DOI: 10.34185/1991-7848.itmm.2026.01.084

МЕТОДОЛОГІЯ ПІДГОТОВКИ ДАТАСЕТУ ДЛЯ НАВЧАННЯ МОДЕЛЕЙ ВИЯВЛЕННЯ ШАХРАЙСТВА В ЕЛЕКТРОННІЙ КОМЕРЦІЇ

Носов В.О.[1] [ORCID], Островська К.Ю.[2] [ORCID]

[1]*Український державний університет науки і технологій, аспірант, Україна*

[2]*Український державний університет науки і технологій, к.т.н., доцент, Україна*

Анотація. *У дослідженні розглянуто проблему підготовки тренувальних даних для систем виявлення шахрайства в транзакціях електронної комерції на основі методів машинного навчання. За результатами аналізу існуючих відкритих джерел обґрунтовано необхідність створення спеціалізованого набору даних. Запропоновано автоматизований конвеєр об'єднання трьох відкритих наборів даних з платформи Kaggle (IEEE-CIS, Sparkov, Fraudulent E-Commerce) зі збереженням реальних міток шахрайства та збагаченням записів синтетичними атрибутами, адаптованими до специфіки українського платіжного ринку. Опрацьовано методи рівномірної нормалізації часових міток, генерації автентифікаційних даних та розбиття на платіжні системи, формування агрегованих профілів клієнтів та пар для навчання моделі IP Insights. Результатом є набір із 500000 транзакцій за 24 місяці з рівнем шахрайства 3,04%, призначений для навчання конвеєра моделей, до яких входять LightGBM, автоенкодер та IP Insights.*

Ключові слова: *датасет, машинне навчання, транзакція, електронна комерція, LightGBM, автоенкодер, IP Insights*

Вступ

Стрімке зростання обсягів електронної комерції супроводжується збільшенням масштабів платіжного шахрайства. За спільним звітом European Bank Authority та European Central Bank за грудень 2025 року, загальний обсяг шахрайства у Європейському економічному просторі сягнув 4,2 млрд євро у 2024 році (ріст на 17% порівняно з 2023 р.) [1]. Шахрайство з картковими транзакціями становило 1,3 млрд євро (+29%), а 80–85% припало на дистанційні операції. За глобальним опитуванням Visa та Merchant Risk Council серед понад 1000 торговців, 98% респондентів зіткнулися з шахрайством протягом останнього року, а понад 80% відзначили труднощі з ефективним використанням даних та підвищенням точності моделей машинного навчання

(МН) [2]. Ці тенденції підтверджують актуальність як удосконалення алгоритмів протидії шахрайству, так і необхідності підготовки якісних тренувальних даних для систем на основі МН.

У попередній роботі [3] було проведено аналіз існуючих методів МН для виявлення шахрайства, на основі якого сформовано концепцію багатошарової системи виявлення шахрайства в онлайн-транзакціях з використанням конвеєра LightGBM, автоенкодера та IP Insights на базі AWS SageMaker. Одним із етапів подальшої роботи стала підготовка спеціалізованого набору даних. Метою даної роботи є опис методології створення такого набору, адаптованого до специфіки українського платіжного ринку, на основі злиття кількох відкритих наборів із Kaggle.

Виклад основного матеріалу

Як інформаційну базу для формування єдиного репрезентативного набору даних було відібрано три відкриті масиви з платформи Kaggle:

1. IEEE-CIS Fraud Detection (близько 590 тис. транзакцій) – містить достовірні мітки класів шахрайства, оригінальні суми операцій у доларах (USD), мережеві індикатори (поштові домени), а також розширені атрибути цифрових відбитків пристроїв і браузерів [4];
2. Sparkov Data Generation (близько 1,3 млн транзакцій) – синтетично згенерований набір, що забезпечує глибоку деталізацію часових міток, фінансових показників та категоризації витрат [5];
3. Fraudulent E-Commerce (близько 100 тис. записів) – набір, що включає унікальні ідентифікатори користувачів, транзакційні суми та відповідну бінарну розмітку аномалій [6].

Перевагою такого об'єднання є збереження автентичних міток класів та ненормалізованих значень сум фінансових операцій із різних джерел.

Проведений аналіз альтернативних відкритих наборів даних у цій предметній області виявив їхню методологічну невідповідність завданням даного дослідження. Зокрема, набір IBM AMLSim [7] орієнтований на специфіку виявлення відмивання коштів (Anti-Money Laundering), тоді як популярний Credit Card Fraud Dataset [8] містить виключно анонімізовані ознаки,

перетворені за допомогою методу головних компонент (PCA), що позбавляє дані семантичного контексту.

Зважаючи на різні часові межі та обсяги вихідних масивів, їхнє лінійне масштабування могло б призвести до статистичних викривлень. З метою запобігання цьому застосовано алгоритм рівномірного перерозподілу, який ґрунтується на ранговому перетворенні (rank transformation): відсортованим за оригінальними мітками транзакціям послідовно присвоєно дати в діапазоні 01.01.2023–31.12.2024 зі збереженням початкового часу доби. Суть підходу полягає у заміні оригінальних часових міток на рівновіддалені значення у цільовому діапазоні із збереженням порядкового рангу кожної транзакції. Для унеможливлення небажаної кластеризації за джерелом походження щоденні пули транзакцій додатково перемішуються.

Формат даних структуровано у вигляді EMV 3D-Secure [9] транзакцій. Оскільки у відкритих масивах відсутні специфічні поля цього протоколу, їх згенеровано синтетично на основі ймовірнісних розподілів. Відтворено такі технічні параметри, як тип автентифікації (challenge, frictionless, stand-in processing), канал ініціації транзакції (browser, application, 3DS requestor initiated), платіжна система (Visa, MasterCard, Prostir), динаміка версій 3DS, структура типових мерчантів та еквайрерів. Для врахування специфіки українського ринку, були використані типові українські мерчанти та банки еквайри, встановлено загальне домінування гривні (UAH) на рівні 70%.

Для формування стійких дискримінативних ознак додатково змодельовано контекстні атрибути: IP-адреси з урахуванням маршрутизації 30% шахрайського трафіку через VPN або датацентри, узгоджені профілі пристроїв (Android, iOS і т.д.) із кореляцією операційної системи та браузера, використання одноразових email-доменів.

Окрім базового набору, архітектура конвеєра із алгоритмів МН вимагає формування спеціалізованих структур даних. Так, набір агрегованих клієнтських профілів містить поведінкові ознаки карток, зокрема кількість і суми операцій за різні проміжки часу, унікальність торговців, IP та пристроїв, а також частоту шахрайства. Ці обчислені метрики утворюють вхідний вектор

для автоенкодера та LightGBM. Паралельно для алгоритму аналізу мережевої поведінки IP Insights підготовлено набір пар «клієнт–IP».

Фінальне хронологічне розбиття датасету у пропорції 70/15/15 імітує реальні умови експлуатації системи, де навчання відбувається на історичних подіях, а прогнозування – на майбутніх. Додатково сформована валідаційна вибірка слугує для калібрування порогів ризику, тоді як тестова для об'єктивної оцінки здатності моделі до генералізації.

Основна зведена статистика по поточній версії датасету представлена у табл.1.

Таблиця 1

Статистика по набору даних

Параметр	Значення
Загальна кількість транзакцій	500000
Загальна кількість шахрайських транзакцій	15178 (3.04%)
Загальна кількість легітимних транзакцій	484822 (96.96%)
Унікальних карток	227651
Унікальних продавців	114
Унікальних IP адрес	470682
Основна валюта	UAH (70%)
Основна операційна система	Android (69,9%)
Основний браузер	Chrome (48,8%)
Payment Network Operator	Visa 50%, Mastercard 40%, Prostir 10%

Висновки

Описано методологію підготовки набору даних транзакцій для навчання конвеєра моделей виявлення шахрайства в електронній комерції. Запропонований підхід дозволив об’єднати кілька відкритих наборів даних зі збереженням реальних міток шахрайства, збагатити дані синтетичними атрибутами автентифікації та характеристиками українського платіжного ринку, забезпечити реалістичну кореляцію згенерованих полів із мітками шахрайства та сформувати додаткові набори для навчання моделей IP Insights й автоенкодера.

Подальша робота передбачає навчання та оптимізацію конвеєра МН на сформованому наборі, порівняльний аналіз з альтернативними алгоритмами та інтеграцію з хмарною інфраструктурою AWS SageMaker.

ЛІТЕРАТУРА / REFERENCE

1. Joint EBA-ECB report on payment fraud. 2025. URL: https://www.eba.europa.eu/publications-and-media/press-releases/joint-eba-ecb-report-payment-fraud-strong-authentication-remains-effective-fraudsters-are-adapting
2. Visa Payments & Fraud Report. 2025. URL: https://www.visaacceptance.com/content/dam/documents/campaign/fraud-report/global-fraud-report-2025.pdf
3. Ostrovska K., Nosov V. Machine learning methods for antifraud systems. Системні технології. 2025. Т. 5, вип. 160. С. 156–163. URL: https://doi.org/10.34185/1562-9945-5-160-2025-16
4. IEEE-CIS Fraud Detection. 2019. URL: https://www.kaggle.com/competitions/ieee-fraud-detection/overview
5. Credit Card Transactions Fraud Detection Dataset. 2020. URL: https://www.kaggle.com/datasets/kartik2112/fraud-detection
6. Fraudulent E-Commerce Transactions. 2024. URL: https://www.kaggle.com/datasets/shriyashjagtap/fraudulent-e-commerce-transactions.
7. Anti-Money Laundering Datasets. 2021. URL: https://github.com/IBM/AMLSim
8. Credit Card Fraud Detection. 2018. URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
9. EMV 3D-Secure. 2025. URL: https://www.emvco.com/emv-technologies/3-d-secure/

METHODOLOGY OF DATASET PREPARATION FOR TRAINING E-COMMERCE FRAUD DETECTION MODELS

Valerii Nosov, Kateryna Ostrovska

Abstract. *This study addresses the problem of preparing training data for machine learning-based fraud detection systems in e-commerce transactions. Based on the analysis of existing open sources, the necessity of creating a specialized dataset is justified. An automated pipeline is proposed for merging three open datasets from the Kaggle platform (IEEE-CIS, Sparkov, Fraudulent E-Commerce), preserving real fraud labels and enriching records with synthetic attributes adapted to the specifics of the Ukrainian payment market. Methods for the uniform normalization of timestamps, generation of authentication data, partitioning by payment systems, and the formation of aggregated customer profiles and pairs for training the IP Insights model have been developed. The result is a dataset comprising 500,000 transactions over a 24-month period with a fraud rate of 3.04%, designed to train a model pipeline that includes LightGBM, an autoencoder, and IP Insights.*

Keywords: *dataset, machine learning, transaction, e-commerce, LightGBM, autoencoder, IP Insights.*