Київський національний університет імені Тараса Шевченка

Економічний факультет

Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

Виявлення шахрайства у страхуванні транспортних засобів на основі методів машинного навчання

студентки 4 курсу
спеціальності 051 «Економіка»
ОП «Економічна кібернетика»
денної форми навчання
Раковської Анастасії Андріївни

Науковий керівник:

Доктор економічних наук, професор
Чорноус Галина Олександрівна

Засвідчую, що в цій роботі немає запозичень із праць інших авторів без відповідних посилань

Студент_____

Роботу допущено до захисту перед ЕК
рішенням кафедри економічної кібернетики
від 12 червня 2025 р., протокол № 15
Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна _____

КИЇВ – 2025

РЕФЕРАТ

Кваліфікаційна робота бакалавра містить: 62 ст., 2 рис., 3 табл., 51 джерело.

Ключові слова: автострахування, шахрайство, машинне навчання, стекінг, AdaBoost, збалансований випадковий ліс.

Об’єкт дослідження: процес виявлення шахрайства в страхуванні транспортних засобів.

Мета дослідження: вивчення того, як методи машинного навчання можуть бути ефективно використані для виявлення шахрайства у сфері страхування транспортних засобів.

Методи дослідження: алгоритми машинного навчання з учителем.

Наукова новизна, теоретична значимість дослідження: систематичне порівняння декількох моделей прогнозування, включаючи дерево рішень, випадковий ліс, AdaBoost та гібридний стекінг-ансамбль, розроблений спеціально для виявлення шахрайства у страхуванні транспортних засобів.

Практична цінність: структурований рейтинг, який може допомогти страховим компаніям у виборі найбільш ефективної моделі виявлення шахрайства, адаптованої до їхніх операційних вимог.

RESUME

Taras Shevchenko National University of Kyiv, Faculty of Economics, Department of Economic Cybernetics

Key words: vehicle insurance, fraud, machine learning, stacking, AdaBoost, balanced random forest.

The graduation research of student Anastasiia Rakovska deals with the development and evaluation of machine learning models for vehicle insurance fraud detection.

The work is interesting for insurance companies, data analysts, and researchers in predictive analysis who seek robust methods to identify and prevent insurance fraud.

Pages 62, pictures 2, tables 3, bibliog. 51.

ЗМІСТ

ВСТУП

Шахрайство у сфері страхування транспортних засобів є значним фінансовим тягарем у всьому світі, кидаючи виклик страховикам і впливаючи на усіх споживачів через вищі страхові тарифи та строгі правила. Традиційні методи виявлення шахрайства, які часто покладаються на ручні перевірки та прості статистичні методи, у деяких випадках стали недостатніми через зростаючу кількість та складність шахрайських схем. Саме тому розробка надійних, заснованих на даних методів, що базуються на машинному навчанні, набуває все більшого значення у вирішенні цих проблем. Таким чином, це дослідження має як практичне, так і академічне значення, так як воно демонструє, як передові аналітичні підходи можуть підвищити ефективність виявлення шахрайства.

Об'єктом дослідження є процес виявлення шахрайства в страхуванні транспортних засобів.

Предметом дослідження є математичні методи та методи машинного навчання, що застосовуються для виявлення та прогнозування шахрайства у сфері страхування транспортних засобів.

Основною метою цього дослідження є вивчення того, як методи машинного навчання можуть бути ефективно використані для виявлення шахрайства у сфері страхування транспортних засобів. Були поставлені такі конкретні завдання:

- надати огляд шахрайства у сфері страхування транспортних засобів як фінансової та операційної проблеми у страховому секторі;
- розглянути існуючі методи виявлення страхового шахрайства;
- підготувати до подальшої обробки базу даних про реальні страхові випадки;
- побудувати прогнозуючі моделі машинного навчання для виявлення шахрайства;
- оцінити точність і надійність моделей та запропонувати оптимальні підходи для практичного застосування;

У дослідженні використовуються кількісні методи, зокрема алгоритми машинного навчання з учителем, такі як дерева рішень, випадкові ліси, AdaBoost та ансамблі стекування. Процес дослідження включає попередню обробку даних,

побудову моделей та оцінку їхньої продуктивності за допомогою стандартних метрик, таких як точність, чутливість, оцінка F1 та ROC-AUC.

Наукова новизна цієї роботи полягає в систематичному порівнянні декількох моделей прогнозування, включаючи дерево рішень, випадковий ліс, AdaBoost та гібридний стекінг-ансамбль, розроблений спеціально для виявлення шахрайства у страхуванні транспортних засобів. Практична новизна полягає у представленні структурованого рейтингу, який може допомогти страховим компаніям у виборі найбільш ефективної моделі виявлення шахрайства, адаптованої до їхніх операційних вимог.

Дослідження спирається насамперед на набір даних "Виявлення шахрайства у сфері страхування транспортних засобів" від Kaggle. Теоретичні засади та огляд літератури значною мірою ґрунтуються на наукових статтях, доступних через авторитетні наукові бази даних, включаючи Scopus, Google Scholar та ScienceOS. Ці джерела забезпечують надійну основу для розуміння наукової бази та практичних викликів у сфері виявлення страхового шахрайства.

Ця робота складається з вступу, трьох розділів, висновків та списку використаної літератури. Обсяг роботи: 62 ст., 2 рис., 3 табл., 51 джерело.

РОЗДІЛ 1. Огляд автострахування та виявлення шахрайства у сфері

1.1. Світовий ринок страхування автомобільних засобів

Кількість автомобілів у сучасному світі стрімко зростає і для цього є багато причин. Підвищення ВВП та рівня доходів, особливо у країнах, що розвиваються, суттєво збільшує попит на приватні транспортні засоби. Наприклад, у Китаї еластичність продажів нових автомобілів у відношенні до доходу приблизно дорівнює 2,5. Іншими словами, збільшення рівня доходів на 1% призводить до зростання попиту на нові автомобілі на 2,5%. Це обумовлено кількома факторами. По-перше, розвиток міст, включаючи розширення доріг та міської інфраструктури, стає додатковим стимулом для покупки автомобіля. У містах із добре розвиненою мережею доріг володіння транспортним засобом часто є більш привабливим варіантом для сімей. По-друге, зі зміною доходу особистості змінюються й її цінності та потреби. Для багатьох людей машина – це не лише про комфорт та зручність. Часто володіння дорогим автомобілем асоціюється із вищим соціальним статусом. Люди починають переходити на машини з кращим двигуном, дизайном чи брендом. Таким чином вони купують новий автомобіль, навіть якщо немає такої прямої потреби. Окрім того, національні виробники намагаються вкладатися у виробництво доступних за ціною транспортних засобів, що у поєднанні зі збільшенням рівня доходу ще більше стимулює попит на купівлю автомобілів [1].

Окрім вищезгаданих причин збільшення кількості автомобілів у світі, стимул розвитку автомобільного бізнесу теж є вигідним для національної економіки, адже це дає неабиякий вклад у місцеве підприємництво (АЗС, автомийки, ремонт автомобілів, страховка тощо). Наприклад, у Китаї уряд намагається просувати національне автомобільну індустрію, використовуючи політику, яка стимулює як місцеве виробництво, так і покупки машин національного виробника. До таких мотивуючих заходів може входити зменшення податків або надання субсидій виробникам та споживачам [2].

Збільшення кількості автомобілів тягне за собою один дуже важливий наслідок: підвищений ризик автотранспортних пригод. Велике скупчення машин,

особливо в густонаселених чи популярних місцях, часто призводить до заторів, що підвищує ймовірність як незначних зіткнень, так і серйозних, які можуть перерости у небезпечні для життя інциденти. Ремонт авто зазвичай забирає багато часу та коштів, що може бути непосильним для водіїв та призвести до значних фінансових втрат. Ускладнює ситуацію й те, що новітні автомобілі потребують більших витрат на утримання та ремонт. Це все через те, що вони часто оснащені провідними технологіями такими як сенсори, камери й автоматизовані системи. Замінити бампер із вбудованими сенсорами обійдеться набагато дорожче порівняно зі звичайним [3]. У таких випадках страхування транспортних засобів постає подушкою безпеки, гарантуючи, що неочікувані витрати на ремонт чи відшкодування не будуть непосильними.

Зростання продажів автомобілів, рівня доходів, обов'язковість страхування цивільної відповідальності перед третіми особами та швидка урбанізація значно впливають на розвиток ринку страхування транспортних засобів. У 2019 році світовий ринок автострахування оцінювався у понад $880 млрд [12].

Автострахування – це договір між власником легкового автомобіля, двоколісних транспортних засобів, комерційних вантажівок чи будь-якого іншого транспортного засобу (страхувальником) та страховою компанією (страховиком), що оберігає страхувальника від фінансових втрат у випадку пошкоджень майна й травм під час аварій, нещасних випадків в результаті стихійного лиха або викрадення транспортного засобу. Це часто включає в себе захист від юридичної відповідальності, що виникає внаслідок дорожньо-транспортної пригоди, що призвела до смерті, травм чи пошкодження майна третьої особи [4].

Поняття страхування транспортного засобу виникло напочатку 20 століття, коли автомобілі почали ставати більш поширеним явищем, особливо після того, як у 1908 році з’явилася модель «T Ford». До того часу більшість транспортних засобів залучали коней, що не потребувало страхування такою ж мірою. Але як тільки більше автівок почали з’являтися на дорогах – аварії стали траплятися частіше. Таким чином уже напочатку 1920-х років страхова індустрія визнала необхідність стандартизації полісів автострахування. Зростання автомобільного

руху разом із зростанням частоти аварій спонукало страховиків створювати більш комплексні поліси, щоб задовільнити попит населення.

Після Другої світової війни уряди почали впроваджувати закони, що вимагали від водіїв оформлення автострахування, а особливо покриття відшкодування перед третіми особами. Така ініціатива гарантувала, що жертви аварій отримували відповідні компенсації, навіть якщо винуватець інциденту не володів належною грошовою сумою, необхідною для відшкодування збитків [4].

Таким чином існують різні види автострахування, які класифікуються в залежності від обсягу покриття, наприклад, повне, часткове чи відсутнє покриття певних ризиків, обов'язковості та конкретними ризиками, на які вони поширюються.

Страхування цивільної відповідальності є однією з базових та найпоширеніших форм страхування транспортних засобів і до теперішнього часу. Воно зазвичай покриває відшкодування за тілесні ушкодження, що включає медичні витрати та втрачені заробітні плати інших осіб, що постраждали в результаті аварії, та ремонт пошкодженого майна [5]. У більшості країн світу законне володіння автомобілем вимагає хоча б мінімальної суми страхування відповідальності перед третіми особами.

Іншою важливою формою страхування транспортних засобів є покриття збитків внаслідок зіткнень, що поширюється на автомобіль страхувальника незалежно від того, хто винуватець події. Такий тип автострахування часто вимагають кредитори, якщо автомобіль взято у кредит чи оренду [5].

Захист від нещасних випадків також є обов'язковим у деяких регіонах. Це страхування може покривати медичні витрати, втрачені заробітні плати або інші витрати незалежно від того, хто є винуватцем аварії. Воно може поширюватися на всіх учасників інциденту або ж лише на водія та пасажирів.

Важливою тенденцією на світовому ринку страхування транспортних засобів є боротьба з незастрахованими водіями. Саме тому також існує покриття незастрахованих чи недостатньо застрахованих водіїв, що захищає страхувальника від ситуацій, коли винуватець аварії не має відповідної страховки

[6]. У багатьох країнах влада запроваджує більш жорсткі правила та покарання за незастраховане водіння. Наприклад, у Німеччині транспортні засоби без оформленого страхування знімаються з реєстрації, а водіям таких транспортних засобів загрожують суворі покарання, включаючи великі штрафи та можливе ув'язнення. Очікується, що ця тенденція зменшить кількість незастрахованих транспортних засобів та покращить фінансову стійкість системи страхування [12].

Крім того, існує комплексне страхування транспортних засобів, що також включає в себе вандалізм, крадіжку, стихійні лиха чи наїзд на тварину.

Якщо ж орендований автомобіль або автомобіль у кредит було розбито чи викрадено, існують спеціальні види страхування транспортних засобів, що покривають різницю між фактичною вартістю машини та залишком заборгованості за кредитом. Покриття відшкодування витрат на оренду оплачує витрати протягом того часу, що автомобіль перебуває на ремонті через стразовий випадок [7].

Ці варіанти покриття можна комбінувати різними способами, щоб пристосувати поліс автострахування до потреб та обставин людини.

Страхування відіграє важливу роль у національній та світовій економіці, суттєво сприяючи економічній стабільності та росту. Механізм страхування є унікальним за своєю здатністю не лише захищати майнові інтереси фізичних та юридичних осіб, а й допомагати у вирішенні ширших макроекономічних завдань, що постають перед державою. Забезпечуючи компенсації різних видів збитків – від непередбачуваних природних явищ, виробничих аварій чи інших ризиків – страхування сприяє зменшенню невизначеності та створенню надійного середовища, сприятливого для економічного зростання [9].

Для кращого розуміння масштабів, особливостей та рівня розвитку ринку автострахування доцільно порівняти ключові показники на різних ринках. У таблиці 1.1 наведено дані щодо кількості зареєстрованих транспортних засобів, середніх щорічних витрат на страхування одного власника автомобіля,

загального обсягу ринку автострахування, а також зазначено, чи є наявність страхового полісу обов'язковою вимогою згідно з законодавством.

Таблиця 1.1

Глобальне порівняння ринку страхування транспортних засобів

Країна	**Кількість зареєстрованих транспортних засобів (млн)**	**Середньорічні витрати на автострахування на одну особу**	**Розмір ринку автострахування (млрд USD)**	**Чи є автострахування обов'язковим?**
США	283.40 (2022)	$2,013 (2023)	314.8 (2024)	Так
Німеччина	49.10 (2023)	€304	56.26 (2024)	Так
Японія	78.76 (2023)	¥31,100 (2023)	54.62 (2024)	Так
Україна	13.16 (2024)	UAH 4750	1.13 (2023)	Так

Джерело: розроблено автором на основі [37-46].

Найважливішим економічним наслідком страхування транспортних засобів є забезпечення фінансової стабільності у випадку дорожньо-транспортних пригод як для окремих осіб, так і для економіки в цілому. За допомогою впровадження обов'язкового мінімального страхового покриття уряди гарантують, що власники автомобілів матимуть змогу понести фінансову відповідальність за збитки чи травми, які вони завдали іншим, зменшуючи економічний тягар аварій для постраждалих від аварій та держави [8].

Таким чином однією із ключових економічних переваг страхування є механізм страхового пулу, що дозволяє розподіляти фінансові ризики, пов'язані з дорожньо-транспортними пригодами, між великою групою застрахованих осіб [8]. Ця система мінімізує економічний шок для будь-якої окремої особи, так як страхові компанії покривають витрати, спричинені внаслідок нещасних випадків, за рахунок внесків, сплачених усіма страхувальниками.

Обов'язкове страхування транспортних засобів також сприяє підвищенню ефективності ринку, зменшуючи витрати, пов'язані з дорожньо-транспортними пригодами [8]. Коли фізичні особи зобов'язані мати страховку, загальний економічний вплив аварій пом'якшується. Це пов'язано з тим, що компенсація, яку надають страховики, забирає фінансовий тягар, з яким стикаються потерпілі,

зменшує потребу в державному фінансуванні медичних послуг та судових позовів, пов’язаних з дорожньо-транспортними пригодами, та забезпечує більш передбачуване економічне середовище для водіїв. Такий процес сприяє загальному економічному благополуччю, пропонуючи фінансову подушку, яка допомагає як у мікро-, так і в макроекономічному контексті.

У країнах із добре розвиненими страховими системами цей сектор постає макроекономічним стабілізатором, гарантуючи, що економіка може справлятися з кризами більш ефективно. Наприклад, у разі стихійних лих чи економічних спадів страхова індустрія допомагає зменшити навантаження на державні ресурси, надаючи альтернативний спосіб компенсації і таким чином дозволяючи спрямовувати кошти на інші національні пріоритети [9].

Більше того, індустрія страхування транспортних засобів відіграє важливу роль в економіці, надаючи робочі місця та вкладаючи свій внесок у національний економічний розвиток. Страховий сектор створює вакансії у сферах андеррайтингу, врегулювання збитків, продажу та обслуговування клієнтів. У країнах з обов’язковим страхуванням транспортних засобів цей сектор займає значну частину національної економіки. Крім того, наявність стабільної та широко розповсюдженої системи страхування підвищує довіру громадян, що, в свою чергу, стимулює автомобільний ринок, сприяючи подальшому економічному зростанню [8].

Пропонуючи значні соціальні виплати, такі як компенсація за виробничі травми та відновлення майна після катастроф, страхування сприяє відновленню засобів до існування людей та економіки в цілому, таким чином відіграючи роль у соціальній політиці. Ця підтримка має вирішальне значення для збереження економічної стабільності у складні часи.

Більше того, страховики захищають економічну взаємозалежність між підприємствами, страхуючи ланцюги поставок, які стають все більш вразливими через ускладнення технологічних компонентів. Такий захист гарантує, що збої в частині ланцюга поставок не призведуть до масштабних економічних наслідків.

Розвиток страхового ринку також призводить до посилення уваги до інвестиційної функції страхових компаній. У світі зростає тенденція позиціонування страхових компаній як інституційних інвесторів, так як вони залучають капітал та спрямовують його в довгострокові інвестиції, сприяючи таким чином загальному розвитку економіки. Ця тенденція висвітлює, як страховий сектор стає важливим джерелом фінансування для інших секторів, стимулюючи інвестиції в національний розвиток [9].

Більше того, страхування робить внесок й у загальну фінансову інфраструктуру, пов’язуючи страховий ринок з банківською системою, ринками капіталу та державним бюджетом. Страховики сприяють економічній стабільності, діючи як «захисники капіталу». Вони менш вразливі до короткострокових проблем з ліквідністю порівняно з іншими фінансовими установами, а перестрахування додатково стабілізує їхню схильність до збитків, розподіляючи або диверсифікуючи передані ризики [10]. Фінансові ресурси, утворені в результаті страхових внесків, часто інвестуються в цінні папери, нерухомість та інші ринки капіталу, зміцнюючи фінансову систему та сприяючи руху капіталу. Цей взаємозв’язок підкреслює невід’ємну роль страхового ринку в ширшій фінансовій екосистемі [9].

Страхування уможливлює реалізацію будівельних проєктів та заходів, що сприяють економічному зростанню. Воно полегшує отримання кредитів, надаючи кредиторам можливість пропонувати фінансування для великих бізнесів за нижчими відсотковими ставками [10].

Ще однією причиною росту ринку страхування транспортних засобів є підвищене використання та вдосконалення стратегій цифрового маркетингу, що дозволяє розширити доступ та спростити оформлення страхових полісів для споживачів. У свою чергу, така доступність сприяє росту ринку автострахування, охоплюючи ширшу аудиторію та залучаючи молоде, більш технічно підковане покоління, яке починає використання транспортних засобів в особистих та професійних цілях. Ця нова демографічна група глибше розуміє необхідність страхування автомобілів. Звертаючись до цього покоління через соціальні мережі

та цифровий контент, страховики сприяють усвідомленню важливості страхування автотранспорту, тим самим підвищуючи зацікавленість та ухвалення страхових полісів. Таким чином, впровадження онлайн платформ та цифрових інструментів не лише дозволяє збільшити клієнтську базу страховиків, а й відкриває новий шлях для зростання галузі. Підвищення автострахових внесків – на 12,4% у 2023 році – відображає ріст тенденцій в споживчому інтересі та розширення ринку. Використання цифрових маркетингових стратегій відіграло ключову роль у стимулюванні цього зростання, надаючи споживачам спрощений спосіб отримувати інформацію про автострахові продукти та купувати їх [11].

Новітні цифрові інструменти, такі як електронні договори страхування, мобільні додатки та дистанційне звітування про страхові випадки, трансформують традиційні методи взаємодії між страховиками та клієнтами. Діджиталізація бізнес-процесів дозволяє страховикам пропонувати послуги більш ефективно, з можливістю використовувати онлайн платформи, які спрощують порівняння та оформлення страхових полісів [13]. Це також спрощує процес врегулювання збитків, роблячи його швидшим та ефективнішим [12].

Більше того, інтеграція страхових технологій (InsurTech) розширює межі страхової індустрії [11]. Такі платформи використовують блокчейн, штучний інтелект та технології біг-дата, щоб пропонувати більш адаптивні та гнучкі страхові продукти, наприклад, поліси на основі використання й телематики, які є особливо привабливими для молодого покоління. Телематичні технології дозволяють страховикам базувати страхові внески на фактичних показниках водія, таких як пробіг, швидкість та характер гальмування [12]. Такий персоналізований підхід до ціноутворення не лише робить страхування транспортних засобів більш доступним, але й допомагає страховикам краще оцінювати та зменшувати ризики, що зрештою підвищує економічну ефективність галузі [11]. Цей підхід також винагороджує безпечне водіння нижчими страховими внесками.

Оскільки стартапи InsurTech використовують новітні технології також для автоматизації управління страховими виплатами, це призводить до зменшення

залежності від традиційних брокерів та посередників, зменшуючи як витрати страховиків, так і страхові внески для страхувальників. Глобальні інвестиції в InsurTech були значними: у 2022 році інвестиції піднялися на 210% порівняно з попереднім роком, що відображає зростаючу довіру галузі до цифрових рішень [13].

Іншою важливою особливістю ринку страхування транспортних засобів є поширення автоматизованих автомобілів. Згідно з прогнозами, такі транспортні засоби призведуть до революції на ринку автомобілів, що, у свою чергу, матиме значний вплив і на страхову галузь. [14] За своєю конструкцією безпілотні автомобілі мінімізують людську помилку, тому очікується, що рівень аварійності, спричинений цим фактором, значно знизиться. Згідно з актуарними оцінками, рівень аварійності в США може знизитися до 80% до 2040 року, що призведе до зменшення кількості нещасних випадків, травм і матеріальних збитків [15]. Оскільки ринок рухається в сторону автоматизації, страховики повинні адаптувати свої продукти, моделі ризиків та операційні процеси для того, щоб залишатися конкурентоспроможними та здатними справлятися із новими викликами.

Одним з найважливіших викликів для страховиків є зміна приписування відповідальності від людини-водія до безпілотних систем водіння. Оскільки транспортні засоби стають все більше автоматизованими, страховикам доведеться внести правки у розподілення відповідальності у разі аварій. Традиційних моделей, що фокусуються на ризиках, пов’язаних з людськими вчинками, та базуються на таких факторах, як історія водіння, вік та тип машини, більше не буде достатньо. Безпілотні технології уможливлюють випадки, коли автономний транспортний засіб виходить з ладу і спричиняє аварію, не перебуваючи при цьому під контролем людини. Саме тому деякі експерти вважають, що відповідальність за дорожньо-транспортні пригоди за участю безпілотних автомобілів перейде до виробників, розробників програмного забезпечення та інших зацікавлених сторін, які беруть участь у роботі та виготовленні автономних систем транспортних засобів [14]. Таким чином

необхідно розробити нову нормативну базу, що охоплюватиме системні несправності та чітко визначатиме, хто несе відповідальність у разі аварій за участю високоавтоматизованих або повністю автономних транспортних засобів.

Така тенденція переходу відповідальності від водіїв до виробників, ймовірно, стимулюватиме попит на страхування відповідальності за якість продукції [15]. Ці страхові поліси захищають виробників чи продавців від фінансових втрат через дефекти їхньої продукції, які спричиняють шкоду чи збитки.

Деякі автовиробники розпочали пропонувати пакетні страхові продукти в рамках купівлі автомобіля. Наприклад, такі компанії, як Tesla, вже прийняли практику пропозиції їхнього власного страхування (Tesla Insurance) при покупці транспортного засобу, особливо для електричних та автономних моделей [15].

Впровадження моделей страхування без вини виділяється як ще одне потенційне рішення для врегулювання збитків, пов’язаних з дорожньо-транспортними пригодами. За такої системи страховики могли б негайно виплачувати компенсацію потерпілим і стягувати витрати з відповідальних сторін шляхом суброгації, усуваючи необхідність у тривалих судових розглядах [15].

Безпілотні автомобілі також створюють нові ризики, що не можуть бути покритими традиційними страховими продуктами. Такі ризики включають, наприклад, вразливість кібербезпеки, так як автоматизовані транспортні засоби можуть бути вразливими до кібератак, злому або витоку даних, а також збої програмного забезпечення, коли системи, що керують автомобілем, виходять з ладу [14]. Такі ситуації можуть призвести до нещасних випадків або крадіжок. Страховим компаніям необхідно буде розробити нові продукти, які спеціально враховуватимуть ці ризики. Сюди будуть входити страхування кібер-ризиків для автономних систем та покриття відповідальності виробників і постачальників технологій.

Доступ до бортових даних – ще одна сфера, що викликає занепокоєння. Дані, що генеруються автоматизованими автомобілями, включаючи поведінку водія,

продуктивність системи та стан транспортного засобу, будуть критично важливими для страховиків задля кращої оцінки ризику та прийняття цінових рішень, причому такі дані мають збиратися та аналізуватися в режимі реального часу. Однак вони часто контролюються виробниками оригінального обладнання, що призводить до юридичних і комерційних конфліктів щодо того, хто має доступ до цих даних [14]. Вирішення таких конфліктів є критично важливим для страховиків, адже воно дозволяє точніше оцінювати ризики та встановлювати страхові внески, що відображатимуть характеристики транспортного засобу в режимі реального часу.

Окрім того, зараз набирають обертів попутні перевезення та так звані «каршерінги», що потенційно може призвести до зменшення кількості приватних транспортних засобів і, відповідно, зниження попиту на традиційні страхові поліси індивідуальних автомобілів. Таким чином, страховикам, можливо, прийдеться переорієнтовуватися на моделі, пристосовані до комерційних автопарків. У такому випадку їм необхідно буде адаптуватися до переходу до більш динамічних моделей ціноутворення, де ціна страхування відображає фактичне використання транспортних засобів, особливо в контексті послуг спільного пересування [14].

1.2. Проблема шахрайства у страхуванні транспортних засобів

Збільшення кількості автомобілів у світі також корелюється з ростом числа випадків шахрайства у страхуванні транспортних засобів. Цьому сприяє й те, що викривлення деталей чи наслідків аварії та інсценування інциденту є відносно легкими для реалізації без негайного виявлення. Більше того, враховуючи те, що обсяг світового ринку страхування транспортних засобів оцінюється в 973,33 млрд доларів США в 2025 році і, за прогнозами, досягне близько 1 796,61 млрд доларів США до 2034 року[19](Рис. 1.1), ринок автострахування є привабливою мішенню для шахраїв. Великі фінансові ставки роблять цей сектор особливо вразливим до шахрайських дій, оскільки окремі особи намагаються скористатися значними сумами грошей, що знаходять в обігу. Все це є причинами того, що

страхове шахрайство в автотранспортному секторі є широко розповсюдженою проблемою, яка завдає суттєвої економічної шкоди страховикам, страхувальникам та суспільству в цілому. За оцінками експертів, шахрайські заяви становлять значну частину виплат страховиків. Дослідження показують, що приблизно 10-20% заяв на відшкодування є шахрайськими. [18] За оцінками, шахрайство в усіх страхових сферах краде щонайменше 308,6 мільярдів доларів на рік в американських споживачів [20]. Проблема є багатогранною і включає в себе як фальсифікацію страхових випадків, так і навмисне викривлення інформації про аварії та збитки. Шахрайство в галузі страхування транспортних засобів часто пов'язане з інсценуванням аварій, сфабрикованими пошкодженнями та завищенням розміру збитків. Види таких злочинів обмежені хіба фантазією шахраїв, проте зазвичай їх поділяють на заяви з перебільшенням (м’яке шахрайство) та навмисне (жорстке) шахрайство [17], й обидва типи мають значний вплив на страхову індустрію.

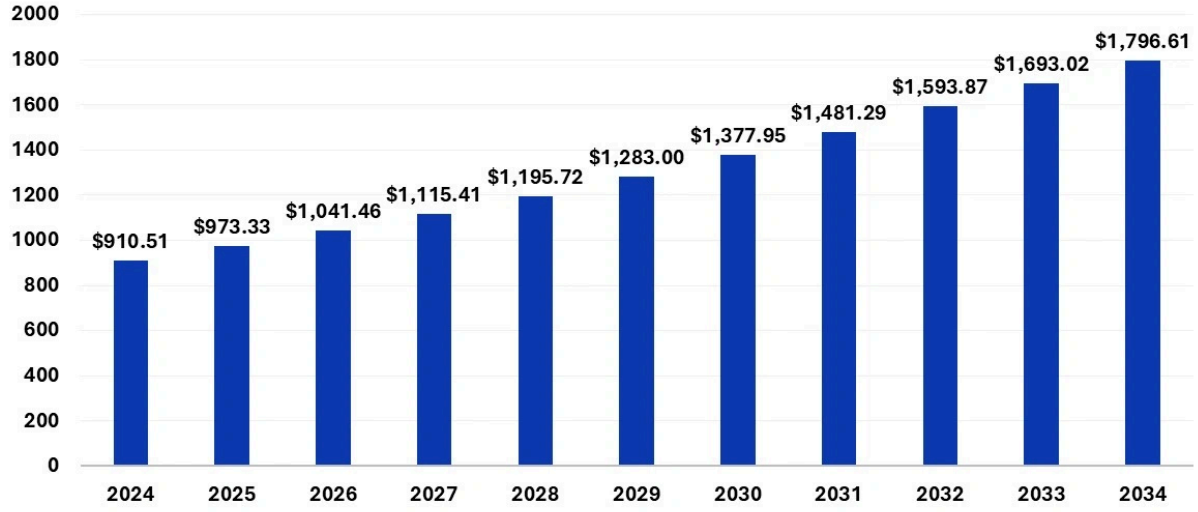

Рис. 1.1. Розмір ринку страхування транспортних засобів та прогноз на 2025-2034 роки (мільярди доларів США).

Джерело: [19].

Однією з найпоширеніших форм шахрайства є вигадування або фальсифікація обставин автомобільної аварії. У таких випадках шахраї маніпулюють або фабрикують деталі автопригоди. Шахраї можуть створювати фальшиві поліцейські звіти, підробляти документи, наприклад, фотографії чи

оцінки пошкоджень транспортного засобу, або інсценувати аварії, використовуючи раніше пошкоджені автомобілі. У деяких випадках можуть залучатися інші особи, у тому числі корумповані працівники поліції, для фальсифікації офіційних документів або надання неправдивих свідчень, щоб підкріпити свою позицію [16]. Такий вид шахрайства може бути важко виявити, оскільки він часто передбачає добре організоване планування та подання фальшивих документів, які на перший погляд можуть здаватися легальними.

Інший поширений вид шахрайства – коли учасники дорожньо-транспортної пригоди заздалегідь домовляються про інсценування інциденту задля отримання страхового відшкодування. Це може включати контрольовані зіткнення в певних місцях або в певний час, часто за відсутності свідків, які могли б підтвердити факт події. У таких випадках аварія планується, а її учасники спільно працюють над тим, щоб представити сфабриковані докази пошкоджень або травм [16]. Сюди входять надання транспортним засобам пошкодженого вигляду та створення фальшивих медичних висновків про травми. Проблематичність шахрайства за попередньою змовою полягає в тому, що воно здатне імітувати законну аварію, що значно ускладнює його виявлення страховиками без ретельного розслідування чи експертного аналізу.

У випадках, коли аварія дійсно сталася, деякі особи можуть намагатися подати шахрайську заяву, перебільшуючи розмір збитків або втрат, щоб отримати більше грошей, ніж мають право на відшкодування. Це може включати в себе неправдиві заяви про додаткове пошкодження автомобіля, внесені до заяви неіснуючі предмети, наприклад, електроніка, або завищення вартості необхідного ремонту [16]. Деякі шахраї можуть навіть включати до заяви вже існуючі збитки, які не є частиною поточної аварії. Такі перебільшення погіршують фінансовий тягар для страховиків, оскільки в кінцевому підсумку вони можуть несвідомо оплатити збитки або медичні рахунки, які виходять за межі того, що насправді було пошкоджено. Завищення розміру відшкодування є однією з найпоширеніших форм м’якого шахрайства.

Деякі шахраї можуть заявити про викрадення власного автомобіля, хоча насправді вони його продали, покинули або сховали з метою отримання особистої вигоди. Цей тип шахрайства стається у випадках, коли власник хоче позбутися автомобіля, але при цьому бажає отримати відшкодування його вартості через страховий поліс. Також поширеним явищем є інсценування вандалізму, коли транспортному засобу навмисно завдають шкоди чи підпалюють, щоб отримати компенсацію від страховика [16]. Таке шахрайство особливо складно виявити без ретельної перевірки, оскільки навмисні пошкодження дуже легко переплутати з нещасним випадком чи зловмисницькою діяльністю третьої особи.

У деяких випадках шахрайства особи фальсифікують травми, які не були отримані в результаті аварії, часто у співпраці з медичними працівниками або експертами з дорожньо-транспортних пригод, що можуть допомогти зі складанням фальшивих звітів про травми та діагнози, гарантуючи, що їхні заяви будуть підкріплені сфабрикованими доказами [16]. Деякі шахраї можуть також сфальсифікувати історію хвороби або довгострокові наслідки нібито отриманих травм. Це може стосуватися як фізичних ушкоджень, так і емоційних збитків, наприклад, психологічної травми. Шахрайські заяви про тілесні ушкодження завдають автостраховим компаніям збитків у розмірі від 6,8 до 9,3 мільярдів доларів на рік. Ці претензії, як правило, стосуються болю в шиї або спині, що зазвичай не передбачає поїздки до лікарні і є важко перевірити [18].

Часом заявники надають неправдиву або оманливу інформацію ще з моменту реєстрації, щоб отримати несправедливу перевагу. Страхові тарифи часто базуються на профілях ризику. Наприклад, ризик інциденту на дорозі для молодого та недосвідченого водія або водія з історією аварій є потенційно вищим, що призводить до вищих страхових внесків [17]. Таким чином, внесення фальшивих даних про свій вік, водійський досвід чи історію водіння дозволяє шахраям платити менше за страховку, експлуатуючи систему за рахунок чесних страхувальників. Іноді шахраї стверджують, що їхній транспортний засіб має особливі засоби безпеки, що знижує ризик і, відповідно, ціну страхового внеску.

Аналогічно, спотворення інформації про модель та вартість транспортного засобу є ще одним способом маніпулювати страховими внесками чи виплатами. У деяких випадках шахрайство трапляється, коли страхувальники роблять неправдиві заяви про модифікацію транспортного засобу, намагаючись збільшити компенсацію внаслідок аварії. Страхувальники щорічно здійснюють шахрайство на 35,1 мільярда доларів, надаючи неправдиву інформацію у своїх заявах, щоб отримати кращий страховий тариф [18].

Наслідки подання неправдивої чи сфальсифікованої інформації є далекосяжними як для страховиків, так і для страхувальників. Після виявлення та розкриття шахрайства страховики повинні повторно оцінити заяви, що часто призводить до затримки або відмови у виплаті страхового відшкодування [17].

Крім того, поява автономних транспортних засобів сприяє новим видам шахрайства у галузі автострахування. Очікується, що автоматизовані автомобілі зменшать кількість аварій та страхових випадків, але вони також створюють унікальні можливості для шахрайських дій [20].

Залежність автомобілів від програмного забезпечення, датчиків, GPS та підключення до Інтернету робить їх вразливими до зломів та кібератак. Зловмисники можуть отримати несанкціонований доступ до систем автомобіля чи віддалено отримати контроль над транспортним засобом, щоб спричинити аварію або несправність чи маніпулювати даними, що призведе до шахрайських вимог про відшкодування збитків чи травм.

Перехід автомобілів між автономним і ручним режимами також відкриває нові можливості для шахраїв [20]. Маніпулюючи часом цих переходів, вони можуть створювати сценарії, в яких аварія виглядає як наслідок збою системи, навіть якщо справжньою причиною є людська помилка.

Оскільки безпілотні автомобілі оснащені датчиками та камерами, які генерують дані, існує ризик того, що шахраї можуть інсценувати аварії за участю безпілотних автомобілів. Крім того, дані, згенеровані безпілотниками, такі як показання акселерометра і телематичні дані, можуть бути змінені для створення неправдивої розповіді, що зніматиме провину з водія [20]. Такі маніпуляції

ускладнюють визначення відповідальності та збільшують ймовірність подання неправдивих позовів. Це постає серйозною проблемою для страховиків, оскільки дані можуть виглядати як такі, що підтверджують шахрайські заяви.

Новітні технології автоматизованих транспортних засобів з вдосконаленим штучним інтелектом і численними компонентами підвищує ризик шахрайства в ланцюжку поставок [20]. Такі компоненти, як датчики і камери, можуть бути фальсифіковані або підроблені, що призводить до дефектів транспортних засобів і шахрайських страхових виплат.

Хоча шахрайські дії часто пов'язані зі злочинними намірами, мотиви такої поведінки є багатогранними і залежать від різноманітних психологічних, економічних та соціальних факторів (рис. 1.2). Вивчення цих факторів дає змогу зрозуміти, що шахрайство – це не просто питання опортунізму, а наслідок ширших системних проблем у страховій індустрії та суспільстві в цілому. Розуміння того, чому споживачі вдаються до шахрайських дій, є важливим для розробки ефективних заходів для запобігання та пом'якшення наслідків шахрайства.

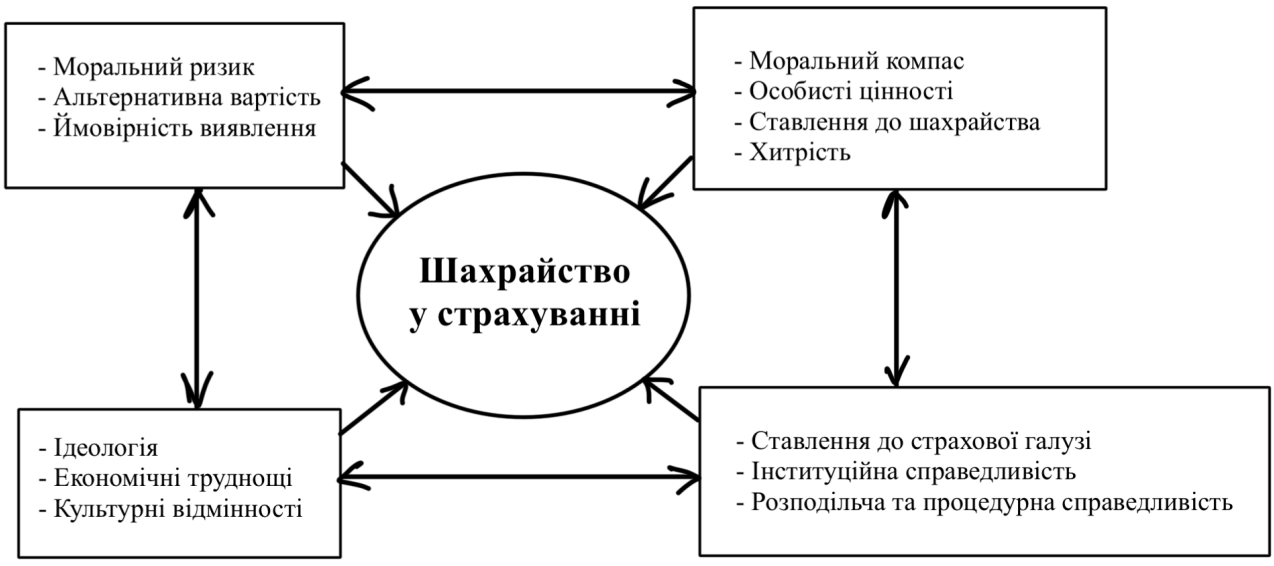

Рис. 1.2. Причини страхового шахрайства.

Джерело: створено автором на основі [22].

Частина людей, які стикаються з фінансовими труднощами, розглядають шахрайство як «злочин без жертв». Деякі шахраї можуть намагатися

дегуманізувати страховиків, оскільки це «лише велика компанія», або переконувати себе в тому, що страхова компанія може дозволити собі ці збитки. До того ж вони часто недооцінюють ймовірність викриття. Це класичний приклад економічної раціоналізації, коли люди порівнюють потенційну фінансову вигоду з мінімальними наслідками та низьким ризиком бути спійманими. [22]

До того ж у деяких громадах або соціальних групах шахрайська поведінка у страхуванні вважається «нормальною» та прийнятною [22]. Спостерігаючи за шахрайською поведінкою інших людей або чуючи про поширені випадки шахрайства в ЗМІ, люди можуть втратити чутливість до цієї проблеми. Більше того, часом споживачі вважають, що страховики стягують надмірно високі тарифи або несправедливо поводяться зі страховими випадками, і тому можуть виправдовувати шахрайські дії як спосіб відновлення справедливості. Таке відчуття особливо поширене серед споживачів, які вважають, що страхові компанії діють насамперед заради прибутку, а не добробуту клієнтів.

Суспільні норми теж відіграють певну роль. Деякі експерти говорять про «синдром ринкової аномії», що означає руйнування довіри до інституцій та зростання поведінки, яка керується власними інтересами та пошуком прибутку [22]. У таких умовах люди можуть відчувати, що шахрайство виправдане або навіть заохочується культурними та економічними системами, які ставлять індивідуальний успіх вище за колективні моральні норми.

Ці всі фактори можна також пояснити трикутником шахрайства [23]. Це схема, розроблена кримінологом Дональдом Кресі, яка використовується для пояснення причин, що стоять за рішенням особи вчинити злочин. Трикутник шахрайства складається з трьох компонентів: можливість, стимул і раціоналізація. У контексті шахрайства у сфері страхування транспортних засобів можливість проявляється у відсутності суворих процесів перевірки заяв або нагляду, що призводить до того, що шахрайські можуть дії залишитися непоміченими. Мотивом можуть поставати, наприклад, фінансові труднощі через особисті обставини, що змушують розглядати шахрайські заяви як вихід із

ситуації. Раціоналізацією часто постає виправдання своїх дій тим, що страхова компанія не зазнає значної шкоди або що це спосіб встановити справедливість.

Проте такі неправомірні дії несуть за собою серйозні наслідки. Страхове шахрайство призводить до значних фінансових втрат для страхових компаній, впливаючи на їхню здатність надавати доступне страхування споживачам. Страховики повинні підвищувати тарифи, щоб компенсувати збільшені витрати, пов'язані з шахрайством, а в деяких випадках це впливає на всіх користувачів, оскільки витрати на шахрайство розподіляються між страхувальниками [17]. Фактично, шахрайство є одним з основних чинників зростання страхових внесків у секторі автострахування [16].

Наприклад, у Сполучених Штатах Америки у 2003 році шахрайські заяви потенційно призводили до додаткових витрат в сумі 950 доларів на рік на сім'ю через те, що такі випадки збільшували вартість страхування для населення [17].

Більше того, страхове шахрайство має хвильовий ефект на економіку в цілому, оскільки призводить до зростання операційних витрат страхових компаній [16]. Для боротьби з шахрайством ці компанії повинні інвестувати в системи виявлення шахрайства, судові витрати, розслідування страхових випадків та використовувати більш досконалу аналітику даних, витрачати більше часу на перевірку заяв, що призводить до збільшення загальної вартості ведення бізнесу для страхових компаній. Ці витрати теж включаються в структуру ціноутворення, що є ще однією причиною зростання страхових цін.

Експертами було підраховано, що в Австралії напочатку 2000-х років шахрайство призводило до втрат у галузі автострахування в сумі від 3 до 3,5 мільярдів австралійських доларів щорічно, а за деякими оцінками – до 9 мільярдів австралійських доларів. Це робило шахрайство в страхування одним із найдорожчих злочинів у країні, оскільки воно призводить до прямих фінансових втрат, які становили близько третини усіх злочинів [17].

Широке розповсюдження шахрайства також призводить до економічної неефективності використання коштів компаній. Коли шахрайські заяви порушують нормальну роботу страхового ринку, вони забирають ресурси, які

могли б бути використані для інших продуктивних цілей, таких як інновації або розробка більш доступних варіантів страхового покриття. Крім того, шахрайство часто підштовхує страховиків до прийняття більш жорстких критеріїв оцінки ризиків, підвищення страхових франшиз або обмеження варіантів страхового покриття [16]. Це, в свою чергу, знижує рівень задоволеності страхувальників та обмежує доступ до бюджетного страхування для певних груп споживачів, що ще більше поглиблює економічну нерівність.

Деякі експерти прогнозують, що якщо шахрайство не буде належним чином вирішено, потреба у вищих цінах може стати непідйомною, що вплине як на конкурентоспроможність страхових компаній, так і на фінансовий добробут споживачів [16]. Це у свою чергу знизить купівельну спроможність страхувальників, що негативно вплине на економіку в цілому.

Шахрайство також завдає соціальної шкоди, оскільки підриває довіру населення до системи страхування. Широке розповсюдження шахрайства може призвести до негативного сприйняття страхової галузі, що може підірвати бажання споживачів співпрацювати зі страховиками, купувати нові поліси або подавати заяви на відшкодування, маючи на це законні підстави. Споживачі не відчуватимуть лояльності до свого страховика, що призводить до збільшення відтоку клієнтів. Це також вплине на довгострокову прибутковість страховиків, оскільки їм, ймовірно, доведеться витрачати більше коштів на утримання клієнтів та долати негативні наслідки підірваної репутації. У крайніх випадках це може навіть підштовхувати людей до вибору неформальних чи нерегульованих схем страхування, створюючи подальшу нестабільність на страховому ринку [16].

РОЗДІЛ 2. Методи виявлення шахрайства в автострахуванні

2.1. Традиційні методи виявлення шахрайства в автострахуванні

Так як шахрайські заяви продовжують обтяжувати страхові компанії, та піднімати тарифи для чесних страхувальників, виявлення обману є критично важливою функцією в цій сфері. Проте виявлення шахрайства часто є проблематичним через обмеженість фізичних доказів порівняно з іншими видами злочинів [17].

До появи інструментів машинного навчання та передових технологій виявлення шахрайства в страхуванні транспортних засобів в основному покладалося на традиційні методи у виконанні людей. Ці підходи ґрунтувалися насамперед на людських судженнях, досвіді та методах розслідування. Сюди, наприклад, входили оцінка поведінкових моделей, претензій, інтерв'ю, перевірка послідовності заявників, спільний нагляд з боку страхових агентів, персоналу ремонтних майстерень та експертів з відшкодування збитків від нещасних випадків та виявлення «червоних прапорців» за допомогою ручного аналізу.

Одним із ключових традиційних методів було використання «червоних прапорців», або індикаторів, які вказують на ймовірність шахрайства. Вони не є остаточним доказом обману, але слугують попереджувальними знаками для заяв, які заслуговують на подальшу перевірку [17]. До поширених «червоних прапорців» належать кілька страхових полісів у різних компаніях, надання недостовірної особистої інформації, відсутність поліцейського звіту або нещасний випадок, що стався за підозрілих обставин, наприклад, пізно вночі в неміській місцевості. Додаткові тривожні ознаки, наприклад, включають повідомлення про викрадення транспортних засобів, які потім швидко повертаються чи знайдені спаленими, або незвично агресивну поведінку страхувальника чи його готовність взяти на себе провину. У першому випадку страхувальник може намагатися уникнути допитів чи перевірки, а у другому – швидко закрити справу.

Іншим широко використовуваним методом є аналіз заяв. Страховики оцінюють послідовність, деталізацію та мову письмових або усних заяв, поданих

страхувальником. Невідповідності між первинними та подальшими заявами, нечіткі часові рамки або емоційна відстороненість можуть свідчити про можливий обман. Аналітики також з'ясовують, чи не брали участь у написанні однієї і тієї ж заяви кілька людей, що може свідчити про змову, наприклад, коли член сім'ї намагається приховати провину іншого. Також вивчався вибір конкретних слів, наприклад, опис руху як «паркування» або «задній хід», оскільки певні дієслова статистично корелювали з випадками шахрайства [17].

Ще одним поширеним традиційним механізмом виявлення шахрайства є візуальний огляд пошкоджень експертами з відшкодування збитків та автомайстернями. Досвідчені фахівці часто виявляють невідповідності між заявленими пошкодженнями та речовими доказами на транспортному засобі [24]. Наприклад, невідповідність у розташуванні або характері пошкоджень може свідчити про спроби шахрайства, наприклад, включення вже наявних несправностей. Крім того, традиційним інструментарієм були поліцейські звіти, свідчення свідків та адміністративні деталі, такі як оперативність подання заяви або місце крадіжки.

Інтерв'ю один-на-один та оцінка історії водіння також належать до поширених традиційних інструментів. Експерти з врегулювання збитків часто покладаються на неформальні судження під час обговорень зі страхувальниками та семінарів для оцінки достовірності [24]. На таких семінарах вони зазвичай можуть виявити непрямі спроби шахрайства, такі як нечіткі або оманливі описи нещасних випадків, завдяки своєму практичному досвіду в цій галузі.

На основі усіх даних та деталей деякі експерти виділяють типологію для класифікації заяв страхувальників:

- Адекватні: чіткі, детальні та фактологічно послідовні;
- Неадекватні: надмірно детальні, емоційні або театральні;
- Невиразні: розпливчасті або уникаючі у ключових аспектах;
- Флегматичні: дуже короткі або байдужі, не містять критичної інформації.

Ця класифікація використовується для того, щоб допомогти слідчим швидко оцінити, які заяви, ймовірно, є правдивими, а які вимагають більш глибокого розслідування [17].

Ще одним важливим традиційним методом є бюджетні обмеження, встановлені страховиками. Страхові компанії накладають суворі фінансові рамки на витрати та оціночні вартості на кожен ремонт, що не дає можливості перебільшувати необхідні витрати або включати не пов'язані з аварією послуги. Цей метод контролю не завжди ефективний у випадку великих аварій, проте може обмежити масштаби опортуністичного шахрайства [24].

Важливу роль відіграють також колективна відповідальність і моральні принципи. У той час як деякі ремонтні майстерні вирішують задовольнити шахрайські вимоги заради лояльності клієнтів або грошової вигоди, інші активно відмовляються від таких дій, керуючись етичними нормами або побоюючись репутаційних втрат. Превентивна поведінка, як-от відмова в обслуговуванні або вимога фотодоказу пошкоджень перед ремонтом, є традиційними стримуючими факторами проти шахрайства [24].

Традиційні системи нагляду – включаючи перевірки страховими представниками та неформальні комунікаційні мережі між компаніями – також використовуються для виявлення та покарання шахраїв. Однак незважаючи на наявність цих систем, багато шахрайських заяв все ще залишаються невиявленими або незареєстрованими, частково через те, що шахрайство вважається «злочином без жертв», а також через те, що випадки притягнення до відповідальності залишаються рідкісними [24].

Ці методи, хоч і були ефективними в деяких випадках, були обмежені суб'єктивністю та нездатністю ефективно обробляти великі обсяги даних, забирали багато часу та мали обмежену масштабованість, що підкреслювало потребу в більш досконалих інструментах виявлення шахрайства.

Саме тому з часом страхові компанії почали використовувати статистико-економетричні методи. Сюди входять такі техніки, як логістична регресія, лінійний дискримінантний аналіз та багатофакторні логістичні моделі, які

використовують статистичні показники [25]. Ці підходи спрямовані на оцінку ймовірності шахрайських заяв за допомогою добре встановлених математичних зав’язків між незалежними та залежними змінними.

Багато експертів вважають, що ці традиційні методи не тільки концептуально простіші та легші для інтерпретації, але й часто більш економічно ефективні, ніж сучасні моделі на основі штучного інтелекту [25]. Однією з ключових причин є те, що традиційні методи вимагають менше ресурсів для впровадження та підтримки. Наприклад, їх можна застосовувати за допомогою стандартного статистичного програмного забезпечення без потреби у високій обчислювальній потужності або великих обсягах даних.

Статистико-економетричні моделі аналізують історичні дані та виявляють закономірності або аномалії, які корелюють з шахрайською поведінкою. Деякі з них теж використовують «червоні прапорці» – тривожні індикатори, такі як незвично високі витрати на ремонт або невідповідності в даних про страхові випадки. Дослідники також вирішують загальні проблеми, такі як дисбаланс даних, застосовуючи методи надмірної та недостатньої вибірки, щоб забезпечити належне представлення випадків шахрайства [25].

Одним з найпотужніших статистичних алгоритмів є наївний Байєс – це імовірнісний класифікатор, заснований на теоремі Байєса, який обчислює ймовірність належності екземпляра до певного класу на основі попередніх знань і спостережуваних доказів. Перевагами алгоритму є простота, обчислювальна ефективність і теоретична обґрунтованість, особливо в задачах класифікації високої вимірності [26].

Основним недоліком наївного Байєса є його припущення про умовну незалежність ознак – це означає, що він розглядає кожну вхідну змінну як незалежну від інших, враховуючи цільовий клас. Ця особливість і є причиною того, чому алгоритм називається «наївним». І хоча це припущення рідко повністю виконується в реальних наборах даних, алгоритм все одно часто працює добре, особливо коли швидкість і масштабованість є пріоритетами. Алгоритм наївного Байєса класифікує дані, обчислюючи апостеріорну

ймовірність кожного класу за формулою Байєса $\mathrm{P}(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}$, де A – це клас, а B – сукупність ознак спостереження, та обираючи той клас, для якого ця ймовірність є максимальною [27].

Алгоритм наївного Байєса демонструє переваги у швидкості обробки, низькому використанні пам'яті та простоті реалізації, що робить його практичним варіантом при роботі з великими наборами страхових даних або в умовах обмеженості обчислювальних ресурсів. Він особливо ефективний при роботі з текстовими даними або категоріальними атрибутами та не є чутливим до пропущених значень у тренувальному наборі даних [32].

Крім того, завдяки своїй прозорій та імовірнісній природі, алгоритм наївного Байєса пропонує інтерпретовані результати, що є цінним у таких сферах, як страхування, де прозорість рішень є важливою для аудиту та дотримання вимог законодавства. Її імовірнісні результати дозволяють страховикам встановлювати регульовані пороги тривоги.

Проте наявність кореляції або залежностей між ознаками може знизити точність передбачень цієї моделі, оскільки припущення незалежності не відповідатиме дійсності. Наприклад, у своєму дослідженні «Виявлення шахрайства в галузі страхування за допомогою алгоритмів дата-майнінгу: дерево рішень, наївний Байєс і метод опорних векторів (на прикладі страхування автомобільних кузовних пошкоджень)» Голейджі й Тарох [33] використали набір даних, що містив інформацію про 360 заяв щодо збитків, з яких 91 заява була шахрайською. Модель машинного навчання досягла показника точності 92.50%, тоді як наївного Байєса досягла показника точності 90.28%. І хоча цей відрив не є великим, він показує, що припущення незалежності змінних не виконується у реальних сценаріях та впливає на результати класифікації наївного Байєса, а у випадку із шахрайством у страхуванні збільшення точності моделі на 2.22% є суттєвим.

Ще одним поширеним статистико-економетричним методом є логістична регресія. Ця модель використовується для прогнозування бінарного результату – шахрайство чи ні – на основі набору вхідних змінних. Логістична регресія

широко застосовується у виявленні страхового шахрайства через свою простоту, інтерпретованість та легкість реалізації, особливо коли набір даних є відносно структурованим, а зв’язки між змінними вважаються лінійними [26].

Щоб зрозуміти, як логістична регресія перетворює вхідні ознаки в оцінки ймовірності, використовується функція logit, яка пов'язує атрибути з log-шансами настання події (шахрайство чи ні):

$\ln\left(\frac{p}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij}$, де:

- p_i представляє ймовірність того, що спостереження є випадком шахрайства:
- використання натурального логарифму робить цю залежність лінійною за коефіцієнтами;
- β_0 вказує на базові log -шанси, коли значення всіх атрибутів дорівнюють нулю;
- кожен з β_j показує, як збільшення ознаки X_{ij} на одну одиницю змінює log-шанси шахрайства.

Розклавши праву частину, ми можемо перетворити log-шанси назад у ймовірності за допомогою логістичної (сигмоїдної) функції $\hat{p}_i = \frac{1}{1+e^{-(\beta_0+\sum_{j=1}^{m}\beta_j X_{ij})}}$. Таке формулювання дозволяє логістичній регресії виводити значення строго між 0 і 1, що робить їх безпосередньо інтерпретованими як ймовірності шахрайства. Лінійна логістична структура також гарантує, що оцінки коефіцієнтів залишаються інтерпретованими: кожен з них відповідає зміні log-шансів на одиницю зміни атрибута [26].

Моделі логістичної регресії особливо корисні, коли якість даних висока, а кількість змінних керована. Тестування різних конфігурацій логістичної регресії, включаючи методи вхідного, прямого і зворотного покрокового відбору, показує, що незалежно від обраного методу, модель демонструє стабільні результати з точністю близько 79%, що свідчить про те, що вона залишається надійним та економічно ефективним інструментом для виявлення заяв про шахрайство [26].

Важливо, що, незважаючи на розвиток машинного навчання, традиційні методи все ще дуже добре працюють у реальних умовах, особливо з точки зору економічної ефективності. Деякі дослідники вважають, що в середньому традиційні статистичні методи забезпечують кращі фінансові результати для страховиків, ніж багато новітніх методів на основі штучного інтелекту [25].

Це свідчить про те, що, особливо для страховиків, які працюють в умовах бюджетних обмежень або на ринках з обмеженими технічними ресурсами, традиційні підходи залишаються практичним і надійним вибором для виявлення шахрайства в секторі автострахування.

Незважаючи на тривале використання та простоту реалізації, традиційні методи виявлення шахрайства, такі як логістична регресія, стикаються з кількома суттєвими проблемами в ефективному виявленні шахрайських заяв у автомобільному страхуванні. Одним з основних перешкод є їхня обмежена здатність моделювати складні та нелінійні взаємозв'язки в даних. Моделі мають обробляти все більше даних з різних джерел – демографія, телематика, поведінка водіїв. Страхове шахрайство часто пов'язане з тонкими закономірностями, взаємозалежностями між змінними та еволюціонуючими тактиками, які не можуть бути адекватно враховані лінійними статистичними моделями. Хоча логістична регресія забезпечує стабільну точність результатів у різних конфігураціях, її чутливість залишається відносно низькою, зі значеннями нижче 45%, що свідчить про те, що значна кількість шахрайських заяв залишається невиявленою [26].

Іншою важливою проблемою є те, що традиційні моделі значною мірою залежать від ручного вибору ознак і перетворень змінних, що вимагає знання предметної області для правильного визначення моделі та уникнення пропусків критичних предикторів (пояснюючих змінних), що підвищує ризик людської упередженості. Вони також часто припускають незалежність між пояснювальними змінними та лінійні зв'язки, що рідко відповідає дійсності в складних та реальних наборах даних [26].

Порівняно новим викликом є поява автономних транспортних засобів. Ці технології змінюють характер аварій і страхових випадків. Традиційні моделі не пристосовані для обробки складних даних, пов'язаних з безпілотними автомобілями, таких як показання датчиків, програмні журнали або сліди рішень, прийнятих штучним інтелектом під час аварії. Крім того, стає важче визначити відповідальність, що ускладнює виявлення шахрайських намірів. Оскільки схеми шахрайства розвиваються паралельно з цими технологіями, традиційним моделям не вистачає гнучкості та адаптивності, необхідних для того, щоб залишатися актуальними в умовах, що швидко змінюються [21].

Зі зростанням обсягу та складності страхових даних традиційні моделі втрачають масштабованість і можуть стати неефективними або надто спрощеними для сучасних потреб виявлення шахрайства. Їх обмежена адаптивність до нових схем шахрайства та нездатність автоматично навчатися на основі закономірностей у великих, багатовимірних даних ще більше знижують їхню ефективність. Ці обмеження підкреслюють, чому, незважаючи на їхню інтерпретованість і економічну ефективність, традиційні методи все частіше доповнюються або замінюються більш складними підходами машинного навчання, які можуть краще справлятися з динамічними, незбалансованими і складними середовищами виявлення шахрайства.

2.2. Методи машинного навчання для виявлення шахрайства в автострахуванні

Оскільки шахрайство в страхуванні транспортних засобів стає все більш витонченим, традиційні методи його виявлення часто не здатні розпізнати тонкі або мінливі схеми. Машинне навчання пропонує потужну альтернативу, дозволяючи системам вчитися на історичних даних і робити точні прогнози щодо підозрілих заяв. Ці алгоритми можуть обробляти великі обсяги структурованих даних, розпізнавати складні взаємозв'язки між змінними та з часом адаптуватися до нових тактик шахрайства. На відміну від систем, заснованих на правилах, моделі машинного навчання постійно вдосконалюються в міру надходження

нових даних, що підвищує їхню здатність виявляти як відомі, так і раніше небачені схеми шахрайства.

Дерево рішень — це непараметричний алгоритм навчання з учителем, який використовується для задач класифікації [26]. Алгоритм дерева рішень є однією з найефективніших моделей завдяки своїй простоті, швидкості та здатності обробляти зашумлені або неповні дані. Модель класифікує дані, створюючи структуру, де внутрішні вузли представляють стан об'єкта чи ситуації та одночасно точки прийняття рішень на основі атрибутів, гілки висвітлюють можливі варіанти розвитку подій, що призводять до результатів цих рішень, а листові вузли представляють кінцеві результати класифікації: в контексті страхування транспортних засобів, шахрайство чи не шахрайство. Алгоритм рекурсивно вибирає найбільш інформативну ознаку для розбиття на підмножини на кожному кроці – змінна, що найкраще відокремлює шахрайстві заяви від нешахрайських, – що максимізує точність класифікації [27].

Однією з основних переваг алгоритму є візуальне представлення шляхів прийняття рішень. Оскільки процес прийняття рішень є прозорим і базується на простих правилах, страховики та аналітики можуть легко відстежити, як була класифікована та чи інша заява. Можливість відстежити та візуалізувати роботу класифікатора є особливо цінною в регуляторному або правовому середовищі, де процес прийняття алгоритмічних рішень є важливим [26]. Таким чином, інтерпретованість дерева рішень і їхня здатність до нелінійного моделювання роблять їх особливо корисними для виявлення шахрайства, де якість даних може коливатися, а патерни можуть бути складними і нестандартними. Дерева рішень швидко навчаються і оцінюють нові дані, використовуючи при цьому мінімальний обсяг пам'яті. Вони є простим, але потужним класифікатором, який добре підходить для реальних сценаріїв. Низькі обчислювальні вимоги і швидке виконання роблять його привабливим рішенням для страхових компаній, які прагнуть покращити виявлення шахрайства, зберігаючи при цьому прозорість моделі [27].

Однак за відсутності обмеження росту дерева рішень можуть почати фіксувати шум, а не закономірності, що лежать в основі даних. Крім того, невеликі зміни в наборі даних можуть призвести до абсолютно різних структур дерев, тому для знаходження надійного універсального рішення необхідне ретельне налаштування, наприклад, обмеження глибини дерева або встановлення кількості класів. Важливим недоліком дерев рішень є також те, що вони не вказують ймовірності того, чи заява шахрайська чи ні, і, відповідно, не висвітлюють різницю між заявами в одному класі [31].

Метод опорних векторів (SVM) – є ще одним потужним алгоритмом машинного навчання, який широко застосовується в різних задачах класифікації, в тому числі для виявлення страхового шахрайства. Метод опорних векторів працює шляхом знаходження оптимальної гіперплощини, яка розділяє класи – в контексті страхування транспортних засобів, шахрайські та нешахрайські заяви – з максимально можливою різницею [29]. Проте метод опорних векторів може бути застосований не лише до лінійних проблем. Однією з сильних сторін SVM є використання ядрових (kernel) функцій, які перетворюють дані з початкового простору в неявний простір ознак високої вимірності. Це перетворення дозволяє методу опорних векторів ефективно вирішувати нелінійні задачі класифікації, роблячи дані лінійно відокремлюваними в новому просторі [33].

Таким чином, метод опорних векторів є дієвим алгоритмом для виявлення шахрайства, особливо у поєднанні з розширеною попередньою обробкою даних і налаштуванням параметрів. Його міцна теоретична основа, гнучкість завдяки ядровим функціям та ефективність у просторах високої вимірності роблять його сильним кандидатом для подальшого розвитку аналітики страхового шахрайства [27]. Проте варто зауважити, що алгоритм вимагає складних обчислень, тому навіть найшвидші інтерпретації можуть працювати повільно [33].

І хоча SVM є широко використовуваним алгоритмом для виявлення шахрайства, багато досліджень показують низьку ефективність цієї моделі. Наприклад, у тому ж дослідженні Голейджі й Тароха SVM показав точність лише 30,28%. В іншому дослідженні «Прогнозування виявлення страхового

шахрайства за допомогою алгоритмів машинного навчання» Рухсара, Бангяла, К. Нісара та С. Нісара [27] SVM досяг точності 73%, що є кращим, але все ще не достатньо високим показником. Можливими причинами низьких результатів можуть бути вибір неправильного типу ядра, проблеми з налаштуванням гіперпараметрів або особливості структури даних, які не були оптимальними для SVM без додаткової попередньої обробки чи балансування даних [34]. Таким чином, лише за умови належної оптимізації цей математично строгий метод може слугувати ефективним компонентом гібридних або ансамблевих систем виявлення шахрайства.

Нейронні мережі, натхненні взаємопов'язаними нейронами людського мозку, пропонують ще один потужний підхід до виявлення шахрайства в автострахуванні. У типовій нейронній мережі кожна вхідна змінна (атрибут) представлена вузлом у вхідному шарі. Потім ці вхідні дані проходять через один або кілька прихованих шарів, де кожен нейрон обчислює зважену суму своїх вхідних даних, застосовує нелінійну функцію активації і пересилає результат. Така багатошарова структура дозволяє мережі вловлювати складні, нелінійні взаємозв'язки між ознаками, які простіші моделі можуть упустити. Після цього один нейрон у вихідному шарі, що містить сигмоїдну функцію активації, виробляє ймовірнісну оцінку, яка вказує на шанси того, що дана заява є шахрайською [26].

Навчання відбувається за допомогою алгоритму зворотного поширення. Коли передбачення мережі відрізняється від істинного класу, ця помилка поширюється назад через шари, а вага кожного зв'язку коригується пропорційно до його внеску в помилку і попередньо визначеної швидкості навчання. Ітеративне застосування цього процесу протягом багатьох повторів поступово мінімізує загальну помилку прогнозування мережі. Хоча існують різні архітектури, найчастіше використовується багатошаровий персептрон (MLP), завдяки своїй гнучкій, але простій конструкції. Деякі дослідження також вивчають мережі з радіально-базисними функціями (RBF), які відрізняються використанням локалізованих гауссових активаційних функцій у прихованому

шарі. У порівняльних експериментах на наборах даних про страхове шахрайство MLP зазвичай перевершують RBF-мережі, демонструючи вищу точність, чутливість та площу під ROC-кривою [26].

Незважаючи на свою здатність передбачати, нейронні мережі потребують ретельного налаштування кількості прихованих шарів, нейронів і швидкості навчання. Вони також потребують великої бази навчальних даних, щоб уникнути перенавчання.

Одномодельні класифікатори є основою багатьох алгоритмів виявлення шахрайства, оскільки вони вивчають вплив окремих атрибутів заяви – суми транзакції, часу подання, демографічних даних страхувальника та текстових описів – на бінарну класифікацію заяви. Оскільки кожне передбачення ґрунтується на одному алгоритмі, ці моделі легко інтерпретувати та перевіряти: аналітик може простежити, як саме вхідні дані призводять до прийняття рішення про шахрайство. Їх простота також приносить практичну користь. Навчання та передбачення, як правило, відбуваються швидко, що робить можливим прогноз в режимі реального часу, а використання ресурсів залишається незначним. Проте, з іншого боку, одномодельним підходам може бути складно врахувати комплексні взаємодії елементів або підтримувати стабільність, коли дані зашумлені або дуже незбалансовані, тому вони часто слугують прозорими базовими лініями або компонентами в рамках більших гібридних систем.

Ансамблеві класифікатори долають багато обмежень, притаманних окремим одномодельним класифікаторам, об'єднуючи кілька моделей в єдину систему прийняття рішень. Хоча ансамблі вимагають більших обчислювальних і реалізаційних зусиль – як для навчання, так і для пояснення комбінованих прогнозів – вони стабільно забезпечують вищу точність і надійність на зашумлених, багатовимірних або сильно незбалансованих наборах даних. Об'єднуючи кілька слабших моделей, ансамблі досягають балансу між точністю та стійкістю, якого не можуть досягти одномодельні класифікатори.

Випадковий ліс (Random Forest) – приклад широко використовуваного ансамблевого методу навчання, що будує кілька дерев рішень і об'єднує їхні

результати. Кожне з дерев приймає своє рішення, а система обирає той варіант, який найбільше повторюється. Випадковий ліс є найвідомішим методом бутстрепової агрегації. Об’єднання результатів кількох дерев рішень дає точніше передбачення, що допомагає зробити результат більш надійним [30]. Проте це робить Random Forest схильним до перенавчання. З цієї причини кожне дерево тренується на випадковій підмножині даних та використовує лише частину атрибутів. Таким чином, кожне дерево навчається на основі дещо іншого сприйняття даних, що підвищує точність і стабільність та зменшує вплив шумних значень [26].

Випадковий ліс ефективно справляється з великими наборами даних і здатний обробляти шум та пропущені значення. Беручи до уваги кілька дерев рішень, навчених на різних вибірках даних, модель може нейтралізувати вплив зашумлених або екстремальних значень, які часто зустрічаються в реальних наборах даних про страхові виплати [27]. І хоча алгоритм часто має високу чутливість, модель гарантує високу загальну точність і здатність запам'ятовувати, що робить її надійним інструментом для виявлення шахрайських заяв в страхуванні транспортних засобів. Наприклад, у дослідженні «Застосування методів дата майнінгу для виявлення шахрайства в автомобільному страхуванні» Бангчанга, Вонґсая та Сіммачана [35] випадковий ліс показав точність передбачень 97,91%.

Проте варто зауважити, що при роботі з незбалансованими наборами даних, де один клас значно переважає інший – як це часто буває з шахрайством, де шахрайських випадків значно менше, – багато методів машинного навчання, включаючи стандартний випадковий ліс, можуть ігнорувати цей дисбаланс, що призводить до низької ефективності для міноритарного класу. У таких випадках рекомендується використовувати техніки, що дозволяють збалансувати дані, перед застосуванням алгоритмів класифікації [36].

Adaboost, скорочено від Adaptive Boosting (адаптивне підсилення), – це потужна техніка ансамблевого навчання, яка покращує продуктивність слабких класифікаторів, об'єднуючи їх у сильну прогностичну модель. Вона належить до

методів бустингу (підсилення). У контексті виявлення страхового шахрайства Adaboost показує дуже багатообіцяючі результати.

Алгоритм працює, об'єднуючи кілька слабких навчальних моделей – часто неглибоких дерев рішень – в один сильний класифікатор. Спочатку всім спостереженням надаються рівні ваги. Під час навчання Adaboost збільшує вагу неправильно класифікованих атрибутів, змушуючи наступну слабку модель більше зосереджуватися на цих прикладах [28]. Такий ітеративний перерозподіл ваг дозволяє Adaboost постійно покращувати роботу класифікації, зменшуючи упередженість і підвищуючи точність з часом.

На відміну від випадкових лісів, Adaboost особливо корисний у проблемах незбалансованої класифікації, таких як виявлення шахрайства, коли шахрайських заяв набагато менше, ніж легітимних. Його адаптивний механізм допомагає зосередитися на прикладах класу меншості, які інші моделі можуть пропустити [27].

Сила Adaboost полягає в його гнучкості та інкрементальному процесі навчання. Ітеративно фокусуючись на неправильно класифікованих прикладах і об'єднуючи результати роботи декількох слабших моделей, він зменшує як упередженість, так і дисперсію. Adaboost особливо ефективний тоді, коли базові класифікатори є «слабкими» – їхня точність лише трохи вища за випадкове вгадування. Саме це робить його розумним вибором для швидкого, адаптивного навчання на структурованих страхових даних [27].

Однією з ключових переваг Adaboost є також те, що він не потребує складного налаштування параметрів, що робить його більш доступним і менш чутливим до неправильного вибору гіперпараметрів порівняно з деякими іншими ансамблевими методами. Однак, хоча він демонструє відмінну точність, Adaboost може бути менш інтерпретованим, ніж окремі дерева рішень, і дещо більш обчислювально інтенсивним через свою ітеративну природу.

У статті «Прогнозування шахрайства в автомобільному страхуванні за допомогою класичних алгоритмів та алгоритмів машинного навчання», автори

показали, що поєднання AdaBoost з різними базовими деревами рішень може дещо підвищити показники ефективності моделей [26].

Шахрайство в автострахуванні продовжує зростати в масштабах і складності, що робить необхідним удосконалення сучасних обчислювальних підходів для його точного виявлення. Методи машинного навчання забезпечують гнучкість і аналітичну потужність, необхідні для виявлення тонких, прихованих або неочевидних закономірностей у великих і часто незбалансованих наборах даних. Розглянуті класифікатори пропонують широкий спектр можливостей – від моделей, що піддаються інтерпретації, до більш надійних ансамблевих методів. Їх застосування для виявлення шахрайства вимагає не лише технічної коректності, але й ретельного узгодження між характером даних, цілями страховика та компромісом між точністю, інтерпретованістю та обчислювальною ефективністю. У табл. 2.1 продемонстровано переваги та недоліки моделей, на які варто звернути увагу при виборі необхідного алгоритму. І хоча жодна модель не пропонує універсального рішення, їх сукупність створює основу для інтелектуальних систем виявлення шахрайства.

Таблиця 2.1

Порівняння моделей машинного навчання

Модель	**Переваги**	**Недоліки**
Дерево рішень	• візуальне представлення, інтерпретація; • низькі обчислювальні вимоги.	• схильність до перенавчання без обмежень глибини; • чутливі до змін у базі даних; • не генерує ймовірності приналежності до класу.

Продовження табл. 2.1

SVM	• дієвий у високовимірних просторах; • строга теоретична база.	• висока обчислювальна складність; • потребує підбору ядра та гіперпраметрів.
Нейронні мережі	• здатність моделювати складні нелінійні взаємозв'язки; • автоматичне вивчення патернів.	• потребують великі навчальні бази даних; • велика ресурсозатратність; • висока складність налаштування гіперпараметрів.
Випадковий ліс	• більш надійний результат; • автоматично обробляє шум та відсутні значення.	• схильність до низької ефективності для міноритарного класу; • великий обсяг пам’яті та час навчання при великій кількості дерев.
Adaboost	• добре справляється з незбалансованими наборами даних; • не потребує складного налаштування параметрів.	• менш інтерпретований; • обчилювально затратний; • чутливий до шумових або неправильно маркованих об’єктів.

Джерело: розроблено автором.

РОЗДІЛ 3. Реалізація і порівняння методів машинного навчання у виявленні шахрайства в автострахуванні.

3.1. Опис датасету та обробка даних

Для проведення практичного аналізу та оцінки ефективності методів машинного навчання у виявленні шахрайства у страхуванні транспортних засобів я вирішила використати датасет Kaggle «Виявлення шахрайства у страхуванні транспортних засобів» (англ. «Vehicle Insurance Claim Fraud Detection») [47]. Цей набір даних містить різноманітні характеристики, пов'язані зі страховими умовами, страхувальниками та деталями заявлених інцидентів. Мета – виявити, чи є заявка на отримання компенсації шахрайською. Даний датасет забезпечує всебічну основу для моделювання виявлення шахрайства та порівняння ефективності різних алгоритмів машинного навчання.

Датасет складається із 15420 заяв на отримання компенсації та 33 характеристик. Ці характеристики включають:

- Month: Місяць, коли стався страховий випадок.
- WeekOfMonth: Тиждень місяця, коли стався страховий випадок.
- DayOfWeek: День тижня, коли стався страховий випадок.
- Make: Виробник транспортного засобу, залученого у страховий випадок.
- AccidentArea: Місцевість, де сталася аварія.
- DayOfWeekClaimed: День тижня, коли страхову заяву було фактично подано страхувальником або опрацьовано страховиком.
- MonthClaimed: Місяць, в якому страхову заяву було фактично подано страхувальником або опрацьовано страховиком.
- WeekOfMonthClaimed: Тиждень місяця, в якому страхову заяву було фактично подано страхувальником або опрацьовано страховиком.
- Sex: Стать страхувальника.
- MaritalStatus: Сімейний стан страхувальника.
- Age: Вік водія.
- Fault: Вказує, чия вина в аварії.

- PolicyType: Тип страхового полісу.
- VehicleCategory: Категорія транспортного засобу.
- VehiclePrice: Ціна транспортного засобу.
- PolicyNumber: Унікальний ідентифікатор страхового полісу.
- RepNumber: Унікальний ідентифікатор страхового представника, який обробляє заяву.
- Deductible: Сума, яку власник полісу повинен сплатити з власної кишені до того, як страхова компанія покриє решту витрат (страхова франшиза).
- DriverRating: Рейтинг водія, часто заснований на історії водіння або інших факторах.
- Days_Policy_Accident: Кількість днів з моменту оформлення полісу до моменту аварії.
- Days_Policy_Claim: кількість днів з моменту оформлення полісу до моменту подачі заяви на відшкодування.
- PastNumberOfClaims: Кількість страхових випадків, які раніше були заявлені страхувальником.
- AgeOfVehicle: Вік транспортного засобу, що фігурує в заяві про відшкодування.
- AgeOfPolicyHolder: Вік страхувальника.
- PoliceReportFiled: Вказує, чи був поданий поліцейський звіт про аварію.
- WitnessPresent: Вказує, чи був присутній свідок на місці аварії.
- AgentType: Тип страхового агента, який обслуговує поліс.
- NumberOfSuppliments: Кількість додаткових документів або претензій, пов'язаних з основною заявою.
- AddressChange_Claim: Вказує, чи була змінена адреса страхувальника на момент подання заяви.
- NumberOfCars: Кількість автомобілів, застрахованих за полісом.
- Year: Рік, в якому було подано або оброблено заяву про відшкодування.
- BasePolicy: Тип базового полісу.

- FraudFound_P: Вказує, чи було виявлено шахрайство в страховій заяві.

Цей комплексний набір даних дозволяє дослідити потенційні закономірності, кореляції та патерни, які можуть допомогти у виявленні шахрайських заяв за допомогою різних методів машинного навчання. Великий об’єм даних, що охоплює демографічні, специфічні для транспортних засобів, пов'язані зі страховими випадками та ситуаційні особливості, дозволяє застосувати багатогранний підхід до виявлення шахрайства, який відображає реальні складнощі в страховій галузі.

Перш ніж застосовувати будь-які моделі машинного навчання, важливо оцінити чистоту та якість набору даних. Це включає виявлення відсутніх значень, невідповідностей і потенційних відхилень, які можуть вплинути на продуктивність моделі. Належна попередня обробка даних гарантує, що набір даних є надійним і точно відображає основну проблему. Важливі кроки на цьому етапі також включають перетворення категоріальних змінних у числову форму, масштабування або нормалізацію числових характеристик та усунення будь-яких дисбалансів у даних. Така попередня обробка є визначальною для підготовки даних до подальшого аналізу, а також для підвищення продуктивності та інтерпретованості моделей.

Після проведення первинної оцінки даних я виявила, що датасет є достатньо чистим, оскільки не містить жодних пропущених значень і дублікатів. Атрибут «PolicyNumber» функціонує виключно як ідентифікаційний номер і не має жодного прогностичного значення для виявлення шахрайства, тому я вирішила вилучити його з подальшого аналізу. «RepNumber» вказує лише на те, який агент опрацьовував заявку. Цей ідентифікатор насправді не має незалежної прогностичної сили і діятиме як випадковий шум, що може зашкодити узагальненню та призвести до перенавчання. Атрибут «Year» також варто вилучити, так як немає сенсу прогнозувати майбутні випадки шахрайства на основі того, в якому році вони трапляються. Після детальнішого дослідження я помітила, що «PolicyType» – це комбіноване значення двох інших стовпців: «VehicleCategory» та «BasePolicy». Ці атрибути не несуть в собі ніякої додаткової

інформації, а зменшення кількості стовпців може пришвидшити навчання та прогноз моделі. Тому я вирішила видалити ці два стовпці і залишити лише «PolicyType».

Особливе спостереження виникло щодо характеристик «DayOfWeekClaimed», «MonthClaimed» та «Age», оскільки їхні мінімальні значення дорівнюють нулю, що не є реалістичним для віку страхувальника та дати і вказує на необхідність спеціальної обробки під час процесу очищення датасету. Після детальнішого дослідження даних я виявила, що датасет містить лише одну заяву, в якій і день, і місяць позначені нулем. Цей рядок, ймовірно, є аномалією, спричиненою помилкою при введенні даних. Видалення таких одиночних аномалій навряд чи призведе до змін чи суттєвого зменшення розміру вибірки – втрата одного запису з 15420 спостережень є незначною. Крім того, виведення правильного дня або місяця з інших полів є малоймовірним та ненадійним. Проте такий підхід не спрацює із стовпцем «Age», так як аж 319 заяв містять вік, позначений нулем. Видалення цих рядків призвело б до вилучення приблизно 2% всього набору даних, що могло б викривити віковий розподіл і негативно вплинути на здатність моделей розпізнавати певні патерни. У таких випадках нульові значення варто заміняти середнім значенням або медіаною. Середні значення можуть бути коливатися через аномальні значення. Натомість медіана є надійним показником, на який не надто впливають екстремальні значення, тому заміна недостовірного віку на медіану зберігає основну тенденцію вікового розподілу.

Важливим спостереженням є також те, що більшість ознак є категоріальними змінними, що суттєво впливає на процес підготовки даних для навчання, так як моделі у моєму дослідженні сприймають лише числові та бінарні значення.

«AccidentArea», «Fault», «Sex», «PoliceReportFiled», «WitnessPresent» і «AgentType» мають лише два унікальних значення, що робить їх ідеальними кандидатами для перетворення у тип boolean (використовувати лише значення 0 та 1).

При перетворенні категоріальних характеристик з трьома та більше унікальними значеннями важливо враховувати, чи мають вони природний порядок (порядкові), чи ні (номінальні), оскільки від цього залежить спосіб кодування.

«PastNumberOfClaims», «NumberOfSuppliments», «VehiclePrice», «AgeOfVehicle», «Days_Policy_Accident», «Days_Policy_Claim», «AgeOfPolicyHolder», «AddressChange_Claim» та «NumberOfCars» є порядковими атрибутами. Щоб перевести ці характеристики в числову форму, достатньо просто замінити кожну категорію цілим числом, яке відображає її ранг. Таким чином, кожне окреме число відповідає одній категорії, а більші числа завжди означають вищий ранг категорії. Позначення кожної категорії таким чином гарантує, що модель розпізнає правильну ієрархію, а не розглядає їх як непов’язаний текст.

«Make», «MaritalStatus», «PolicyType», «Month», «DayOfWeek», «DayOfWeekClaimed» і «MonthClaimed» є номінальними атрибутами – кожна категорія є незалежним класом без рейтингу чи ієрархії, тому проста заміна категорій на числа не спрацює. One-hot encoding (унітарне кодування) — це метод перетворення номінальних категорій у двійковий формат, що підходить для алгоритмів машинного навчання. Цей метод створює нові стовпчики для кожної категорії, де 1 означає, що категорія присутня, а 0 — що її немає [48]. Такий підхід гарантує, що модель не зробить жодних упереджених припущень про порядок між категоріями і вчиться на чистих бінарних ознаках. Важливим моментом є те, що унітарне кодування створює стільки ж нових стовпців, скільки унікальних категорій має атрибут. Таким чином після перетворення вищеперерахованих характеристик загальна кількість стовпців збільшилася до 91.

Так як кількість атрибутів суттєво пливає на швидкість моделі та використання пам’яті, оптимально було б вилучити стовпці, які не несуть в собі багато важливої інформації. Для кожного бінарного стовпця, створеного унітарним кодуванням, обчислюється сума його значень — кількість рядків, в яких з'являється ця категорія. Якщо сума значень стовпця дорівнює п'яти або

менше, це означає, що категорія зустрічається щонайбільше в п’яти спостереженнях з понад п'ятнадцяти тисяч, що робить її невагомою і вказує на те, що вона навряд чи допоможе моделі розпізнати шахрайство. Такі стовпці можуть бути вилучені, що дозволить зменшити розмір бази даних без втрати важливої інформації.

3.2. Реалізація методів машинного навчання

Перед початком навчання моделі набір даних варто поділити на дві окремі підмножини: тренувальну та тестову. Я вирішила 80 % спостережень віднести до тренувальної множини, а 20 % — до тестової. Тренувальна множина використовується для підбору параметрів моделі, тоді як тестова множина залишається повністю прихованою до самого кінця. Дуже важливо розділити дані до початку навчання, інакше модель може опосередковано запам’ятати інформацію з тестового набору під час навчання, що призведе до нереалістично високих оцінок продуктивності. Відокремлення 20% даних гарантує, що оціночні показники моделі правильно відтворюють її здатність до передбачень та узагальнення на дійсно небачених прикладах, тоді як збереження 80% даних для навчання забезпечує достатню кількість прикладів, щоб модель могла виявити закономірності, особливо з огляду на відносну рідкість шахрайських заяв. Окрім того, встановлення фіксованого значення параметра «random_state» при виконанні поділу датасету забезпечує однаковість результатів при повторних запусках і можливість об’єктивно порівнювати показники різних моделей.

Дерева рішень легко справляються як з числовими, так і з закодованими категоріальними ознаками, не вимагаючи масштабування, через що ця модель чудово підходить для даного датасету. Після того, як категоріальні атрибути перетворені в числову форму, дерево рішень автоматично визначає найбільш інформативні розбиття, щоб розрізнити законні та потенційно шахрайські заяви. Для побудови дерева рішень я використовую бібліотеку «scikit-learn» («sklearn»), що надає реалізації алгоритмів машинного навчання з учителем та без, а також інструменти для попередньої обробки, вибору моделі та оцінювання [49]. Для

побудови було використано стандартну реалізацію алгоритму дерева рішень CART (Classification And Regression Trees). При створенні моделі знову важливо вказувати фіксований параметр «random_state». Після цього, модель вивчає навчальні дані.

Однак виявлення шахрайства за своєю природою є незбалансованим — шахрайство присутнє лише в невеликій частці заяв — тому звичайне дерево рішень, як правило, майже скрізь прогнозує «відсутність шахрайства», гарантуючи високу точність моделі. Саме тому для порівняння я вирішила застосувати метод SMOTE (англ. Synthetic Minority Oversampling Technique) — що можна розшифрувати як метод синтетичного наддисбалансу меншості. Це метод, який використовується в машинному навчанні для вирішення проблем незбалансованих наборів даних, особливо в задачах бінарної класифікації. Незбалансовані набори даних виникають, коли один клас має значно менше зразків, ніж інший [50].

Таким чином я застосувала алгоритм SMOTE до початкової тренувальної множини, щоб синтетично збільшити розмір меншого класу (шахрайства). Завдяки цьому методу тепер у мене є змога побудувати нове дерево рішень на основі нового збалансованого набору даних, в якому кількість шахрайських заяв дорівнює кількості легітимних, що допомагає пом'якшити проблему класового дисбалансу і має покращити здатність моделі виявляти шахрайські заяви.

Попри популярність методу, навіть після балансування одне дерево рішень, як правило, досягає обмеженої точності, визначаючи меншу частину реальних випадків шахрайства або генеруючи неприйнятно високий рівень хибних спрацьовувань. Тому я також вирішила спробувати алгоритм випадкового лісу.

Для побудови випадкового лісу я використала ту ж саму бібліотеку «sklearn». Проте незбалансовані дані впливатимуть на роботу випадкового лісу теж. У цьому випадку я вирішила використати варіацію алгоритму – збалансований випадковий ліс. На противагу стандартному алгоритму, класифікатор збалансованого випадкового лісу з бібліотеки «imblearn» автоматично змушує кожне дерево навчатися на приблизно рівній кількості прикладів обох класів.

Тренуючи кожне дерево на цій збалансованій підмножині, ліс в цілому вчиться розпізнавати патерни, пов'язані з класом меншості, замість того, щоб за замовчуванням прогнозувати клас більшості. Цей процес значно покращує запам’ятовування шахрайства без потреби в синтетичному створенні даних або ручному коригуванні вагових коефіцієнтів класів, що призводить до більш надійної моделі, придатної для узагальнення та передбачень.

Важливим є те, що модель не завжди показує найкращі результати при налаштуванні за замовчуванням. Саме тому я використала GridSearchCV для пошуку найоптимальнішої комбінації гіперпараметрів, зокрема, кількості дерев (n_estimators), максимальної глибини дерева (max_depth), мінімальної кількості заяв, необхідних для розбиття вузла (min_samples_split), мінімальної кількості заяв на лист (min_samples_leaf) та кількості ознак, що враховуються при кожному розбитті (max_features). GridSearchCV — це потужний інструмент у «scikit-learn», який дозволяє здійснювати пошук найкращої комбінації параметрів, що забезпечує найвищу продуктивність моделі. Використовуючи показник чутливості (recall) як таргет та виконуючи триразову перехресну перевірку за попередньо визначеною сіткою параметрів, GridSearchCV визначив комбінацію гіперпараметрів, яка максимізувала здатність моделі виявляти шахрайські заяви без надмірного перенавчання. Такий перехід від простого класифікатора за замовчуванням до ретельно налаштованого ансамблю часто призводить до значно більш надійної моделі виявлення шахрайства на незбалансованому наборі страхових даних. Отримані від GridSearchCV параметри тепер вставляються у нову модель збалансованого випадкового лісу.

Я також побудувала класифікатор AdaBoost, використовуючи дерева рішень як базові навчальні елементи. Для навчання я знову використала синтетично доповнені дані SMOTE, щоб мінімізувати побічні ефекти дисбалансу бази даних. Важливим є знову використати GridSearchCV за такими гіперпараметрами: кількість раундів навчання (n_estimators), швидкість навчання (learning_rate) та максимальна глибина кожного дерева рішень (base_estimator__max_depth), використовуючи показник чутливості як метрику оцінювання.

Було вирішено не будувати модель методу опорних векторів: алгоритм може бути повільним та ресурсозатратним у випадку великих баз даних. Після унітарного кодування простір ознак розширився до понад дев'яноста вимірів, що робить пошук розділювальної гіперплощини ще складнішим і схильним до перенавчання. SVM також чутливий до масштабування ознак, а це означає, що перед навчанням кожен числовий стовпець має бути нормалізовано або стандартизовано, що вносить додаткову складність. Через дисбаланс даних метод опорних векторів потребує ретельного налаштування параметрів та порогових значень, щоб уникнути простого віднесення майже всіх заяв до класу більшості.

Таке ж рішення було прийнято щодо нейронних мереж: даний датасет є відносно невеликим, включаючи лише кілька тисяч заяв, з яких шахрайськими є лише кілька сотень. Нейронні мережі зазвичай потребують великих обсягів високоякісних тренувальних даних для ефективного передбачення та уникнення перенавчання; без достатньої кількості прикладів шахрайства модель може запам'ятовувати шум, а не вивчати загальні закономірності. Крім того, певні характеристики в наборі даних потребують масштабування, що ще більше ускладнює попередню обробку.

Крім того, було вирішено об’єднати алгоритми випадкового лісу та AdaBoost в єдину, більш надійну модель, що тепер буде наділена їхніми взаємодоповнюючими перевагами. Гібридний ансамбль стекування використовує два базові навчальні модулі: AdaBoost, навчений на SMOTE-збалансованих даних, і збалансований випадковий ліс, навчений на оригінальному незбалансованому наборі даних. Їхні прогнози потім подаються в невеликий оптимізований метакласифікатор випадкового лісу, який інтегрує сигнали від обох моделей для отримання остаточного прогнозу шахрайства. Для обох базових навчальних модулів було використано гіперпараметри, отримані GridSearchCV при побудові попередніх моделей. GridSearchCV було використано й на цьому етапі задля налаштування гіперпараметрів метакласифікатора. Такий багаторівневий підхід зберігає високу чутливість кожного алгоритму, водночас зменшуючи загальну похибку завдяки прогностичному консенсусу.

3.3. Порівняння ефективності моделей та аналіз результатів

Порівнюючи два варіанти дерева рішень — одне, навчене на початкових незбалансованих даних, а інше, навчене після застосування SMOTE, — видно, що жодна з моделей не досягає високих показників виявлення шахрайства, але дерево, яке навчається на SMOTE даних, має дещо кращі результати в цілому. Звичайне дерево рішень досягає 88,78% точності, але лише 0,542 показника ROC AUC (площа під ROC-кривою). Показник ROC AUC 0,5 вказує на випадкове вгадування, а 1,0 — на ідеальний розподіл. Звичайне дерево рішень знаходить лише 29 справжніх випадків шахрайства (recall ≈ 14,6 %), в той час як інші 170 заяв помилково класифікуються як легітимні.

Дерево рішень, навчене на збалансованих даних SMOTE, має дещо нижчу загальну точність (88,13%), але вищу збалансовану точність та ROC AUC (≈ 55,5%). Модель правильно ідентифікувала більше шахрайств (recall ≈ 18,1%), але ціною більшої кількості хибних спрацьовувань. Оскільки SMOTE штучно дозволяє моделі навчатися на однаковій кількості випадків шахрайських і легітимних заяв, версія SMOTE краще фіксує реальні шахрайські патерни (менше помилкових спрацьовувань) і дає вищий показник F1 (≈ 0,164 замість ≈ 0,144).

Звичайне дерево рішень дуже залежить від класу більшості — без SMOTE воно вчиться позначати майже все як нешахрайство. З практичної точки зору, дерево рішень з SMOTE є кращим з двох: його покращена чутливість і збалансована точність роблять його більш корисним для виявлення шахрайства. Проте навіть після застосування SMOTE дерево рішень все ще не може виявляти шахрайство достатньо ефективно з кількох ключових причин. По-перше, SMOTE збільшує кількість заяв класу меншості, але не додає нових, справді інформативних патернів шахрайства. Дерево рішень, будучи моделлю із високою дисперсією, результати якого сильно залежать від конкретного набору навчальних даних, все одно буде чіплятися за випадкові викиди або шум у цих синтетичних прикладах, а не розкривати глибші взаємозв'язки. По-друге, SMOTE може створювати зразки, що неоднозначно належать до обох класів. Одне дерево

не має механізму для згладжування цих неоднозначностей, тому воно або надмірно адаптується до синтетичних заяв (перенавчається), або повертається до поведінки, подібної до класу більшості, у суміжних регіонах. Крім того, дерева рішень виконують поділ по одній ознаці за раз, що ускладнює їх здатність захоплювати складні взаємодії між кількома атрибутами, а саме такі взаємодії часто характеризують реальне шахрайство. Таким чином, висока дисперсія, шумні синтетичні дані, велика розмірність і обмежена гнучкість поділу сприяють тому, що показники чутливості (recall) і точності (precision) дерева рішень залишаються скромними навіть із SMOTE даними, незважаючи на зростання збалансованої точності.

Стандартний випадковий ліс навчився передбачати "відсутність шахрайства" майже скрізь, даючи 93,58% точності, але лише 50,25% збалансованої точності. Він ідентифікує лише 1 справжній випадок шахрайства (recall ≈ 0,5%) з 199, ніколи не генеруючи жодного хибного позитивного результату (precision = 100 %). Іншими словами, навчаюсь на даних, більшість яких складають легітимні заяви, ліс фактично ігнорує рідкісні випадки шахрайства і позначає майже кожну заяву як нешахрайську — звідси оманливо висока загальна точність, але жахливий показник чутливості.

На противагу цьому, як і передбачалося, збалансований випадковий ліс досягає загальної точності у 62,03%, проте значно вищої збалансованої точності та AUC (≈ 76,9%). Після налаштування гіперпараметрів і застосування збалансованої вибірки для кожного дерева, модель уловлює 187 шахрайств (recall ≈ 93,97%), проте 1159 хибних спрацьовувань (precision ≈ 13,9%). Такий різкий стрибок в ефективності стався завдяки тому, що під час навчання алгоритм надавав кожному дереву приблизно однакову кількість прикладів шахрайства і легітимних заяв, що дозволило ансамблю вивчити справжні патерни шахрайства замість того, щоб за замовчуванням підлаштовуватися під клас більшості. Хоча загальна точність моделі знизилися через високий рівень хибних спрацьовувань, її здатність виявляти майже кожну шахрайську заяву, що відображається у високому показнику чутливості recall і збалансованому AUC, робить її набагато

більш придатною для виявлення шахрайства, ніж звичайний випадковий ліс, який просто не здатен ідентифікувати більшість рідкісних випадків шахрайства.

Ансамбль AdaBoost, навчений на збалансованих даних SMOTE, досягає помітно іншого профілю продуктивності, ніж окремі дерева рішень. Його загальна точність є значно нижчою (60,9%), але збалансована точність у 76,8% показує, що AdaBoost краще справляється з класифікацією незбалансованих даних. Зокрема, модель досягає надзвичайно високого показника чутливості recall — 94,97% (189 правильно класифікованих шахрайств і лише 10 пропущених). Однак такий результат досягається ціною великої кількості хибних спрацьовувань, що призводить до низького показника precision (13,65%). Отриманий показник F1 на рівні 0,239 вказує на те, що модель поєднує агресивне виявлення шахрайства та неминучу хибну класифікацію великої кількості легітимних заяв.

Гібридний ансамбль стекування поєднує два базові навчальні елементи: модель AdaBoost, навчену на збалансованих SMOTE даних, і збалансований випадковий ліс, навчений на вихідному незбалансованому наборі даних. Ці моделі тоді подаються в невеликий метакласифікатор випадкового лісу. Цей ансамбль досягає загальної точності 61,38% при збалансованій точності 77,49%, демонструючи свою здатність розпізнавати обидва класи. Найважливішим є те, що модель стекування охоплює 95,98% всіх випадків шахрайства (191 істинно позитивний результат та лише 8 хибно негативних), демонструючи виняткову чутливість. Однак це супроводжується 1183 хибними спрацьовуваннями, що знижує показник precision до 13,90%.

Порівняння ефективності дерева рішень, випадкового лісу, AdaBoost та гібридної моделі в одній таблиці дозволяє чітко побачити їхні переваги та недоліки (табл. 3.1). Просте дерево рішень, навіть збалансоване зі SMOTE, не може піднятися вище помірного рівня продуктивності: його висока загальна точність маскує низьку збалансовану точність і дуже низький показник recall, що робить модель погано придатною для проблеми, де виявлення класу меншості (шахрайства) є критично важливим. Натомість обидва ансамблеві методи,

збалансований випадковий ліс і AdaBoost, демонструють кращі результати: кожен з них досягає збалансованої точності і значення ROC AUC близько 0,77, а рівень показника recall перевищує 93%. Ці результати показують, що використання ансамблю декількох слабких учнів перевершує будь-яке одиночне дерево.

Проте, саме гібридний ансамбль стекування досягає найоптимальніших результатів. Шляхом скеровування моделі AdaBoost, навченої на даних SMOTE, і збалансованого випадкового лісу, навченого на оригінальному наборі даних, через невеликий метакласифікатор випадкового лісу, модель стекування підвищує збалансовану точність та ROC AUC до 77,49%, — незначний, але суттєвий ріст у порівнянні з іншими моделями. Найважливішим є те, що модель охоплює 95,98% фактичних випадків шахрайства, що є найвищим показником recall, зберігаючи при цьому найвищий показник F1.

Табл. 3.1

Характеристики ефективності розроблених моделей

Модель	**Дерево рішень SMOTE**	**Збалансований випадковий ліс**	**AdaBoost**	**Гібридна модель стекінгу**
Ключові метрики моделі				
Точність (Accuracy)	0,8813	0,6203	0,6089	0,6138
Збалансована точність (Balanced Accuracy)	0,5553	0,7690	0,7676	0,7749
F1	0,1644	0,2421	0,2386	0,2428
Чутливість (Recall)	0,1809	0,9397	0,9497	0,9598
Точність класифікації (Precision)	0,1506	0,1389	0,1365	0,1390

Продовження табл. 3.1

Площа під ROC-кривою (ROC AUC)	0,5553	0,7690	0,7676	0,7749
Матриця невідповідностей				
Істинно негативні	2682	1726	1689	1702
Хибно позитивні	203	1159	1196	1183
Хибно негативні	163	12	10	8
Істинно позитивні	36	187	189	191

Джерело: розроблено автором.

У випадку страхування транспортних засобів ціна пропуску фактично шахрайської заяви може бути значною, що призводить не лише до прямих фінансових втрат, але й до довгострокової репутаційної шкоди, оскільки невиявлене шахрайство сприяє подальшому зловживанню системою. На противагу цьому, хибно класифікована легітимна заява означає лише додаткову перевірку, втрати від якої є значно меншими та яку страхові компанії часто можуть подолати завдяки оптимізованим робочим процесам розслідувань та командам із запобігання шахрайству. Таким чином, хоча гібридний ансамбль стекування дає більше хибних спрацьовувань, ніж, наприклад, AdaBoost, стабільна перевага за збалансованими показниками точності робить його найнадійнішим вибором. Виявлення майже кожного випадку шахрайства, навіть якщо це означає ручну перевірку більшої кількості помічених випадків, є набагато кращим, ніж пропуск шахрайських заяв, які залишаться непоміченими. На практиці такий підхід допомагає захистити резервні фонди, зберегти цілісність ціноутворення для чесних страхувальників і посилити позицію страховика щодо протидії шахрайству, оскільки постійна ретельна перевірка на зловмисницькі дії піднімає планку для потенційних шахраїв.

Отже, з практичної точки зору, страховим компаніям, які шукають єдину модель для виявлення потенційного шахрайства, варто адаптувати гібридний ансамбль стекування. Його багаторівнева структура залучає вивчені шахрайські патерни обох моделей та згладжує їхні індивідуальні недоліки.

ВИСНОВКИ

У цьому дослідженні було розглянуто шахрайство у страхуванні транспортних засобів як фінансову та операційну проблему, з якою стикаються страховики. Шахрайські заяви суттєво впливають на страхову галузь, збільшуючи операційні витрати, завищуючи тарифи для усіх споживачів та ускладнюючи управління ризиками для страховиків.

Дослідження виявило, що страхове шахрайство зростає не лише за кількістю, але й за складністю. Оскільки шахрайські схеми стають все більш витонченими, традиційні методи виявлення часто виявляються недостатніми для того, щоб виявити кожен такий випадок.

Було розглянуто існуючі методи виявлення шахрайства, що використовуються у страховій галузі, включаючи ручні перевірки, базові статистичні методи та новітні досягнення на основі машинного навчання. Стало зрозуміло, що методи машинного навчання пропонують суттєве покращення як точності, так і ефективності, що підкреслює необхідність продовження досліджень та впровадження цих технологій.

Для практичної частини мого дослідження було використано реальний набір даних про заявки на страхову компенсацію, ретельно підготовлений шляхом попередньої обробки, очищення, відбору ознак і балансування даних за допомогою SMOTE. Ця підготовка дозволила наступним етапам моделювання отримати надійні та змістовні результати.

Використовуючи цей підготовлений набір даних, було побудовано кілька моделей прогнозування – дерева рішень, випадковий ліс, AdaBoost, а також ансамбль стекування, що поєднує прогнози AdaBoost і випадкового лісу за допомогою невеликого оптимізованого випадкового лісу в якості метакласифікатора. Ансамбль стекування забезпечив дещо кращі результати класифікації в цілому, зокрема покращивши збалансовану точність і чутливість, які мають вирішальне значення для правильної ідентифікації реальних випадків шахрайства.

Після систематичного порівняння продуктивності цих моделей було виявлено, що гібридні моделі, особливо методи стекування з ретельно налаштованими параметрами, є найбільш перспективними для практичного застосування. Страхові компанії можуть використовувати ці результати, впроваджуючи подібні гібридні підходи для ефективного виявлення шахрайства, тим самим заощаджуючи ресурси та підвищуючи операційну ефективність. Включення нових потоків даних, таких як телематика в реальному часі, публікації з соціальних мереж або текстовий аналіз описів заявок збільшить кількість атрибутів, на яких тренується модель.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Linn, J., & Shen, C. (2021). The effect of income on vehicle demand: Evidence from China's new vehicle market. Resources for the Future. https://media.rff.org/documents/VehicleDemandWorkingPaper.pdf
2. Huang, L. (2024). Quantitative analysis of private car ownership. Financial Engineering and Risk Management. Claius Scientific Press. https://www.clausiuspress.com/assets/default/article/2024/09/30/article_1727708274.pdf
3. Insurance Information Institute. (2016). More accidents, larger claims drive costs higher. https://www.iii.org/sites/default/files/docs/pdf/auto_rates_wp_092716-62.pdf
4. Helm, R. E. (1968). Motor vehicle liability insurance: A brief history. St. John's Law Review. https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=3541&context=lawreview
5. MoneyGeek. (2025). Car Insurance Coverage: Types and How It Works. https://www.moneygeek.com/insurance/auto/coverage/
6. EBSCO. Vehicle Insurance: Research Starter. https://www.ebsco.com/research-starters/law/vehicle-insurance
7. IMARC Group. (2024). Motor Insurance Market: Global Industry Trends, Share, Size, Growth, Opportunity, and Forecast 2024-2029. https://www.imarcgroup.com/motor-insurance-market
8. Hakobyan, G. M. (2024). Analysis of international experience in the introduction and development of mandatory insurance types. Región y Mundo. https://geopolitika.am/dir/wp-content/blogs.dir/1/files/2024_1_123_140.pdf
9. Галущак В.В. (2024). Роль і місце страхування в економічному зростанні країни. Причорноморські економічні студії. http://bses.in.ua/journals/2024/90_2024/15.pdf

10. Weisbart, S. (2018). How insurance drives economic growth. Insurance Information Institute. https://www.iii.org/sites/default/files/docs/pdf/insurance-driver-econ-growth-053018.pdf
11. Zubaidah, M., Ristiawan, B., & Sari, D. (2024). The effect of digital marketing on increasing interest in purchasing motor vehicle policies in Generation Z. Journal of Marketing and Digital Strategies. https://jurnal.unpal.ac.id/index.php/jm/article/view/1111
12. Н. Приказюк, Т. Моташко (2020). Автострахування: сучасні тренди. http://publications.lnu.edu.ua/collections/index.php/economics/article/view/3025/3252
13. Suruchanu, A., Kraus, N., & Kraus, K. (2023). Innovative-digital development of the national motor insurance market in the context of convergence with the European insurance market. European Academic Journal. https://www.journal.eae.com.ua/index.php/journal/article/view/229
14. Kester, J. (2022). Insuring future automobility: A qualitative discussion of British and Dutch car insurers' responses to connected and automated vehicles. Journal of Risk and Insurance. https://www.sciencedirect.com/science/article/pii/S2210539522001249
15. Lin, X.; Lee, C.-Y.; Fan, C.K. (2025). Exploring the Impacts of Autonomous Vehicles on the Insurance Industry and Strategies for Adaptation. World Electric Vehicle Journal. https://www.proquest.com/openview/78b935b2603d34f0edbf5a1e9b95527b/1?cbl=5046847&pq-origsite=gscholar
16. Višća, N. (2024). Multidisciplinary approach and complexity of proving fraud in insurance. https://doi.ub.kg.ac.rs/doi/zbornici/10-46793-xxiiivestacenje-241v/
17. Lincoln, R.; Wells, H.; Petherick, W. (2003). An Exploration of Automobile Insurance Fraud. Humanities & Social Sciences papers. https://d1wqtxts1xzle7.cloudfront.net/44655683/An_Exploration_of_Automobile_Insurance_F20160412-14111-im0yw9-libre.pdf?1460463277=&response-

content-disposition=inline%3B+filename%3DAn_Exploration_of_Automobile_Insurance_F.pdf

18. Bishop, L. (2024, September 19). Insurance fraud statistics. ValuePenguin. https://www.valuepenguin.com/auto-home-insurance-fraud
19. Vehicle insurance market size, share, and trends 2025 to 2034. Precedence Research (2025). https://www.precedenceresearch.com/vehicle-insurance-market
20. Insurance fraud: Impacts on premiums, claim costs, and the public. American Academy of Actuaries. (2020). https://www.actuary.org/sites/default/files/2024-09/casualty-brief-insurance-fraud.pdf
21. Edwards, N., Needham, R., & Davis, R. (2024). The advent of autonomous vehicles may increase fraud risk. Kennedys Law. https://kennedyslaw.com/en/thought-leadership/article/2024/the-advent-of-autonomous-vehicles-may-increase-fraud-risk/
22. Ribeiro, R.; Silva, B.; Pimenta, C.; Poeschl G. (2019). Why do consumers perpetrate fraudulent behaviors in insurance? Springer Nature. https://www.fep.up.pt/docentes/cpimenta/textos/pdf/Ribeiro_et_al-2019-Crime,_Law_and_Social_Change.pdf
23. Tawtf. (2024). The psychology of insurance fraud: Understanding why it happens. LinkedIn. https://www.linkedin.com/pulse/psychology-insurance-fraud-understanding-why-happens-tawtf/
24. Macedo, AM, Viana Cardoso, C, Marques Neto, JS & Amaral da Costa Brás da Cunha, C. (2021). Car insurance fraud: The role of vehicle repair workshops. International Journal of Law, Crime and Justice. https://www.sciencedirect.com/science/article/abs/pii/S175606162100001X?via%3Dihub
25. Benedek, B., & Nagy, B. Z. (2023). Traditional versus AI-based fraud detection: Cost efficiency in the field of automobile insurance. Financial and Economic Review. https://hitelintezetiszemle.mnb.hu/en/fer-22-2-st3-benedek-nagy

26. Nordin, S.-Z. S., Wah, Y. B., Haur, N. K., Hashim, A., Rambeli, N., & Abdul Jalil, N. (2024). Predicting automobile insurance fraud using classical and machine learning models. International Journal of Electrical and Computer Engineering. https://www.researchgate.net/publication/377878599_Predicting_automobile_insurance_fraud_using_classical_and_machine_learning_models
27. Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering and Technology. https://search.informit.org/doi/abs/10.3316/informit.263147785515876
28. Boost Your Models with AdaBoost Explained. DigitalOcean (2025). https://www.digitalocean.com/community/tutorials/adaboost-optimizer
29. What are support vector machines (SVMs)? IBM (2023). https://www.ibm.com/think/topics/support-vector-machine
30. What is random forest? IBM. https://www.ibm.com/think/topics/random-forest
31. Gepp, A., Wilson, J. H., Kumar, K., & Bhattacharya, S. (2012). A comparative analysis of decision trees vis-a-vis other computational data mining techniques in automotive insurance fraud detection. Journal of Data Science. https://pure.bond.edu.au/ws/portalfiles/portal/12803749/Gepp_a_comparative_analysis_of_decision_trees_paper.pdf
32. Bhowmik, R. (2011). Detecting Auto Insurance Fraud by Data Mining Techniques. Journal of Emerging Trends in Computing and Information Sciences. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=45e6de6fbff3ccf40e93941c72f12e14bf8e5747
33. Goleiji, L., Tarokh M. J. (2015). Fraud Detection in the Insurance using Decision Tree, Naïve Bayesian and Support Vector Machine Data Mining Algorithms (Case Study- Automobile's Body Insurance). Majlesi Journal of Multimedia Processing.

https://web.archive.org/web/20180411053445id_/http://journals.iaumajlesi.ac.ir/mp/index/index.php/mjmm/article/viewFile/210/104

34. Gondalia, D., Gurav, O., Joshi, A., Joshi, A., & Selvan, S. (2022). Automobile insurance claim fraud detection using Random Forest and ADASYN. International Research Journal of Engineering and Technology (IRJET). https://www.irjet.net/archives/V9/i5/IRJET-V9I523.pdf
35. Bangchang, K. N., Wongsai, S., & Simmachan, T. (2023). Application of data mining techniques in automobile insurance fraud detection. https://dl.acm.org/doi/pdf/10.1145/3613347.3613355
36. Uddin, M.; Ansari, M.F.; Adil, M.; Chakrabortty, R.K.; Ryan, M.J. (2023). Modeling Vehicle Insurance Adoption by Automobile Owners: A Hybrid Random Forest Classifier Approach. Processes. https://www.mdpi.com/2227-9717/11/2/629
37. Number of Registered Vehicles. CEIC. https://www.ceicdata.com/en/indicator/number-of-registered-vehicles
38. Мінфін. https://minfin.com.ua/ua/2024/03/20/123560792/
39. The true cost of auto insurance in 2025. Bankrate. https://www.bankrate.com/insurance/car/the-true-cost-of-auto-insurance/
40. Cost of Living in Germany – Updated for 2025. Study in Germany. https://www.studying-in-germany.org/cost-of-living-in-germany/
41. Average annual household expenditure on voluntary automobile insurance premiums per household in Japan from 2014 to 2023. Statista. https://www.statista.com/statistics/1290038/japan-annual-household-expenses-voluntary-automobile-insurance-premiums/
42. Cost of compulsory car insurance has almost doubled in Ukraine. Ukraine open for business. https://open4business.com.ua/en/cost-of-compulsory-car-insurance-has-almost-doubled-in-ukraine/
43. NAIC Releases 2023 Market Share Data. NAIC. https://content.naic.org/article/naic-releases-2023-market-share-data

44. German Car Insurance Market Size & Share Analysis - Growth Trends & Forecasts (2025 - 2030). Mordor Intelligence. https://www.mordorintelligence.com/industry-reports/germany-motor-insurance-market
45. Japanese Motor Insurance Market Size & Share Analysis - Growth Trends & Forecasts (2025 - 2030). Mordor Intelligence. https://www.mordorintelligence.com/industry-reports/japan-motor-insurance-market
46. STATISTICS: UKRAINE, FY2023: motor insurance remains the leader among insurance classes in terms of premiums. Xprimm.com. https://www.xprimm.com/STATISTICS-UKRAINE-FY2023-motor-insurance-remains-the-leader-among-insurance-classes-in-terms-of-premiums-articol-34-21643.htm
47. Shivam Bansal. Vehicle Insurance Claim Fraud Detection [Data set]. Kaggle. https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection/data
48. One Hot Encoding in Machine Learning. Geeksforgeeks (2025). https://www.geeksforgeeks.org/ml-one-hot-encoding/#
49. What is Scikit-Learn (Sklearn)? IBM (2025). https://www.ibm.com/think/topics/scikit-learn
50. SMOTE for Imbalanced Classification with Python. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/
51. Performing Feature Selection with gridsearchcv in Sklearn. Geeksforgeeks (2024). https://www.geeksforgeeks.org/performing-feature-selection-with-gridsearchcv-in-sklearn/

Календарний план виконання кваліфікаційної роботи бакалавра

№	Етапи роботи	Терміни виконання	Відмітка керівника про виконання
1	Вибір та затвердження теми кваліфікаційної роботи	До 20 грудня	Виконано вчасно
2	Формулювання завдань та структури кваліфікаційної роботи	До 10 січня	Виконано вчасно
3	Огляд літератури та аналіз існуючих методів виявлення страхового шахрайства	До 1 лютого	Виконано вчасно
4	Попередня обробка бази даних	До 20 лютого	Виконано вчасно
5	Розробка методології дослідження: вибір алгоритмів машинного навчання та критеріїв оцінки	До 5 березня	Виконано вчасно
6	Реалізація моделей (дерева рішень, випадкові ліси, AdaBoost)	До 25 березня	Виконано вчасно
7	Побудова та налаштування гібридної стекінгової моделі	До 10 квітня	Виконано вчасно
8	Аналіз та інтерпретація отриманих результатів: визначення найкращої моделі та її обмежень	До 20 квітня	Виконано вчасно
9	Формування висновків	До 1 травня	Виконано вчасно
10	Оформлення основної частини кваліфікаційної роботи відповідно до вимог	До 20 травня	Виконано вчасно
11	Подання роботи до попереднього захисту	До 1 червня	Виконано вчасно
12	Внесення правок за результатами попереднього захисту	До 12 червня	Виконано вчасно
13	Редагування, перевірка на плагіат, підготовка до захисту	До 17 червня	Виконано вчасно
14	Захист роботи	25 червня	Виконано вчасно

Науковий керівник: Чорноус Галина Олександрівна

Студентка: Раковська Анастасія Андріївна