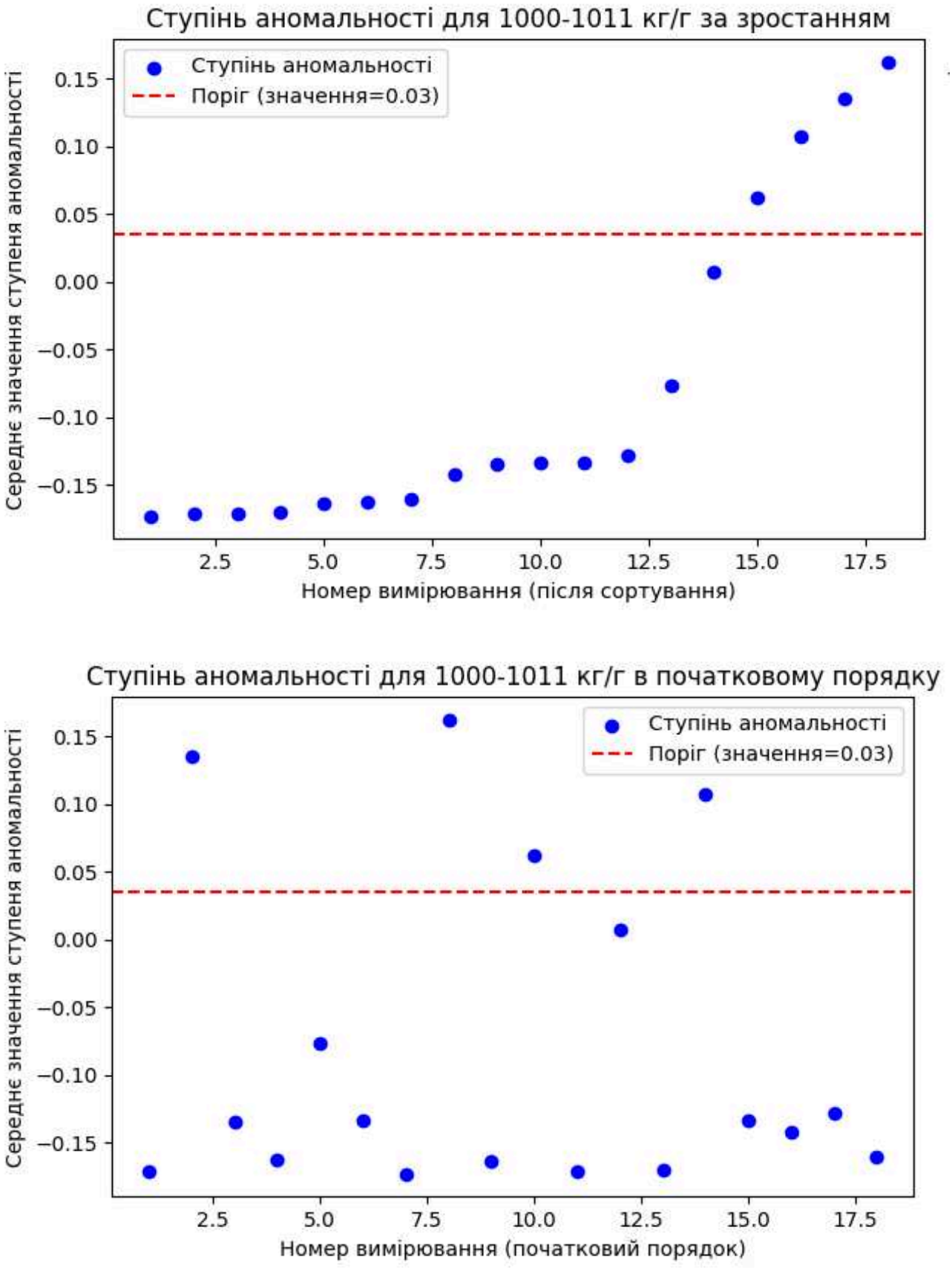

Рисунок 4.17 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 1 т/г витратоміра №3.

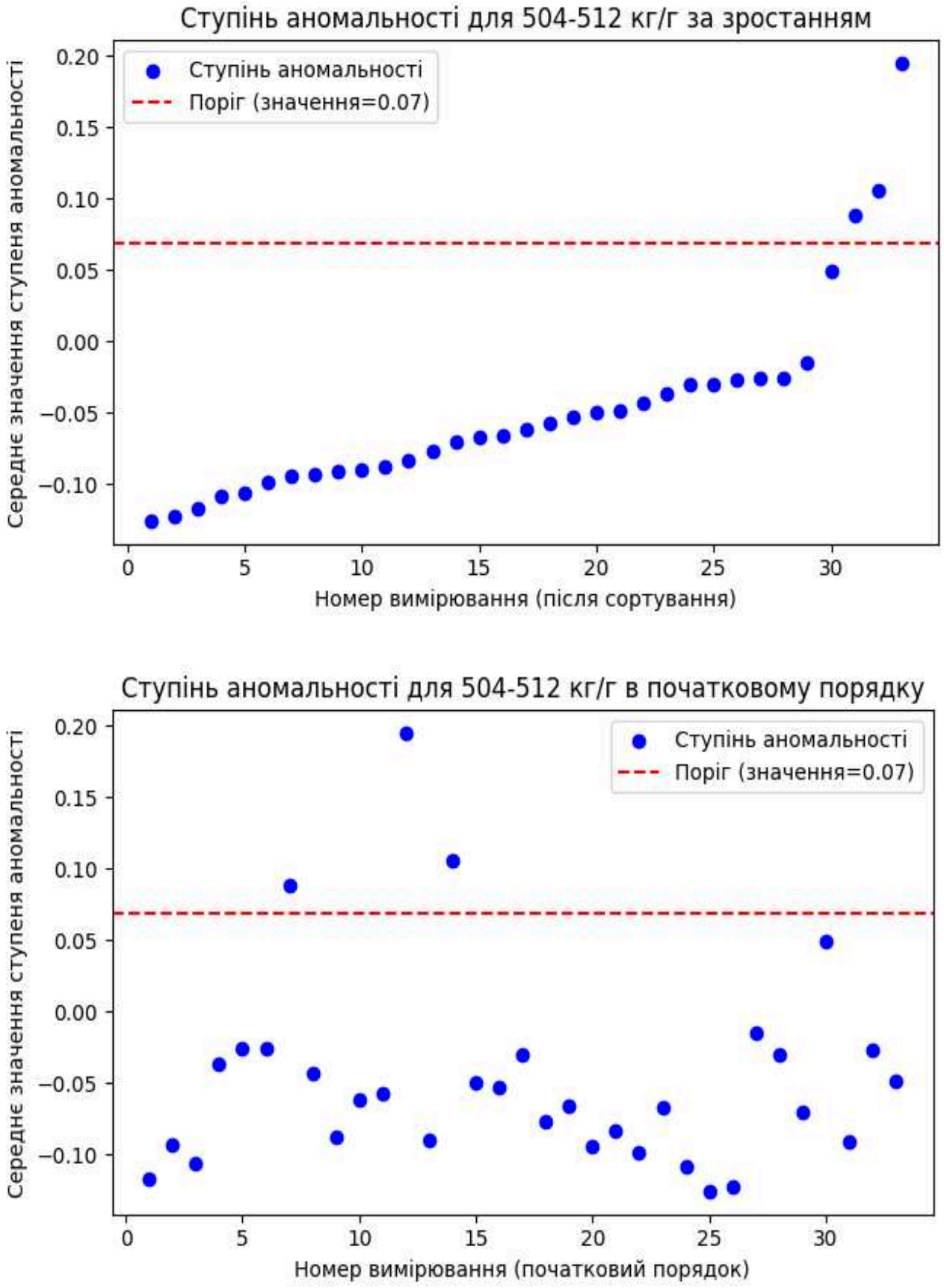

Рисунок 4.18 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 0,5 т/г витратоміра №3.

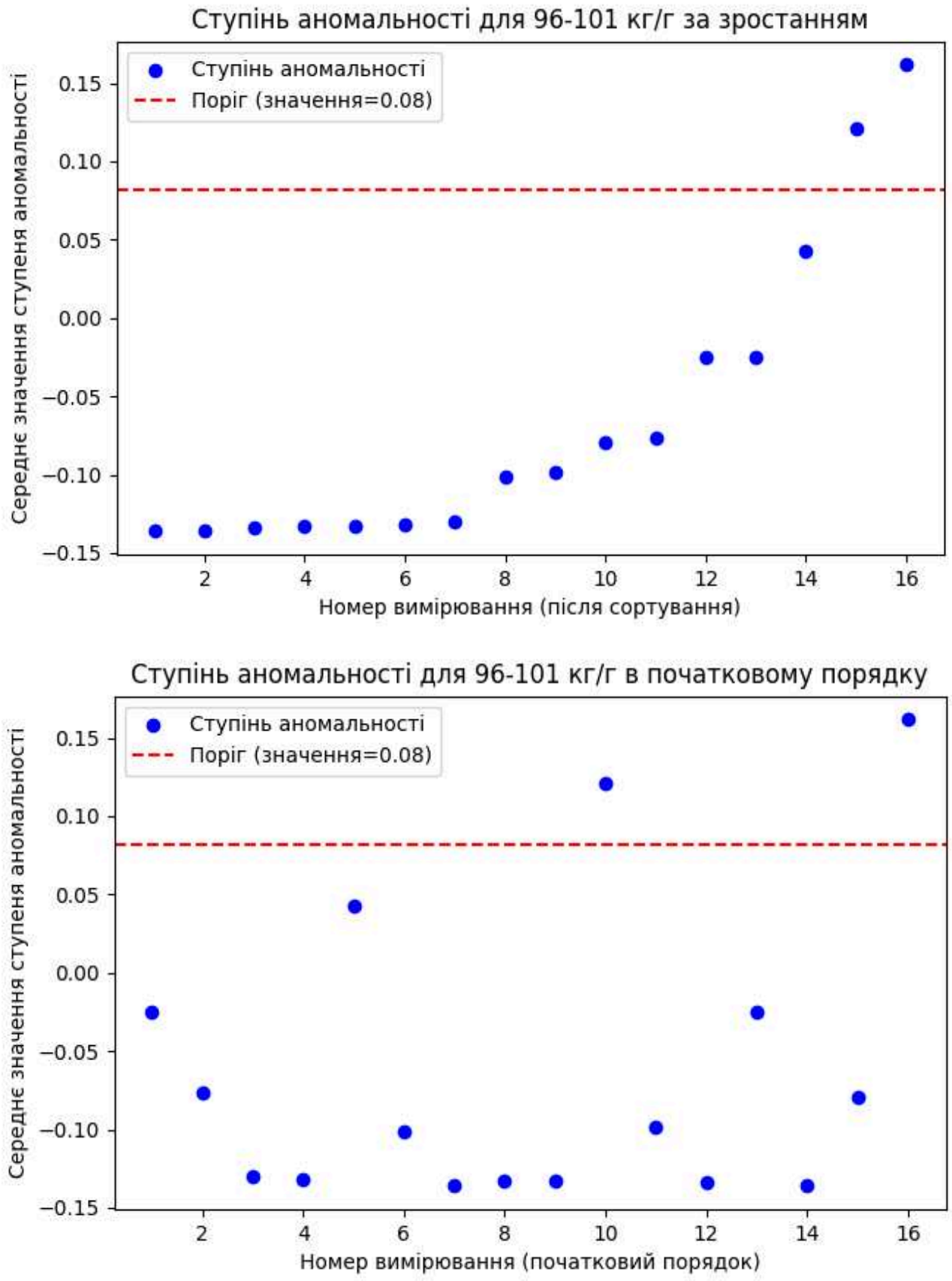

Рисунок 4.19 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 0,1 т/г витратоміра №3.

Візуальний аналіз результатів запусків моделі при різних значеннях параметра "кількість припущених аномалій" дозволяє чітко ідентифікувати області скупчення нормальних точок і точки, що відрізняються за ступенем аномальності. У деяких вибірках виразно спостерігаються аномалії (викиди), які розташовані у верхній частині графіка та мають високі значення ступеня аномальності порівняно з основною масою точок. Зазвичай такі аномальні точки розташовані в діапазоні від 0 до 0,25 за шкалою аномальності, що свідчить про ефективність підходу усереднення значень ступеня аномальності для різних значень параметра "кількість припущених аномалій". Це усереднення дозволяє зменшити вплив зміни масштабу і стабілізує відлік відносно нуля.

Для вибірок з однорідними даними, де немає вираженого розмежування між нормальними та аномальними точками, значення ступеня аномальності зазвичай обмежені рівнями до 0,05 і нижче нуля. Це свідчить про відсутність значних відхилень, що дозволяє вважати всі точки такими, що не містять аномалій. Враховуючи це, доцільно встановити порогове значення для виявлення аномалій на рівні, вищому за 0,05, оскільки значення нижче цього порогу зазвичай відповідають нормальним точкам. Це узгоджується з логікою роботи моделі, яка класифікує точки з низькими значеннями ступеня аномальності як нормальні при малих значеннях параметра "кількість припущених аномалій".

На основі статистичних розрахунків у розділі 3.6, де було визначено, що максимальна кількість викидів у вибірці становить 33%, можна умовно вважати, що близько 60% значень кожної вибірки є нормальними точками, що не містять аномалій. Це дозволяє при сортуванні значень ступеня аномальності за зростанням виділити перші 60% значень як нормальні, які формують базу для визначення нормального поведінкового патерну.

Для оцінки "нормальної" відстані між точками в цій базі пропонується розрахувати максимальну відстань серед перших 60% відсортованих значень. Це максимальне значення відстані можна вважати репрезентативним для нормальної поведінки вибірки. Оскільки аномалії (викиди) характеризуються

значними відхиленнями від нормальних значень, доцільно використовувати подвоєне значення цієї максимальної відстані як порогове для виявлення аномалій. Якщо відстань між точкою і її найближчими сусідами перевищує встановлений поріг, точку можна вважати аномальною. Водночас порогове значення повинно бути більшим за 0,05 за шкалою ступеня аномальності, що забезпечує коректність класифікації та дозволяє уникнути помилкових виявлень аномалій у випадках однорідної структури даних.

Запропоновані порогові значення, такі як 60% відсортованих значень вибірки для виділення нормальних точок та подвоєне значення максимальної відстані для визначення аномалій, можуть коригуватися з часом. Постійне застосування алгоритму дозволить накопичувати статистичні дані, які можуть бути використані для подальшого уточнення порогів відповідно до особливостей конкретного набору даних. Такий підхід підвищує адаптивність алгоритму, що дозволяє враховувати відмінності у структурі та характеристиках нових вибірок, а також покращує загальну ефективність виявлення аномалій.

У новому алгоритмі також змінено орієнтацію значень порівняно зі стандартною моделлю Isolation Forest. Тепер аномальні значення розташовані вище нуля, а нормальні — нижче. Така орієнтація полегшує візуальний аналіз результатів, знижуючи перенасиченість графіка від'ємними значеннями та роблячи його більш зрозумілим і наочним.

4.5 Алгоритм на основі моделі ізольованого лісу

Зважаючи на можливість застосування моделі ІЛ для роботи з невеликими вибірками, важливо врахувати, що вибірка все ж не повинна бути надто малою. Зокрема, вибірки обсягом 3-5 значень є недостатніми, оскільки результати, отримані на таких малих обсягах даних, можуть бути некоректними. Таким чином, модель ІЛ рекомендовано застосовувати для калібрування високоточних витратомірів або під час проведення міжнародних звірень, де забезпечується

мінімально необхідна кількість вимірювань — щонайменше 10 значень. Це забезпечує достатній обсяг даних для адекватного аналізу та виявлення викидів, відповідаючи вимогам точності, необхідної для метрологічних досліджень.

Використання графічного відображення аномалій у вигляді графіка зі шкалою аномальності дозволяє візуалізувати структуру даних і побачити, як кожна точка відрізняється від інших. Це забезпечує можливість визначення кількості потенційних викидів у вибірці. Завдяки такому підходу можна візуально оцінити, які точки суттєво відхиляються від основного скупчення значень, що значно полегшує процес ідентифікації аномалій на ранньому етапі.

Для аналізу змін структури даних при додаванні нових значень було проведено серію запусків алгоритму ІЛ на вибірці з масовою витратою 25 т/г. Спочатку вибірка складалася з 5 вимірювань, після чого послідовно додавалися нові значення з подальшим повторним запуском моделі після кожного додавання. Це дозволило відстежити, як змінюється структура даних і як нові точки впливають на загальну оцінку аномальності вибірки.

Результати запусків представлені на рисунках 4.20–4.30, де кожен графік демонструє зміну ступеня аномальності та розподіл точок після кожного додавання нового вимірювання. Такий підхід забезпечує ініціалізацію процесу вимірювання та дозволяє провести попередній аналіз для оцінки кількості викидів.

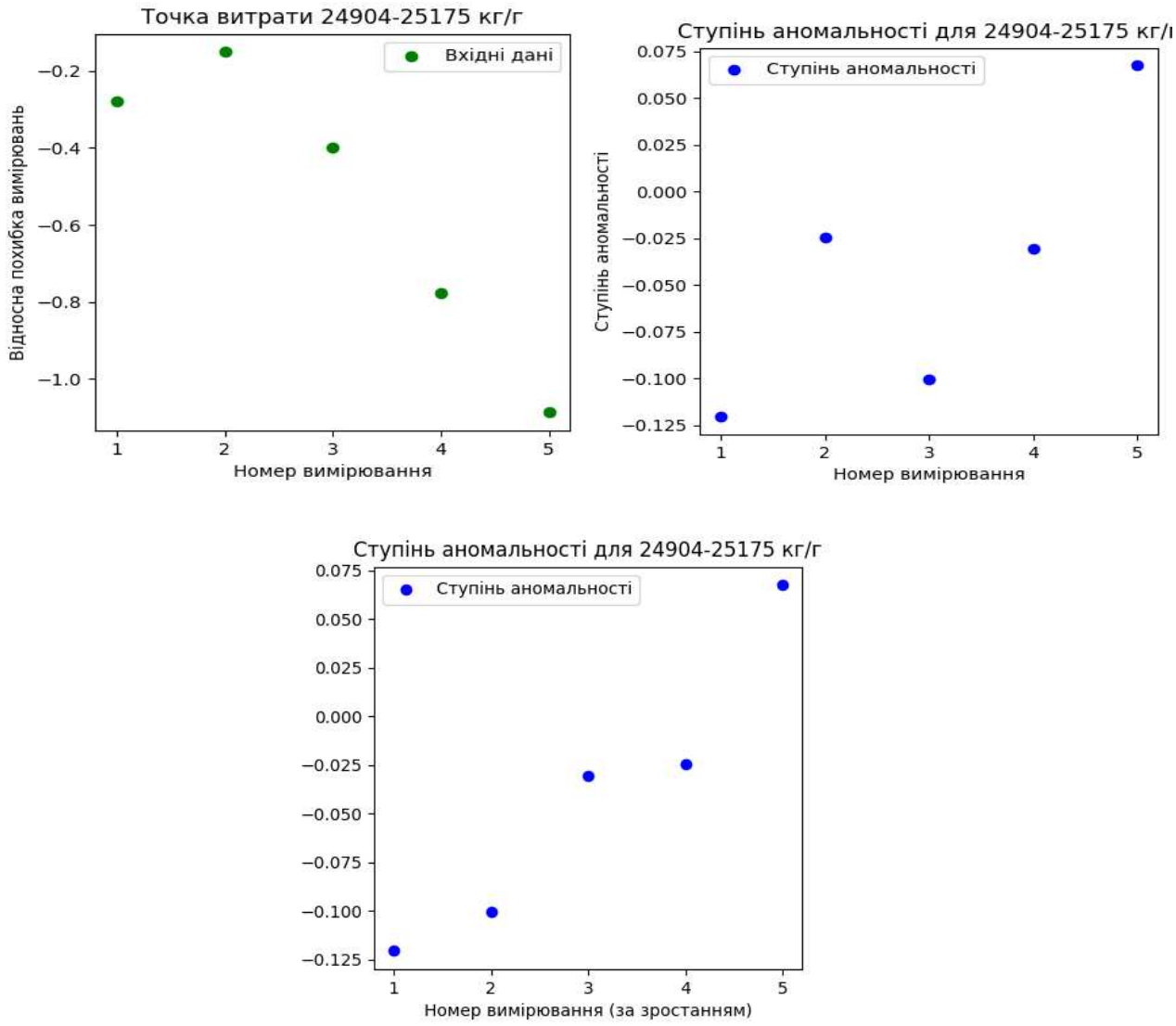

Рисунок 4.20 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 5).

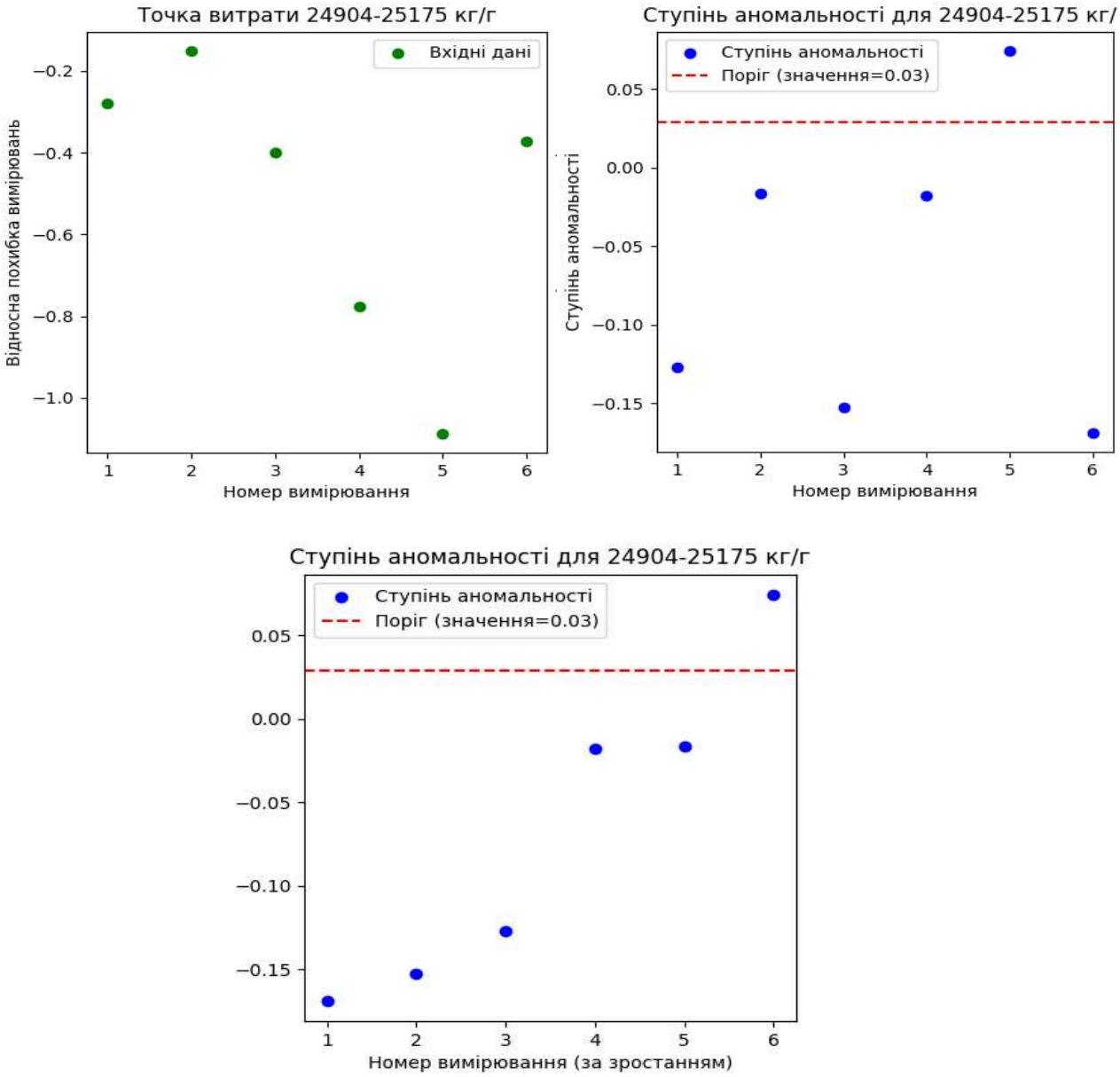

Рисунок 4.21 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 6).

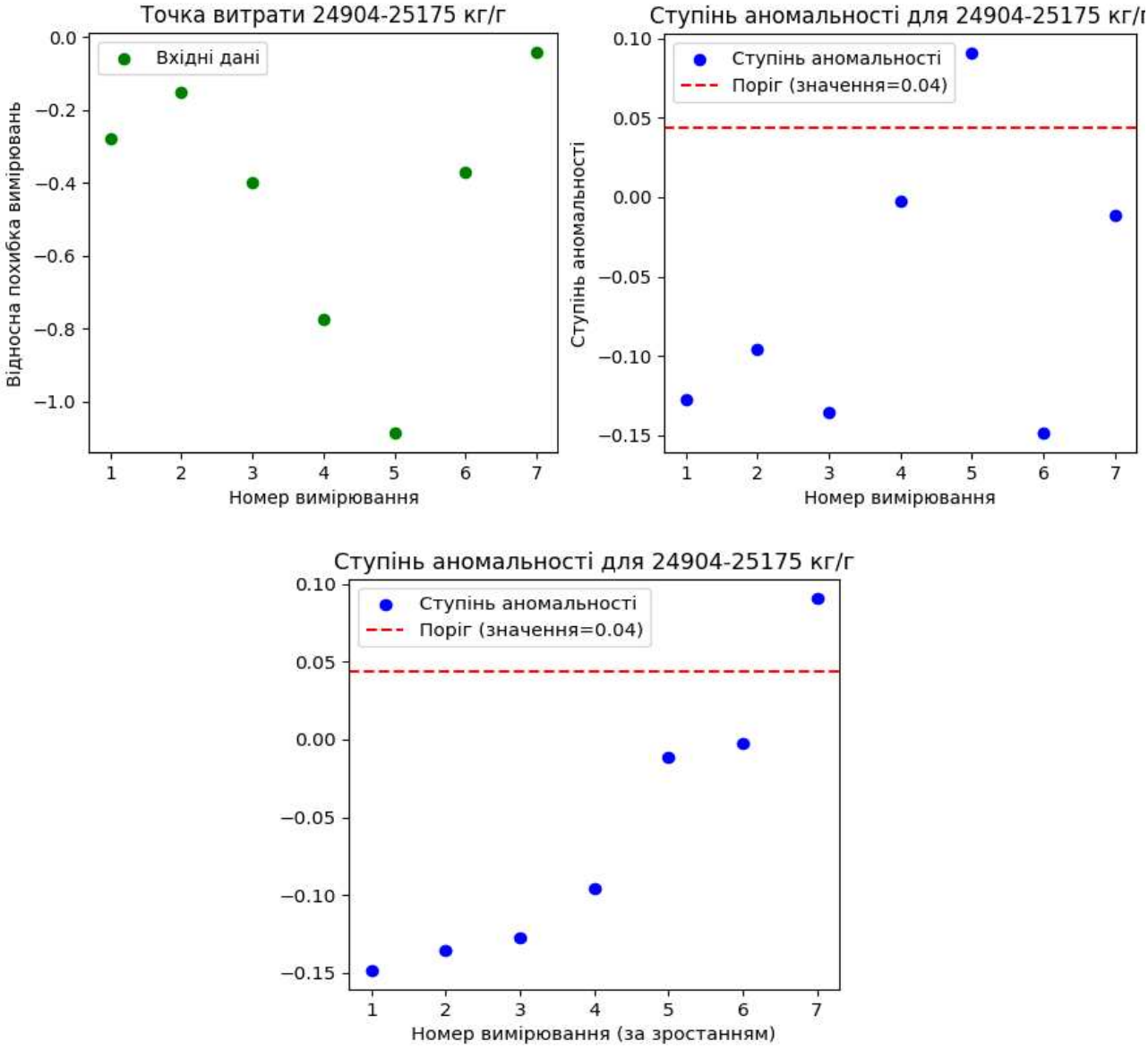

Рисунок 4.22 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 7).

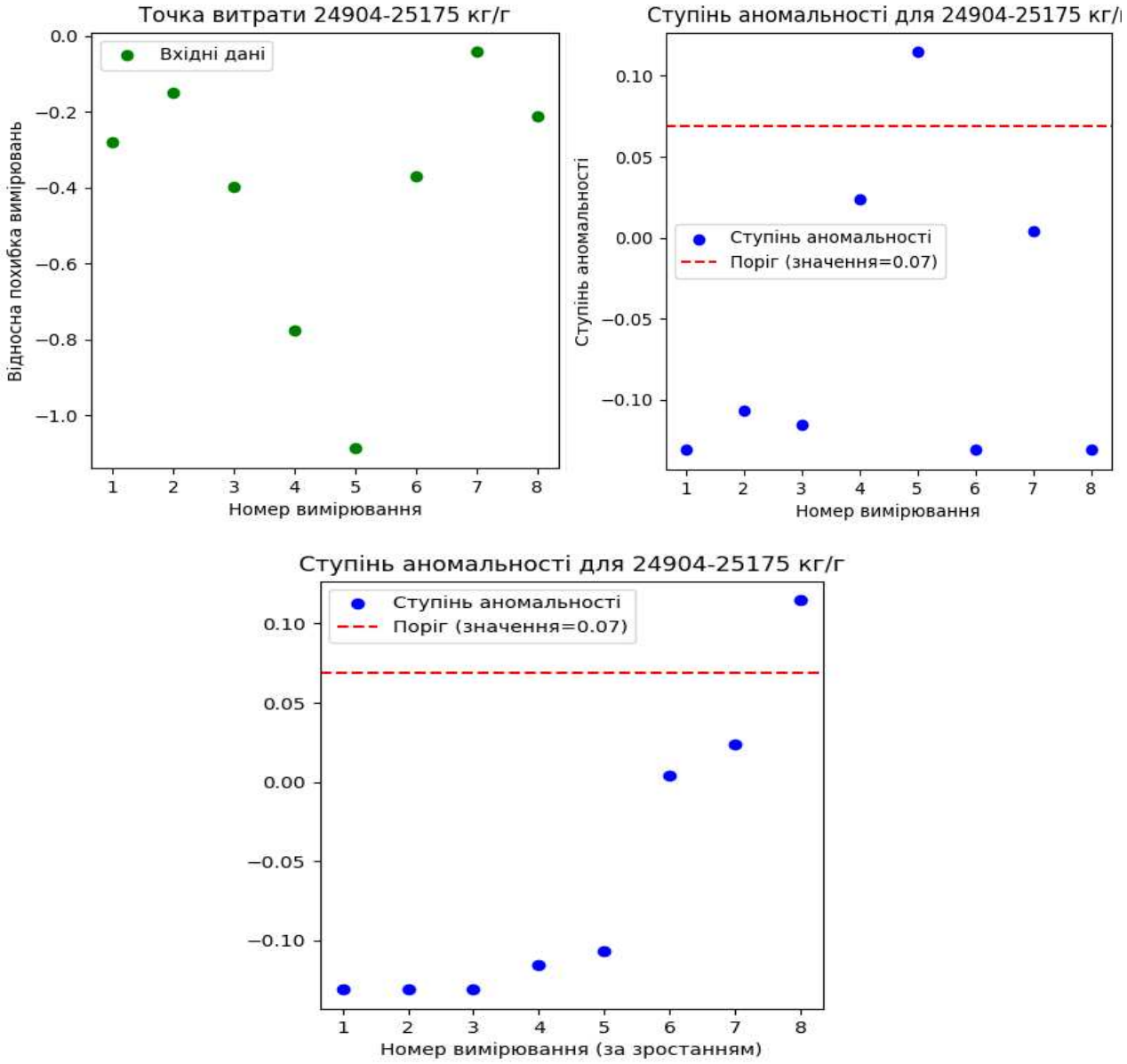

Рисунок 4.23 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 8).

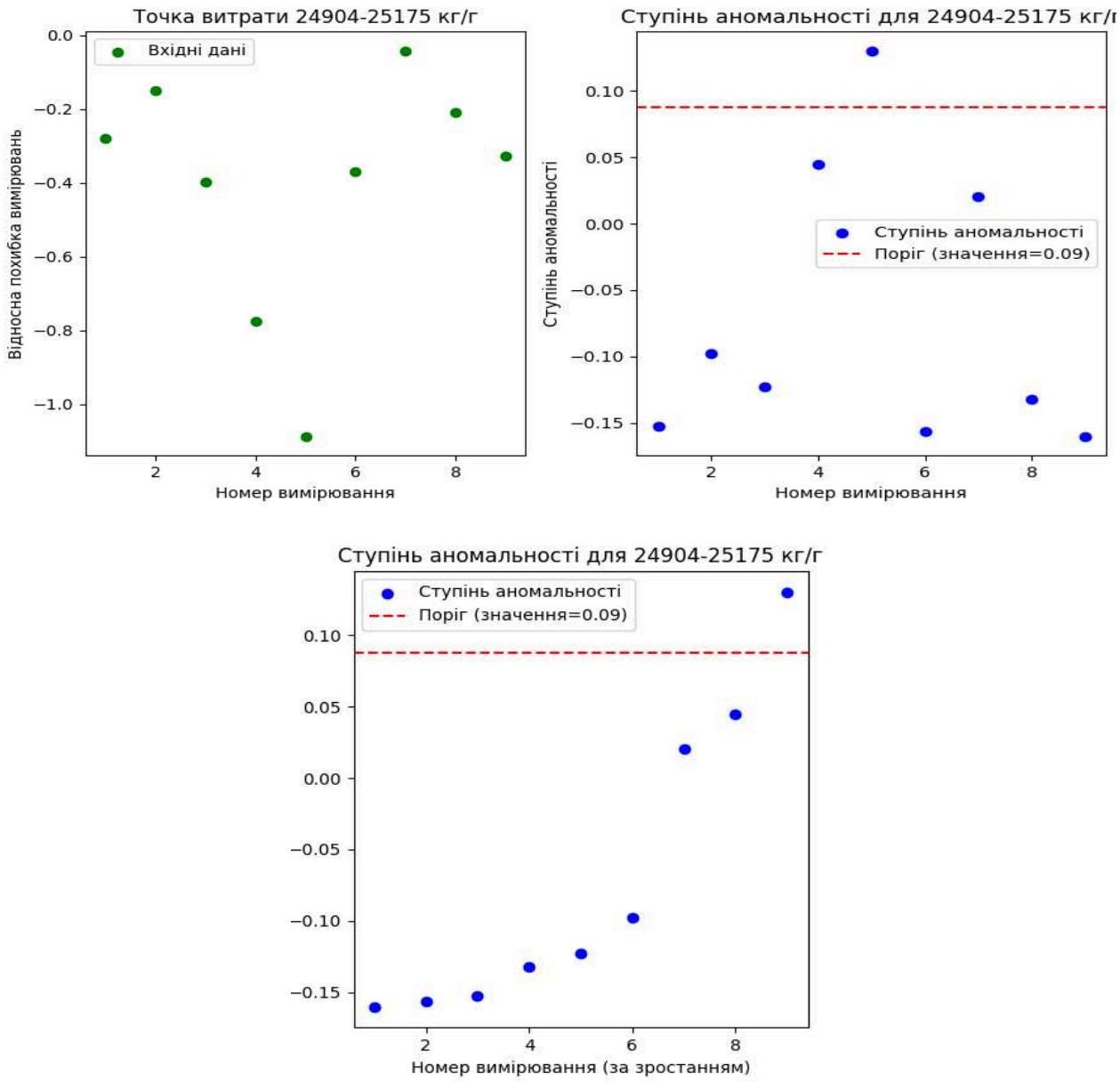

Рисунок 4.24 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 9).

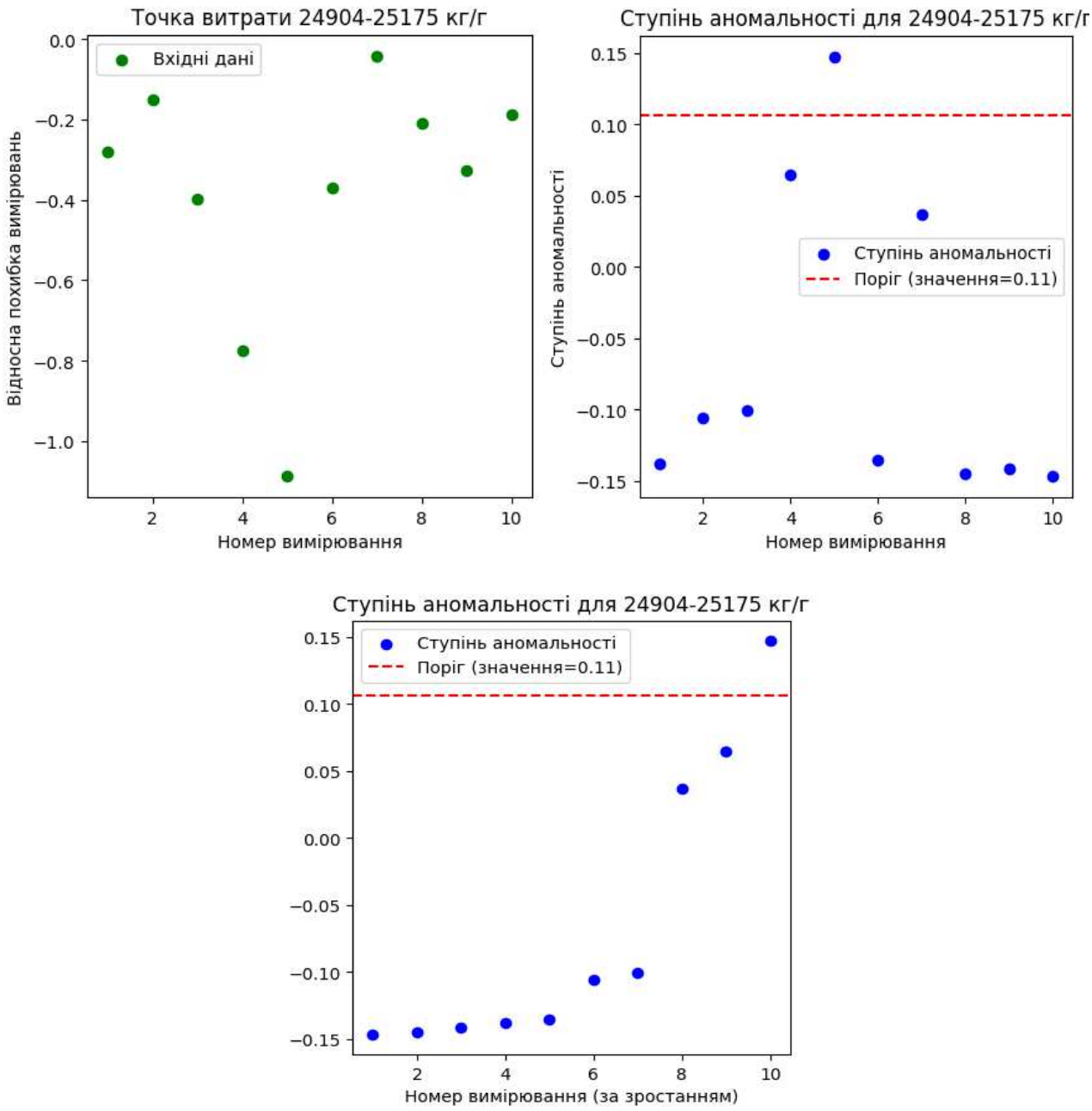

Рисунок 4.25 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 10).

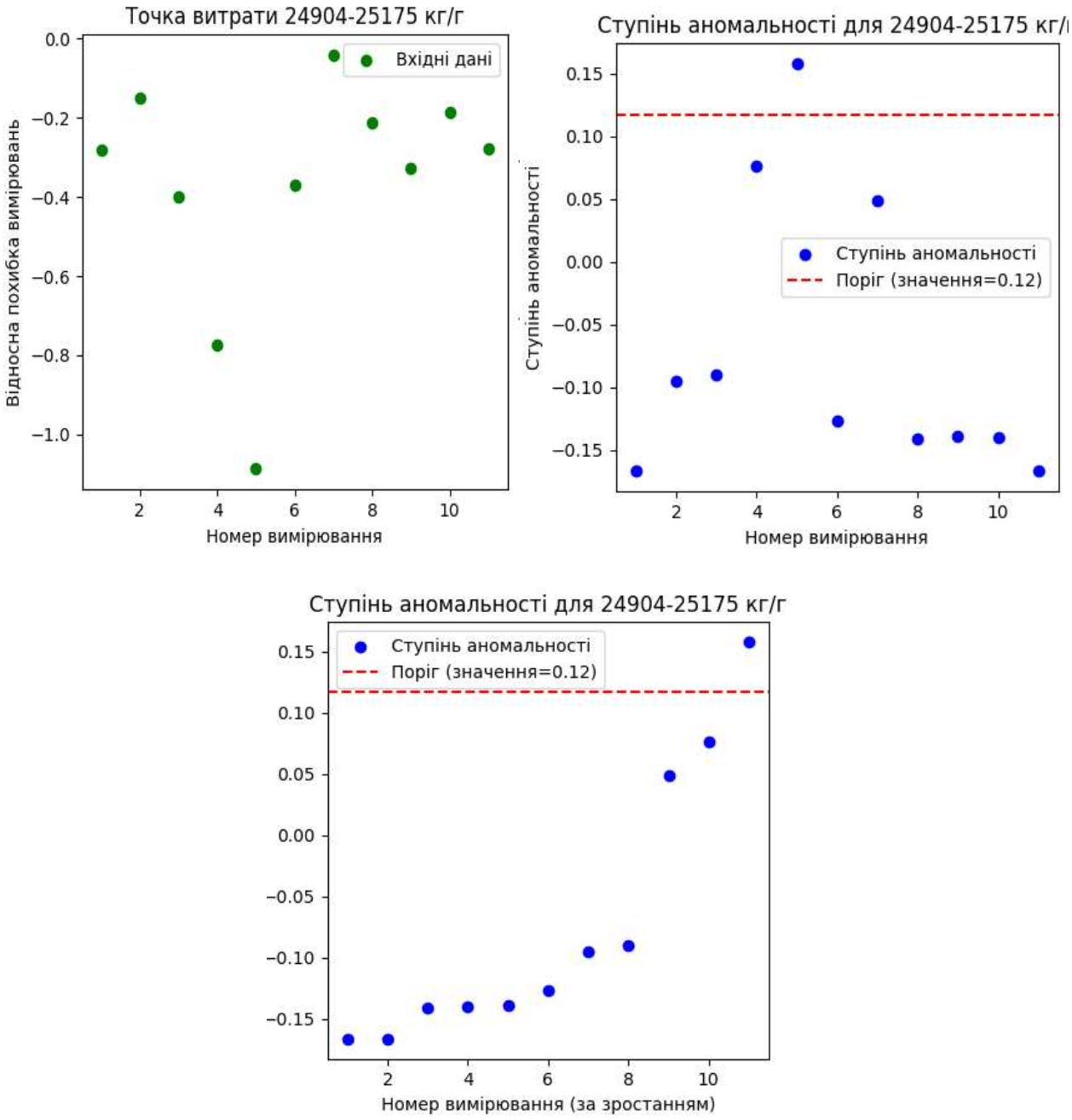

Рисунок 4.26 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 11).

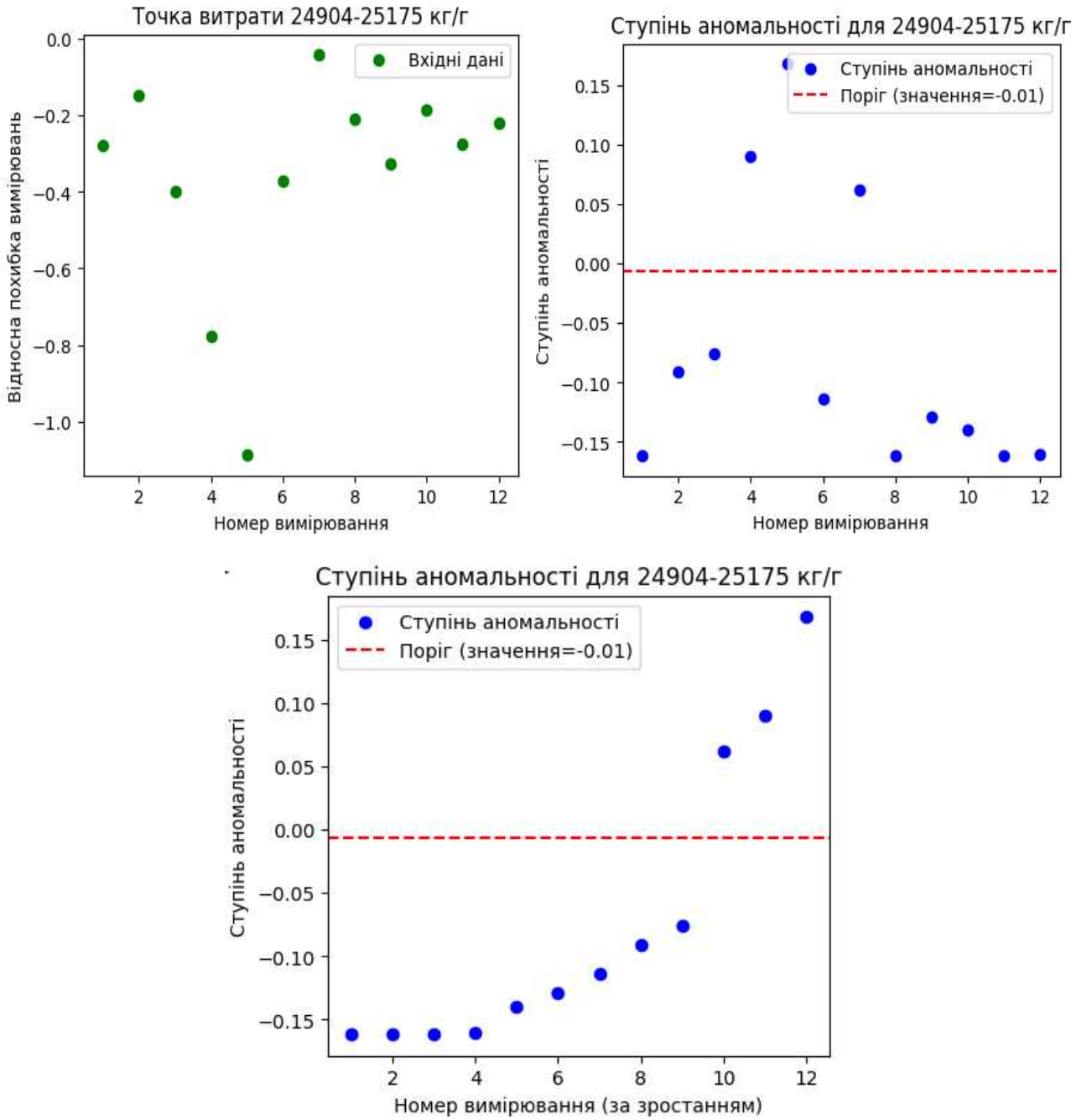

Рисунок 4.27 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 12).

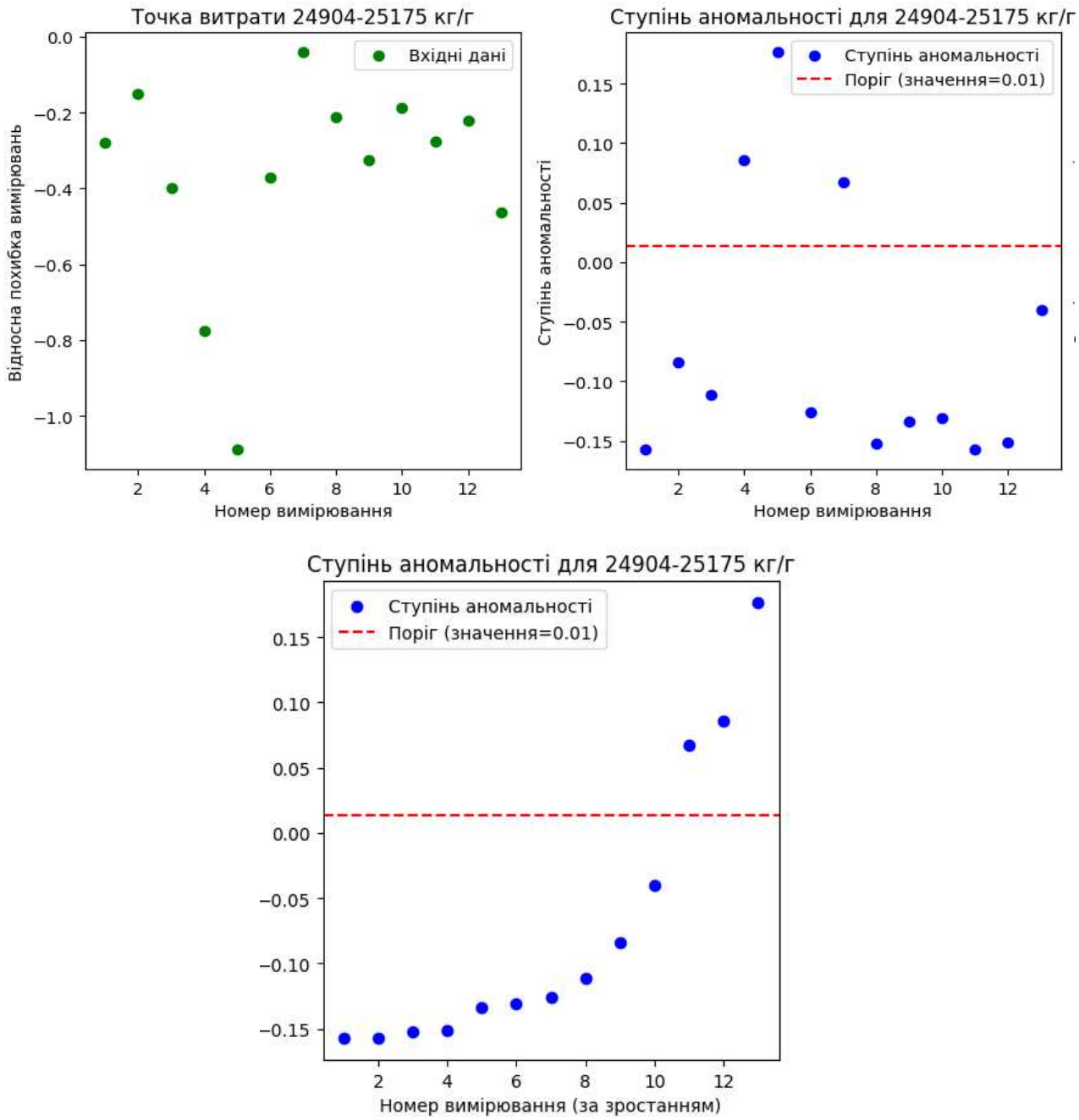

Рисунок 4.28 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 13).

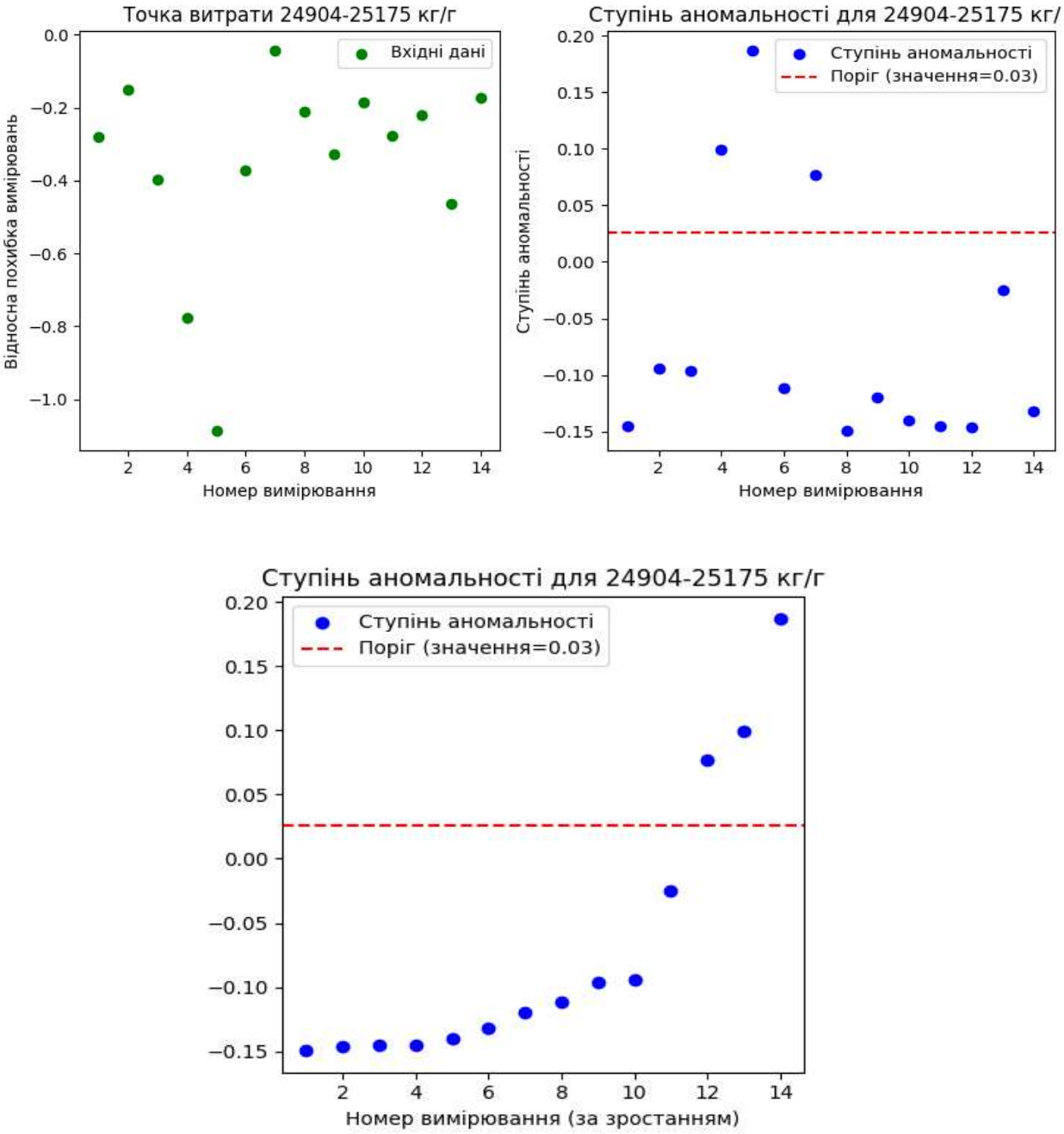

Рисунок 4.29 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 14).

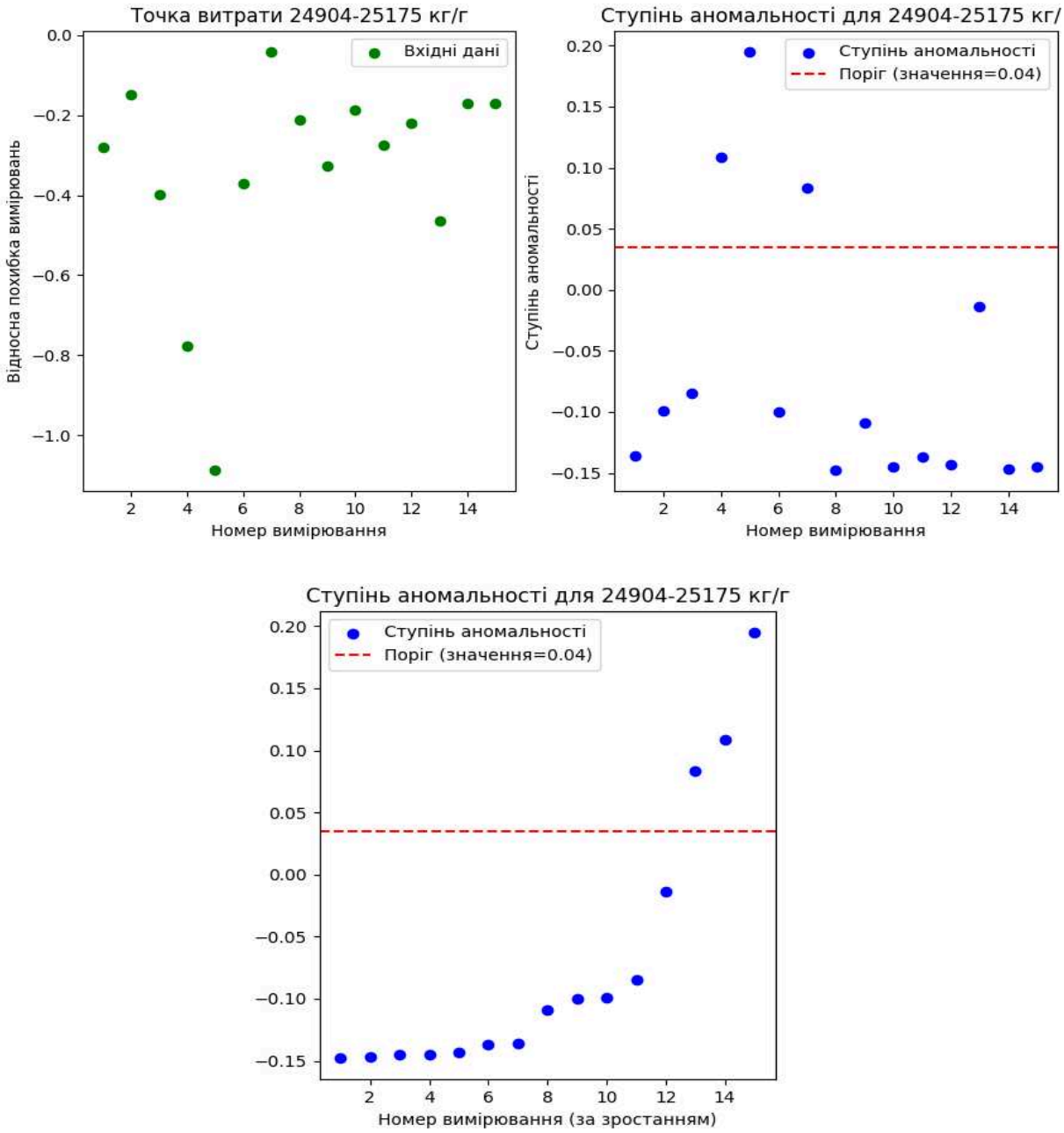

Рисунок 4.30 – Результат запуску алгоритму на основі моделі ІЛ для точки витрати 25 т/г (кількість вимірювань — 15).

При аналізі результатів алгоритму ІЛ для виявлення аномалій важливо звертати увагу на такі ключові аспекти:

1. **Шкала аномальності.** Значення на шкалі аномальності, що перевищують 0, а особливо 0,5, можуть свідчити про потенційні викиди, навіть якщо алгоритм не визначив їх як аномальні. Такі значення потребують ретельного аналізу, оскільки вони можуть сигналізувати про відхилення,

яке не досягло порогового рівня, але має підвищений ризик бути аномалією. Шкала аномальності забезпечує можливість оцінки як явних, так і прихованих потенційних викидів, що є важливим для попереднього аналізу;

2. **Візуальне відокремлення точок від основного скупчення.** Значне відокремлення окремих точок від основної групи значень може бути індикатором аномалій. Якщо результати мають однорідну структуру, і не спостерігається яскраво вираженого відокремлення точок, це свідчить про стабільність вимірювань і низький рівень аномальності. У випадках, коли певні точки значно відхиляються від основного скупчення, варто звернути увагу на ці точки як на потенційні викиди і провести додатковий аналіз;
3. **Динаміка змін при додаванні нових вимірювань.** Порівняння структури аномальності у вибірках при додаванні нових точок дозволяє оцінити, чи є потенційні викиди стійкими або зникають при збільшенні обсягу вибірки. У випадках, коли значення аномальності знаходяться на межі порогового значення, доцільно проводити додаткові вимірювання для підтвердження або спростування гіпотези щодо їхньої аномальності;
4. **Порогові умови для виявлення аномалій.** Визначення порогового значення базується на двох умовах: значення аномальності повинно перевищувати 0,5, а також повинно бути перевищення вдвічі максимального відстані у перших 60% нормальних точок. У разі, коли відстані у 60% точок є великими, можливо, вибірка є нерівномірною і потребує додаткового аналізу. Якщо ж відстані є невеликими, це свідчить про однорідність вибірки і підвищує довіру до відсутності аномалій.

У міжнародній практиці, під час звірень у напрямку масової витрати рідини, зазвичай виконують 11 повторних вимірювань для певного рівня витрати рідини. При 11 вимірюваннях на точці витрати 25 т/г (рис. 4.26) алгоритм ІЛ визначив один явний викид. Крім того, дві точки мають значення аномальності, що перевищують 0, але не досягли значного рівня аномальності (понад 0,5). Ці

значення знаходяться на межі порогового рівня, тому на цьому етапі їх не можна однозначно класифікувати як аномалії, але вони потребують додаткової перевірки.

На рис. 4.27 представлено результати після додаткового вимірювання, що збільшує кількість спостережень до 12. На цьому етапі алгоритм визначив уже три аномальні точки: один явний викид, а також два потенційних, які досягли порогового значення і були класифіковані як аномалії. Ці три точки значно відокремлені від основного скупчення, що вказує на необхідність подальшого їх аналізу.

Після отримання таких результатів було прийнято рішення провести ще два додаткових вимірювання, збільшивши загальну кількість спостережень до 14. Результати після цих вимірювань підтвердили, що три точки залишаються аномальними, а їхні значення на шкалі аномальності ще більше відокремилися від основного скупчення. Це надає впевненість у їхній природі як викидів, тому ці три значення виключаються з набору даних, залишаючи 11 значень, які необхідні для звірень.

Аналогічним чинном було проведено запуски алгоритму для всіх інших вибірок при 11 вимірюваннях та аналогічно проведені необхідні додаткові вимірювання для виключення викидів та отримання 11 необхідних 11 вимірювань після виключення викидів.

Для оцінки точності вимірювань і перевірки, чи змінюється розкид похибок, тобто чи підвищується стабільність та надійність результатів, було розраховано стандартну невизначеність за типом А до та після виключення викидів. Стандартна невизначеність за типом А показує ступінь розсіювання похибок навколо середнього значення і є індикатором випадкових відхилень у вимірюваннях. Чим меншим є розкид похибок, тим стабільніші та надійніші результати вимірювань, що є важливим для досягнення високої точності.

Для розрахунку стандартної невизначеності вимірювання за типом А використовується статистичний метод обчислення розсіювання значень навколо середнього. Стандартна невизначеність за типом А визначається через стандар-

тне відхилення вимірювань і дає змогу оцінити випадкові коливання результатів (формула 2.2).

Результати розрахунку представлені в таблиці 4.2

Таблиця 4.2 - Стандартна невизначеність вимірювань за типом А

Стандартна невизначеність вимірювань за типом А,%	Точка масової витрати								
	Витратомір №1			Витратомір №2			Витратомір №3		
	45т/г	25т/г	5 т/г	5 т/г	2,5т/г	1 т/г	1 т/г	0,5т/г	0,1т/г
До виключення викидів	0.0330	0.0911	0.0202	0.0202	0.0312	0.0239	0.0325	0.0236	0.0607
Після виключення викидів	0.0225	0.0308	0.0202	0.0177	0.0312	0.0202	0.0131	0.0212	0.0433
Кількість викидів (розмір вибірки – 11)	1	3	0	1	0	1	2	1	2

Значення таблиці відображені на рис 4.31.

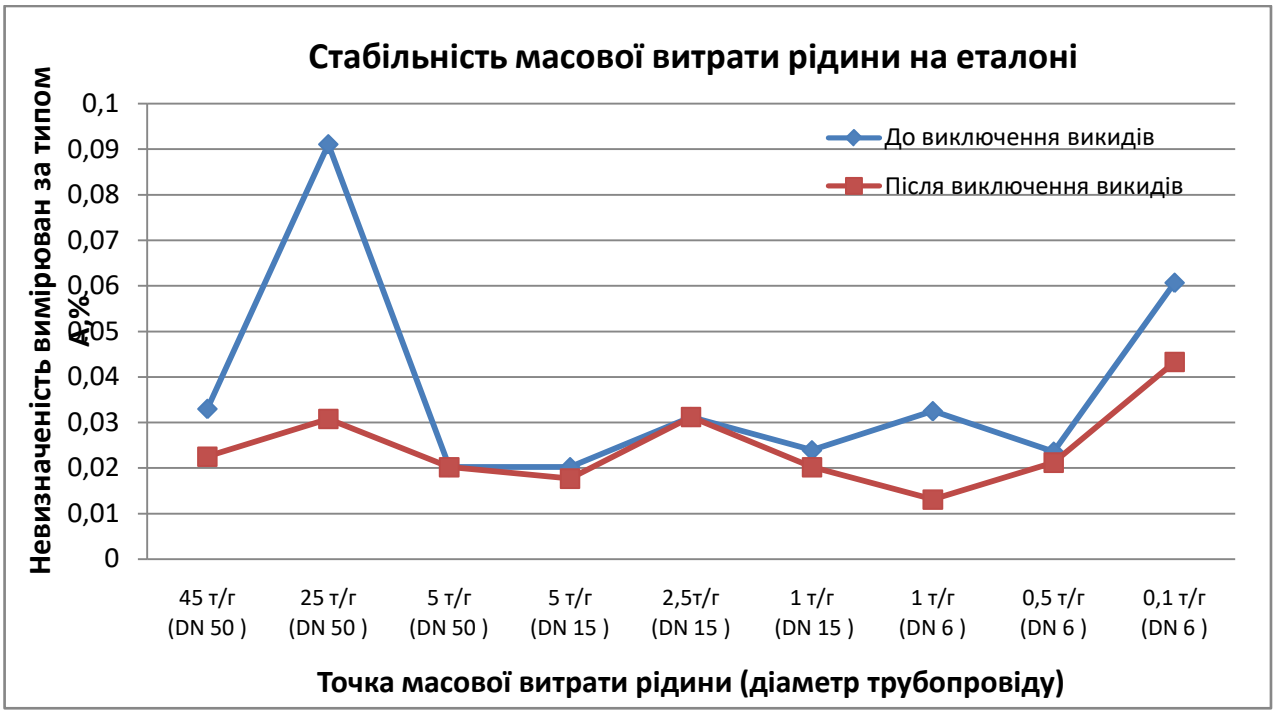

Рисунок 4.31 - Розрахунок стандартної невизначеності за типом А.

Після розрахунку стандартної невизначеності за типом А у вибірках, де було виявлено та видалено викиди, спостерігається значне зниження рівня невизначеності. Це є показником того, що видалення аномальних значень

суттєво зменшило розкид похибок навколо середнього значення, роблячи результати вимірювань більш стабільними та надійними [108].

Таким чином, якщо після виключення викидів невизначеність знижується, це підтверджує, що метод виявлення аномалій сприяє стабілізації даних і робить результати вимірювань менш схильними до випадкових флуктуацій.

Додатково, порівняння невизначеності до і після видалення викидів дозволяє оцінити ефективність моделі ІЛ у виявленні аномалій. Якщо очищення даних від викидів призводить до зменшення невизначеності, це свідчить про те, що модель ІЛ ефективно виявляє та виключає відхилення, які могли б спотворити середнє значення і знизити точність вимірювань. У цьому випадку зниження стандартної невизначеності є прямим індикатором підвищення якості та точності вимірювань завдяки застосуванню моделі.

Отже, аналіз змін у невизначеності за типом А після видалення викидів дозволяє підтвердити доцільність використання моделі ІЛ. Цей підхід не лише ідентифікує аномальні значення, але й знижує загальну похибку вимірювань, забезпечуючи більш стабільні, точні та якісні дані. У кінцевому підсумку, це підвищує надійність отриманих результатів і сприяє кращому розумінню вимірювань, роблячи їх менш залежними від випадкових впливів.

4.6 Висновок до четвертого розділу

Розділ 4 зосереджувався на експериментальному дослідженні параметрів моделі ІЛ та її адаптації для виявлення викидів у метрологічних даних. Проведений аналіз підтвердив, що базова модель ІЛ здатна ідентифікувати викиди, однак для забезпечення оптимальних результатів необхідне налаштування її параметрів з урахуванням особливостей кожної вибірки даних. Зокрема, важливими є такі параметри, як "кількість припущених викидів" та “кількість дерев у лісі”, оскільки вони впливають на баланс між точністю виявлення викидів та кількістю помилкових спрацьовувань.

Аналіз параметрів моделі також продемонстрував, що використання фіксованих порогових значень, таких як "0" для класифікації викидів, є недостатньо точним через варіативність вибірок. Застосування підходів, заснованих на відстанях між точками та їхніх відхиленнях, дозволяє адаптувати модель до характеристик конкретної вибірки, підвищуючи точність і достовірність виявлення викидів. Проте встановлення універсального порогу для всіх типів даних є неможливим через унікальні особливості кожної вибірки. Таким чином, модель ІЛ потребує індивідуального налаштування для кожного набору даних.

Для забезпечення стабільності шкали ступеня аномальності при різних запусках моделі було запропоновано підхід із використанням змінних значень параметра "кількість припущених викидів" та подальшого усереднення результатів. Це дозволяє сформувати єдину координатну систему для подальшого аналізу, зберігаючи відносне положення точок у просторі викидів. Такий підхід підвищує стабільність і точність аналізу, зменшуючи залежність від вибраних параметрів чутливості.

Модель була адаптована для завдань калібрування коріолісових витратомірів та проведеню міжнародних звірень, де обсяг даних дозволяє провести точний аналіз. Застосування алгоритму дозволило не лише візуалізувати структуру даних через графік ступеня аномальності, але й оцінити кількість потенційних викидів у вибірках та загальну стабільність даних.

Після виключення викидів було розраховано стандартну невизначеність за типом А для всіх вибірок. У вибірках, де були знайдені викиди, спостерігалося суттєве зниження стандартної невизначеності, що свідчить про зменшення варіативності даних. У вибірках витратами 25 т/г та 1 т/г невизначеність знизилася з 0,09% до 0,03% і з 0,03% до 0,01% відповідно. Таке значне зниження — у 3 та 2,5 рази відповідно — вказує на те, що знайдені викиди суттєво впливали на варіативність даних, тобто були надмірними похибками, які спотворювали результати вимірювань.

Отримані результати підтверджують, що застосування моделі ІЛ сприяло підвищенню достовірності результатів шляхом зниження стандартної невизна-

ченості за типом А завдяки виключенню викидів. Це дозволило мінімізувати вплив випадкових похибок на дані, що суттєво покращило точність і надійність метрологічних вимірювань. Модель ІЛ також забезпечує автоматизацію процесу виявлення викидів, ефективно адаптуючись до специфічних особливостей кожної вибірки.

ВИСНОВОК

У дисертаційній роботі вирішено науково-практичну задачу вдосконаленні процесів ідентифікації та виключення викидів із даних вимірювань, для зниження стандартної невизначеності вимірювань за типом А Комплексний підхід до дослідження дозволив досягти поставлених задач та отримати кількісні результати, що підтверджують ефективність запропонованих методів.

1. У роботі виконано **аналіз існуючих методів виявлення викидів**, включаючи статистичні методи та методи ML. Показано, що методи IQR і MAD є ефективними для малих вибірок із невідомим розподілом, однак вони поступаються гнучкістю та адаптивністю моделі ІЛ, яка краще працює із специфічними характеристиками вибірок.

2. Проведено дослідження та модернізацію вимірювальної системи, що включало заміну насосного обладнання та впровадження частотного перетворювача для стабілізації витрати рідини. Це дозволило знизити стандартну невизначеність на всьому діапазоні вимірювань, найбільше зниження вдбулось :

- На діапазоні витрати 0,5–2,5 т/г з 0,1% до 0,03%.

- На точці витрати 25 т/г з 0,21% до 0,085%.

Ці зміни забезпечили покращену стабільність витрати рідини протягом вимірювань.

3. Проведено експериментальне дослідження моделі ІЛ для задач виявлення викидів у метрологічних даних. Було налаштовано ключові параметри моделі, що дозволило адаптувати її до конкретних вибірок. Модель ІЛ продемонструвала ефективність, знижуючи стандартну невизначеність за тпом А на всіх точках витрати де було виявлено викии, найбільший вплив викидів на варіативність даних було помічено :

- в точці витрати 1 т/г — з 0,03% до 0,01%.

- в точці витрати 25 т/г — з 0,09% до 0,03%.

Це свідчить про значне покращення якості даних після виключення викидів.

У ході дослідження стабільності витрати було помічено, що на діапазонах витрат 1 т/г та 25 т/г виявлялися не лише значні нестабільності, але й викиди у вигляді надмірних похибок. Це вказує на необхідність подальшого аналізу причин виникнення таких явищ, оскільки характеристики стабільності витрати та наявності аномалій взаємопов’язані. Вимірювання проводиться протягом певного проміжку часу, і витрата рідини змінюється протягом цього часу, що відображається у вигляді флуктуацій витрати та надмірних похибок у даних.

Результати роботи створюють основу для подальших досліджень причин нестабільності на цих діапазонах витрат, аналізу факторів, що сприяють виникненню аномалій, а також розробки методів стабілізації витрати рідини для підвищення точності вимірювань. Отримані дані підтверджують доцільність і ефективність застосування моделі ІЛ для задач виявлення аномалій у метрології, що сприяє вдосконаленню процесів вимірювання та обробки даних.

СПИКОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. V. Aschepkov, "Methods of machine learning in modern metrology," *Measuring Equipment and Metrology*, vol. 85, no. 1, pp. 57–60, 2024, doi: 10.23939/istcmtm2024.01.
2. M. Vallejo, C. Espriella, J. Gómez-Santamaría, A. Ramírez-Barrera, and E. Delgado-Trejos, "Soft metrology based on machine learning: a review," *Measurement Science and Technology*, vol. 31, no. 3, pp. 1–16, 2019, doi: 10.1088/1361-6501/ab4b39.
3. P. Timoney et al., "Implementation of machine learning for high-volume manufacturing metrology challenges," *Proceedings Metrology, Inspection, and Process Control for Microlithography XXXII*, vol. 10585, 2018, doi: 10.1117/12.2300167..
4. S. Mueller, *Introduction to Machine Learning with Python*, 2016-2017. [Online]. Available: https://library-it.com/wp-content/uploads/2021/01/a_myuller_s_gvido_vvedenie_v_mashinnoe.pdf
5. Y.O. Hryshkun, S.M. Kravchenko, A.Yu. Levchenko, and Yu.I. Lysogor, "Machine Learning Methods," *Znanstvena misel journal*, no. 39, p. 55, 2020. [Online]. Available: https://www.znanstvena-journal.com/wp-content/uploads/2020/10/Znanstvena-misel-journal-%E2%84%9639-2020-VOL.1.pdf#page=55
6. S. Bilson, A. Thompson, D. Tucker, and J. Pierce, "A machine learning approach to automation and uncertainty evaluation for self-validating thermocouples," *NIST.SP.2100-05*, Digest Conference ITS10, 2023.
7. J.V. Pierce, O. Ongrey, G. Machin, and S.D. Sweeney, "Self-Validating Thermocouples Based on Fixed Points at High Temperatures," *Metrologia*, vol. 47, no. 1, 2010, doi: 10.1088/0026-1394/47/1/L01.
8. Rousseeuw, P. and Leroy, A., *Robust Regression and Outlier Detection*, John Wiley, 1987.

9. В.С. Попукайло, "Знаходження аномальних вимірювань при обробці малого об’єму," *Технологія та конструктування в електронній апаратурі*, no. 4-5, pp. 42-46, 2016. [Online]. Available: http://nbuv.gov.ua/ujrn/tkea_2016_4-5_8
10. Hawkins, D., *Identification of Outliers*, Chapman and Hall, London, 1980.
11. Knorr, E. and Ng, R., "A unified approach for mining outliers," in *Proceedings of the KDD Conference*, 1997, pp. 219–222.
12. VLDB Journal: Very Large Data Bases, "Unified approaches to outliers," vol. 8, no. 3–4, pp. 237–253.
13. Ramaswamy, S., Rastogi, R., and Shim, K., "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.
14. O.M. Vasilevsky, "Concept of Metrological Support in Industry 4.0," *Information Technologies and Computer Engineering*, vol. 48, no. 2, 2020, doi:10.31649/1999-9941-2020-48-2-37-44.
15. Є.Т. Володарський and Л.О. Кошева, *Статистична обробка даних: Навч. посібник*. Київ: НАУ, 2008, p. 103..
16. W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.
17. М. В. Карташов, *Імовірність, процеси, статистика*, Київ, 2007, p. 504.
18. M. Y. Krajnyuk, "Deep Learning for Measurements: Application of Neural Networks in Solving Real Problems," in *Metrological Aspects of Decision Making in the Conditions of Work at Technogenic Hazardous Objects*. [Online]. Available: https://er.chdtu.edu.ua/bitstream/ChSTU/462/1/sbornik_konf_2023.pdf.
19. Н. Б. Долішня, "Порівняльний аналіз теоретичних методів статистичної обробки експериментальних даних," *Вісник Вінницького національного медичного університету*, no. 1, 2019, doi: 10.31393/reports-vnmedical-2019-23(1)-17.

20. H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," in *Information and Decision Sciences: Proceedings of the 6th International Conference on FICTA*, Springer Singapore, 2018, pp. [online access unavailable].
21. H. Fitriyah and A. S. Budi, "Outlier Detection in Object Counting Based on Hue and Distance Transform Using Median Absolute Deviation (MAD)," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, Lombok, Indonesia, 2019, pp. 217–222, doi: 10.1109/SIET48054.2019.8985993.
22. Я.С. Довженко, В.В. Скляров, and К.В. Пилипенко, "Підготовка державного первинного стандарту України для одиниць твердості за шкалами Роквелла та Супер Роквелла (ДЕТУ 02-04-99) для міжнародних порівнянь," *Метрологія та прилади*, no. 5, pp. 6-12, 2012.
23. E. Batista and P. Lau, "EURAMET regional key comparison EURAMET.M.FF-K4.b: Volume intercomparison at 20 L," *Metrologia*, vol. 46(1A), 2009, doi: 10.1088/0026-1394/46/1A/07013.
24. Malengo, E. Batista, R. Arias, L. Mićić, A. Bošnjaković, M. Mirjana, E. Piluri, G. Svendsen, M. Huu, A. Sarevska et al., "Final report on EURAMET project 1395/EURAMET.M.FF-K4.1.2016: volume comparison at 20 L," *Metrologia*, vol. 57(1A), 2020, doi: 10.1088/0026-1394/57/1A/07021.
25. M. Huovinen and E. Frahm, "EURAMET.M.FF-S13 final report," *Metrologia*, vol. 59(1A), 2022, doi: 10.1088/0026-1394/59/1A/07010.
26. J. Geršl and L. Lojek, "Final report on EURAMET project No. 1046: Intercomparison of water flow standards using electromagnetic flowmeters," *Metrologia*, vol. 50(1A), 2013, doi: 10.1088/0026-1394/50/1A/07002.
27. E. Batista, "Final report on EUROMET key comparison EUROMET.M.FF-K4 for volume intercomparison of 100 ml Gay-Lussac pycnometer," *Metrologia*, vol. 43(1A), 2006, doi: 10.1088/0026-1394/43/1A/07009.
28. M. Benkova, E. Frahm, K. Romieu, H. Warnecke, O. Büker, S. Haack, B. Akselli, V. Mazur, C. Berkmann, and G. Zygmantas, "Comparisons of standards

for liquid flow rates under static load changes," *Metrologia*, vol. 61(1A), 2024, doi: 10.1088/0026-1394/61/1A/07003.

29. K. Lee, S. Chun, Y. Terao, N. Thai, C. Yang, M. Tao, and M. Gutkin, "Final report of the APMP water flow key comparison: APMP.M.FF-K1," *Metrologia*, vol. 48(1A), 2011, doi: 10.1088/0026-1394/48/1A/07003.
30. J. Man, R. Arias, Y. Terao, Y. Lee, G. Ligong, V. Tulasombut, T. Chan, N. Thai, R. Steyn, and H. Sampath, "Final report on APMP key comparison of volume of liquids at 20 L and 100 mL: APMP.M.FF-K4," *Metrologia*, vol. 48(1A), 2011, doi: 10.1088/0026-1394/48/1A/07001.
31. S. Chun and N. Furuichi, "Final report of the APMP water flow supplementary comparison (APMP.M.FF-S1)," *Metrologia*, vol. 59(1A), 2022, doi: 10.1088/0026-1394/59/1A/07004.
32. Morales, D. Malta, F. Kornblit, R. Ramírez, R. Arias, and S. Trujillo, "Final report on regional supplementary comparison SIM.M.FF-S5: Volume of liquids at 50 mL," *Metrologia*, vol. 49(1A), 2012, doi: 10.1088/0026-1394/49/1A/07012.
33. Jacques, S. Juarez, J. Maldonado, and V. Bean, "NORAMET intercomparison of volume standards at 50 mL and 100 mL (SIM.M.FF-S1)," *Metrologia*, vol. 40(1A), 2003, doi: 10.1088/0026-1394/40/1A/07001.
34. S. Trujillo, J. Maldonado, M. Vega, E. Santalla, A. Sica, D. Cantero, M. Salazar, A. Morales, P. Solano, and L. Rodríguez, "SIM.M.FF-S7: Final report on SIM/ANDIMET supplementary comparison for volume of liquids at 100 mL and 100 μL," *Metrologia*, vol. 53(1A), 2016, doi: 10.1088/0026-1394/53/1A/07014.
35. E. Frahm, R. Arias, M. Maldonado, J. Vargas, J. Mendoza, A. Arredondo, and M. Silvosa, "Supplementary comparison SIM.M.FF-S9.2016 for water flow measurement," *Metrologia*, vol. 61(1A), 2024, doi: 10.1088/0026-1394/61/1A/07001.
36. M. Benkova, E. Frahm, K. Romieu, H. Warnecke, O. Büker, S. Haack, B. Akselli, V. Mazur, et al., "Comparisons of standards for liquid flow rates under

static load changes," *Metrologia*, vol. 61(1A), 2024, doi: 10.1088/0026-1394/61/1A/07003.

37. R. Arias, M. Maldonado, J. Wright, C. Jacques, C. Lachance, P. Lau, H. Többen, and G. Cignolo, et al., "Results of the key comparison CCM.FF-K4 for volume of liquids at 20 L and 100 mL," *Metrologia*, vol. 43(1A), 2006, doi: 10.1088/0026-1394/43/1A/07005.
38. Z. Zelenka, R. Arias, M. Maldonado, E. Batista, W. Jintao, A. Malengo, D. Malta, et al., "BIPM/CIPM key comparison CCM.FF-K4.1.2011. Final report for volume of liquids at 20 L and 100 mL," *Metrologia*, vol. 52(1A), 2015, doi: 10.1088/0026-1394/52/1A/07011.
39. J. Paik, K. Lee, P. Lau, R. Engel, A. Loza, Y. Terao, and M. Reader-Harris, "Final report on CCM.FF-K1 for water," *Metrologia*, vol. 44(1A), 2007, doi: 10.1088/0026-1394/44/1A/07005.
40. N. Furuichi, R. Arias, C-T. Yang, S. Chun, T. Meng, I. Shinder, E. Frahm, O. Büker, Chr. Mills, et al., "Final report 'Key comparison CCM.FF-K1.2015 – water flow: 30 m³/h ... 200 m³/h'," *Metrologia*, vol. 59(1A), 2022, doi: 10.1088/0026-1394/59/1A/07013.
41. ДСТУ 4403:2005, "Метрологія. Державна повірочна схема для засобів вимірювання обємної та масової витрати рідини й обєму та маси рідини, що протікає по трубопроводу."
42. В.О. Ащепков, "Дослідження метрологічних характеристик державного первинного еталона одиниці об'ємної та масової витрати рідини при підготовці до участі у міжнародних звіреннях," *Український метрологічний журнал*, no. 1 (77), pp. 31–37, 2024, doi: 10.24027/2306-7039.1.2024.300937.
43. А. М. Слізков and Л. А. Дмитренко, "Оцінювання невизначеності вимірювання результатів кількісних вимірювань," *Вісник*, 2012. [Online]. Available: http://lib.khnu.km.ua/pdf/visnyk_tup/2012/(187)2012-2-t.pdf#page=219. [Accessed: Nov. 18, 2024].

44. WGFF, "Guidelines for CMC Uncertainty and Calibration Report Uncertainty," technical report, 2013, 6 p. [Online]. Available: http://www.bipm.org/utils/en/pdf/ccm-wgff-guidelines.pdf

45. M.G. Cox, "The evaluation of key comparison data," *Metrologia*, vol. 39, no. 6, pp. 589–595, 2002, doi: 10.1088/0026-1394/39/6/10.

46. M.G. Cox, "The evaluation of key comparison data: determining the largest consistent subset," *Metrologia*, vol. 44, no. 3, pp. 187–200, 2007, doi: 10.1088/0026-1394/44/3/005.

47. K. Jurado, S. Ludvigson, and S. Ng, "Measuring Uncertainty," *American Economic Review*, vol. 105, no. 3, pp. 1177–1216, 2015, doi: 10.1257/aer.20131193.

48. Zakharov, M. Serhiienko, and T. Chunikhina, "Measurement uncertainty evaluation by kurtosis method at calibration of a household water meter," in *XXX International Scientific Symposium "Metrology and Metrology Assurance" (MMA)*, 2020, pp. 83–86, doi: 10.1109/MMA49863.2020.9254260.

49. G. A. Susto, A. Beghi, and S. McLoone, "Anomaly detection through on-line isolation forest: An application to plasma etching," in *Proceedings of the 28th Annual SEMI Advanced Semiconductor Manufacturing Conference*, May 2017, pp. 89–94.

50. Alghushairy, Omar, et al. "A review of local outlier factor algorithms for outlier detection in big data streams." *Big Data and Cognitive Computing* 5.1 (2020)

51. Khan, Kamran, et al. "DBSCAN: Past, present and future." *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE, 2014.

52. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*, pp. 413–422, Pisa, Italy, Dec. 2008, doi: 10.1109/ICDM.2008.17.

53.Zhou, Chong, and Randy C. Paffenroth. "Anomaly detection with robust deep autoencoders." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017.

54.Li, Kun-Lun, et al. "Improving one-class SVM for anomaly detection." *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*. Vol. 5. IEEE, 2003.

55. S. Buschjäger, P.-J. Honysz, and K. Morik, "Generalized Isolation Forest: Some Theory and More Applications Extended Abstract," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 793–794, doi: 10.1109/DSAA49011.2020.00120.

56.F. Liu, K. Ting, and Z. Zhou, "Isolation Forest," in *IEEE International Conference on Data Mining*, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.

57.S. Kebir and K. Tabia, "Anomaly Detection in Real Scarce Data: A Case Study on Monitoring Elderly's Physical Activity and Sleep," in *IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, 2023, pp. 385–392, doi: 10.1109/BIBE60311.2023.00069.

58.B. Yu, Y. Yu, J. Xu, G. Xiang, and Z. Yang, "MAG: A Novel Approach for Effective Anomaly Detection in Spacecraft Telemetry Data," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 3891–3899, 2024, doi: 10.1109/TII.2023.3314852.

59.Z. Li, P. Wang, Z. Wang, and D. Zhan, "FlowGANAnomaly: Flow-Based Anomaly Network Intrusion Detection with Adversarial Learning," *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 58–71, 2022, doi: 10.23919/cje.2022.00.17380.

60.L. Barbieri, M. Brambilla, M. Stefanutti, C. Romano, N. Carlo, and M. Roveri, "A Tiny Transformer-Based Anomaly Detection Framework for IoT Solutions," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 462–478, 2023, doi: 10.1109/OJSP.2023.3333756.

61. N. Guo, C. Lin, H. Yan, J. Zang, and M. Xiong, "Real-Time Pantograph Anomaly Detection Using Unsupervised Deep Learning and K-Nearest Neighbor Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–13, 2024, doi: 10.1109/TIM.2024.3370747.
62. M. Occorso, M. An, R. Olsen, and V. Perry, "Anomaly Detection as a Data Reduction Approach for Test Event Analysis at the Edge," in *IEEE International Conference on Big Data (BigData)*, 2023, pp. 3863–3867, doi: 10.1109/BigData59044.2023.10386215.
63. H. Xiang, X. Zhang, M. Dras, A. Beheshti, W. Dou, and X. Xu, "Deep Optimal Isolation Forest with Genetic Algorithm for Anomaly Detection," in *IEEE International Conference on Data Mining (ICDM)*, 2023, pp. 678–687, doi: 10.1109/ICDM58522.2023.00077.
64. В.О. Ащепков, "Використання моделі ISOLATION FOREST для виявлення аномалій у даних вимірювань," *Сучасний стан наукових досліджень та технологій в промисловості*, no. 1 (27), pp. 98–113, 2024, doi: 10.30837/ITSSI.2023.26.
65. J. Yang, S. Rahardja, and P. Fränti, "Outlier detection: how to threshold outlier scores?" in *AIIPCC '19: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, Article No. 37, pp. 1–6, doi: 10.1145/3371425.33714.
66. P.J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.
67. L. Sunitha, M. Balaraju, J. Sasikiran, and E.V. Ramana, "Automatic outlier identification in data mining using IQR in real-time data," *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, vol. 3, no. 6, 2014.
68. C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the

median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

69. V. Aschepkov, "Methods for outlier detection in metrological studies," *Measuring Equipment And Metrology*, vol. 85, no. 3, pp. 25–29, 2024, doi: 10.23939/istcmtm2024.03.025.
70. V. Aschepkov, "Improving the efficiency of processing of measurement results using the machine learning method," in *Theses of Reports 1st European Competition of Young Best Metrologists in Ukraine*, Ivano-Frankivsk, Ukraine, June 24–28, 2024, pp. 6–8.
71. M. Gupta, J. Gao, C. Aggarwal, and J. Han, *Outlier Detection for Temporal Data*, Morgan & Claypool Publishers, 2014.
72. T.V. Pollet and L. van der Meij, "To remove or not to remove: the impact of outlier handling on significance testing in testosterone data," *Adaptive Human Behavior and Physiology*, vol. 3, no. 1, pp. 43–60, 2017.
73. J.W. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection," in *International Conference on Fuzzy Systems and Data Mining (FSDM)*, in *Frontiers in Artificial Intelligence and Applications (FAIA)*, vol. 309, no. 2, pp. 208–215, 2018.
74. V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
75. K. Kaur and A. Garg, "Comparative study of outlier detection algorithms," *International Journal of Computer Applications (IJCA)*, vol. 147, no. 9, 2016.
76. Atkinson, "Fast very robust methods for the detection of multiple outliers," *Journal of the American Statistical Association*, vol. 89, pp. 1329–1339, 1994.
77. V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, 1994.
78. D. Hawkins, *Identification of Outliers*, Chapman and Hall, London, 1980.
79. E. Knorr and R. Ng, "A unified approach for mining outliers," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 219–222, 1997.

80. M. Hariri, M.C. Kind, and R.J. Brunner, "Extended Isolation Forest," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1479–1489, 2021, doi: 10.1109/TKDE.2019.2947676.

81. H. Li, L. Zhu, and Z. Yang, "A SCiForest based semi-supervised learning method for the seismic interpretation of channel sand-body," *Journal of Applied Geophysics*, vol. 167, pp. 51–62, 2019, doi: 10.1016/j.jappgeo.2019.04.019.

82. M. Jiang, S. Han, and H. Huang, "Anomaly Detection with Score Distribution Discrimination," in *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 2023, pp. 984–996, doi: 10.1145/3580305.3599258.

83. Z. Ding and L. Xing, "Improved software defect prediction using Pruned Histogram-based isolation forest," *Reliability Engineering & System Safety*, vol. 204, p. 107170, 2020, doi: 10.1016/j.ress.2020.107170.

84. H. Ma, B. Ghojogh, M.N. Samad, D. Zheng, and M. Crowley, "Isolation Mondrian Forest for Batch and Online Anomaly Detection," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 3051–3058, doi: 10.1109/SMC42975.2020.9283073.

86. K. Ting, B. Xu, T. Washio, and Z. Zhou, "Isolation Distributional Kernel: A New Tool for Kernel based Anomaly Detection," in *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, August 2020, pp. 198–206, doi: 10.1145/3394486.3403062.

87. R.T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, pp. 144–155, Morgan Kaufmann, San Francisco, 1994.

88. S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *Information Systems*, vol. 26, no. 1, pp. 35–58, 2001.

89. R.T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers, San Francisco, pp. 144–155, 1994.

90.L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.

91.Ankerst, Mihael, et al. "OPTICS: Ordering points to identify the clustering structure." *ACM Sigmod record* 28.2 (1999): 49-60.

92.Agrawal, K. P., et al. "Development and validation of OPTICS based spatio-temporal clustering technique." *Information Sciences* 369 (2016): 388-401.

93.Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." *International Journal* 1.6 (2013): 90-95.

94.S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000.

95.M. Breunig, H. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.

96.E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pp. 392–403, 1998.

97.L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.

98.E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *The VLDB Journal: The International Journal on Very Large Data Bases*, vol. 8, no. 3–4, pp. 237–253, 2000.

99.C. Aggarwal, *Outlier Analysis*, Springer, 2017, doi: 10.1007/978-3-319-47578-3.

100.D.M. Hawkins, *A Practical Guide to Heavy Tailed Data*, Wiley, 2010.

101.L. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

102.S. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

103.P.J. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.

104.M. Breuning, H. Kriegel, R.T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.

105.E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pp. 392–403, 1998.

106.G. Huang, Z. Zhang, and W. Yang, "Outlier Detection Method based on Improved Two-step Clustering Algorithm and Synthetic Hypothesis Testing," in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pp. 915–919, 2019.

107.T. R. Bandaragoda, K. M. Ting, D. Albrecht, F. T. Liu, and J. R. Wells, "Efficient anomaly detection by isolation using nearest neighbour ensemble," in *Proceedings of the IEEE International Conference on Data Mining Workshop*, 2014, pp. 698–705.

108.Ащепков В.О., Бяллович Д.Ю., Скляров В.В. Вплив порогових значень на стандартну невизначеність типу А при вимірюваннях масової витрати рідини // *Український метрологічний журнал*. 2024. №3 (30) doi: 10.24027/2306-7039.3.2024.312469.