«Системні технології» 3 (164) 2026 «System technologies»
DOI 10.34185/1562-9945-3-164-2026-03
УДК 004.85:004.62

В.О. Носов, К.Ю. Островська

АВТОМАТИЗОВАНИЙ КОНВЕЄР ФОРМУВАННЯ ДАТАСЕТУ ДЛЯ НАВЧАННЯ МОДЕЛЕЙ ВИЯВЛЕННЯ ШАХРАЙСТВА

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, Credit Card Transactions Fraud Detection Dataset, 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, синтетичні дані

Постановка проблеми. Стрімкий розвиток електронної комерції як одного з ключових секторів глобальної економіки неминуче супроводжується пропорційним зростанням масштабів платіжного шахрайства. Відповідно до спільного звіту [1] European Banking Authority (EBA) та European Central Bank (ECB), загальний обсяг транзакційного шахрайства в Європейському економічному просторі за 2024 рік сягнув 4.2 млрд євро, продемонструвавши зростання на 17% порівняно з попереднім роком. Зокрема, втрати від шахрайства з платіжними картками склали 1.3 млрд євро (зростання на 29%), причому переважна більшість таких інцидентів припадає саме на онлайн-операції (80–85% за обсягом).

Шахрайські схеми дедалі частіше експлуатують винятки з правил обов'язкової автентифікації клієнта (Strong Customer Authentication) та застосовують методи соціальної інженерії щодо користувачів. Масштабність проблеми підтверджується результатами глобального опитування [2], проведеного компанією Visa спільно з Merchant Risk Council – 98% опитаних представників роздрібної торгівлі стикалися щонайменше з одним типом шахрайства протягом року. При цьому понад 80% торговців відзначають суттєві труднощі в ефективному використанні накопичених даних для підвищення точ-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ності систем машинного навчання (МН), через що вдосконалення цих інструментів є однією з пріоритетних задач галузі.

Оскільки ефективність моделей МН для задач класифікації та виявлення аномалій критично залежить від якості та репрезентативності тренувальної вибірки, виникає науково-практична проблема доступу до релевантних даних. З огляду на сувору конфіденційність реальних транзакційних даних фінансових установ, для навчання моделей переважно використовуються відкриті набори даних, які часто характеризуються обмеженою схемою атрибутів.

Аналіз останніх досліджень і публікацій. У попередній роботі [3] було запропоновано концепцію багатошарової системи виявлення шахрайства в онлайн-транзакціях з використанням конвеєра моделей: класифікатора LightGBM, автоенкодера для виявлення аномалій та алгоритму IP Insights для аналізу мережевої поведінки. Навчання такого конвеєра потребує спеціалізованого набору даних, який одночасно містить транзакційні атрибути для класифікатора, поведінкові профілі клієнтів для автоенкодера та пари «entity–IP» для IP Insights. Саме тому якість та структура тренувальних даних є ключовим фактором успішності всієї системи.

Проблематиці підготовки та стандартизації даних для систем виявлення шахрайства присвячено ряд досліджень. У роботі «Fraud Dataset Benchmark and Applications» [4] систематизовано наявні відкриті датасети та запропоновано єдину методологію їх оцінювання, включаючи стандартизовані тренувально-тестові розбиття та спільні домовленості іменування ознак. Дослідники підкреслюють, що датасети для виявлення шахрайства мають специфічні властивості, які відрізняють їх від інших табличних еталонних наборів даних: екстремальний дисбаланс класів (частка шахрайства може сягати 0.01%), висока кардинальність ознак (IP-адреси, номери телефонів), змагальний характер задачі (шахраї адаптують поведінку) та критична роль конструювання ознак (feature engineering) на основі агрегатів та збагачення даних.

В іншому дослідженні [5] розкрили проблему генерації синтетичних даних для систем виявлення шахрайства з використанням дифузійних моделей. Автори продемонстрували, що синтетичне збагачення датасетів, зокрема набору IEEE-CIS Fraud Detection [6], суттєво покращує якість класифікації на незбалансованих даних. Ця робота підтверджує перспективність підходу синтетичної генерації атрибутів, який застосовано у даному дослідженні для збагачення записів полями протоколу EMV 3D-Secure (3DS2) [12].

Аналіз літератури виявив, що більшість публічних наборів мають суттєві обмеження, серед яких анонімізовані ознаки без семантичного контексту, відсутність інформації про пристрій або середовище, орієнтацію на ринки окремих країн. Це обумовило необхідність розробки спеціалізованого датасету, що стало предметом даного дослідження.

Мета дослідження. Метою дослідження є створення спеціалізованого набору даних транзакцій електронної комерції шляхом інтеграції та синтетичного збагачення відкритих джерел. Розроблений датасет адаптовано до специфіки українського платіжного ринку, структуровано відповідно до стандарту 3DS2 та призначено для нав-

чання багатошарової антишахрайської системи на основі конвеєра моделей LightGBM, автоенкодера та IP Insights.

Викладення основного матеріалу дослідження.

Вибір та характеристика вихідних наборів даних. Формування датасету базується на об'єднанні трьох відкритих масивів з платформи Kaggle, відібраних за критеріями наявності реальних або реалістичних міток шахрайства, достатнього обсягу та різноманітності атрибутів:

- IEEE-CIS Fraud Detection (IEEE-CIS) [6] (~590 тис. транзакцій, рівень шахрайства ~3.5%) – це датасет з реальними мітками шахрайства, створений у межах змагання Kaggle за участю Vesta Corporation. Складається з основної таблиці (суми, домени пошти, 339 анонімізованих ознак) та ідентифікаційної (браузер, пристрій). Головною перевагою масиву є наявність реальних даних про цифрові відбитки пристроїв (user agent). Серед суттєвих обмежень: відсутність абсолютних дат (лише зміщення у секундах), глибока анонімізація більшості ознак, відсутність інформації про торговця, орієнтація на ринок США та висока частка пропусків (приблизно 75%) у даних ідентифікації;
- Credit Card Transactions Fraud Detection Dataset (CCTFDD) [7] (~1.3 млн транзакцій, рівень шахрайства ~0.6%) – синтетичний набір, згенерований за допомогою інструменту Sparkov Data Generation [8] на основі статистичних розподілів реальних транзакцій. Його сильними сторонами є великий обсяг, наявність повних часових міток (2019–2020) та псевдоанонімізованих (хешованих) номерів карток. Обмеження включають синтетичне походження, орієнтацію на ринок США та повну відсутність автентифікаційних даних, інформації про пристрій та браузер;
- Fraudulent E-Commerce Transactions (FECT) [9] (~100 тис. записів, рівень шахрайства ~10%) – це датасет, сфокусований безпосередньо на e-commerce сценаріях. Ключова цінність цього масиву полягає у підвищеному рівні шахрайства, що забезпечує значну кількість позитивних прикладів для збалансування загальної вибірки. Обмеження: відносно невеликий обсяг, нестандартні назви колонок, відсутність інформації про пристрій і параметри автентифікації.

Об'єднання трьох наборів дозволяє компенсувати їхні індивідуальні недоліки, оскільки IEEE-CIS забезпечує реальні дані пристроїв, CCTFDD великий обсяг з повними датами, а FECT підвищену частку позитивних прикладів. Ключовою перевагою є збереження автентичних міток класів та оригінальних сум з усіх джерел.

Проведений аналіз альтернативних відкритих масивів даних довів їхню методологічну невідповідність завданням даного дослідження. Наприклад, датасет IBM AMLSim [10] орієнтований на задачі виявлення відмивання коштів, тоді як широковідомий Credit Card Fraud Dataset [11] містить виключно 28 анонімізованих PCA-компонент, що повністю позбавляє дані семантичного контексту.

Конвеєр уніфікації даних. Оскільки кожне джерело має власний формат, набір колонок та способи кодування, розроблено автоматизований конвеєр обробки даних

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

(див. рис. 1). Конвеєр реалізовано як скрипт з окремими функціями-завантажувачами для кожного джерела.

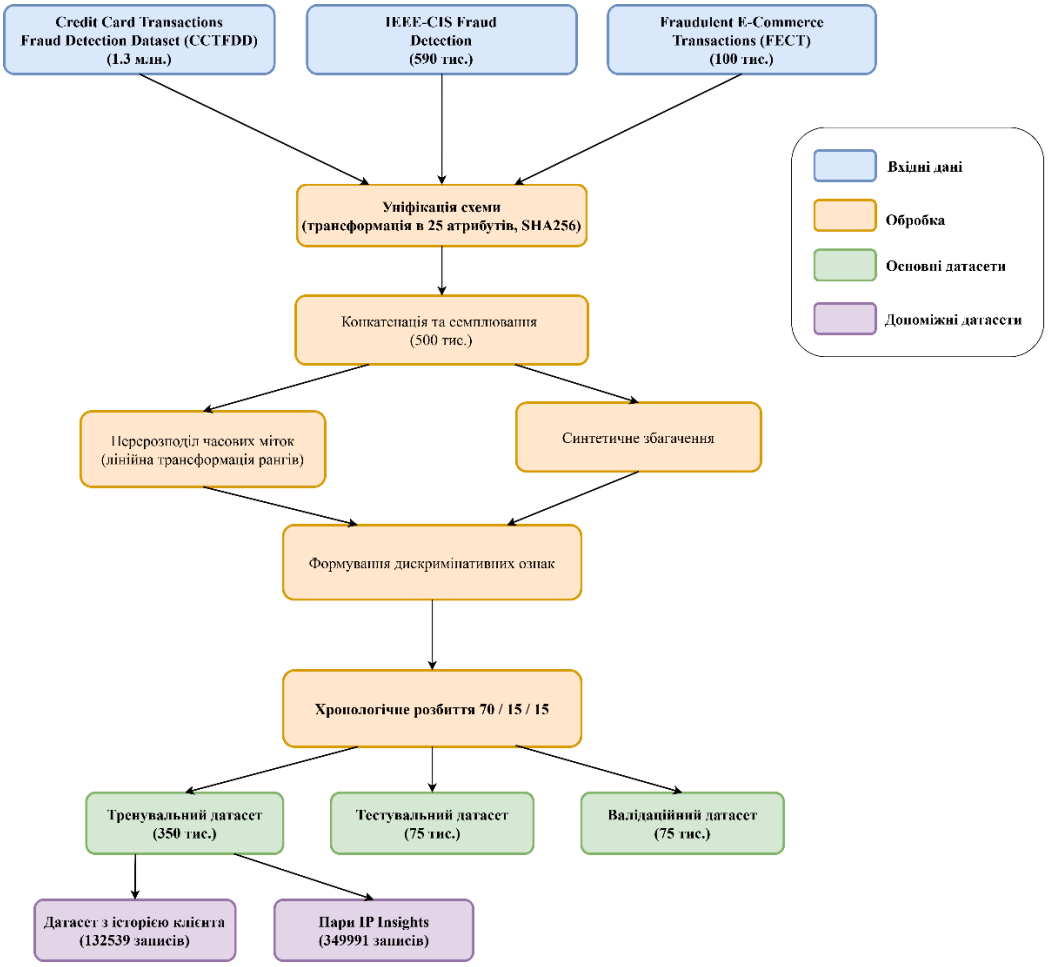

Рисунок 1 – Схема конвеєра уніфікації та злиття даних

Загальна послідовність етапів:

- завантаження та зіставлення полів кожного набору до єдиної цільової структури з 25 атрибутів;
- псевдоанонімізація ідентифікаторів карток через SHA256 хешування;
- конкатенація трьох уніфікованих наборів;
- стратифіковане семплювання до 500000 записів (метод відбору, що зберігає пропорційний розподіл кожного джерела та співвідношення шахрайських і легітимних транзакцій у вибірці);
- рівномірний перерозподіл часових міток;
- синтетичне збагачення атрибутами 3DS2 [12];
- диференціація розподілів (закладення залежностей між ознаками та міткою шахрайства);
- хронологічне розбиття та формування допоміжних наборів.

Рівномірний перерозподіл часових міток. Вихідні масиви охоплюють різні часові діапазони (CCTFDD: 2019–2020, IEEE-CIS: зміщення в секундах, FECT: довільний період) та мають суттєво різну кількість записів. Просте лінійне масштабування чи

пряме об'єднання призвело б до штучної кластеризації даних – в одні місяці домінували б транзакції з одного джерела, в інші з іншого, що викликало б нерелевантні зміщення рівня шахрайства (dataset shift) та перенавчання моделей на часові артефакти.

Для забезпечення рівномірної щільності транзакцій у цільовому часовому діапазоні (01.01.2023–31.12.2024) застосовано метод нормалізації на основі лінійного відображення порядкових рангів (квантильного перетворення). Після злиття всіх масивів транзакції сортуються за оригінальними часовими мітками (зберігаючи їхній відносний хронологічний порядок), після чого кожній транзакції з рангом i серед N записів присвоюється нова дата за формулою:

$$t'(i) = t_{start} + \frac{i}{N} \times (t_{end} - t_{start}), \quad (1)$$

де $t'(i)$ – нове значення дати для транзакції з порядковим рангом i;

t_{start} – початкова дата цільового часового діапазону;

t_{end} – кінцева дата цільового часового діапазону;

i – порядковий номер (ранг) транзакції у масиві після сортування за оригінальним часом ($0 \leq i \leq N$);

N – загальна кількість записів у об'єднаному наборі даних.

Хоча такий підхід нівелює глобальні макроекономічні сплески (наприклад, святкові розпродажі), він гарантує ідеально збалансоване тренувальне середовище. При цьому оригінальний час доби строго зберігається, що дозволяє моделям засвоювати критично важливі внутрішньодобові закономірності шахрайської активності. Для остаточного усунення ризику кластеризації за джерелом походження, транзакції в межах кожного згенерованого дня додатково перемішуються за допомогою детермінованого генератора псевдовипадкових чисел.

Синтетичне збагачення атрибутами. Як цільову схему транзакцій обрано протокол 3DS2 [12]. Такий вибір обґрунтовано кількома причинами. По-перше, стандарт визначає чітко структурований набір параметрів транзакції, який охоплює всі ключові ознаки, необхідні для побудови антишахрайської системи. По-друге, формалізованість протоколу забезпечує однозначну інтерпретацію кожного поля, що спрощує як генерацію синтетичних атрибутів, так і подальший аналіз результатів моделей. По-третє, розроблювана система сфокусована безпосередньо на етапі автентифікації платежу, де приймається рішення про схвалення або відхилення транзакції, а 3DS2 є стандартом, що регулює цей етап у сучасних онлайн-платежах. Оскільки жоден із відкритих наборів не містить потрібних атрибутів, їх згенеровано синтетично на основі актуальних ринкових даних:

- платіжні системи – за даними Національного банку України [13], станом на початок 2026 року частка Visa серед активних карток становила 50.7%, Mastercard 49%, національна система Простір близько 0.3%. У датасеті використано спрощений розподіл: Visa 50%, Mastercard 40%, Простір 10%. Частка системи Простір свідомо за-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

вищена порівняно з реальними даними для забезпечення достатнього обсягу вибірки для навчання моделі;

- тип автентифікації – стандарт 3DS2 передбачає кілька типів обробки транзакцій: challenge (з додатковою верифікацією), frictionless (без взаємодії з користувачем) та stand-in processing (обробка на стороні платіжної системи);
- операційні системи (ОС) та браузери – за даними StatCounter [14], частка Android на мобільному ринку України у 2026 році становила 65–70%, iOS 20–35%. У датасеті Android на рівні 70%, iOS 20%, Windows 5%, macOS 3%, Linux 2%, з кореляцією браузерів до ОС;
- еквайрери – розподіл відображає структуру ринку інтернет-еквайрингу України. У датасеті серед еквайрерів: Privat24 (20%), Liqpay (18%), WayForPay (15%), Platon (15%), MonoBank (7%), Fondy (5%) та інші;
- валюти та торговці – домінування гривні (UAH, 70%) відповідає специфіці українського e-commerce. Решту обсягу розподілено між USD (10%), EUR (10%) та іншими валютами. Датасет включає 114 торговців за 16 категоріями, серед яких приблизно 85% українські та 15% іноземні.

Загальний опис атрибутів. За результатами роботи запропонованого конвеєра формуються три взаємопов'язані масиви даних, кожен з яких призначений для відповідного компонента багатошарової моделі МН. Основний набір транзакцій (тренувальна, валідаційна та тестова вибірки) має єдину уніфіковану структуру, що складається з 25 атрибутів (див. таблиця 1). Ця схема концептуально відповідає специфікації розширеного повідомлення протоколу 3DS [12] і охоплює ключові вектори аналізу: фінансовий, географічний, мережевий та апаратний.

Таблиця 1

Структура основного набору транзакцій

Атрибут	Опис
id	Унікальний ідентифікатор транзакції
device_channel	Автентифікаційний канал
auth_type	Тип автентифікації
merchant_id, merchant_mcc, merchant_name, merchant_country_code	Ідентифікатор, категорія, назва та країна торговця
product_category	Категорія товарів
purchase_amount, purchase_currency, amount_usd	Сума в оригінальній валюті, тип валюти, нормалізована сума в USD
pno, card_bin	Платіжна система та BIN картки
card_id	Хешований ідентифікатор картки
billing_country_code	Країна покупки
email_domain	Домен електронної пошти
sender_browser, sender_browser_version	Браузер та його версія

sender_device_ip	IPv4-адреса пристрою
sender_device_model, sender_os, sender_os_version	Модель пристрою, операційна система та версія
three_ds_version	Версія протоколу 3DS
acq_id	Ідентифікатор еквайрера
is_recurring	Ознака рекурентного платежу
timestamp, transaction_hour	Дата/час транзакції та година
is_fraud	Цільовий ідентифікатор шахрайства

Для аналізу історичної поведінки користувачів формується допоміжний масив агрегованих клієнтських профілів (див. таблиця 2). Він містить розраховані статистичні та поведінкові метрики для кожної унікальної картки. Зазначені ознаки відіграють подвійну роль, оскільки вони утворюють вхідний вектор для навчання автоенкодера (який моделює закономірності легітимної поведінки та генерує оцінку похибки відновлення), а також слугують критично важливими додатковими параметрами для фінального класифікатора LightGBM.

Таблиця 2

Структура датасету клієнтських профілів

Атрибут	Опис
client_id	Хешований ідентифікатор картки
transactions_count_last_day	Кількість транзакцій за останню добу
transactions_amount_last_day	Сума транзакцій за останню добу
days_since_last_transaction	Кількість днів з останньої операції
avg_transaction_amount	Середня сума транзакції за весь період
max_transaction_amount_30d	Максимальна сума транзакції за 30 днів
unique_merchants_7d	Кількість унікальних торговців за 7 днів
unique_ips_7d	Кількість унікальних IP-адрес за 7 днів
unique_devices_7d	Кількість унікальних пристроїв за 7 днів
fraud_count_90d	Кількість шахрайських операцій за 90 днів

Третій масив є спеціалізованим набором пар «entity–IP», що має мінімалістичну структуру з двох колонок без заголовків (див. таблиця 3). Такий формат відповідає вхідним вимогам алгоритму Amazon SageMaker IP Insights, який вивчає латентні векторні представлення та обчислює оцінку аномальності.

Таблиця 3

Структура масиву пар «entity-IP»

Атрибут	Опис
entity_id	Хешований ідентифікатор картки
ip_address	IPv4-адреса пристрою

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

З метою дотримання методологічної чистоти та запобігання data leakage, обидва допоміжні набори обчислюються та формуються виключно на основі тренувальної вибірки (див. рис. 1).

Статистичні характеристики результуючого датасету. Основна зведена статистика по поточній версії датасету представлена у таблиці 4.

Таблиця 4

Статистика по отриманому 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%)
Платіжні системи	Visa 50%, Mastercard 40%, Простір 10%

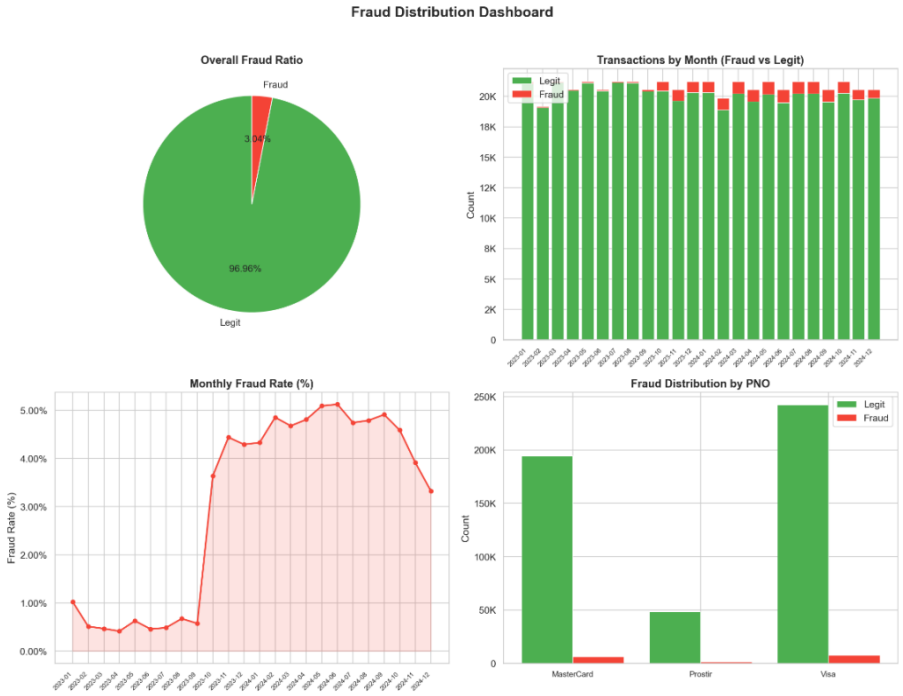

Рисунок 2 – Розподіл шахрайства за місяцями та платіжними системами

Досягнутий рівномірний розподіл транзакцій за місяцями (у середньому ~20800 записів/міс.) підтверджує коректність застосованого алгоритму квантильного перетворення часових міток. Крім того, середня сума шахрайських транзакцій ($460–580) суттєво перевищує аналогічні показники для легітимних операцій ($45–190), що формує дискримінативний сигнал для ефективного розділення класів моделями МН. Для візуальної перевірки підготовлених даних було побудовано низку аналітичних дашбордів.

Зокрема, на рис. 2 представлено динаміку розподілу шахрайства за місяцями та платіжними системами.

Як видно з рис. 2, запропонований підхід забезпечив рівномірний розподіл транзакцій та дотримання цільових пропорцій розподілу платіжних систем. Певне зростання рівня шахрайства з жовтня 2023 року об'єктивно пояснюється специфікою вихідних даних, оскільки у першій половині часового діапазону домінують транзакції з масиву CCTFDD (~0.6%), а в другій з масиву IEEE-CIS (~3.5%). Для більш глибокого розуміння атрибутивного складу датасету на рис. 3 візуалізовано розподіл за типами автентифікації та географічно-валютними ознаками відповідно.

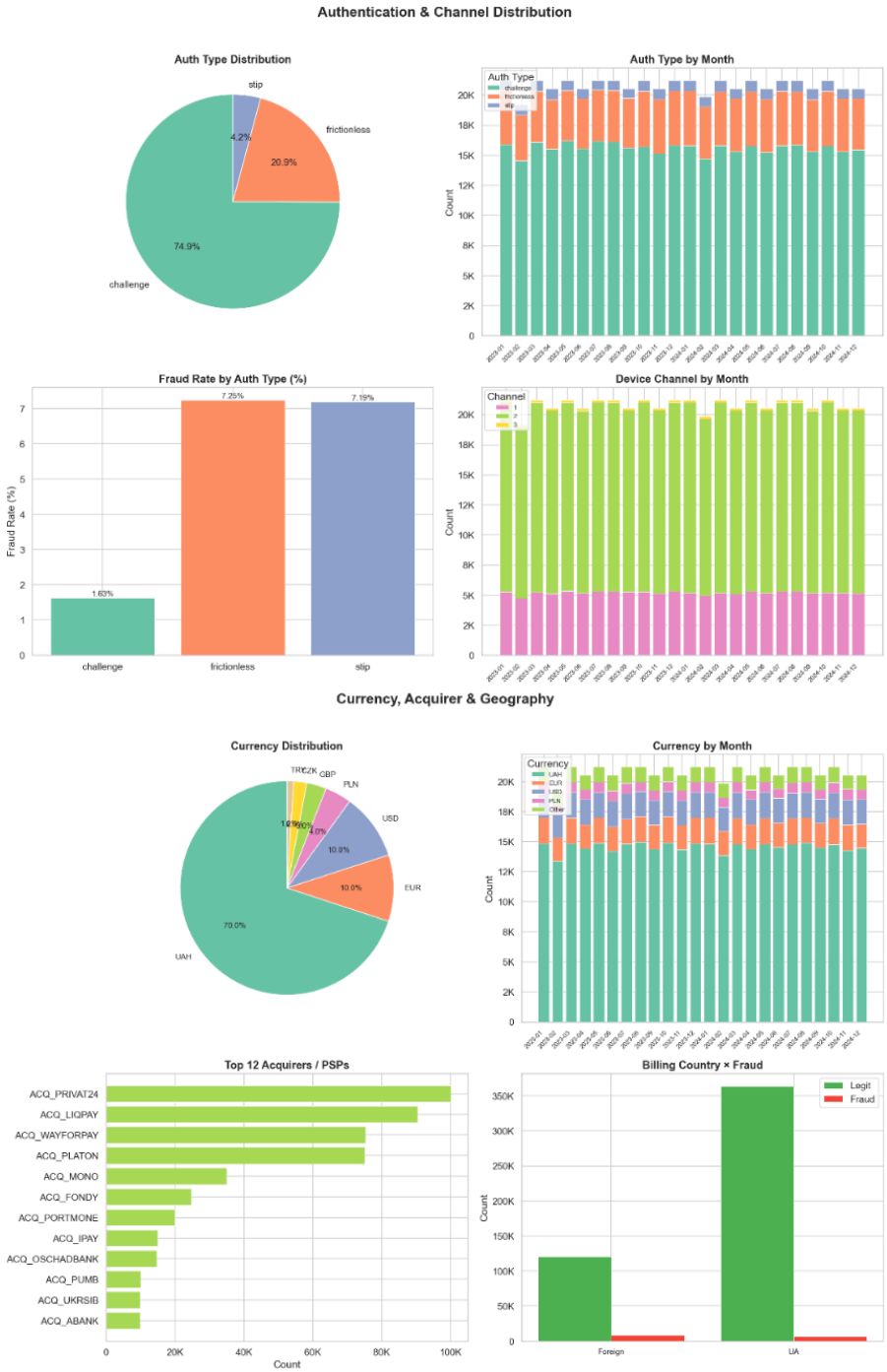

Рисунок 3 – Розподіл транзакцій за атрибутами

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

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.

Висновки. Описано методологію підготовки набору даних транзакцій для навчання конвеєра моделей виявлення шахрайства в електронній комерції. Запропонований підхід дозволив інтегрувати кілька відкритих наборів даних зі збереженням автентичних міток шахрайства, збагатити дані синтетичними атрибутами автентифікації з урахуванням специфіки українського платіжного ринку, а також сформувати спеціалізовані масиви для навчання моделей IP Insights та автоенкодера.

Подальша робота передбачає вдосконалення датасету шляхом додавання поведінкових часових рядів, а також безпосереднє навчання та оптимізацію розробленого конвеєра МН, його порівняльний аналіз з альтернативними алгоритмами та розгортання у хмарному середовищі AWS SageMaker.

ЛІТЕРАТУРА

1. Joint EBA-ECB report on payment fraud. European Banking Authority, European Central Bank. 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 (дата звернення: 15.03.2026).
2. Global eCommerce Payments & Fraud Report. Visa Acceptance Solutions, Merchant Risk Council. 2025.
URL: https://www.visaacceptance.com/content/dam/documents/campaign/fraud-report/global-fraud-report-2025.pdf (дата звернення: 16.03.2026).
3. Ostrovska K., Nosov V. Machine learning methods for antifraud systems. Системні технології. 2025. Т. 5, вип. 160. С. 156–163. DOI: 10.34185/1562-9945-5-160-2025-16.
4. Grover P., Xu J., Tittelfitz J. et al. Fraud Dataset Benchmark and Applications. Amazon Science. 2022. DOI: 10.48550/arXiv.2208.14417.
5. Pushkarenko Y., Zaslavskyi V. Synthetic Data Generation for Fraud Detection Using Diffusion Models. Information Systems and Innovative Technologies in Professional Activity (ISIJ). 2024. Vol. 55, No. 2. P. 185–198. DOI: 10.11610/isij.5534.
6. IEEE-CIS Fraud Detection: Kaggle Competition. 2019.
URL: https://www.kaggle.com/competitions/ieee-fraud-detection/overview (дата звернення: 15.03.2026).
7. Credit Card Transactions Fraud Detection Dataset : Kaggle Dataset. 2020. URL: https://www.kaggle.com/datasets/kartik2112/fraud-detection (дата звернення: 15.03.2026).
8. Sparkov Data Generation: GitHub Repository.
URL: https://github.com/namebrandon/Sparkov_Data_Generation (дата звернення: 15.03.2026).
9. Fraudulent E-Commerce Transactions: Kaggle Dataset. 2024.
URL: https://www.kaggle.com/datasets/shriyashjagtap/fraudulent-e-commerce-transactions (дата звернення: 15.03.2026).
10. Anti-Money Laundering Datasets (IBM AMLSim): GitHub Repository. 2021.
URL: https://github.com/IBM/AMLSim (дата звернення: 15.03.2026).

11. Credit Card Fraud Detection Dataset. Machine Learning Group, Université Libre de Bruxelles. 2018. URL: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (дата звернення: 15.03.2026).
12. EMV 3-D Secure Protocol and Core Functions Specification v2.2.0. EMVCo. 2025. URL: https://www.emvco.com/emv-technologies/3-d-secure/ (дата звернення: 20.03.2026).
13. Visa випередила Mastercard за кількістю карток в обігу в Україні. Forbes Ukraine. 2025. URL: https://forbes.ua/news/visa-viperedila-mastercard-za-kilkistyu-kartok-v-obigu-v-ukraini-27052025-30063 (дата звернення: 18.03.2026).
14. Mobile Operating System Market Share Ukraine. StatCounter. 2024. URL: https://gs.statcounter.com/os-market-share/mobile/ukraine (дата звернення: 20.03.2026).

REFERENCES

1. European Banking Authority & European Central Bank. (2025). *Joint EBA-ECB report on payment fraud*. 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 Acceptance Solutions & Merchant Risk Council. (2025). *2025 Global eCommerce Payments & Fraud Report*.
https://www.visaacceptance.com/content/dam/documents/campaign/fraud-report/global-fraud-report-2025.pdf
3. Ostrovska, K., & Nosov, V. (2025). Machine learning methods for antifraud systems. *System technologies, 5*(160), 156–163. https://doi.org/10.34185/1562-9945-5-160-2025-16
4. Grover, P., Xu, J., Tittelfitz, J., Cheng, A., Li, Z., Zablocki, J., Liu, J., & Zhou, H. (2022). Fraud Dataset Benchmark and Applications. *Amazon Science*.
https://doi.org/10.48550/arXiv.2208.14417
5. Pushkarenko, Y., & Zaslavskyi, V. (2024). Synthetic Data Generation for Fraud Detection Using Diffusion Models. *Information Systems and Innovative Technologies in Professional Activity (ISIJ), 55*(2), 185–198. https://doi.org/10.11610/isij.5534
6. IEEE-CIS Fraud Detection. (2019). *Kaggle*. https://www.kaggle.com/competitions/ieee-fraud-detection/overview
7. Credit Card Transactions Fraud Detection Dataset. (2020). *Kaggle*.
https://www.kaggle.com/datasets/kartik2112/fraud-detection
8. Sparkov Data Generation. *GitHub*.
https://github.com/namebrandon/Sparkov_Data_Generation
9. Fraudulent E-Commerce Transactions. (2024). *Kaggle*.
https://www.kaggle.com/datasets/shriyashjagtap/fraudulent-e-commerce-transactions
10. Anti-Money Laundering Datasets (IBM AMLSim). (2021). *GitHub*.
https://github.com/IBM/AMLSim
11. Credit Card Fraud Detection Dataset. (2018). Machine Learning Group, Université Libre de Bruxelles. *Kaggle*. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
12. EMVCo. (2025). *EMV 3-D Secure Protocol and Core Functions Specification v2.2.0*. https://www.emvco.com/emv-technologies/3-d-secure/
13. Visa vyperedyla Mastercard za kilkistiu kartok v obihu v Ukraini [Visa overtook Mastercard by number of cards in circulation in Ukraine]. (2025). *Forbes Ukraine*.
https://forbes.ua/news/visa-viperedila-mastercard-za-kilkistyu-kartok-v-obigu-v-ukraini-27052025-30063 [in Ukrainian].

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

14. StatCounter. (2024). *Mobile Operating System Market Share Ukraine*. https://gs.statcounter.com/os-market-share/mobile/ukraine

Received 30.03.2026.
Accepted 01.04.2026.
Published 30.04.2026

Automated pipeline for building a fraud detection training dataset

The study addresses the problem of preparing training data for machine learning-based fraud detection systems in e-commerce transactions. Due to the strict confidentiality of real transaction data, researchers often rely on publicly available datasets that typically suffer from limited attribute schemas, anonymized features, and a focus on specific national markets. An analysis of existing open datasets revealed the necessity of creating a specialized dataset, as none of the available sources provide a sufficient combination of realistic fraud labels, semantic transparency of features, and domain-specific attributes required for training a multi-component fraud detection system.

An automated pipeline for integrating three open Kaggle datasets (IEEE-CIS, Credit Card Transactions Fraud Detection Dataset, Fraudulent E-Commerce) is proposed. The pipeline preserves authentic fraud labels and original transaction amounts while enriching records with synthetic attributes adapted to the specifics of the Ukrainian payment market. The methods developed include: uniform normalization of timestamps based on quantile rank transformation to eliminate dataset shift artifacts while preserving intra-day patterns, synthetic generation of authentication attributes according to the EMV 3D-Secure 2.0 standard with payment network distributions based on National Bank of Ukraine statistics, formation of aggregated client behavioral profiles, and generation of "entity-IP" pairs for IP Insights model training. Both auxiliary datasets are derived exclusively from the training subset to prevent data leakage.

The resulting dataset comprises 500000 transactions spanning 24 months with a fraud rate of 3.04%, designed for training a model pipeline that includes LightGBM, an autoencoder, and IP Insights. The chronological split simulates real-world deployment conditions where models are trained on historical events and evaluated on future ones.

Keywords: dataset, machine learning, transaction, e-commerce, LightGBM, autoencoder, IP Insights, EMV 3D-Secure, fraud detection, synthetic data.

Носов Валерій Олександрович – аспірант кафедри інформаційних технологій і систем Українського державного університету науки і технологій.
ORCID: https://orcid.org/0009-0003-5841-8995
Островська Катерина Юріївна – к.т.н., доцент, доцент кафедри інформаційних технологій і систем Українського державного університету науки і технологій.
ORCID: https://orcid.org/0000-0002-9375-4121

Nosov Valerii Oleksandrovych – postgraduate student of the Department of Information Technology and Systems of Ukrainian State University of Science and Technology.
ORCID: https://orcid.org/0009-0003-5841-8995
Ostrovska Kateryna Yuriivna – Ph.D., Associate Professor of the Department of Information Technology and Systems of Ukrainian State University of Science and Technology.
ORCID: https://orcid.org/0000-0002-9375-4121