УДК 004.896:004.032.26

ВИЯВЛЕННЯ АНОМАЛІЙ У ЛОГАХ ДОСТУПУ RFID СИСТЕМ ІЗ ЗАСТОСУВАННЯМ АЛГОРИТМУ ISOLATION FOREST НА СИНТЕТИЧНОМУ ДАТАСЕТІ

Морозова Г. С.[1]; Пархоменко О. Ю.[2]
[1,2] Національний університет кораблебудування імені адмірала Макарова
[1,2] Україна, Миколаїв
[1] ORCID ID 0009-0005-9149-7378; [2] ORCID ID 0000-0002-7940-7414

Анотація. *У роботі досліджено використання алгоритму Isolation Forest для виявлення аномалій у логах RFID-систем на синтетичному датасеті (5000 записів). Експериментально отримано ідентифікацію 250 аномалій (5%) з метриками Precision, Recall та F1-score 0,804, що підтверджує ефективність моделі. Теоретично обґрунтовано доцільність unsupervised learning для таких задач. Результати можуть бути застосовані для підвищення безпеки в реальних системах контролю доступу.*
Ключові слова: *виявлення аномалій; RFID-системи; Isolation Forest; машинне навчання; синтетичний датасет.*

Вступна частина. Сучасні системи контролю доступу на основі RFID-технологій відіграють ключову роль у забезпеченні безпеки об’єктів, дозволяючи автоматизувати процеси ідентифікації та реєстрації користувачів. Однак, з ростом складності таких систем виникає проблема виявлення аномальної поведінки, яка може свідчити про несанкціонований доступ, зловживання або технічні збої. Аномалії в логах доступу, такі як незвичайний час входу, нетипова локація чи атипова тривалість сесії, вимагають ефективних методів аналізу для своєчасного реагування [1].

Мета дослідження полягає в оцінці ефективності алгоритму Isolation Forest для виявлення аномалій у логах доступу RFID-систем на основі синтетично згенерованого датасету. Завдання включають генерацію реалістичного датасету, що моделює нормальну та аномальну поведінку користувачів, тренування моделі Isolation Forest, аналіз результатів за допомогою метрик (Precision, Recall, F1-score) та візуалізацію розподілу аномалій. Дослідження спрямоване на демонстрацію практичної застосовності алгоритму в інформаційних системах, сприяючи розвитку алгоритмів та моделей для інтелектуального контролю доступу.

Основна частина. Для моделювання системи контролю доступу був згенерований синтетичний датасет, що імітує логи доступу RFID-карт. Датасет складається з 5000 записів, де кожен запис містить такі атрибути: ідентифікатор користувача (user_id), час доступу (timestamp), локацію (location) та тривалість сесії (duration) у секундах. Нормальна поведінка моделюється як доступи у робочий час (8-18 годин, понеділок-п’ятниця), з тривалістю сесії, що підпорядковується нормальному розподілу з середнім 300 секунд та стандартним відхиленням 100 секунд (обмежено від 60 до 600 секунд). Локації розподіляються з ймовірностями: Офіс А (40%), Офіс В (30%), Серверна (20%), Вхід (10%).

Аномалії становлять 5% від загальної кількості записів (250 аномалій) і вводяться випадково: або зміна часу на нічний (0-7 або 19-23 години), або екстремальна тривалість (менше 60 секунд або більше 600 секунд), або незвичайна локація («Заборонена зона» або «Невідома локація»). Генерація датасету здійснюється за допомогою бібліотек pandas та numpy у Python.

Для виявлення аномалій застосовується алгоритм Isolation Forest з бібліотеки scikit-learn [2, 3]. Алгоритм базується на випадковому поділі даних у бінарних деревах, де аномалії ізолюються швидше через меншу щільність [4, 5]. Параметр contamination встановлено на 0.05, що відповідає очікуваній частці аномалій. Препроцесинг включає витягнення фіч: година дня (hour), день тижня (day_of_week), кодована локація (location_encoded) та тривалість. Фрагмент коду для моделі:

```
from sklearn.ensemble import IsolationForest
from sklearn.preprocessing import LabelEncoder
data['hour'] = data['timestamp'].dt.hour
data['day_of_week'] = data['timestamp'].dt.dayofweek
le = LabelEncoder()
data['location_encoded'] = le.fit_transform(data['location'])
features = data[['hour', 'day_of_week', 'duration', 'location_encoded']]
iso_forest = IsolationForest(contamination=0.05, random_state=42)
data['anomaly'] = iso_forest.fit_predict(features)
data['anomaly'] = data['anomaly'].map({1: 0, -1: 1})
```

Експеримент проводився у середовищі Python з використанням бібліотек scikit-learn для моделі, matplotlib та seaborn для візуалізації. Після генерації датасету та тренування моделі Isolation Forest були отримані результати класифікації: 4750 записів позначено як нормальна поведінка та 250 як аномалії, що відповідає заданій частці.

Аналіз середніх значень фіч за статусом показує відмінності: аномальні доступи мають дещо нижчу середню годину (12,492 проти 13,028), менший середній день тижня (2,728 проти 3,067), але значно вищу середню тривалість (815,98 секунд проти 298,40 секунд), що свідчить про ефективність моделі у виявленні екстремальних сесій.

Таблиця 1. Таблиця середніх значень

Статус	Середня година	Середній день тижня	Середня тривалість
Нормальна	13,028	3,066737	298,398705
Аномальна	12,492	2,728000	815,982246

Візуалізація результатів дозволяє глибше зрозуміти розподіл аномалій.

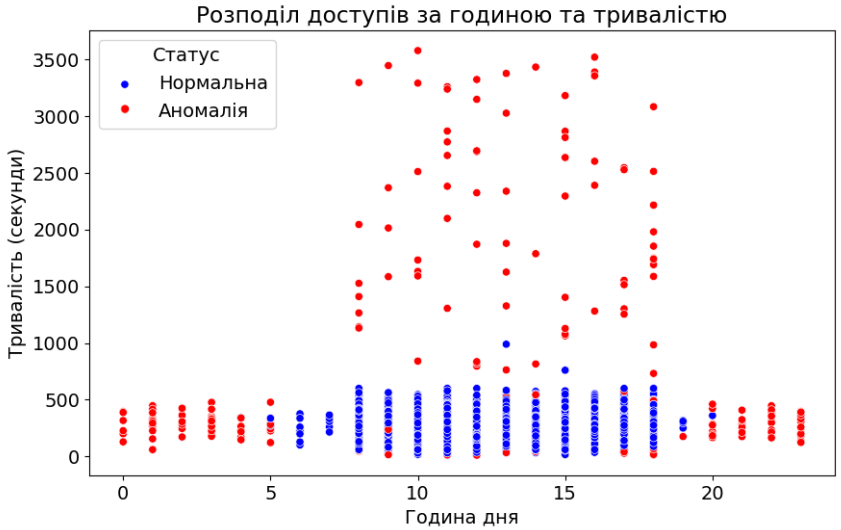

Рис. 1. Розподіл доступів за годиною та тривалістю

На рис. 1 scatter plot, де нормальні точки (сині) зосереджені в робочих годинах з помірною тривалістю, тоді як аномалії (червоні) виділяються високою тривалістю або нетиповими годинами. Це підтверджує, що модель ефективно ізолює відхилення.

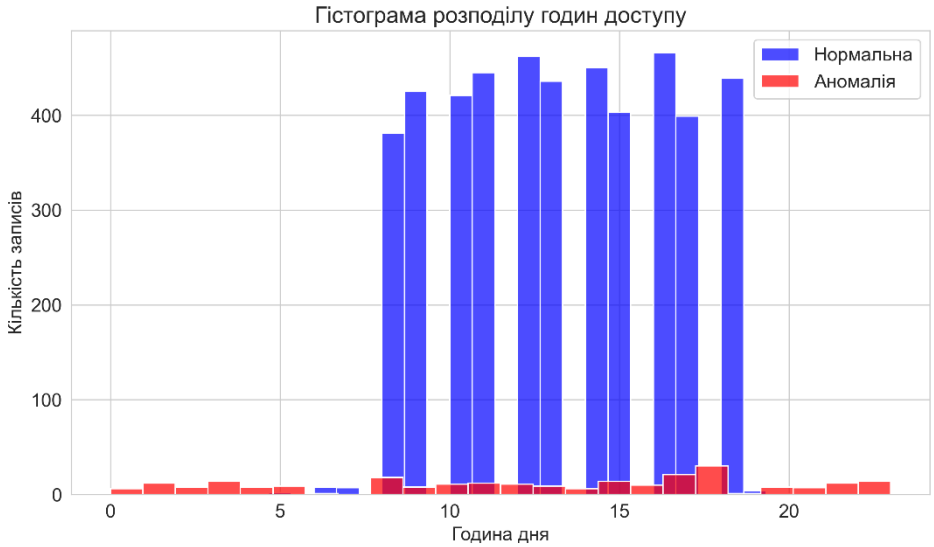

Рис. 2. Гістограма розподілу годин доступу

Гістограма розподілу годин доступу (рис. 2) представляє гістограму, де нормальні доступи переважно в діапазоні 8-18 годин, з піком близько 13-14 години, тоді як аномалії розподілені ширше, включаючи нічні години. Це ілюструє чутливість алгоритму до часових патернів.

Для оцінки якості моделі, оскільки датасет синтетичний і містить відомі справжні аномалії, були обчислені метрики на основі порівняння передбачених і справжніх міток. Оцінка якості моделі за метриками показала: Precision = 0,804, Recall = 0,804, F1-score = 0,804. Це означає, що 80,4% виявлених аномалій є справжніми (висока точність), а модель захоплює 80,4% усіх справжніх аномалій (висока повнота). Збалансоване значення F1-score свідчить про відсутність значного перекосу між помилковими спрацьовуваннями та пропущеними аномаліями. Аналіз матриці помилок (TN=4701, FP=49, FN=49, TP=201) підтверджує низьку кількість як хибнопозитивних, так і хибнонегативних класифікацій, що є критичним для систем безпеки, де пропуск небезпечної події (FN) може мати серйозні наслідки. Отже, Isolation Forest демонструє високу ефективність і надійність у виявленні аномалій навіть на синтетичних даних, що обґрунтовує його потенціал для використання в реальних RFID-системах.

Таблиця 2. Матриця помилок (Confusion Matrix)

	Передбачено нормальна	**Передбачено аномалія**
Справжня нормальна	4701	49
Справжня аномалія	49	201

Матриця помилок показує, що модель має низьку кількість помилкових позитивів (49) та помилкових негативів (49), з високою кількістю справжніх позитивів (201) та справжніх негативів (4701), що свідчить про добру загальну продуктивність.

Висновки. Проведене моделювання продемонструвало високу ефективність алгоритму Isolation Forest для виявлення аномалій у логах доступу RFID-систем на синтетичному датасеті обсягом 5000 записів. Конкретні результати включають ідентифікацію 250 аномалій (5% від

загальної кількості), з точністю (Precision) 0,804, повнотою (Recall) 0,804 та F1-score 0,804, що підтверджує збалансовану продуктивність моделі. Аналіз середніх значень показав, що аномальні доступи характеризуються середньою тривалістю 815,98 секунд проти 298,40 секунд для нормальних, а також розподілом у нічні години, що узгоджується з синтетичними даними. Матриця помилок виявила 49 помилкових позитивів і 49 помилкових негативів, що свідчить про низьку кількість помилок при класифікації. Однак, враховуючи синтетичну природу датасету, результати можуть варіюватися на реальних даних через наявність шуму чи складніших патернів, на що слід зважати при інтерпретації. Перспективи подальших досліджень включають інтеграцію з реальними датасетами, комбінування з іншими алгоритмами, такими як Autoencoders, та розширення на мультимодальні дані для підвищення точності в системах інтелектуального контролю доступу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Alqahtani, H., & Maple, C. (2024). A secure framework for the Internet of Things anomalies using machine learning. *Discover Internet of Things*, 4(1), Article 29. https://doi.org/10.1007/s43926-024-00088-z

[2] Bertalanic, B., Hribar, J., & Fortuna, C. (2025). Anomaly detection using machine learning and adopted digital twin concepts in radio environments. *Scientific Reports*, 15, Article 2759. https://doi.org/10.1038/s41598-025-02759-5

[3] Aljumah, A. (2024). An optimized isolation forest based intrusion detection system for heterogeneous and streaming data in the industrial Internet of Things (IIoT) networks. *Discover Applied Sciences*, 6(10), Article 416. https://doi.org/10.1007/s42452-024-06165-w

[4] Cao, Y., Xiang, H., Zhang, H., Zhu, Y., & Ting, K. M. (2024). Anomaly detection based on isolation mechanisms: A survey. *arXiv preprint* arXiv.2403.10802. https://doi.org/10.48550/arXiv.2403.10802

[5] Loconte, D., Ieva, S., Pinto, A., Loseto, G., Scioscia, F., & Ruta, M. (2024). Federated learning-based anomaly detection with isolation forest in the IoT-edge continuum. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(12), Article 3702995. https://doi.org/10.1145/3702995

Morozova G. S., Parkhomenko O. Yu.

Detection of anomalies in RFID system access logs using the Isolation Forest algorithm on a synthetic dataset

Abstract. *The study explores the Isolation Forest algorithm for anomaly detection in RFID logs using a synthetic dataset (5000 records). Experimental results identified 250 anomalies (5%) with Precision, Recall, and F1-score of 0,804, confirming model efficiency. The theoretical basis for unsupervised learning in such tasks is substantiated. Results can enhance security in real access control systems.*

Keywords: *anomaly detection; RFID systems; Isolation Forest; machine learning; synthetic dataset.*