ПРИНЦИПОВІ КОНЦЕПЦІЇ ТА СТРУКТУРУВАННЯ РІЗНИХ РІВНІВ ОСВІТИ З ОПТИКО-ЕЛЕКТРОННИХ ІНФОРМАЦІЙНО-ЕНЕРГЕТИЧНИХ ТЕХНОЛОГІЙ

УДК 004.89

Д.І. УГРИН, Ю.О. УШЕНКО, Ю.Я. ТОМКА, С.В. ПАВЛОВ, М.В. ТАЛАХ, Л.І. Д'ЯЧЕНКО, К.П. ГАЗДЮК

ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ В КОНТЕКСТІ ПІДГОТОВКИ ФАХІВЦІВ В СФЕРІ ТРАНСФЕРУ ТЕХНОЛОГІЙ ТА БЕЗПЕКИ ФІНАНСОВИХ ТРАНЗАКЦІЯХ

Чернівецький національний університет ім. Ю. Федьковича, Коцюбинського, 2, м. Чернівці, Україна, Вінницький національний технічний університет, Україна

Анотація. Стаття присвячена актуальній проблемі виявлення шахрайських аномалій у фінансових транзакціях за допомогою методів машинного навчання. В умовах стрімкої цифрової трансформації фінансових систем та зростання обсягів транзакцій традиційні методи виявлення шахрайства стають неефективними, що підкреслює нагальну потребу впровадження автоматизованих та адаптивних рішень. Дослідження базується на поетапному підході, що включає підготовку та обробку даних, побудову та навчання моделей класифікації, а також оцінку їх ефективності. Було проведено порівняльний аналіз семи популярних алгоритмів машинного навчання: лінійної регресії, дерев рішень, випадкового лісу (Random Forest), нейронних мереж, градієнтного бустингу (Gradient Boosting), XGBoost та SVC. Ключові результати дослідження показали, що ансамблеві методи демонструють найвищу ефективність у виявленні шахрайства: Random Forest, Gradient Boosting та XGBoost виявилися найбільш доцільними для задач виявлення шахрайства, демонструючи стабільно високі результати. Це особливо важливо з огляду на типовий дисбаланс класів (невелика кількість шахрайських транзакцій порівняно з легітимними) у реальних фінансових даних. Ефективність моделей значно перевершує інші розглянуті алгоритми, що вказує на їхню здатність виявляти складні, неочевидні закономірності в даних. Було підтверджено критичну важливість правильного налаштування гіперпараметрів моделей та урахування дисбалансу класів для досягнення максимальної точності та повноти виявлення шахрайських транзакцій. Це дозволяє уникнути перенавчання на домінуючому класі та підвищити чутливість системи до рідкісних, але важливих шахрайських випадків. Практична значущість дослідження полягає в тому, що запропонований підхід дозволяє фінансовим установам значно підвищити операційну ефективність, мінімізувати фінансові втрати та зміцнити довіру клієнтів. Впровадження таких систем забезпечує комплексний та адаптивний захист фінансової системи у сучасному динамічному цифровому середовищі. Результати дослідження підтверджують ефективність машинного навчання як потужного інструменту для боротьби з фінансовим шахрайством.

Ключові слова: машинне навчання, фінансове шахрайство, прогнозування аномалій, ансамблеві методи, Random Forest, Gradient Boosting, XGBoost, трансфер технологій, дисбаланс класів, фінансові транзакції.

Abstract. The article is devoted to the topical problem of detecting fraudulent anomalies in financial transactions using machine learning methods. In the context of rapid digital transformation of financial systems and growth in transaction volumes, traditional methods of fraud detection are becoming ineffective, which highlights the urgent need to implement automated and adaptive solutions. The research is based on a step-by-step approach that includes data preparation and processing, building and training classification models, and evaluating their effectiveness. A comparative analysis of seven popular machine learning algorithms was conducted: linear regression, decision trees, random forest, neural networks, gradient boosting, XGBoost, and SVC. The key findings of the study showed that ensemble methods demonstrate the highest effectiveness in detecting fraud: Random Forest, Gradient Boosting, and XGBoost proved to be the most suitable for fraud detection tasks, demonstrating consistently high results. This is especially important given the typical class imbalance (a small number of fraudulent transactions compared to legitimate ones) in real financial data. The effectiveness of the models significantly outperforms the other algorithms considered, indicating their ability to detect complex, non-obvious patterns in the data. The critical importance of correctly configuring model hyperparameters and accounting for class imbalance to achieve maximum accuracy and completeness in detecting fraudulent transactions has been confirmed.

This avoids overfitting on the dominant class and increases the system's sensitivity to rare but important fraudulent cases. The practical significance of the study lies in the fact that the proposed approach allows financial institutions to significantly improve operational efficiency, minimize financial losses, and strengthen customer trust. The implementation of such systems provides comprehensive and adaptive protection of the financial system in today's dynamic digital environment. The results of the study confirm the effectiveness of machine learning as a powerful tool for combating financial fraud.
Keywords: machine learning, financial fraud, anomaly prediction, ensemble methods, Random Forest, Gradient Boosting, XGBoost, transfer of technology, class imbalance, financial transactions.

DOI: 10.31649/1681-7893-2025-50-2-13-29

ВСТУП

У сучасному світі фінансові системи зазнають стрімкої цифрової трансформації, що супроводжується зростанням обсягів транзакцій і ускладненням фінансових операцій. Водночас ці зміни призводять до зростання ризиків, пов'язаних із шахрайськими діями, які становлять серйозну загрозу як для фінансових установ, так і для їхніх клієнтів. Шахрайство у фінансовому секторі набуває все більш витончених форм, використовуючи сучасні технології для обходу традиційних систем захисту.

За даними Асоціації сертифікованих фахівців із розслідування шахрайства (ACFE), щорічні втрати від фінансового шахрайства сягнули понад 5% доходу організацій у всьому світі. У звіті за 2024 рік зазначено, що 53% випадків шахрайства були пов'язані з факторами пандемії COVID-19, а середній розмір збитків уперше зріс з 2016 року. Зокрема, зловмисники дедалі частіше використовують криптовалюти для приховування слідів злочину, а також діють у регіонах, де фінансовий контроль є слабшим. У середньому типовий випадок шахрайства триває близько 12 місяців, перш ніж його буде виявлено, що свідчить про критичну необхідність більш ефективних засобів моніторингу.

У таких умовах традиційні методи виявлення шахрайства — базовані на статичних правилах і ручному аналізі — вже не відповідають сучасним викликам. Вони не здатні своєчасно опрацьовувати великі обсяги даних і розпізнавати складні аномальні шаблони поведінки користувачів. Саме тому все більшої актуальності набуває впровадження технологій машинного навчання та штучного інтелекту, які дозволяють створювати автоматизовані, адаптивні системи для прогнозування шахрайських транзакцій у режимі реального часу.

Ця стаття присвячена розробленню підходу до виявлення фінансових аномалій із використанням алгоритмів машинного навчання. Метою дослідження є формування ефективної системи, здатної ідентифікувати підозрілі транзакції на ранніх етапах, мінімізуючи тим самим фінансові втрати, репутаційні ризики та витрати на юридичні процеси. Крім того, автоматизація таких процесів дозволяє значно підвищити операційну ефективність, відповідати сучасним нормативним вимогам і зміцнити довіру клієнтів до фінансових послуг.

Таким чином, застосування передових аналітичних технологій створює передумови для комплексного захисту фінансової системи, забезпечуючи її стійкість у динамічному цифровому середовищі та трансферу технологій.

1. ОГЛЯД ЛІТЕРАТУРИ ТА АНАЛОГІВ ДОСЛІДЖЕНЬ

Цифрова трансформація фінансового сектору докорінно змінила характер здійснення транзакцій, зробивши їх значно зручнішими, швидшими та доступнішими для кінцевих користувачів. У сучасній фінансовій екосистемі активно розвиваються цифровий банкінг, мобільні платіжні системи, криптовалюти та інші фінансово-технічні рішення, які сприяють зростанню інтегрованості та взаємозалежності фінансових процесів. Однак разом із цими перевагами зростають і ризики, зокрема у сфері фінансового шахрайства, яке становить серйозну загрозу стабільності фінансових установ і довірі споживачів.

Фінансове шахрайство охоплює широкий спектр злочинної діяльності, зокрема викрадення персональних даних, фішинг, шахрайські транзакції з використанням платіжних карток, відмивання грошей тощо. Згідно з дослідженнями [1], ефективна боротьба з цими загрозами є складним завданням, особливо з урахуванням стрімкого зростання обсягів транзакцій і постійної еволюції шахрайських схем.

Традиційні методи виявлення шахрайства — такі як системи на основі правил, фільтри ризику та ручна перевірка транзакцій — зазвичай ґрунтуються на фіксованих критеріях (сума, геолокація, частота транзакцій тощо). Однак в умовах високої динаміки цифрових операцій ці підходи демонструють обмежену ефективність [2, 3]. Часто виявлення порушень відбувається постфактум, що призводить до фінансових втрат і зниження довіри клієнтів. Поширення спроб шахрайства з оплатою протягом років продемонстровано на рисунку 1.

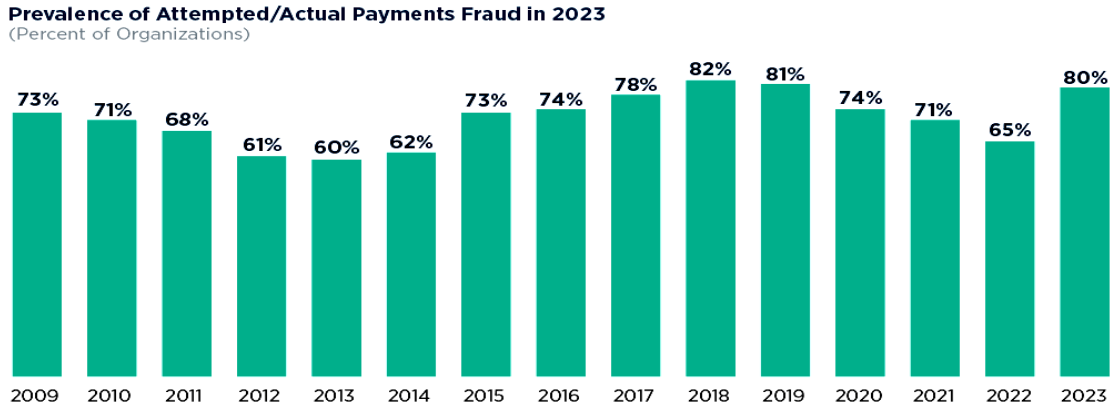

Рисунок 1 – Поширення спроб шахрайства з оплатою протягом років

З появою сучасних технологій — зокрема машинного навчання (ML), штучного інтелекту (AI) та великих даних (Big Data) — з'явилися нові можливості для виявлення та запобігання складним схемам шахрайства. Алгоритми ML дозволяють здійснювати аналіз транзакцій у режимі реального часу, виявляти аномалії та адаптуватися до змін у поведінці зловмисників [4]. Значну роль відіграє також поведінкова аналітика, яка дає змогу формувати профілі клієнтів і фіксувати відхилення від типових патернів взаємодії [5].

Крім того, технології блокчейн і розподілені реєстри (DLT) підвищують прозорість фінансових процесів та ускладнюють реалізацію шахрайських дій завдяки неможливості підробки або зміни даних у реєстрі без сліду [6].

У відкритому доступі повноцінні програмні реалізації систем виявлення фінансового шахрайства трапляються рідко. Більшість таких рішень розробляються в межах комерційних контрактів для конкретних фінансових установ або підприємств. Дані, що використовуються для навчання моделей, зазвичай є конфіденційними й адаптованими під специфіку конкретного бізнесу, що унеможливлює їх широке використання в академічних або відкритих проєктах.

Найбільш наближеними аналогами є дослідницькі моделі, оприлюднені на платформах на кшталт Kaggle, де дослідники працюють з відкритими датасетами (наприклад, Credit Card Fraud Detection Dataset), виконуючи повний цикл обробки: від очищення й нормалізації до побудови моделей і візуалізації результатів. Такі проєкти дозволяють оцінити ефективність різних алгоритмів в умовах наявного набору даних, але їх практична інтеграція у реальні бізнес-процеси обмежена через відмінності в обсягах, структурі та динаміці даних.

Наукова література свідчить про зростаючий інтерес до застосування методів штучного інтелекту в боротьбі з фінансовим шахрайством. У роботах [7, 8] досліджено застосування нейронних мереж, дерев рішень, наївного баєсівського класифікатора та ансамблевих моделей. Автори [9] запропонували використання рекурентних нейронних мереж (RNN) для обробки послідовностей транзакцій, що забезпечило значно кращі результати в порівнянні з класичними алгоритмами.

Дослідники також акцентують на необхідності постійної адаптації моделей до нових форм шахрайства [9]. З цією метою пропонуються гібридні системи, які поєднують наглядові та безнаглядові методи навчання, що дає змогу виявляти нові, ще не класифіковані патерни зловживань.

Згідно з аналітичними звітами (ACFE, 2022; PwC, 2023), боротьба з шахрайством дедалі більше стає міждисциплінарною задачею, яка охоплює не лише сфери ризик-менеджменту та IT, але й маркетинг, клієнтське обслуговування та стратегічне управління. Сучасні організації повинні забезпечувати інтеграцію аналітики ризиків у всі бізнес-процеси, створюючи злагоджені команди, що об'єднують спеціалістів із різних галузей.

Традиційно оцінка вартості фінансового шахрайства включає розрахунок прямих і непрямих витрат: втрат від шахрайських дій, витрат на програмне забезпечення, заробітної плати аналітиків, витрат на юридичну підтримку та альтернативні витрати, які виникають через порушення довіри клієнтів. Візуалізацію цього підходу подано на схемі нижче (рис. 2).

Рисунок 2 – Схема проведення розрахунку втрат від шахрайства

Ця схема ілюструє загальну оцінку вартості шахрайства для організації, об'єднуючи різні компоненти, що складають фінансовий тягар від шахрайських дій. Вона включає чотири основні складові:

1. Фактичні втрати від шахрайства – це прямі фінансові збитки, яких зазнала організація внаслідок шахрайських дій (наприклад, незаконне зняття коштів, підробка транзакцій, викрадення даних тощо).

2. Ресурси, чисельність персоналу та заробітна плата (ЗП) – витрати на людські ресурси, задіяні у виявленні, розслідуванні та запобіганні шахрайству, включаючи їхню заробітну плату, робочий час та інші витрати, пов'язані з управлінням ризиками.

3. Засоби та працівники для захисту від шахрайства – інвестиції у програмне забезпечення, технічні системи безпеки, аналітичні інструменти, а також оплата праці фахівців із кібербезпеки та комплаєнсу.

4. Відсоток прибутку, витрачений на шахрайство – це частина загального прибутку компанії, що фактично йде на покриття витрат, пов'язаних із шахрайством — як прямих, так і непрямих.

Ця схема демонструє, що вартість шахрайства не обмежується лише прямими втратами, а охоплює весь спектр витрат: на персонал, технології, засоби захисту та втрачений прибуток. Саме тому організації прагнуть впроваджувати ефективні системи виявлення та попередження шахрайства, щоб мінімізувати ці сукупні витрати.

2. ПОСТАНОВКА ЗАДАЧІ ТА КЛЮЧОВІ АСПЕКТИ

Метою дослідження є розробка ефективної системи в контексті трансферу технологій для виявлення шахрайських транзакцій на основі методів машинного навчання. Для цього необхідно виконати комплекс заходів, що охоплюють етапи підготовки даних, побудови та налаштування моделей, а також оцінювання їхньої ефективності.

1. Підготовка та обробка даних.

Першим кроком є формування якісного вхідного набору даних, що включає:

1.1. Очищення даних — усунення дублікатів, некоректних записів, обробка пропущених значень;

1.2. Нормалізація числових ознак — наприклад, масштабування сум транзакцій для забезпечення стабільної роботи алгоритмів;

1.3. Кодування категоріальних змінних — застосування технік, таких як one-hot або label encoding;

1.4. Формування нових ознак (feature engineering) — зокрема, виявлення часових, географічних або поведінкових патернів, які можуть бути інформативними для класифікації.

2. Побудова моделей машинного навчання.

Наступним етапом є вибір та налаштування моделей, здатних розпізнавати аномальні транзакції. Для вирішення задачі класифікації доцільно протестувати кілька підходів, серед яких:

2.1. Логістична регресія;
2.2. Дерева рішень;
2.3. Ансамблеві методи (Random Forest, Gradient Boosting, XGBoost);
2.4. Штучні нейронні мережі.

Оскільки дані можуть бути суттєво дисбалансованими (наприклад, шахрайські транзакції становлять лише 1% від загальної кількості), потрібно застосовувати спеціальні підходи до навчання моделей — коригування ваг класів, oversampling або undersampling.

Крім того, для досягнення оптимальних результатів буде проведено гіперпараметричну оптимізацію, яка дозволяє підвищити точність та стабільність моделей.

3. Оцінювання ефективності моделей.

Якість побудованих моделей перевіряється за допомогою низки метрик:
3.1. Точність (accuracy);
3.2. Повнота (recall);
3.3. Точність прогнозу (precision);
3.4. F1-міра;
3.5. ROC-AUC.

Для глибшого аналізу буде здійснено порівняння продуктивності моделей на наборах даних із різним співвідношенням класів, зокрема 50/50, 99/1 та 83/17. Це дозволить оцінити стійкість системи до різного ступеня дисбалансу та адаптивність до реальних сценаріїв.

4. Вибір і підготовка даних для експерименту.

Оскільки доступ до реальних фінансових транзакцій обмежений, як джерело буде використано відкриті набори даних із платформи Kaggle. При виборі буде враховано такі критерії, як наявність міток про шахрайство, структура полів (категоріальні та числові ознаки), розмір вибірки тощо. Після завантаження дані пройдуть етап попередньої обробки, і тільки після цього використовуватимуться для навчання моделей.

3. МАТЕРІАЛИ ДОСЛІДЖЕННЯ ТА МЕТОДИ

У межах дослідження розроблено метод машинного навчання, призначений для виявлення шахрайських аномалій у фінансових транзакціях. Його мета — зменшення фінансових втрат, яких зазнають фінансові установи, підприємства та фізичні особи через шахрайські дії. Запропонований підхід реалізується поетапно, із застосуванням сучасних інструментів та бібліотек Python.

Етап 1. Підготовка вхідних даних. На першому етапі здійснюється завантаження датасету з використанням бібліотеки pandas. Необхідною передумовою для побудови ефективної моделі є якісна підготовка даних — видалення пропусків, дублювань, а також очищення від нерелевантних атрибутів. Отримані чисті дані структуруються у вигляді DataFrame.

Етап 2. Масштабування та нормалізація даних. Для забезпечення коректної роботи алгоритмів машинного навчання важливо привести числові значення до одного масштабу. Зокрема, нормалізується ознака Amount за допомогою RobustScaler із модуля sklearn.preprocessing, який враховує медіану та інтерквартильний розмах, що робить його стійким до викидів. Масштабування особливо важливе для моделей, чутливих до різниці у діапазонах ознак — таких як логістична регресія, метод опорних векторів, нейронні мережі, а також алгоритми, що використовують градієнтний спуск.

Крім того, із наборів даних вилучаються непридатні або нерелевантні ознаки, такі як ID, географічні атрибути або інші службові поля, що не несуть значущої інформації для задачі класифікації.

Етап 3. Формування навчальної та тестової вибірки. Масив підготовлених даних поділяється на тренувальну та тестову вибірки у співвідношенні 80/20 за допомогою функції train_test_split з параметрами test_size=0.2, random_state=42 та stratify=y. Такий розподіл дозволяє:
3.1. Уникнути перенавчання моделі;
3.2. Здійснити коректне налаштування гіперпараметрів;
3.3. Об'єктивно оцінити ефективність класифікації на раніше «небачених» даних;
3.4. Зберегти збалансоване співвідношення класів.

Етап 4. Побудова та навчання моделей класифікації. Для класифікації транзакцій використовується бінарний підхід (шахрайська/не шахрайська). В залежності від алгоритму (наприклад, логістична регресія, дерева рішень, випадковий ліс, градієнтний бустинг), моделі налаштовуються індивідуально: обираються відповідні параметри, такі як кількість дерев, глибина дерева, швидкість навчання тощо. Підбір параметрів здійснюється з урахуванням якості класифікації та обчислювальних ресурсів.

Етап 5. Оцінка ефективності моделі. На завершальному етапі проводиться аналіз якості побудованих моделей. Використовуються такі метрики:

5.1. Матриця похибок (Confusion Matrix) — дозволяє оцінити кількість правильних та хибних передбачень для кожного класу (TP, FP, FN, TN);

5.2. Звіт про класифікацію (Classification Report) — включає такі метрики, як precision, recall, F1-score;

5.3. AUC-ROC крива — демонструє здатність моделі розрізняти класи незалежно від порогу прийняття рішень;

5.4. Accuracy — частка правильних передбачень. Зазначимо, що через можливий дисбаланс класів ця метрика розглядається додатково.

Запропонована методика є універсальною й може бути адаптована до інших фінансових наборів даних з аналогічними задачами виявлення шахрайства.

У дослідженні було проаналізовано ефективність семи популярних алгоритмів машинного навчання для виявлення аномалій у фінансових транзакціях. Кожна модель має власні переваги та обмеження, що обумовлює їхню придатність до конкретних типів даних:

1. Лінійна регресія (Linear Regression). Незважаючи на свою простоту, ця модель може застосовуватись для бінарної класифікації. Вона дозволяє враховувати вагу класів, що є важливим для незбалансованих наборів. Найкращі результати показав saga, тоді як liblinear був ефективнішим для малих вибірок. Модель обмежена у здатності відображати складні нелінійні залежності, а також потребує попереднього масштабування ознак.

2. Дерева рішень (Decision Tree). Інтерпретована та швидка модель, що не потребує масштабування ознак. У дослідженні продемонструвала схильність до перенавчання, тому було обрано обмеження глибини (до 5 рівнів) та встановлені пороги для мінімального розбиття вузлів і зразків у листі. Модель чутлива до змін у даних і поступається ансамблевим методам за точністю.

3. Випадковий ліс (Random Forest). Ансамблевий підхід на основі дерева рішень, що суттєво зменшує ризик перенавчання. Показав стабільно високі результати на всіх наборах. У моделі використовувалися 50 дерев глибиною до 10 рівнів, із мінімумом 10 зразків на листі. Перевагою є також можливість визначення важливості ознак.

4. Нейронна мережа (MLP Classifier). Багатошарова модель, що добре працює зі складними та великорозмірними даними. Основні переваги — гнучкість і здатність моделювати складні взаємозв’язки. Однак модель потребує ретельного налаштування архітектури, значних обчислювальних ресурсів та великої кількості навчальних даних. Складна для інтерпретації.

5. Gradient Boosting. Ансамблевий метод із послідовним навчанням дерев, що дозволяє коригувати помилки попередніх моделей. Продемонстрував високу точність при помірній глибині (до 3) та кількості дерев (100). Значення learning rate встановлено на рівні 0.1 для балансу між швидкістю та якістю. Перевагою є стійкість до незбалансованих даних завдяки підтримці ваг класів.

6. XGBoost. Оптимізована версія градієнтного підсилення з високою продуктивністю. Підтримує обробку пропущених значень і має вбудовану регуляризацію, що знижує ризик перенавчання. Використовувався з тими ж параметрами, що й Gradient Boosting, і підтвердив свою ефективність на різних наборах даних.

7. SVC (Support Vector Classifier). Один із найточніших, але й найресурсоємніших алгоритмів. Найкраще працює на незбалансованих наборах, де класова нерівність компенсується відповідними вагами. Навчання займало значний час (до години), однак модель демонструвала добру здатність розмежовувати складні класи.

Ці результати підтверджують доцільність використання ансамблевих методів (Random Forest, Gradient Boosting, XGBoost) для задач виявлення шахрайства, а також демонструють важливість налаштування параметрів і урахування дисбалансу класів для підвищення точності моделей.

Розподіл даних на навчальну та тестову вибірки є фундаментальним етапом у процесі машинного навчання, який дозволяє об'єктивно оцінити здатність моделі до узагальнення на нових, раніше не бачених даних. Цей підхід передбачає поділ повного набору даних на дві частини:

1. Навчальна вибірка використовується для побудови моделі. Вона становить зазвичай 70–80% усіх даних, що забезпечує моделі широкий спектр прикладів для виявлення закономірностей у транзакціях різного типу. Чим більший обсяг навчальної вибірки, тим більше патернів — як легальних, так і шахрайських — може вивчити модель, що, у свою чергу, підвищує її ефективність під час передбачення.

2. Тестова вибірка — це решта даних (переважно 20–30%), яка не використовується під час навчання. Вона слугує для перевірки продуктивності моделі на нових прикладах, що дозволяє виявити її здатність до генералізації. Такий підхід відповідає концепції навчання з учителем, коли модель спочатку "вчиться", а потім її "перевіряють" на незалежному наборі.

У процесі поділу даних можуть бути задані додаткові параметри:

- test_size=0.2 — визначає частку даних, відведену для тестування (20%);
- stratify=y — забезпечує пропорційне представлення класів у навчальній і тестовій вибірках, що є критично важливим для незбалансованих наборів;
- random_state=42 — фіксує порядок випадкового розбиття, забезпечуючи відтворюваність результатів та запобігаючи можливому перекосу у вибірках.

Ці заходи спрямовані на мінімізацію перенавчання — ситуації, коли модель демонструє високу точність на тренувальних даних, але втрачає ефективність при роботі з новими.

Оцінка якості моделі здійснюється, зокрема, за допомогою звіту про класифікацію, який відображає основні метрики ефективності. Однією з ключових є точність (Precision) — це частка правильно передбачених позитивних випадків серед усіх передбачених як позитивні. Високе значення точності свідчить про незначну кількість хибнопозитивних результатів. Обчислюється за формулою:

$$Precision = \frac{TP}{TP + TF} \tag{1}$$

де TP — істинно позитивні, TF — хибнопозитивні передбачення.

Recall (чутливість) — це показник, що відображає здатність моделі виявляти всі фактичні позитивні випадки. Він обчислюється як частка правильно класифікованих позитивних результатів серед усіх реальних позитивних прикладів. Високе значення чутливості свідчить про незначну кількість хибнонегативних передбачень. Формально обчислюється за формулою:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

де TP — істинно позитивні, FN — хибнонегативні результати.

F1-міра — це гармонічне середнє між точністю ($precision$) та чутливістю ($recall$), яке дозволяє досягти балансу між цими двома метриками. Вона особливо корисна в умовах незбалансованих даних, коли важливо враховувати як хибнопозитивні, так і хибнонегативні передбачення. Розраховується за допомогою наступної формули:

$$\mathrm{F1} = \left(\frac{2}{recall^{-1} + precision^{-1}}\right) = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

Support — це кількість реальних прикладів кожного класу, що містяться в тестовому наборі. Цей показник допомагає краще інтерпретувати значення метрик (точності, чутливості тощо) в контексті обсягу даних для кожного класу.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve) — показник, що відображає здатність моделі розрізняти між класами при різних порогових значеннях.

ROC-крива показує співвідношення між рівнем виявлення позитивних класів (True Positive Rate) та рівнем хибнопозитивних спрацьовувань (False Positive Rate) при зміні порогу класифікації.

AUC (площа під кривою) — числовий показник, що відображає загальну якість класифікатора:

1.0 — ідеальна класифікація,

0.5 — модель не здатна розрізняти класи (випадкове вгадування),

<0.5 — продуктивність гірша, ніж випадкова.

Точність (Accuracy) — загальна частка правильно класифікованих прикладів серед усіх прикладів. Хоча ця метрика є простою для розуміння, вона може бути оманливою при незбалансованих даних. Наприклад, якщо 95% зразків належать до одного класу, модель, яка завжди передбачає саме його, досягне 95% точності, але не матиме реальної корисності для виявлення менш представленого класу.

Формула обчислення точності:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + TP + FN} \quad (4)$$

де TP — істинно позитивні, TN — істинно негативні, FN — хибнонегативні передбачення.

Оцінювання моделей за допомогою метрик, таких як матриця похибок, AUC-ROC та точність (Accuracy), дає змогу всебічно проаналізувати ефективність алгоритмів, визначити їхні переваги та обмеження. Такий підхід дозволяє обґрунтовано обирати оптимальні методи й інструменти, що формують основу для побудови ефективної системи виявлення шахрайських транзакцій, здатної мінімізувати фінансові ризики як для бізнесу, так і для користувачів.

4. ВИБІР НАБОРІВ ДАНИХ ДОСЛІДЖЕННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ШАХРАЙСЬКИХ АНОМАЛІЙ У ФІНАНСОВИХ ТРАНЗАКЦІЯХ

У даному дослідженні було використано три різні набори даних для аналізу та прогнозування шахрайських аномалій у фінансових транзакціях із застосуванням методів машинного навчання (рис. 3).

Перший набір даних містить інформацію про транзакції з кредитних карток за визначений період часу. Загальна кількість записів становить 248 807, з яких лише 492 є випадками шахрайства. Таким чином, спостерігається суттєвий дисбаланс класів: приблизно 99% — легітимні транзакції, 1% — шахрайські. У структурі цього набору представлено 28 анонімізованих ознак, позначених як V0–V28, зміст яких не розкривається з міркувань конфіденційності. Доступними залишаються лише два параметри: «Час» та «Сума», що обмежує можливості повного аналізу транзакцій.

Другий набір даних, був штучно згенерований, однак має рівномірний розподіл класів: по 560 863 транзакцій у кожному класі — шахрайському та легітимному. Цей набір використовується переважно для тренування моделей, оскільки він дозволяє уникнути впливу дисбалансу класів під час навчання. Структурно він подібний до першого: також містить ознаки V0–V28, проте відсутній параметр «Час», натомість з'являється атрибут «ID», який не має інформативної цінності для навчання моделей. У зв'язку з цим стовпець «ID» видаляється з набору перед обробкою, щоб запобігти виникненню помилок, спричинених розбіжностями у структурі даних.

	# V26	# V27	# V28 abc	# Amount Transaction amount	# Class 1 for fraudulent transactions, 0 otherwise
	7.52	-2.6 3.52	-22.6 31.6	-15.4 33.8	0 25.7k	0 1
3528	-0.189114843888824	0.133558376740387	-0.0210530534538215	149.62	0
8143	0.125894532368176	-0.00898309914322813	0.0147241691924927	2.69	0
35251	-0.139096571514147	-0.0553527940384261	-0.0597518405929204	378.66	0
2038	-0.221928844458407	0.0627228487293033	0.0614576285006353	123.5	0
19756	0.502292224181569	0.219422229513348	0.215153147499206	69.99	0
37034	0.105914779097957	0.253844224739337	0.0810802569229443	3.67	0
659	-0.257236845917139	0.0345074297438413	0.00516776890624916	4.99	0

Рисунок 3.а – Обрані три набори даних для дослідження шахрайських аномалій у фінансових транзакціях (зокрема, а) з інформацією про транзакції з кредитних карток за визначений період часу; б) штучно згенерований, однак має рівномірний розподіл класів; в) наближений до реальних умов)

# V26	# V27	# V28	# Amount The transaction amount	# Class Binary label indicating whether the transaction is fraudulent (1) or not (0)
-8.23 5.62	-10.5 113	-39 77.3	50 24k	0 1
-0.43482398075374323	-0.08123010860166756	-0.1510454864555771	17982.1	0
0.2965027040269081	-0.24805205857704948	-0.0645119229692018	6531.37	0
-0.3128945478459945	-0.30025803544491	-0.24471822939465607	2513.54	0
-0.5159502587046078	-0.16531649030384224	0.048423625556267516	5384.44	0
1.071125586437274	0.023711598569375818	0.419117267018846	14278.97	0
0.2526698696704796	0.06668092979208433	0.09581150731982829	6901.49	0

б)

dob Date of Birth of Credit Card Holder	trans_num Transaction Number	# merch_lat Latitude Location of Merchant	# merch_long Longitude Location of Merchant	# is_fraud Whether Transaction is Fraud (1) or Not (0)
17-07-1989 1% 21-06-1978 1% Other (14057) 97%	14383 unique values	19 67.5	-167 -88.6	0 1
09-11-1939	a3806e984cec6ac0096d8184c64ad3a1	65.654142	-164.722603	1
09-11-1939	a59185fe1b9ccf21323f581d7477573f	65.468863	-165.473127	1
09-11-1939	86ba3a888b42cd3925881fa34177b4e0	65.347667	-165.914542	1
09-11-1939	3a068fe1d856f0ecedbed33e4b5f4496	64.445035	-166.080207	1
09-11-1939	891cdd1191028759dc20dc224347a0ff	65.447094	-165.446843	1
09-11-1939	ef010a5f4f570d306a05	64.088838	-165.104078	1

в)

Рисунок 3 (продовження) – Обрані три набори даних для дослідження шахрайських аномалій у фінансових транзакціях (зокрема, а) з інформацією про транзакції з кредитних карток за визначений період часу; б) штучно згенерований, однак має рівномірний розподіл класів; в) наближений до реальних умов)

Третій набір даних був включений до дослідження з кількох причин. По-перше, попередні два набори, хоч і придатні для технічного аналізу, містять обмежену кількість реальних транзакційних параметрів через високий рівень анонімізації. По-друге, необхідно було перевірити працездатність моделей на даних, які ближчі до реальних умов. Третій набір містить 5 100 записів зі співвідношенням класів приблизно 83% до 17% (легітимні та шахрайські транзакції відповідно), що є більш наближеним до першого набору, проте відрізняється за структурою. Це є природним, зважаючи на складність отримання реальних фінансових транзакцій з відкритих джерел у стандартизованому форматі. Через структурну відмінність, для цього набору здійснюється окреме навчання моделей, що дозволяє забезпечити коректність прогнозування.

У подальшому описі для зручності ідентифікації ці набори даних позначатимуться відповідно до їх співвідношення класів: 99:1, 50:50 та 83:17.

5. ДОСЛІДЖЕННЯ ТА ПРАКТИЧНА ПЕРЕВІРКА МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Проведення досліджень та практична перевірка моделей машинного навчання є ключовим етапом у створенні ефективних систем прогнозування шахрайських аномалій у фінансових транзакціях. Метою цього етапу є не лише підтвердження ефективності обраних методів, але й їхнє практичне випробування в умовах, максимально наближених до реальних. У фінансовому секторі точність та надійність прогнозів мають критичне значення, оскільки своєчасне виявлення шахрайських дій може значно мінімізувати фінансові втрати та підвищити довіру клієнтів до банківських та фінансових установ.

У дослідженні використовувався підхід, що заснований на роботі з різними наборами даних, які характеризуються різним рівнем дисбалансу класів. Це відображає реальну ситуацію у фінансових транзакціях, де переважна більшість операцій є легальними, а шахрайські випадки становлять незначний відсоток. Для всебічної оцінки ефективності моделей було обрано три сценарії співвідношення класів: 50%/50%, 83%/17% та 99%/1%. Такий підхід дозволив оцінити здатність моделей виявляти шахрайські транзакції навіть у тих випадках, коли вони є вкрай рідкісними.

Проведені дослідження допомогли ідентифікувати оптимальні моделі для впровадження у фінансових установах. Ці моделі здатні не тільки знижувати фінансові ризики, але й підвищувати ефективність виявлення шахрайства, мінімізуючи кількість хибних спрацьовувань. Це, в свою чергу, призведе до зменшення витрат на забезпечення безпеки та гарантуватиме стабільну роботу системи в умовах реальних фінансових потоків.

Нижче представлено детальний аналіз продуктивності кожної моделі на різних наборах даних.

Модель Logistic Regression для трьох наборів:

1. Набір 50%/50%. Модель показала високу точність у 96% та збалансовані показники precision та recall, що свідчить про її ефективність на збалансованих даних.

2. Набір 99%/1%. Точність моделі різко знизилася до 10%, демонструючи практично нульову ефективність у виявленні шахрайських транзакцій (клас 1) при сильному дисбалансі даних. Хоча модель ідеально передбачала всі шахрайські транзакції, вона хибно класифікувала значну кількість легальних операцій як шахрайські. Це може свідчити про потенційне перенавчання, витік даних або недостатньо ретельний підбір параметрів.

3. Набір 83%/17%. Точність становила 87%, проте F1-score для класу 1 значно знизився (0.62), вказуючи на часті помилки у виявленні шахрайства.

Нижче приведено таблицю 1 та знімок екрану (рис.4), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму Logistic Regression на наборі даних 83%/17%.

Таблиця 1 – Таблиця з результатами Logistic Regression у наборі 83%/17%

Logistic Regression Набір даних 83%/17%				
	Precision	Recall	F1-score	Support
0	0.97	0.88	0.92	2521
1	0.51	0.81	0.62	369
accuracy	0.87	0.87	0.87	0.87
Macro avg	0.74	0.85	0.77	2890
Weighted avg	0.91	0.87	0.88	2890

На рисунку 4 забражено консольний вивід LogisticRegression у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for LogisticRegression:
              precision    recall  f1-score       support
0              0.970435  0.885363  0.925949  2521.000000
1              0.510169  0.815718  0.627737   369.000000
accuracy       0.876471  0.876471  0.876471     0.876471
macro avg      0.740302  0.850541  0.776843  2890.000000
weighted avg   0.911667  0.876471  0.887873  2890.000000

Confusion Matrix for LogisticRegression:
[[2232  289]
 [  68  301]]

ROC-AUC Score for LogisticRegression: 0.9269
Accuracy for LogisticRegression: 0.8765
```

Рисунок 4 – Консольний вивід LogisticRegression у наборі 87%/13%

Модель Decision Tree:

1. Набір 50%/50%. Модель показала високі показники precision та recall (96% для обох класів) і загальну точність 96%, що свідчить про її добру працездатність на рівномірних даних.

2. Набір 99%/1%. Модель зберегла високу точність для класу 0, але precision для класу 1 значно знизився. Загальна точність становила 58%, що вказує на проблему з дисбалансом.

3. Набір 83%/17%. Загальна точність моделі була високою – 97%. F1-score для класу 1 становив 0.88, що краще, ніж у Logistic Regression, але все ще демонструє зниження recall.

Нижче приведено таблицю 2 та знімок екрану (рис.4), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму Decision Tree на наборі даних 83%/17%.

Таблиця 2 – Таблиця з результатами Decision Tree у наборі 83%/17%

Decision Tree Набір даних 83%/17%.				
	Precision	Recall	F1-score	Support
0	0.98	0.98	0.98	2521
1	0.91	0.86	0.88	369
accuracy	0.97	0.97	0.97	0.97
Macro avg	0.94	0.92	0.93	2890
Weighted avg	0.97	0.97	0.98	2890

На рисунку 5 зображено консольний вивід Decision Tree у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for DecisionTree:
              precision    recall  f1-score       support
0              0.980315  0.987703  0.983995  2521.000000
1              0.911429  0.864499  0.887344   369.000000
accuracy       0.971972  0.971972  0.971972     0.971972
macro avg      0.945872  0.926101  0.935669  2890.000000
weighted avg   0.971519  0.971972  0.971655  2890.000000

Confusion Matrix for DecisionTree:
[[2490   31]
 [  50  319]]

ROC-AUC Score for DecisionTree: 0.9861
Accuracy for DecisionTree: 0.9720
```

Рисунок 5 – Консольний вивід DecisionTree у наборі 87%/13%

Модель Random Forest:

1. Набір 50%/50%: Продемонстрував вражаючі результати: точність 98% та високі значення F1-score для обох класів.

2. Набір 99%/1%: Незважаючи на точність 71%, precision для класу 1 був близьким до 0, що свідчить про відсутність розпізнавання шахрайських транзакцій.

3. Набір 83%/17%: Точність залишалася на рівні 90%, однак F1-score для класу 1 знизився до 0.44, вказуючи на проблеми з виявленням меншого класу.

Нижче приведено таблицю 3 та знімок екрану (рис. 5), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму Random Forest на наборі даних 83%/17%.

Таблиця 3 – Таблиця з результатами Random Forest у наборі 83%/17%.

Random Forest Набір даних 83%/17%.				
	Precision	Recall	F1-score	Support
0	0.90	0.99	0.95	2521
1	0.90	0.28	0.44	369
accuracy	0.90	0.90	0.90	0.90
Macro avg	0.94	0.64	0.69	2890
Weighted avg	0.91	0.90	0.88	2890

На рисунку 6 зображено консольний вивід Random Forest у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for RandomForest:
              precision    recall  f1-score       support
0              0.905789  0.999207  0.950207  2521.000000
1              0.981651  0.289973  0.447699   369.000000
accuracy       0.908651  0.908651  0.908651     0.908651
macro avg      0.943720  0.644590  0.698953  2890.000000
weighted avg   0.915475  0.908651  0.886046  2890.000000

Confusion Matrix for RandomForest:
[[2519    2]
 [ 262  107]]

ROC-AUC Score for RandomForest: 0.9789
Accuracy for RandomForest: 0.9087
```

Рисунок 6 – Консольний вивід Random Forest у наборі 87%/13%

Модель Neural Network (MLP Classifier):

1. Набір 50%/50%: Показав майже ідеальні показники з точністю 98% та F1-score для обох класів на рівні 0.99.

2. Набір 99%/1%. Виникла значна проблема з precision для класу 1, хоча загальна точність 83% залишалася прийнятною.

3. 83%/17%. Точність впала до 80%, а F1-score для класу 1 становив лише 0.53, що свідчить про недостатню адаптацію до дисбалансу.

Нижче наведено таблицю 4 та знімок екрану (рис. 6), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму MLP Classifier на наборі даних 83%/17%.

Таблиця 4 – Таблиця з результатами MPL Classifier у наборі 83%/17%.

MPL Classifier Набір даних 83%/17%.				
	Precision	Recall	F1-score	Support
0	0.97	0.80	0.87	2521
1	0.38	0.84	0.53	369
accuracy	0.80	0.80	0.80	0.90
Macro avg	0.67	0.82	0.70	2890
Weighted avg	0.89	0.90	0.83	2890

На рисунку 7 зображено консольний вивід Neural Network у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for NeuralNetwork:
              precision    recall  f1-score       support
0              0.973064  0.802459  0.879565  2521.000000
1              0.385943  0.848238  0.530508   369.000000
accuracy       0.808304  0.808304  0.808304     0.808304
macro avg      0.679504  0.825349  0.705037  2890.000000
weighted avg   0.898099  0.808304  0.834997  2890.000000

Confusion Matrix for NeuralNetwork:
[[2023  498]
 [  56  313]]

ROC-AUC Score for NeuralNetwork: 0.8380
Accuracy for NeuralNetwork: 0.8083
```

Рисунок 7 – Консольний вивід MPL Classifier у наборі 83%/17%

Модель Gradient Boosting:

1. Набір 50%/50%. Експеримент продемонстрував високу ефективність на рівномірному наборі даних з точністю 97%.

2. Набір 99%/1%. Загальна точність становила 47% через високий дисбаланс, що призвело до значного зниження F1-score для класу 1.

3. Набір 83%/17%. Точність досягла 99%. Модель добре обробляла дисбалансовані дані, хоча F1-score для класу 1 трохи знизився до 0.97.

Нижче наведено таблицю 5 та знімок екрану (рис. 7), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму Gradient Boosting у наборі даних 83%/17%.

Таблиця 5 – Таблиця з результатами GradientBoosting у наборі 83%/17%

Gradient Boosting Набір даних 83%/17%.				
	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	2521
1	0.98	0.95	0.97	369
accuracy	0.99	0.99	0.99	0.90
Macro avg	0.99	0.99	0.98	2890
Weighted avg	0.99	0.9	0.99	2890

На рисунку 8 зображено консольний вивід Gradient Boosting у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for GradientBoosting:
              precision    recall  f1-score       support
0              0.993291  0.998413  0.995846  2521.000000
1              0.988764  0.953930  0.971034   369.000000
accuracy       0.992734  0.992734  0.992734     0.992734
macro avg      0.991028  0.976171  0.983440  2890.000000
weighted avg   0.992713  0.992734  0.992678  2890.000000

Confusion Matrix for GradientBoosting:
[[2517    4]
 [  17  352]]

ROC-AUC Score for GradientBoosting: 0.9992
Accuracy for GradientBoosting: 0.9927
```

Рисунок 8 – Консольний вивід Gradient Boosting у наборі 87%/13%

Модель XGBoost:

1. Набір 50%/50%: Модель показала високий рівень точності 97% з добре збалансованими метриками для обох класів.

2. Набір 99%/1%. Модель продемонструвала найкращий результат серед усіх моделей для набору 99%/1%, досягнувши точності 99% та F1-score для класу 1 на рівні 0.96.

3. Набір 83%/17%. Модель зберегла високу точність 99%, добре адаптуючись до частково дисбалансованих даних.

Нижче наведено таблицю 6 та знімок екрану (рис. 8), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму XGBoost у наборі даних 83%/17%.

Таблиця 6 – Таблиця з результатами XGBoost у наборі 83%/17%

XGBoost Набір даних 83%/17%				
	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	2521
1	0.98	0.95	0.96	369
accuracy	0.99	0.99	0.99	0.99
Macro avg	0.98	0.97	0.98	2890
Weighted avg	0.99	0.99	0.99	2890

На рисунку 9 зображено консольний вивід XGBoost у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for XGBoost:
              precision    recall  f1-score       support
0              0.992897  0.998017  0.995450  2521.000000
1              0.985955  0.951220  0.968276   369.000000
accuracy       0.992042  0.992042  0.992042     0.992042
macro avg      0.989426  0.974618  0.981863  2890.000000
weighted avg   0.992010  0.992042  0.991980  2890.000000

Confusion Matrix for XGBoost:
[[2516    5]
 [  18  351]]

ROC-AUC Score for XGBoost: 0.9996
Accuracy for XGBoost: 0.9920
```

Рисунок 9 – Консольний вивід XGBoost у наборі 87%/13%

Модель SVC:

1. Набір 50%/50%. Модель показала середню продуктивність з точністю 52%, що свідчить про її низьку здатність до виявлення обох класів.

2. Набір 99%/1%. Модель демонструє різке падіння F1-score для класу 1, точність лише 55%.

3. Набір 83%/17%. Точність моделі становила 81%, але F1-score для класу 1 був лише 0.10, що вказує на слабке розпізнавання меншого класу.

Нижче наведено таблицю 7 та знімок екрану (рис. 9), в якому зображено згадані в попередніх розділах метрики точності та результати навчання алгоритму SVC у наборі даних 83%/17%.

Таблиця 7 – Таблиця з результатами SVC у наборі 83%/17%

SVC Набір даних 83%/17%				
	Precision	Recall	F1-score	Support
0	0.87	0.92	0.89	2521
1	0.13	0.87	0.1	369
accuracy	0.81	0.81	0.81	0.99
Macro avg	0.50	0.50	0.49	2890
Weighted avg	0.77	0.81	0.79	2890

На рисунку 10 зображено консольний вивід SVC у наборі 87%/13% з відповідними метриками.

```
Full Classification Report for SVC:
              precision    recall  f1-score      support
0              0.873087  0.927806  0.899615  2521.000000
1              0.137441  0.078591  0.100000   369.000000
accuracy       0.819377  0.819377  0.819377     0.819377
macro avg      0.505264  0.503199  0.499808  2890.000000
weighted avg   0.779158  0.819377  0.797519  2890.000000

Confusion Matrix for SVC:
[[2339  182]
 [ 340   29]]

ROC-AUC Score for SVC: 0.5299
Accuracy for SVC: 0.8194
```

Рисунок 10 – Консольний вивід SVC у наборі 87%/13%

Також було проведено ті самі дослідження на двох інших наборах даних. Результати цих досліджень будуть винесені в таблицю, в якій буде показано показник точності та ROC-AUC score для кожного застосованого алгоритму. Нижче приведено таблицю з результатами алгоритмів у наборі даних 50%/50%.

Таблиця 8 – Результати досліджень у наборі 50%/50%

Алгоритми	ROC-AUC	Accuracy
Logistic Regression	0.9931	0.9606
Decision Tree	0.9833	0.9601
Random Forest	0.9993	0.9839
MLP Classifier	1.0000	0.9998
Gradient Boosting	0.9983	0.9793
XGBoost	0.9985	0.9772
SVC	0.7396	0.5214

У наступній таблиці наведено результати досліджень у наборі 99%/1%.

Таблиця 9 – Результати досліджень у наборі 99%/1%

Алгоритми	ROC-AUC	Accuracy
Logistic Regression	0.9815	0.1066
Decision Tree	0.8326	0.5820
Random Forest	0.9641	0.7169
MLP Classifier	0.9840	0.8322
Gradient Boosting	0.9720	0.4734
XGBoost	0.9748	0.4821
SVC	0.0644	0.5512

Загалом, результати дослідження дозволяють сформулювати такі висновки щодо ефективності застосованих моделей. Алгоритми XGBoost та Gradient Boosting продемонстрували найвищу продуктивність за всіх варіантів дисбалансу класів, що свідчить про їхню високу адаптивність. Натомість Logistic Regression та Support Vector Classifier (SVC) виявилися неефективними у випадках значного дисбалансу, втрачаючи здатність точно класифікувати шахрайські транзакції. Методи, що базуються на деревоподібних структурах, загалом краще справляються з частково незбалансованими даними, однак вимагають ретельного налаштування при роботі з сильно зміщеними вибірками. Загалом, більшість протестованих моделей підтвердили свою репутацію в контексті виявлення шахрайських аномалій.

ВИСНОВКИ

У статті досліджено застосування машинного навчання в контексті трансферу технологій для прогнозування шахрайських аномалій у фінансових транзакціях. Мета дослідження передбачала створення ефективного дослідження, яке б ідентифікувало підозрілі транзакції на ранніх етапах, мінімізуючи фінансові втрати, репутаційні ризики та юридичні витрати. За даними Асоціації сертифікованих фахівців із розслідування шахрайства (ACFE), щорічні втрати від фінансового шахрайства сягнули понад 5% доходу організацій у всьому світі, а середній розмір збитків у 2024 році вперше зріс з 2016 року.

Традиційні методи виявлення шахрайства, що базуються на статичних правилах та ручному аналізі, вже не відповідають сучасним викликам через зростання обсягів даних та ускладнення фінансових операцій. Вони не здатні своєчасно обробляти великі обсяги даних і розпізнавати складні аномальні шаблони поведінки. У зв'язку з цим, впровадження технологій машинного навчання та штучного інтелекту набуває все більшої актуальності, оскільки вони дозволяють створювати автоматизовані, адаптивні системи для прогнозування шахрайських транзакцій у режимі реального часу.

У дослідженні було розроблено поетапний підхід до виявлення шахрайських аномалій за допомогою машинного навчання, який включає підготовку та обробку даних, побудову та навчання моделей класифікації, а також оцінку їх ефективності. Було проаналізовано ефективність семи популярних алгоритмів машинного навчання, таких як лінійна регресія, дерева рішень, випадковий ліс, нейронні мережі, Gradient Boosting, XGBoost та SVC.

Дослідження показало, що ансамблеві методи, такі як Random Forest, Gradient Boosting та XGBoost, є найбільш доцільними для задач виявлення шахрайства. Ці моделі продемонстрували стабільно високі результати на різних наборах даних, зокрема, при роботі з незбалансованими даними, що характерно для реальних фінансових транзакцій. Важливість правильного налаштування параметрів та врахування дисбалансу класів для підвищення точності моделей також була підтверджена.

Застосування запропонованого підходу дозволяє фінансовим установам значно підвищити операційну ефективність, мінімізувати фінансові втрати та зміцнити довіру клієнтів, забезпечуючи комплексний захист фінансової системи у динамічному цифровому середовищі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ / REFERENCES

1. Zhou, Y., Shu, K., Liu, H. (2021). Detecting fraud in online transactions using machine learning: A review. *ACM Computing Surveys*, 54(1), 1–36.
2. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569.
3. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602–613.
4. Jurgovsky, J., Granitzer, G., Ziegler, K., Calabretto, S., Portier, P. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100, 234–245.
5. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, 18(1), 30–55.
6. Chen, Y., Xu, X., Liu, J., Hu, J. (2020). Blockchain-based financial fraud detection: A review. *IEEE Access*, 8, 111697–111707.
7. Phua, C., Lee, V., Smith, K., Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119* [Electronic resource]. Access mode: https://arxiv.org/abs/1009.6119
8. Bauder, R., Khoshgoftaar, T. M. (2021). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Information Science and Systems*, 9(1), 1–15.
9. Association of Certified Fraud Examiners (ACFE). (2022). *Report to the Nations: 2022 Global Study on Occupational Fraud and Abuse*. Austin, Texas: ACFE. 80 c. [Electronic resource]. Access mode: https://www.acfe.com/report-to-the-nations/2022/
10. PricewaterhouseCoopers (PwC). (2023). *Global Economic Crime and Fraud Survey 2023*. 44 c. [Electronic resource]. Access mode: https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html

11. International Monetary Fund (IMF). Fraud in Financial Institutions [Electronic resource]. – Access mode: https://www.imf.org/external/index.htm.
12. Scikit-learn Documentation. Machine Learning in Python [Electronic resource]. – Access mode: https://scikit-learn.org/stable/documentation.html.
13. XGBoost Documentation. Extreme Gradient Boosting [Electronic resource]. – Access mode: https://xgboost.readthedocs.io/.
14. Google AI Blog. Fighting Fraud with Machine Learning [Electronic resource]. – Access mode: https://ai.googleblog.com/.
15. Kaggle. Fraud Detection Datasets [Electronic resource]. – Access mode: https://www.kaggle.com/.
16. Bishop, C. M. Pattern Recognition and Machine Learning [Electronic resource]. – Access mode: https://www.springer.com/gp/book/9780387310732.
17. Міністерство фінансів України. Фінансова безпека [Electronic resource]. – Access mode: https://mof.gov.ua/uk/.
18. Financial Times. Combating Financial Fraud with AI [Electronic resource]. – Access mode: https://www.ft.com/.
19. *International Journal of Financial Studies*. AI in Fraud Detection [Electronic resource]. – Access mode: https://www.mdpi.com/journal/ijfs.
20. Visa Inc. Fraud Prevention with Advanced Analytics [Electronic resource]. – Access mode: https://usa.visa.com/.

Надійшла до редакції 20.09.2025р.

УГРИН ДМИТРО ІЛЛІЧ – доктор технічних наук, професор, доцент кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна, ***e-mail:d.ugryn@chnu.edu.ua***
УШЕНКО ЮРІЙ ОЛЕКСАНДРОВИЧ – доктор фізико-математичних наук, професор, завідувач кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна, ***e-mail:y.ushenko@chnu.edu.ua***
ТОМКА ЮРІЙ ЯРОСЛАВОВИЧ – кандидат фізико-математичних наук, доцент кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна**,** ***e-mail:y.tomka@chnu.edu.ua***
ПАВЛОВ СЕРГІЙ ВОЛОДИМИРОВИЧ – д.т.н., професор кафедри біомедичної інженерії та оптикоелектронних систем, Вінницький національний технічний університет, Вінниця, Україна, ***e-mail: psv@vntu.edu.ua***
ТАЛАХ МАРІЯ ВІТАЛІЇВНА – кандидат біологічних наук, доцент кафедри комп'ютерних наук, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна, ***e-mail:m.talah@chnu.edu.ua***
Д'ЯЧЕНКО ЛІЛІЯ ІВАНІВНА – кандидат технічних наук, доцент кафедри програмного забезпечення комп'ютерних систем, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна, ***e-mail: l.dyachenko@chnu.edu.ua***
ГАЗДЮК КАТЕРИНА ПЕТРІВНА – доктор філософії, доцент, завідувач кафедри програмного забезпечення комп'ютерних систем, Чернівецький національний університет ім. Ю. Федьковича, Чернівці, Україна**.** ***e-mail: k.gazdiuk@chnu.edu.ua***

D.I. UHRYN, Yu.O. USHENKO, YU.YA. TOMKA, S.V.PAVLOV, M.V. TALAH, L.I. DYACHENKO, K.P. GAZDIUK

AGILE RISK MANAGEMENT METHODOLOGY FOR DECISION-MAKING IN STARTUP PROJECTS BASED ON STOCK PRICE FORECASTING

Yuriy Fedkovych Chernivtsi National University, Vinnytsia National Technical University, Ukraine