Міністерство освіти і науки України
Вінницький національний технічний університет

ПОЛЬГУЛЬ ТЕТЯНА ДМИТРІВНА

УДК 004.8:044.89

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ ШАХРАЙСТВА ПРИ ІНСТАЛЮВАННІ МОБІЛЬНИХ ДОДАТКІВ З ВИКОРИСТАННЯМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

05.13.06 – інформаційні технології

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата технічних наук

Вінниця – 2020

Дисертацією є кваліфікаційна наукова праця на правах рукопису.

Робота виконана у Вінницькому національному технічному університеті, Міністерство освіти і науки України.

Науковий керівник: доктор технічних наук, професор,
Яровий Андрій Анатолійович,
Вінницький національний технічний університет,
завідувач кафедри комп'ютерних наук

Офіційні опоненти: доктор технічних наук, професор,
Поворознюк Анатолій Іванович,
Національний технічний університет «Харківський політехнічний інститут», професор кафедри обчислювальної техніки та програмування

доктор технічних наук, професор,
Цмоць Іван Григорович,
Національний університет «Львівська політехніка», професор кафедри автоматизованих систем управління

Захист відбудеться «03» квітня 2020 р. о $10^{\underline{00}}$ годині на засіданні спеціалізованої вченої ради Д 05.052.01 у Вінницькому національному технічному університеті за адресою: 21021, Україна, м. Вінниця, вул. Хмельницьке шосе, 95, ГНК, ауд. 210.

З дисертацією можна ознайомитись у бібліотеці Вінницького національного технічного університету за адресою: 21021, Україна, м. Вінниця, вул. Хмельницьке шосе, 95, ГНК.

Автореферат розісланий «28» лютого 2020 р.

Вчений секретар
спеціалізованої вченої ради С. М. Захарченко

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Обґрунтування вибору теми дослідження. У зв'язку з появою на ринку значної кількості мобільних додатків, якими користуються мільярди користувачів, компанії-розробники мобільних додатків звертаються за послугами маркетингових кампаній з метою залучення інсталювань саме до їхнього додатку. Саме така потреба у маркетингових кампаніях стала однією з причин появи шахраїв та їх шахрайських способів інсталювання мобільних додатків. Шахраї, у свою чергу, приводять до компаній-розробників необхідну кількість фейкових (несправжніх) «користувачів» та отримують за це відповідну грошову винагороду. Проте такі «користувачі» ніколи не повертаються у мобільний додаток, оскільки є фейковими, ми ж їх називатимемо шахрайськими.

У наш час вже існують такі відомі види шахрайства при інсталюванні додатків, як мобільне викрадення (mobile hijacking), кліковий спам (click spamming), ферми дій (action farms), а також методи та системи виявлення шахрайства при інсталюванні мобільних додатків такі, як Fraudlogix та Kraken, Adjust, Kochava та TCM Attribution Analytics, Protect360 від Appsflyer, FraudScore та AppMetrica. Проте необхідно зазначити, що лише останні дві використовують інтелектуальну складову, причому система AppMetrica просто ґрунтується на алгоритмах та API системи FraudScore. Але вказані системи-аналоги виконують рейтингування користувачів на основі не всіх, а вибіркових вхідних даних, тому відбувається упущення шахраїв системами. Інші вказані системи використовують існуючі бази з шахрайськими даними (наприклад, IP-адресами), що також призводить до упущення шахраїв, які мають інші властивості, шаблони, поведінку.

Очевидно, що причиною вищевказаних недоліків систем є відсутність єдиного підходу до виявлення шахрайства на основі всіх наявних даних. Також, недоліком існуючих систем є те, що вони розпізнають лише відомі види шахрайства і не можуть розпізнавати нові шахрайські шаблони. А в сучасному світі важливою є можливість системи адаптуватись, тому необхідним є створення відповідних інформаційних технологій, що матимуть змогу самонавчатися.

Все вищенаведене є передумовою актуальності створення інформаційної технології виявлення шахрайства при інсталюванні мобільних додатків, яка б відстежувала та визначала шаблони шахраїв, які непомітні людині. Розв'язанню цієї задачі присвячена дана робота.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційне дослідження проводилось згідно з планами науково-дослідних робіт кафедри комп'ютерних наук Вінницького національного технічного університету, в тому числі в межах: наукового проекту Ф61/199-20150/4711 "Методологія побудови високопродуктивних інтелектуалізованих паралельно-ієрархічних систем на основі сучасних мережевих обчислювальних комплексів з гетерогенною архітектурою" (№ державної реєстрації: 0115U001975, 2015 р.), при виконанні якого автор брала участь як виконавець окремих підрозділів; кафедральної теми 47 К2 "Моделі, методи, технології та пристрої інтелектуальних інформаційних систем управління, економіки, навчання та комунікацій" (2016-2018 р.), при виконанні якої автор брала участь як відповідальний виконавець; кафедральної теми 22 К1 "Розробка спеціалізованих засобів штучного інтелекту на основі інтелектуального

аналізу даних та машинного навчання" (2019 р.), при виконанні якої автор бере участь як виконавець окремих підрозділів.

Мета і завдання дослідження. Метою дисертаційного дослідження є підвищення точності та швидкодії процесів виявлення шахрайства при інсталюванні мобільних додатків.

Основними задачами дослідження є:

– аналіз методів та постановка задачі виявлення шахрайства при інсталюванні мобільних додатків;

– формалізація процесу виявлення шахрайства як аномалії в даних;

– аналіз та класифікація різнорідних даних при інсталюванні мобільних додатків;

– розробка узагальненого методу виявлення шахрайства при інсталюванні мобільних додатків;

– розробка методу подолання різнорідності вхідних даних;

– розробка інформаційної технології виявлення шахрайства при інсталюванні мобільних додатків.

Об’єкт дослідження – процеси виявлення шахрайства при інсталюванні мобільних додатків.

Предмет дослідження – моделі, методи та інформаційні технології виявлення шахрайства як аномалій в даних при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних.

Методи дослідження, що використані в роботі: методи шкалювання під час вирішення задач аналізу та класифікації різнорідних даних при інсталюванні мобільних додатків та розробки методу подолання різнорідності вхідних даних; теорія множин для вирішення задачі формалізації процесу виявлення шахрайства як аномалії в даних, а також методи класифікації, статистичні методи, методи машинного навчання, інтелектуальний аналіз даних, методи кластеризації, нейромережеві методи для вирішення задач розробки узагальненого методу виявлення шахрайства при інсталюванні мобільних додатків та розробки інформаційної технології виявлення шахрайства при інсталюванні мобільних додатків.

Наукова новизна отриманих результатів. В ході розв’язання поставлених задач були отримані наукові результати.

1. Вперше запропоновано метод подолання різнорідності вхідних даних, що являє собою сукупність процедур вибору ознак, зниження розмірності та нормалізації даних, відмінність якого полягає у новій моделі процесу подолання різнорідності даних шляхом шкалювання за інформативністю, що дозволяє всю множину різнорідних даних про користувачів звести до вектору уніфікованих ознак без зменшення діагностичної цінності інформації.

2. Удосконалено модель класифікації користувачів на основі глибинних нейронних мереж у частині зниження розмірності та нормалізації даних згідно запропонованого методу подолання різнорідності даних, яка є основою для створення узагальненого портрету шахрая з метою спрощення процесів їх виявлення.

3. Вперше розроблено узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків, відмінність якого полягає у використанні

запропонованої моделі класифікації користувачів та методу подолання різнорідності вхідних даних, що дозволяє визначити класи користувачів та підвищити точність виявлення шахрайства при інсталюванні мобільних додатків.

Практичне значення отриманих результатів роботи полягає у наступному:

– здійснено класифікацію різнорідних даних, що дозволило спростити процес аналізу різнорідних за метриками, розмірностями і шаблонами даних та автоматизувати його;

– розроблено алгоритм пошуку аномалій в даних, алгоритми процесу подолання різнорідності вхідних даних, алгоритм виявлення шахрая при інсталюванні мобільних додатків, алгоритм створення узагальненого портрету шахрая та алгоритм мінімізації часу виявлення шахраїв на основі розпаралелення обчислювальних процесів, які покладені в основу інформаційної технології, що підвищило точність та швидкодію виявлення шахраїв;

– запропоновано інформаційну технологію виявлення шахрайства при інсталюванні мобільних додатків, яка використовує запропоновані метод виявлення шахрайства при інсталюванні мобільних додатків, метод подолання різнорідності вхідних даних, модель класифікації даних, і, на відміну від існуючих систем Antifraud, дозволила підвищити точність класифікації користувачів до 99,14 %, зокрема точність класифікації шахраїв – до 82,76 %;

– розроблено програмне забезпечення “Mobile App Install Fraud Detection System” для виявлення шахрайства при інсталюванні мобільних додатків.

Результати дисертаційної роботи впроваджені на підприємствах та навчальних закладах: Garuda AI B. V. (м. Схіпгол, Нідерланди) – інформаційна технологія; ТОВ «ВІН ІНТЕРАКТИВ» (м. Вінниця, Україна) – алгоритми подолання різнорідності даних, модель процесу подолання різнорідності даних; ТОВ «4ХайТек» (м. Вінниця, Україна) – модель процесу подолання різнорідності вхідних даних, методика виявлення шахрая при інсталюванні мобільних додатків; ПП «Літсофт» (м. Київ, Україна) – алгоритм подолання різнорідності даних, узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків, методика створення узагальненого портрету шахрая; у навчальний процес кафедри комп’ютерних наук Вінницького національного технічного університету (ВНТУ) – узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків та у навчальний процес кафедри інформатики, програмної інженерії та економічної кібернетики Херсонського державного університету – інформаційна технологія виявлення шахрайства при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних.

Особистий внесок здобувача. Усі результати, які складають основний зміст дисертації, отримані здобувачем самостійно. У роботах [3], [7], [11], [12], [19], [20] здобувачеві належать усі теоретичні та практичні результати. У роботах, опублікованих у співавторстві, здобувачу належать: [1] – розроблено метод виявлення шахрайства, математичну модель процесу шкалювання, алгоритм шкалювання різнорідних масивів, алгоритм обробки отриманих груп однорідних даних, схему процесу виявлення шахраїв, інтелектуальну систему автоматичного виявлення шахраїв, здійснено класифікацію різнорідних даних при інсталюванні мобільних додатків; [4] – здійснено формалізацію процесу виявлення шахрайства

як аномалії в даних, розроблено метод виявлення аномалій при інсталюванні мобільних додатків; [2], [5] – розроблено метод формування портрету шахрая та алгоритм розробки нечіткої моделі для формування портрету шахрая; [6] – запропоновано інтелектуальну інформаційну технологію виявлення шахрайства при інсталюванні мобільних додатків, удосконалено класифікацію користувачів з використанням глибинних нейронних мереж, створення узагальненого портрету шахрая; [8] – розроблено систему виявлення шахрайства при інсталюванні програмних додатків; [9] – запропоновано метод, моделі та алгоритми подолання різнорідності даних для виявлення шахрайства; [10] – розроблено метод та алгоритм аналізу різнорідних даних, математичну модель процесу аналізу різнорідних даних, схему експериментального дослідження виявлення аномалій в різнорідних даних; [13], [14] – розробка програмного забезпечення для модулю збору даних та для модулю визначення схожості користувачів, розробка алгоритму мінімізації часу виявлення шахраїв, алгоритму пошуку аномалій в даних; [15] – запропоновано модель класифікації користувачів; [16] – розроблено метод виявлення шахрайства при інсталюванні мобільних; [17] – запропоновано метод подолання різнорідності вхідних даних; [18] – запропоновано модель визначення шахрайських способів встановлення мобільних додатків.

Апробація матеріалів дисертації. Результати дисертаційної роботи доповідались та обговорювались на 9 науково-технічних конференціях: XLV, XLVI, XLVIII науково-технічних конференціях професорсько-викладацького складу, співробітників та студентів ВНТУ (2016, 2017, 2019 рр.); науково-практичній конференції «Сучасні тенденції розвитку системного програмування» (м. Київ, Національний авіаційний університет, 2016 р.); XIV Міжнародній конференції «Контроль і управління в складних системах (КУСС-2018)» (м. Вінниця, ВНТУ, 2018 р.); V Міжнародній науково-технічній конференції студентів, магістрів та аспірантів «Інформатика, управління та штучний інтелект» (м. Харків, Національний технічний університет «Харківський політехнічний інститут», 2018 р.); 5th International Winter School on Big Data BigDat2019 (м. Кембридж, University of Cambridge, Великобританія, 2019 р.); 577th International Conference on Innovative Engineering Technologies (ICIET) (м. Бангкок, Таїланд, 2019 р.); The 10th International Conference on Dependable Systems, Services and Technologies (DESSERT'2019) (м. Лідс, Великобританія, Leeds Beckett University, 2019 р.), де отримано нагороду «Best Paper Award»; The 14th International conference "Computer sciences and Information technologies" (CSIT 2019) (м. Львів, Україна, 2019 р.); The 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), (м. Метц, Франція, 2019 р.).

Публікації. За темою дисертації опубліковано 20 праць, в тому числі 5 статей надруковано у наукових виданнях, які входять до переліку фахових видань з технічних наук, затверджених МОН України (одна з яких також входить до наукометричної бази даних Scopus). Крім того, 6 статей опубліковано в міжнародних наукових виданнях, п'ять з яких входять до міжнародної наукометричної бази Scopus (три з яких також входять до міжнародної наукометричної бази IEEE Xplore), 7 робіт опубліковано у збірках матеріалів

конференцій (три з яких міжнародні), отримано 2 свідоцтва про реєстрацію авторського права на твір.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків. Основний зміст викладено на 163 сторінках друкованого тексту, містить 63 рисунки, 17 таблиць. Список використаних джерел містить 107 найменувань. Загальний обсяг 245 сторінок.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі наведено загальну характеристику роботи, обґрунтовано актуальність теми, зазначено її зв'язок з науковими програмами, планами та темами, сформульовано мету та напрям досліджень. Визначено основні задачі досліджень, наукову новизну та практичне значення основних результатів, а також відомості про їх впровадження, апробацію та публікації.

У першому розділі на основі огляду та аналізу сучасної літератури з галузі виявлення шахрайства в інформаційних технологіях, в роботі запропоновано розглядати задачу виявлення шахрайства при інсталюванні мобільних додатків як задачу пошуку аномалій в даних, тому що шахрайство визначено як навмисне породження аномалії в даних сторонньою особою (шахраєм) або механізмом з певною метою.

Розглянуто та проаналізовано методи і здійснено постановку задачі виявлення шахрайства при інсталюванні мобільних додатків. Проведено варіантний аналіз існуючих методів пошуку аномалій в даних, усі розглянуті методи можна розділити на методи класифікації, методи кластеризації, статистичні методи, мікс методів на прикладі глибинного навчання. Здійснено аналіз моделей подібності. Показано, що жоден із зазначених методів не може одночасно здійснювати інтелектуальну обробку повної вхідної інформації та вказувати причину, яка дає змогу позначити дані як аномальні.

Виділено такі відомі шахрайські способи під час інсталювання мобільних додатків як кліковий спам (Click Spamming), мобільне викрадення (Mobile Hijacking), ферми дій (Action Farms). Проаналізовано існуючі системи виявлення шахрайства при інсталюванні мобільних додатків, серед яких: система Fraudlogix, Kraken, Adjust, Kochava, TMC Attribution Analytics, FraudShield, Forensiq, Appsflyer, FraudScore, AppMetrica. Проте зазначено, що більшість з них видають результат у вигляді таблиці рейтингування, а не чіткої відповіді; опираються на бази з відомими шахрайськими даними (наприклад, з IP-адресами) та не в повній мірі використовують усі важливі вхідні дані. У зв'язку з цим, не розпізнаються шахраї, які мають інші властивості, шаблони, поведінку. Також, існує проблема неможливості визначення причини класифікації користувача як шахрайського існуючими системами. Тому й виникла необхідність створення інформаційної технології виявлення шахрайства при інсталюванні мобільних додатків, яка могла б відстежувати та визначати навіть нові шаблони шахраїв і мала змогу самонавчатися.

На основі проведеного аналізу визначено задачі дослідження.

У другому розділі формалізовано процес виявлення шахрайства як аномалії в даних з використанням теорії множин, що дозволило визначити властивості

аномальних і неаномальних даних. Це дало змогу визначити властивості даних, які позначають шахраїв, у визначеній предметній області та спростити процес пошуку шахрайства при інсталюванні мобільних додатків.

За рахунок здійсненої формалізації, множина аномальних даних Z розглядається у роботі як група (множина) даних $\langle a_1, a_2, ..., a_q \rangle$, яка входить у множину вхідних даних $A = \langle a_1, a_2, ..., a_n \rangle$, що характеризується множиною властивостей $P_1(a), P_2(a) ..., P_s(a)$, але виходять за задані межі властивостей $O_1(a), O_2(a), ..., O_t(a)$, що притаманні неаномальним даним X множини вхідних даних A, де $X \subseteq A, Z \subseteq A$ та, згідно теореми, $X \cap Z = \varnothing$. У задачі виявлення аномалій у вхідних наборах даних з мобільних додатків, аномальними будемо вважати ті дані (елементи множини), які: не мають властивостей $O_1(a), O_2(a), ..., O_t(a)$, що визначають множину неаномальних даних X; не співпадають по розмірності; не співпадають по властивостям групи даних; не входять в область гранично допустимих значень множини неаномальних даних X. Один з прикладів аномалій – а саме – приклад аномалії, що зображує наявність у вхідному наборі А групи елементів, які не співпадають по властивостям групи даних, подано на рисунку 1.

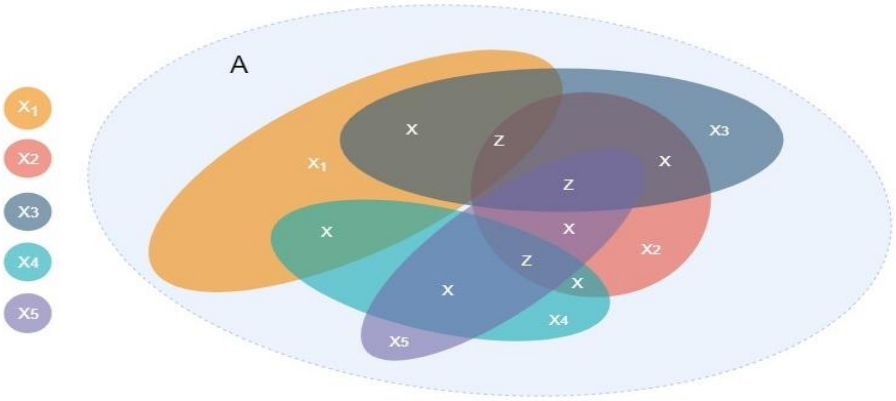

Рисунок 1 – Діаграма Венна, де всі елементи підмножин X_1, X_2, X_3, X_4, X_5 є неаномальними, Z – множина аномальних даних, А – вхідний набір даних

Завдяки здійсненій у роботі формалізації процесу виявлення шахрайства як аномалії в даних, здійснено процес вибору характеристик даних, які дозволяють визначити клас користувача – шахрай чи органічний (не шахрай). Для визначення таких характеристик у роботі здійснено аналіз та класифікацію даних при інсталюванні мобільних додатків, які показали, що дані в таких системах – різнорідні (рис. 2).

У процесі дослідження було показано складність аналізу таких вхідних даних. Саме через таку складність, існуючі системи часто відкидають деякі з таких даних через неможливість їх обробки. Але чим більша кількість вхідних даних, тим більше залежностей та кореляцій можна з них виділити, більше різних шаблонів шахраїв можна буде виявити. Так як більшість систем класифікацій не працюють з такою кількістю різнотипних і різномірних даних, то при дослідженні постало питання як правильно та ефективно використовувати усі різнорідні дані про користувача.

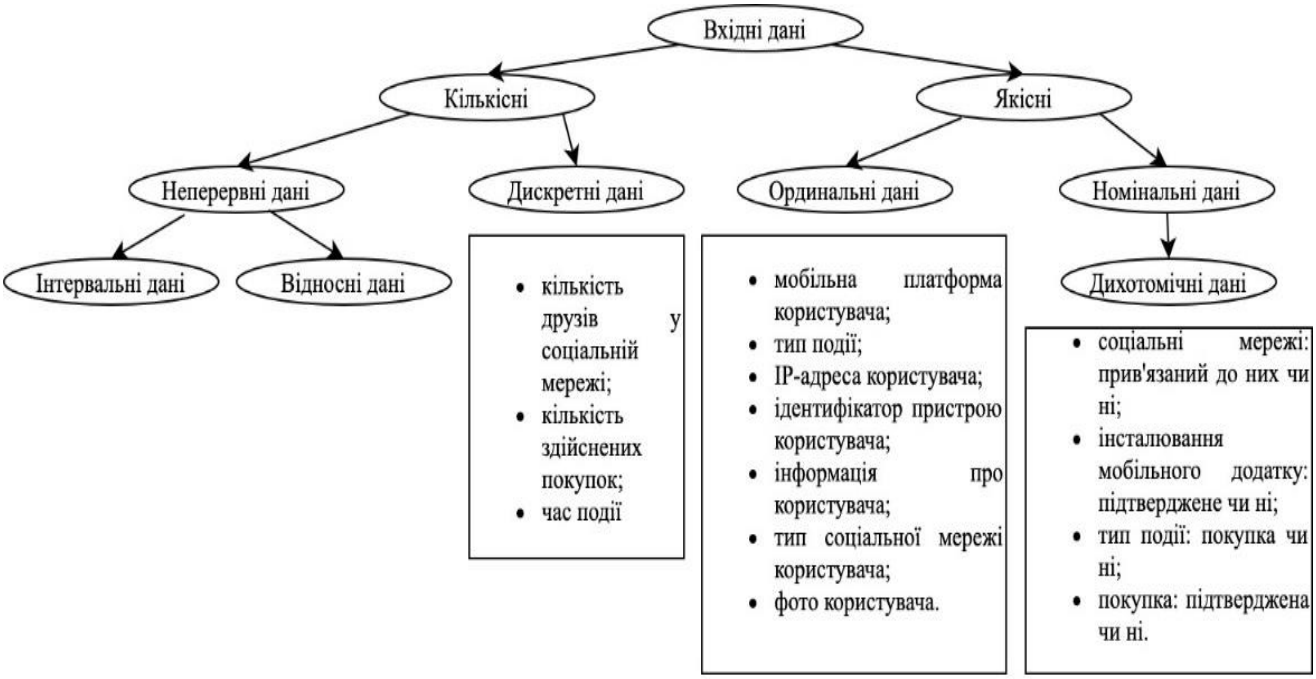

Рисунок 2 – Класифікація різнорідних вхідних даних в системах виявлення шахрайства при інсталюванні мобільних додатків

Для вирішення поставленої цілі та для обробки повного масиву вхідних даних та характеристик цих даних, за допомогою яких можна визначити, чи користувач є шахраєм, чи ні, в роботі було проведено експертне опитування, в якому прийняли участь 13 експертів провідних IT-компаній України, Швейцарії, США, Нідерландів, які мають досвід виявлення шахрайства. За допомогою експертного опитування було виділено характеристики даних, які було розкласифіковано на 2 групи, а саме: характеристики, за якими однозначно можна визначити клас користувача та характеристики, за якими неможливо однозначно визначити клас користувача.

Здійснений аналіз та класифікація даних дозволили розробити узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків, який дає змогу визначити класи користувачів та підвищити точність виявлення шахрайства під час інсталювання мобільних додатків за допомогою вирішення наступних етапів (рис. 3): виявлення характеристик даних користувача; подолання різнорідності даних; побудова моделі класифікації; класифікація; формування бази знань для виявлення шахрайства; формування бази даних шахраїв; інтелектуальний аналіз наявних даних та формування портрету користувача; створення узагальненого портрету шахрая та прогнозування. Таким чином, здійснена класифікація різнорідних даних при інсталюванні мобільних додатків на основі процесу вибору ознак дала змогу спростити процес аналізу різнорідних за метриками, розмірностями і шаблонами даних та автоматизувати його.

Для вирішення задачі подолання різнорідності даних запропоновано метод подолання різнорідності вхідних даних, що являє собою сукупність процедур вибору ознак, зниження розмірності та нормалізації даних. Відмінність цього методу полягає у новій моделі процесу подолання різнорідності даних шляхом шкалювання за інформативністю, що дозволяє всю множину різнорідних даних про користувачів звести до вектору уніфікованих ознак без зменшення

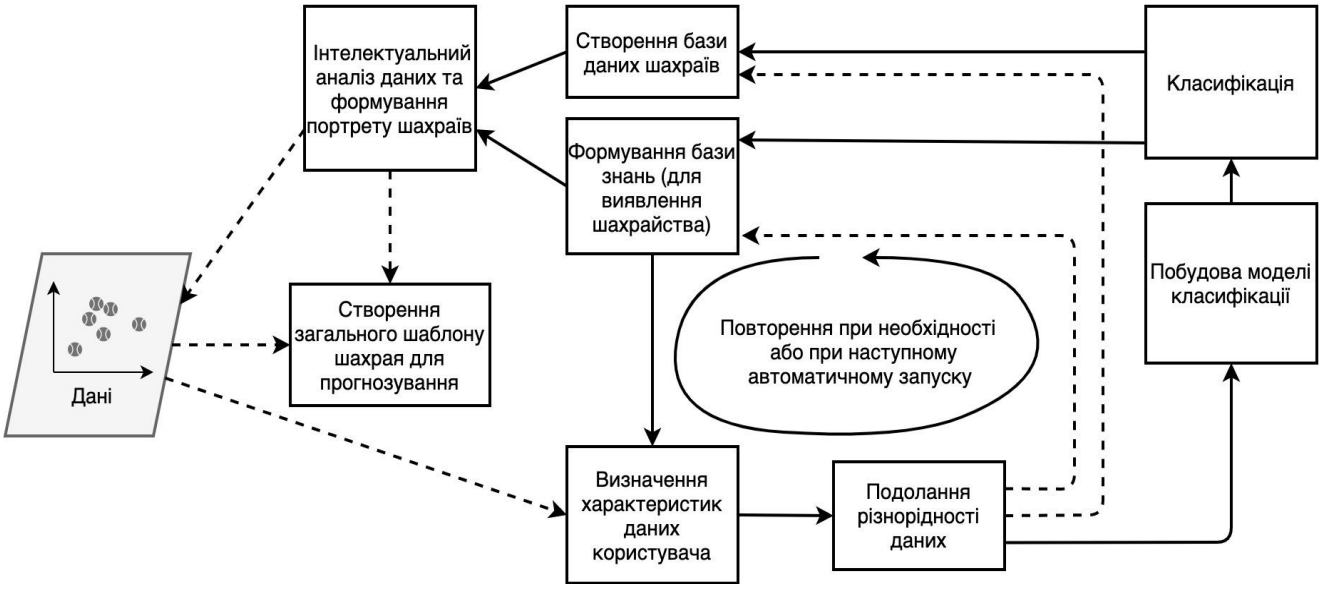

Рисунок 3 – Етапи узагальненого методу виявлення шахрайства при інсталюванні мобільних додатків

діагностичної цінності інформації. При шкалюванні вхідних даних, відповідно запропонованих в роботі шкал, отримано математичну модель процесу шкалювання (1), яка містить коефіцієнти визначеної метрики та має вигляд:

$$\vec{I}\begin{pmatrix} U_1(i_1,i_2,...,i_{s1}) \\ U_2(i_1,i_2,...,i_{s2}) \\ ... \\ U_n(i_1,i_2,...,i_{sn}) \end{pmatrix} \rightarrow$$

$$\rightarrow \begin{cases} \vec{G}_1\begin{pmatrix} U_1(g_{11},...,g_{1r}) \\ U_2(g_{11},...,g_{1r}) \\ ... \\ U_n(g_{11},...,g_{1r}) \end{pmatrix} \rightarrow \begin{cases} \vec{X}\begin{pmatrix} U_1(x_1,...,x_n) \\ U_2(x_1,...,x_n) \\ ... \\ U_n(x_1,...,x_n) \end{pmatrix} \rightarrow F_1(\vec{X}) \rightarrow \vec{X}_1\begin{pmatrix} U_n(k_{01}) \\ U_n(k_{02}) \\ ... \\ U_n(k_{0n}) \end{pmatrix} \\ \vec{B}\begin{pmatrix} U_1(b_1,...,b_k) \\ U_2(b_1,...,b_k) \\ ... \\ U_n(b_1,...,b_k) \end{pmatrix} \rightarrow F_2(\vec{B}) \rightarrow \vec{B}_1\begin{pmatrix} U_n(k_{11}) \\ U_n(k_{12}) \\ ... \\ U_n(k_{1n}) \end{pmatrix} \rightarrow \\ \vec{W}\begin{pmatrix} U_1(w_1,...,w_m) \\ U_2(w_1,...,w_m) \\ ... \\ U_n(w_1,...,w_m) \end{pmatrix} \rightarrow F_3(\vec{W}) \rightarrow \vec{W}_1\begin{pmatrix} U_n(k_{21}) \\ U_n(k_{22}) \\ ... \\ U_n(k_{2n}) \end{pmatrix} \end{cases} \\ \vec{G}_2\begin{pmatrix} U_1(g_{21},...,g_{2l}) \\ U_2(g_{21},...,g_{2l}) \\ ... \\ U_n(g_{21},...,g_{2l}) \end{pmatrix} \rightarrow \end{cases} \quad (1)$$

$$\rightarrow F_4(\vec{X}_1,\vec{B}_1,\vec{W}_1) \rightarrow \vec{D}\begin{pmatrix} U_1(k_{01},k_{11},...,k_{21}) \\ U_2(k_{02},k_{12},...,k_{22}) \\ ... \\ U_n(k_{0n},k_{1n},...,k_{2n}) \end{pmatrix} \rightarrow F_5(\vec{D}) \rightarrow \vec{R}\begin{pmatrix} U_1(C_1) \\ U_2(C_0) \\ ... \\ U_n(C_1) \end{pmatrix},$$

де $\vec{I}\begin{pmatrix} U_1(i_1,i_2,...,i_{s1}) \\ U_2(i_1,i_2,...,i_{s2}) \\ ... \\ U_n(i_1,i_2,...,i_{sn}) \end{pmatrix}$ – інформація з бази даних по кожному з користувачів, а саме вектор, елементами якого є множини усіх визначених ознак по кожному з користувачів $(U_1,U_2,...,U_n)$; $\vec{G}_1$ та $\vec{G}_2$ – вектори різнорідних вхідних даних, поділені на дві групи, а саме так, як визначено на основі експертного опитування; $\vec{X},\vec{B},...,\vec{W}$ – вектори однорідних даних, поділених по типам; $F_1(\vec{X}),F_2(\vec{B}),...,F_3(\vec{W})$ – відповідні функції переведення однорідних даних за певною ознакою у критерій, значення якого від 0 до 1. На виході буде отримано вектори $\vec{X}_1,\vec{B}_1,\vec{W}_1$, що міститимуть дані про користувачів із значенням критерію по відповідній ознаці; $F_4(\vec{X}_1,\vec{B}_1,...,\vec{W}_1)$ – функція об'єднання усіх критеріїв по користувачам у вектор $\vec{D}$; $\vec{D}$ – вектор уніфікованих ознак без зменшення діагностичної цінності інформації; $F_5(\vec{D})$ – функція класифікації користувачів на кластери C_0 (шахраї) та C_1 (органічні користувачі). Результат класифікації буде представлено у вигляді вектору користувачів $\vec{R}$, кожен з користувачів якого міститиме у якості параметру відповідний клас.

На основі запропонованої математичної моделі шкалювання (1) розроблено три алгоритми процесу подолання різнорідності вхідних даних.

Алгоритм 1 подолання різнорідності вхідних даних при інсталюванні мобільних додатків:

1. Отримання даних $\bar{I}\begin{pmatrix} U_1(i_1,i_2,...,i_{s1}) \\ U_2(i_1,i_2,...,i_{s2}) \\ ... \\ U_n(i_1,i_2,...,i_{sn}) \end{pmatrix}$.

2. Визначення типу даних.

3. Перетворення даних, залежно від їх типу, алгоритмами 2 та 3 з метою зведення їх до однорідних даних.

Для виконання пункту 3 алгоритму 1, запропоновано алгоритм 2 шкалювання даних, який дає можливість переходу від різнорідних даних до бінарних, значення яких можуть бути або 0, або 1, за допомогою запропонованих в роботі коефіцієнтів. Тут 0 означатиме, що користувач є шахраєм, а 1 позначатиме користувача як органічного. Бінарність даних після перетворення визначається кінцевою метою задачі – визначити користувач з певними даними є шахраєм чи ні. Тому в першу групу G_1 входять дані, результатом аналізу яких є бінарна відповідь «так» або «ні» (0 або 1). А ті дані, які неможливо привести до бінарного типу, групуються у другу групу даних G_2, для яких у даній роботі розроблений метод (на базі алгоритму 3) інтелектуального аналізу даних з використанням коефіцієнтів схожості та системи правил з сформованою базою знань. Якщо ж даний метод не визначає користувача ні як шахрая, ні як органічного, то даний користувач включається в групу підозрілих користувачів, для аналізу яких застосовуються

алгоритми нечіткої логіки на основі попередньо сформованих правил з бази знань, яка в процесі експлуатації постійно нарощується.

Слід зазначити, що не всі вхідні дані можна відразу однозначно шкалювати. Тому в роботі розроблений *алгоритм 2* шкалювання вхідних даних, який включає такі етапи:

1. Поділ усіх даних $\vec{I}$ на дві групи G_1 і G_2.

1.1. До першої групи G_1 входять дані, за якими однозначно можна визначити значення 0 або 1 за відповідним коефіцієнтом, де 0 визначає користувача як шахрая, а 1 визначає користувача органічним (шкала 1, 2, 3 рис. 4).

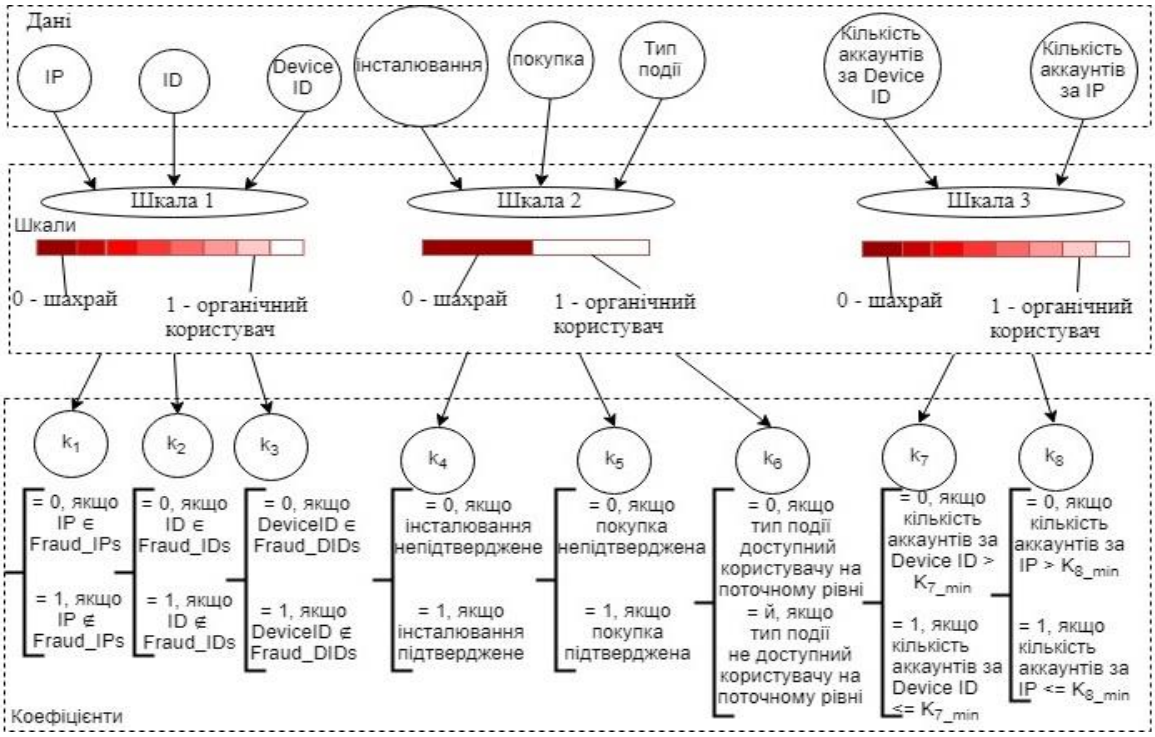

Рисунок 4 – Модель даних першої групи G_1

1.2. До другої групи G_2 входять дані, за якими неможливо однозначно визначити значення коефіцієнту. Так, наприклад, невідомі граничні значення часу між подіями, за допомогою яких можна однозначно визначити чи користувач є шахраєм, чи органічним користувачем. Один із коефіцієнтів визначається на основі типів подій, які виконує користувач. Даний коефіцієнт не може входити до першої групи G_1, оскільки по таким даним не існує чітко визначеної умови, яка не буде змінюватися з часом.

2. Визначення значень коефіцієнтів на основі даних першої групи G_1 та побудова моделі даних групи G_1 (рис. 4).

3. Визначення однозначних шахраїв, органічних та підозрілих користувачів на основі значень коефіцієнтів першої групи G_1, формування їх шаблонів та занесення їх у базу знань.

Результат шкалювання за першою групою представлено у вигляді таблиці, фрагмент якої подано у таблиці 1. На основі таких таблиць створюється база даних шахраїв.

Таблиця 1 – Приклад використання методу подолання різнорідності вхідних даних

Ознака	Коефіцієнт		Умова	Розпізнавання шахрайства
	До шкалювання	Після шкалювання		
IP-адреса	127.0.0.1	1	$IP \notin FRAUD\_deviceIP$	органічний користувач
ID пристрою	355776785678735	0	$deviceID \in FRAUD\_deviceID$	шахрай
ID покупки підтверджена чи ні	6cf9094a-6fbf-4898-bb3e-0cd895b1cafb	0	$purchaseID.isConfirmed = false$	шахрай

4. Визначення характеристик однозначно визначених користувачів, формування їх шаблонів та занесення їх у базу знань.

Для аналізу другої групи даних G_2 в роботі розроблено метод на основі інтелектуального аналізу даних, оснований на *алгоритмі 3* процесу подолання різнорідності вхідних даних, який включає такі етапи:

1. Визначення початкового узагальненого портрету шахрая.

2. Визначення значень коефіцієнтів другої групи даних G_2 на основі правил попередньо сформованої бази знань (шкали 4, 5 рис. 5):

2.1. Отримання шаблонів та характеристик шахраїв з бази даних, що була сформована у процесі аналізу групи G_1.

2.2. Знаходження коефіцієнту схожості між поточним користувачем та шаблоном і характеристиками шахрая (рис. 5).

2.3. Формування масиву значень коефіцієнтів схожості поточного користувача з кожним із однозначно визначених шахрайських користувачів G_{2_fraud} з бази знань.

2.4. Виконання кроків 2.1 – 2.3 для однозначно визначених органічних користувачів G_{2_org} з бази знань.

2.5. Інверсія масиву значень коефіцієнтів схожості поточного користувача з кожним із однозначно визначених органічних користувачів G_{2_org} з бази знань.

2.6. Формування спільного масиву коефіцієнтів $G_2 = G_{2\_fraud} \bigcup \left(1 - G_{2\_org}\right)$ за поточною ознакою на основі масивів, отриманих в кроках 2.3, 2.4, 2.5.

2.7. Формування вектору підозрілих користувачів G_{2_susp}.

3. Формування вектору відшкальованих даних по кожному користувачу за

коефіцієнтами, що об’єднує результати, отримані на кроці 2 алгоритму 2 та кроці 2.7 алгоритму 3. Сформований вектор є вектором уніфікованих ознак без зменшення діагностичної цінності інформації.

4. Віднесення підозрілих користувачів до класів «шахраї» або «органічні користувачі» з використанням нечіткої логіки.

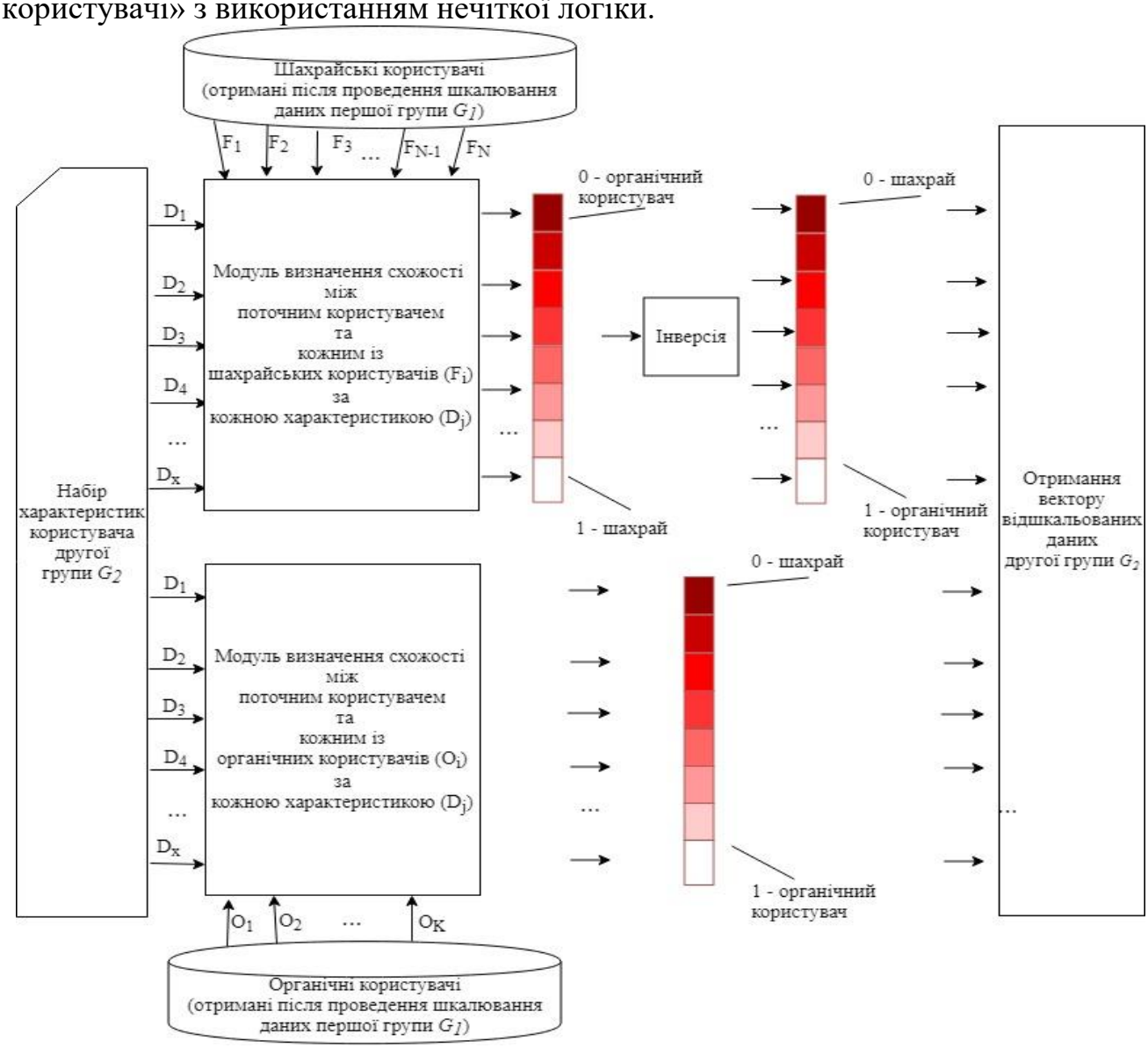

Рисунок 5 – Модель шкалювання даних другої групи G_2

В результаті виконання кроків 1 та 2 алгоритму 2, проведено шкалювання даних групи G_1 та за допомогою коефіцієнтів k_1, k_2, k_3, …, k_8 різнорідні дані приведено до однорідного стану, що дозволяє однозначно визначити шахраїв, якщо значення хоча б одного з коефіцієнтів дорівнює 0.

Алгоритм було використано в програмі, на яку отримано свідоцтво про реєстрацію авторських прав на твір (комп’ютерну програму) № 76347 від 26.01.2018 р.

Зазначимо, що запропоновані метод та алгоритми можна автоматизувати. Також, з кожним повторним використанням алгоритмів, знаходяться все нові характеристики органічних та шахрайських користувачів, які заноситимуться у базу знань. Це дасть змогу удосконалювати подальше виявлення шахрайства.

Як було зазначено, в процесі шкалювання та на основі результатів експертного оцінювання, в роботі було розроблено 17 шкал, частина яких представлена на рисунках 4, 5. Для конвертації наявних даних до значень на розроблених шкалах, у роботі було сформовано 17 коефіцієнтів, які впливають на прийняття рішень при виявленні шахрайства. Аналіз цих коефіцієнтів дозволив здійснити поділ коефіцієнтів на дві групи. Перша група охоплює коефіцієнти, які дозволяють провести попередній аналіз даних, а саме, однозначно з множини користувачів визначити шахрайських користувачів, органічних та підозрілих. А друга група охоплює коефіцієнти, за якими неможливо зробити первинний аналіз. Саме через наявність другої групи коефіцієнтів виникла необхідність у побудові моделі класифікації, яка б на основі отриманої множини однорідних значень всіх наявних даних з використанням алгоритму 1 подолання різнорідності вхідних даних при інсталюванні мобільних додатків та алгоритму 2 шкалювання вхідних даних, дозволяла б визначити клас користувача. Це дає змогу в подальшому автоматизувати процес виявлення шахрайства, створити портрет шахраїв, що містив би навіть нові та непомітні експертам характеристики шахрая.

Отже, отримавши множину однорідних значень всіх наявних даних за допомогою запропонованих шкал, виконано етап класифікації даних з використанням відомої моделі класифікації з метою виявлення даних, які однозначно визначають шахраїв, і даних, які однозначно визначають органічних користувачів. У якості моделі класифікації було обрано повністю зв’язану глибинну нейронну мережу (fully-connected deep neural network – DNN) з трьома прихованими шарами, оскільки штучні глибокі нейронні мережі дозволяють розв’язувати задачі розпізнавання і класифікації з високою точністю.

Удосконалену модель класифікації вхідних даних для виявлення аномалій в Big Data на основі глибинної нейронної мережі з трьома прихованими шарами, що використовується у даній роботі, представлено на рисунку 6. Узагальнену математичну модель процесу класифікації представлено у дисертаційній роботі.

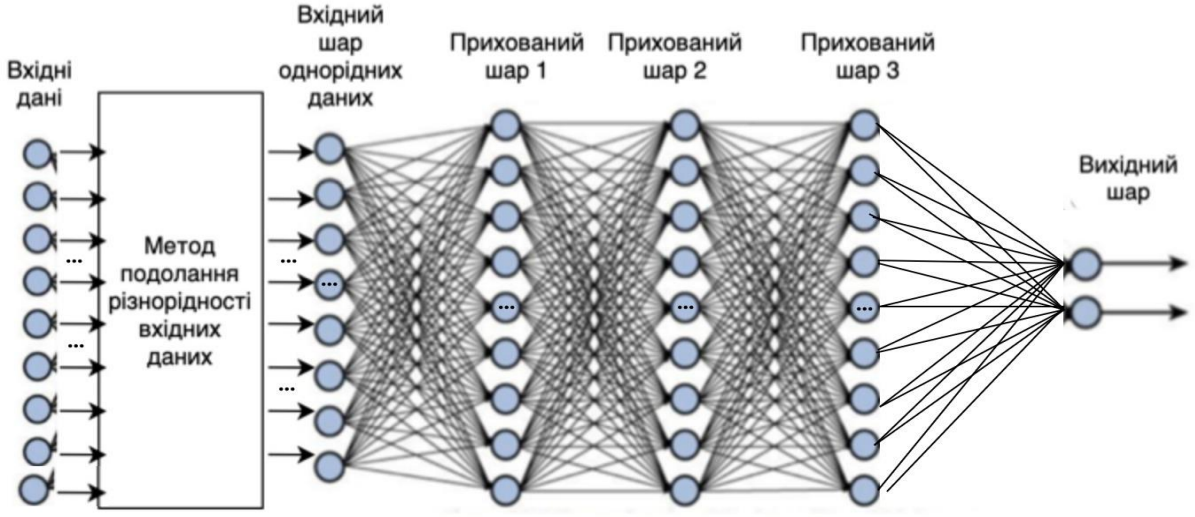

Рисунок 6 – Модель класифікації вхідних даних для виявлення аномалій в Big Data

Удосконалена модель класифікації за рахунок використання розробленого методу подолання різнорідності вхідних даних замість стандартних методів перетворення вхідних даних дозволяє підвищити точність виявлення шахрайства.

У третьому розділі розроблено інформаційну технологію з використанням системного аналізу та моделювання, а саме: побудована логічна, концептуальна моделі інформаційної технології. Здійснено аналіз та розробку структур даних інформаційної технології, розроблено шаблон шахрая для формування портрету шахрая. Як показали дослідження, усі розроблені моделі є необхідними для ефективного функціонування інформаційної технології виявлення шахрайства на основі запропонованого в роботі узагальненого методу, в основі якого лежать моделі, методи та алгоритми виявлення аномалій в даних.

У процесі розробки концептуальної моделі інформаційної технології, виникла необхідність розробити блок адаптації системи до нових ознак через створення узагальненого портрету шахрая. Такий портрет являє собою набір характеристик, що визначають клас користувача – шахрай чи ні – та ознаки, за якими користувача віднесено до цього класу.

У зв'язку з цим в роботі розроблено алгоритми виявлення шахрая при інсталюванні мобільних додатків та алгоритм створення узагальненого портрету шахрая на основі методів, моделей та алгоритмів, розроблених у 2-му розділі, які є основою запропонованої в роботі інформаційної технології, але для використання її у виробництві, в роботі була розроблена та визначена методика виявлення шахрайства. Методика включає в себе послідовне виконання таких основних алгоритмів: алгоритм 1 ініціалізації / корекції моделі виявлення шахрайства і бази знань; алгоритм 2 виявлення шахрайства при наявності ініціалізованої моделі виявлення шахрайства і бази знань.

Алгоритм 1 ініціалізації / коррекції моделі виявлення шахрайства та бази знань складається з:

1. Визначення характеристик даних користувачів по групам G_1 та G_2.

2. Створення узагальненого шаблону шахрая на основі поміченої вибірки, наданої експертами з використанням бази знань, що містить набір правил нечіткої логіки.

3. Подолання різнорідності вхідних даних користувачів:

3.1. Визначення коефіцієнтів за характеристиками групи G_1, з використанням яких також можна здійснити визначення однозначних органічних користувачів, визначення однозначних шахрайських користувачів, визначення неоднозначних користувачів.

3.2. Визначення коефіцієнтів за характеристиками групи G_2 з використанням коефіцієнтів подібності користувача за кожною з характеристик групи G_2 з узагальненим шаблоном шахрая.

3.3. Отримання вектору однорідних вхідних даних, а саме – сукупності коефіцієнтів, що представляють характеристики груп G_1 та G_2.

4. Тренування, валідація, тестування класифікаційної моделі, на вхід якої подаються однорідні дані. Тобто, побудова моделі класифікації, вхідними даними якої є однорідні дані, отримані з використанням розробленого в роботі узагальненого методу подолання різнорідності.

5. Корекція узагальненого портрету шахрая на основі результатів з використанням моделі класифікації та вхідних однорідних даних моделі, отриманих з використанням методу подолання різнорідності даних.

Алгоритм 2 виявлення шахрайства, при наявності ініціалізованої моделі виявлення шахрайства та бази знань, використовується для визначення класу нових користувачів або ж існуючих користувачів за їх новими діями у мобільному додатку. Розглянемо кроки даного алгоритму при появі нової події нового або старого користувача у мобільному додатку. Основні етапи алгоритму 2:

1. Визначення характеристик користувача.

2. Визначення класу користувача з використанням бази знань, а саме – узагальненого шаблону користувача, за визначеними характеристиками.

3. У випадку, якщо існуюча база знань не дозволяє визначити клас поточного користувача, то виконати кроки 3-5 алгоритму 1 та повторити пункт 2 алгоритму 2.

Зазначимо, що наявність та використання узагальненого портрету шахрая дозволить автоматично виявляти шахраїв у різноманітних наборах даних, а також дасть змогу виявляти причину їх появи і здійснювати коригування узагальненого шаблону шахрая. Зважаючи на це, можна зазначити, що задача такого виду відноситься до задач нечіткою логіки.

Основною задачею модуля, який за допомогою нечіткої логіки та розробленого в роботі алгоритму автоматичного виявлення шахрайства виявляє діапазон значень коефіцієнту подібності між набором характеристик користувача та узагальненим шаблоном шахрая по конкретній характеристиці, є визначення користувача шахраєм по даній характеристиці. Матрицю знань предметної області представлено в дисертаційній роботі.

Формалізувавши проблему формування нових правил визначення груп шахрайських та органічних користувачів з використанням нечіткої логіки, запропоновано *алгоритм* автоматичного виявлення шахраїв, який складається з таких основних етапів:

1 Отримання списку користувачів довжиною *userCount* з бази даних.

2 Створення початкового узагальненого шаблону шахрая.

3 Встановлення значення кількості переглянутих алгоритмом користувачів *reviewedUsers* = 0.

4 Перевірити, чи виконується умова *reviewedUsers* < *userCount*.

4.1 Поки умова 4 виконується, то виконуємо наступні дії:

4.1.1 Отримання вхідних даних для поточного користувача.

4.1.2 Визначення коефіцієнтів групи 1 $\overline{G_1}$ для даного користувача.

4.1.3 Хоча б один з коефіцієнтів визначає користувача шахраєм?

4.1.3.1 Якщо так, то:

a Помічення користувача шахраєм (віднесення до класу шахраїв Class0 та U_{fraud}).

b Формування нових правил та додавання їх у базу знань.

4.1.3.2 Якщо ні, то: якщо хоча б один з коефіцієнтів визначає користувача органічним?

a) Помічення користувача органічним (Class1 and Uorg).

b) Формування нових правил та додавання їх у базу знань.

4.1.4 Збільшуємо значення змінної *reviewedUsers* на 1.

4.2 Якщо умова 4 не виконується, то виконуємо наступні кроки:

4.2.1 Визначення вектору коефіцієнтів подібності користувача з шахрайськими користувачами $\overline{G}2_{fraud}$.

4.2.2 Визначення вектору коефіцієнтів подібності користувача з органічними користувачами $\overline{G}2_{org}$.

4.2.3 Формування спільного вектору значень характеристик

$$\overline{G}2 = \overline{G}2_{fraud} \cup \left(1 - \overline{G}2_{org}\right). \quad (2)$$

4.2.4 Формування нових правил та додавання їх у базу знань.

4.2.5 Налаштування моделі класифікації з використанням даних помічених користувачів з вектору $\overline{G}2$ відповідно формулі

$$CV\left(\hat{f}\right) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}\left(x_i\right)\right), \quad (3)$$

де N – кількість частин інформації; L – функція втрат; y – очікуваний результат для входу x; $\hat{f}^{-k}\left(x\right)$ – фітнес-функція; k – поточна частина інформації.

4.2.6 Визначення класів усіх користувачів на основі побудованої моделі класифікації.

4.2.7 Коригування узагальненого шаблону шахрая.

Алгоритм автоматичного виявлення шахраїв ліг в основу методу виявлення аномалій в різнорідних даних при інсталюванні мобільних додатків.

Розробка інформаційної технології містить також: алгоритм створення узагальненого портрету шахрая, алгоритм пошуку аномалій в даних як основу функціонування інформаційної технології за допомогою діаграми activity, здійснено аналіз процесів в інформаційній технології виявлення шахрайства за допомогою UML-діаграми послідовності.

Запропоновано інформаційну технологію виявлення шахрайства при інсталюванні мобільних додатків, яка використовує запропоновані метод виявлення шахрайства при інсталюванні мобільних додатків, метод подолання різнорідності вхідних даних, модель класифікації даних, і, на відміну від існуючих систем Antifraud, дозволила підвищити точність класифікації користувачів до 99,14 %, зокрема точність класифікації шахраїв – до 82,76 %.

У четвертому розділі апробовано розроблену інформаційну технологію. Для проведення експериментальних досліджень з використанням інформаційної технології виявлення шахрайства в роботі було здійснено аналіз та підготовку вхідних даних. Вибірка є поміченою для подальшого аналізу та оцінки обраних критеріїв ефективності. У наборі даних представлено 284807 користувачів та 31 різнорідна ознака по кожному з них, наведено приклади вхідних даних та значення стовпців набору, подано кількість класів у даному поміченому наборі даних та розподіл користувачів по наявним класам. Також подано статистичний звіт по наданому набору даних та гістограми кожної чисельної характеристики. Даний аналіз та підготовка даних була необхідна для виявлення того, на скільки вибірка є репрезентативною, скільки класів вона має, та скільки записів є унікальними.

В роботі розроблено методику проведення експериментальних досліджень розробленої інформаційної технології, основні етапи якої об’єднано в схему, що показана на рисунку 7.

У процесі дослідження виникла необхідність розробки алгоритму мінімізації часу виявлення шахраїв на основі розпаралелення обчислювальних процесів у зв’язку з тим, що вхідні дані відносяться до великих даних (Big Data). Застосування алгоритму дозволило підвищити швидкодію на 9,48 %.

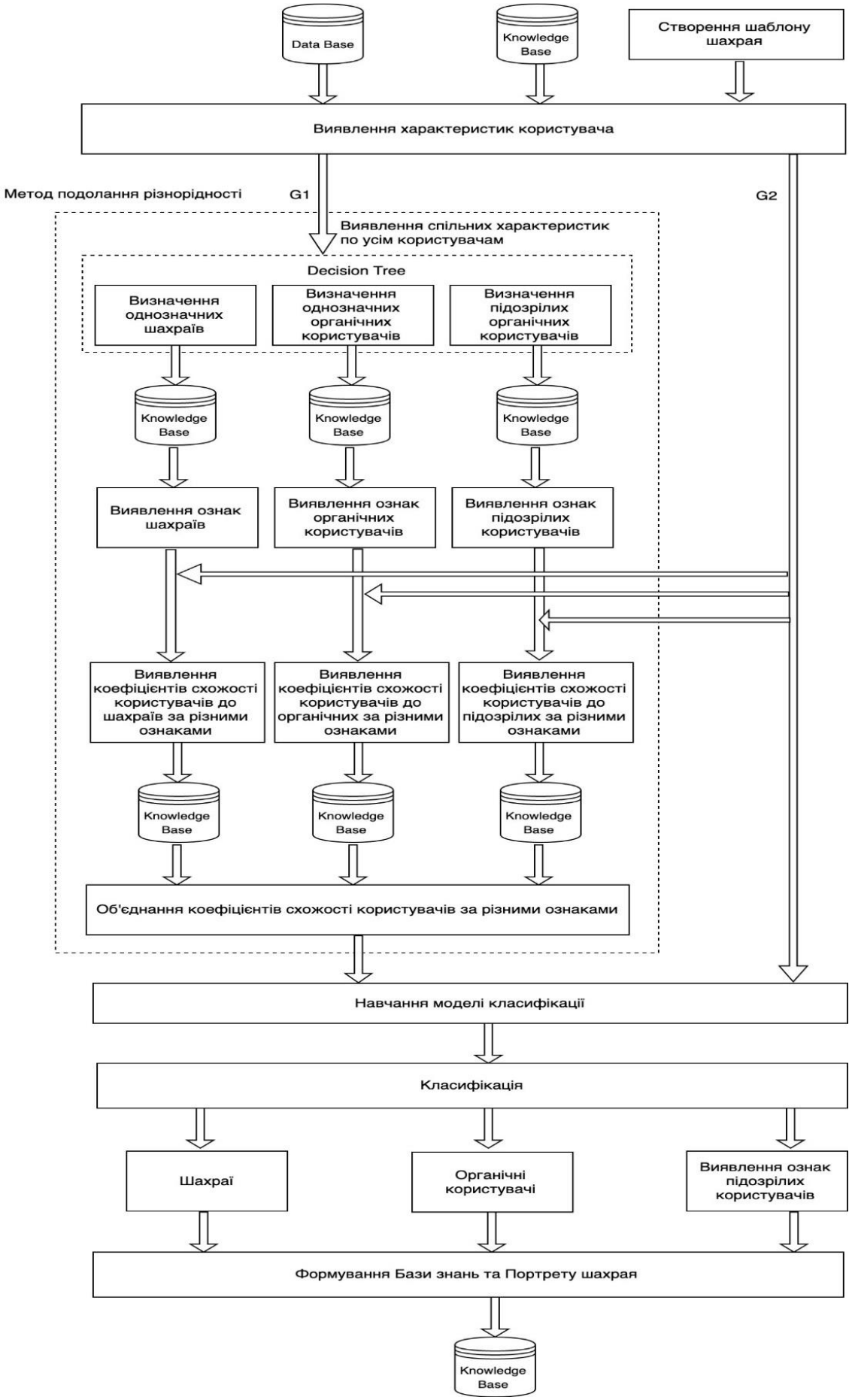

Рисунок 7 – Схема експериментального дослідження виявлення шахрайства як аномалій в різнорідних даних при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних

Доведено адекватність моделей процесу подолання різнорідності вхідних даних та моделі класифікації вхідних даних на вибірці з розробленого мобільного додатку «MobNsters: Mafia War Strategy» (Orneon Ltd.), яку було поділено на три набори даних – тренувальний, перевірочний (валідаційний) та тестовий (контрольний). Сформовано графіки втрат (loss) та точності (accuracy) на етапах навчання та перевірки моделей (рис. 8). Для уникнення перенавчання здійснено навчання моделі у 100 епох. Побудовано нормалізовану та ненормалізовану матрицю невідповідності (рис. 9), на основі якої визначено чотири результати – істинно позитивний, істинно негативний, хибно позитивний, хибно негативний, а також: чутливість моделі – 99,23 %, специфічність – 82,95 %, позитивне прогностичне значення – 99,91 %, негативне прогностичне значення – 36,61 %, помилка першого роду – 17,05 %, помилка другого роду – 0,78%, частка помилкових відхилень – 0,09 %, частка помилкових упущень – 63,39%, загальна точність – 99,14 %, F-міра – 99,57 %, коефіцієнт кореляції Метьюса – 54,78 %, J статистика Йоудена – 82,18 %, позначеність – 36,52 %. Визначено, що точність методу виявлення шахрайства, який працює з даними, отриманими методом подолання різнорідності вхідних даних, рівна 99,14%. Наявних шахраїв метод визначає з точністю 82,76%, загалом точність визначення класу користувачів – 99,14%. Результати роботи програмного забезпечення подано в дисертаційній роботі. Визначено критерії ефективності та здійснено аналіз результатів тестування. Здійснено порівняльний аналіз ефективності розробленого узагальненого методу виявлення шахрайства та існуючих методів, що використовуються для виявлення шахрайства. Точність запропонованої удосконаленої моделі класифікації на тестовій вибірці на 1,14 % вища від найкращої із змодельованих моделей класифікації. Здійснено порівняльний аналіз точності та швидкодії розробленої інформаційної технології та систем-аналогів виявлення шахрайства. Показано, що точність виявлення класу користувача з використанням розробленої технології було підвищено на 1.26%, а точність виявлення шахрая підвищено на 1.36% у порівнянні з системою-аналогом, що дала найкращий результат.

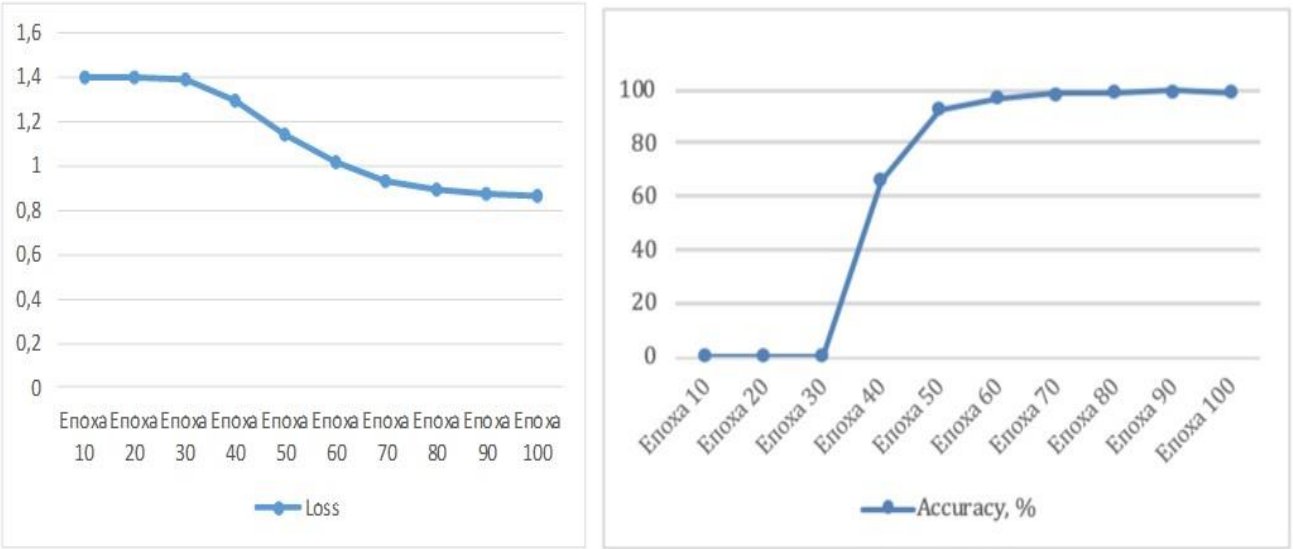

Рисунок 8 – Втрати та точність на етапах навчання та перевірки

Істинно позитивна (ІП) = 56219	**Хибно позитивна** (ХП) = 52
Хибно негативна (ХН) = 438	**Істинно негативна** (ІН) = 253

Рисунок 9 – Матриця невідповідності контрольної вибірки без нормалізації з визначенням кожного значення для подальших обрахунків

Проведено порівняльний аналіз запропонованої інформаційної технології з системами-аналогами FraudLogix, Kochava, Appfslyer (табл. 2, 3).

Таблиця 2 – Порівняльний аналіз швидкодії розробленої інформаційної технології та систем-аналогів виявлення шахрайства

Метод	Підвищення середньої швидкодії виявлення шахрайства на тестовій вибірці запропонованого методу виявлення шахрайства на, %
Fraudlogix	1,78 %
Kochava	2,0082 %
Appsflyer	5,83 %

Таблиця 3 – Порівняльний аналіз ефективності розробленого узагальненого методу виявлення шахрайства та існуючих методів, що використовуються для виявлення шахрайства

Метод	Точність моделі класифікації на тренувальній вибірці	Точність моделі класифікації на тестовій вибірці
Запропонований метод виявлення шахрайства при інсталюванні мобільних додатків	0,9962 %	0,9914 %

На основі запропонованої в роботі інформаційної технології розроблено модульне програмне забезпечення “Mobile App Install Fraud Detection System” на основі алгоритмів, які реалізують розроблені методи, моделі та веб-систему, що впроваджена на підприємстві Garuda AI B.V. (Нідерланди).

У додатках подано документи щодо впровадження результатів роботи; основні лістинги програмного забезпечення; список експертів, які проводили оцінювання; сертифікати на публікації і виступи та список публікацій за темою дисертації.

ВИСНОВКИ

У дисертаційній роботі на основі виконаних теоретичних і експериментальних досліджень розв’язано актуальне наукове завдання підвищення точності та швидкодії процесу виявлення шахраїв шляхом розробки узагальненого методу виявлення шахрайства при інсталюванні мобільних додатків, методу подолання різнорідності вхідних даних, удосконаленої моделі класифікації користувачів на основі глибинних нейронних мереж, алгоритмів процесу подолання різнорідності вхідних даних, алгоритму створення

узагальненого портрету шахрая та алгоритму пошуку аномалій в даних, які покладені в основу запропонованої інформаційної технології.

При цьому отримано такі основні результати:

1. Здійснений аналіз об’єкту дослідження дозволив визначити задачу виявлення шахрайства як одну із задач пошуку аномалій в даних.
2. Здійснений аналіз методів виявлення шахрайства при інсталюванні мобільних додатків та їх класифікація дозволили визначити методи машинного навчання для вирішення поставлених задач.
3. Формалізація процесу виявлення шахрайства як аномалії в даних з використанням теорії множин дозволила визначити властивості аномальних і неаномальних даних у визначеній предметній області та спростити процес пошуку шахрайства при інсталюванні мобільних додатків.
4. Здійснена класифікацію різнорідних даних при інсталюванні мобільних додатків на основі процесу вибору ознак дозволила спростити процес аналізу різнорідних за метриками, розмірностями і шаблонами даних та автоматизувати його.
5. Вперше розроблено узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків, відмінність якого полягає у використанні запропонованої моделі класифікації користувачів та методу подолання різнорідності вхідних даних, що дозволяє визначити класи користувачів та підвищити точність виявлення шахрайства при інсталюванні мобільних додатків.
6. Вперше запропоновано метод подолання різнорідності вхідних даних, що являє собою сукупність процедур вибору ознак, зниження розмірності та нормалізації даних, відмінність якого полягає у новій моделі процесу подолання різнорідності даних шляхом шкалювання за інформативністю, що дозволяє всю множину різнорідних даних про користувачів звести до вектору уніфікованих ознак без зменшення діагностичної цінності інформації.
7. Удосконалено модель класифікації користувачів на основі глибинних нейронних мереж у частині зниження розмірності та нормалізації даних згідно запропонованого методу подолання різнорідності даних, яка є основою для створення узагальненого портрету шахрая з метою спрощення процесів їх виявлення.
8. Розроблено моделі інформаційної технології з використанням технологій системного аналізу та моделювання, а саме: побудована логічна, концептуальна моделі інформаційної технології. Здійснено аналіз та розробку структур даних інформаційної технології, розроблено шаблон шахрая для формування портрету шахрая. Як показали дослідження, усі розроблені моделі є необхідними для ефективного функціонування інформаційної технології виявлення шахрайства на основі запропонованого в роботі узагальненого методу, в основі якого лежать моделі, методи та алгоритми виявлення аномалій в даних.
9. Розроблено алгоритми процесу подолання різнорідності вхідних даних, алгоритм створення узагальненого портрету шахрая та алгоритм пошуку аномалій в даних, які покладені в основу запропонованої інформаційної технології, що підвищило точність та швидкодію процесу виявлення шахраїв.
10. Розроблено алгоритм пошуку аномалій в даних як основа функціонування інформаційної технології за допомогою діаграми activity, а також здійснено аналіз

процесів в інформаційній технології виявлення шахрайства за допомогою UML-діаграми послідовності.

11. Запропоновано інформаційну технологію виявлення шахрайства при інсталюванні мобільних додатків, яка використовує запропоновані метод виявлення шахрайства при інсталюванні мобільних додатків, метод подолання різнорідності вхідних даних, модель класифікації даних, і, на відміну від існуючих систем Antifraud, дозволила підвищити точність класифікації користувачів до 99,14 %, зокрема точність класифікації шахраїв – до 82,76 %.

12. Розроблено методику проведення експериментальних досліджень розробленої інформаційної технології.

13. Розроблено алгоритм мінімізації часу виявлення шахраїв на основі розпаралелення обчислювальних процесів.

14. Доведено адекватність моделей процесу подолання різнорідності вхідних даних та моделі класифікації вхідних даних на вибірці з розробленого мобільного додатку «MobNsters: Mafia War Strategy» (Orneon Ltd.). Зокрема показано, що точність виявлення класу користувача з використанням розробленої технології було підвищено на 1.26%, а точність виявлення шахрая підвищено на 1.36% у порівнянні з системою-аналогом, що дала найкращий результат.

15. На основі запропонованої в роботі інформаційної технології розроблено модульне програмне забезпечення “Mobile App Install Fraud Detection System” з використанням алгоритмів, які реалізують розроблені методи, моделі та веб-систему, що впроваджена на підприємстві Garuda AI B.V. (Нідерланди).

16. Результати дисертаційної роботи та розроблене модульне програмне забезпечення впроваджені на іноземному (Garuda AI B.V.) та українських (ТОВ «ВІН ІНТЕРАКТИВ», ТОВ «4ХайТек», ПП «Літсофт») підприємствах та у навчальний процес (кафедри комп’ютерних наук ВНТУ; кафедри інформатики, програмної інженерії та економічної кібернетики Херсонського державного університету). Впровадження результатів дисертаційних досліджень підтверджено відповідними актами.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

[1] T. Polhul, and A. Yarovyi, “Development of a method for fraud detection in heterogeneous data during installation of mobile applications”, *Eastern-European Journal of Enterprise Technologies*, no. 1/2 (97), 2019. doi: 10.15587/1729-4061.2019.155060

[2] A. Yarovyi, and T. Polhul, “Applied Aspects of Implementation of Intelligent Information Technology for Fraud Detection During Mobile Applications Installation”, *Advances in Intelligent Systems and Computing IV. CCSIT 2019: Advances in Intelligent Systems and Computing, Springer*, Cham, Switzerland, vol 1080, pp. 377-386, 2019. doi: https://doi.org/10.1007/978-3-030-33695-0_26

[3] T. Polhul, “Conceptual Model of an Intelligent System for Detecting Fraud During Mobile Applications Installation”, in *Proc. 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, Leeds, United Kingdom, pp. 167-174, 2019. doi: 10.1109/DESSERT.2019.8770030. Режим доступу: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8770030&isnumber=8770005.

[4] T. D. Polhul, A. A. Yarovyi, R. Romaniuk, P. Komada, N. Askarova "Method of data anomaly detection in the process of mobile applications installation", *Proc. SPIE 11176, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments* 2019, 111761Y; https://doi.org/10.1117/ 12.2536855
[5] T. Polhul and A. Yarovyi, "Method of Fraudster Fingerprint Formation During Mobile Application Installations", in *Proc. 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Metz, France, 2019, pp. 1099-1103. doi: 10.1109/IDAACS.2019.8924369. Режим доступу: https://ieeexplore.ieee.org/ document/8924369.
[6] A. Yarovyi, and T. Polhul, "Intelligent information technology for fraud detection during mobile applications installation", in *Proc. 14th International conference "Computer sciences and Information technologies" (CSIT 2019)*, pp. 1-5, Lviv, 2019. doi: 10.1109/STC-CSIT.2019.8929827. Режим доступу: https://ieeexplore.ieee.org/ document/8929827.
[7] T. Polhul, "Development of an intelligent system for detecting mobile app install fraud", *International Journal of Advances in Electronics and Computer Science (IJAECS)*, vol. 6, no. 7, pp. 13-17, July, 2019.
[8] А. А. Яровий, О. Н. Романюк, І. Р. Арсенюк, та Т. Д. Польгуль, "Виявлення шахрайства при інсталюванні програмних додатків з використанням інтелектуального аналізу даних", *Наукові праці Донецького національного технічного університету. Серія "Інформатика, кібернетика та обчислювальна техніка"*, № 2 (25), с. 126-131, 2017. Режим доступу: http://science.donntu.edu.ua/ wp-content/uploads/2018/03/ikvt_2017_2_site-1.pdf
[9] Т. Д. Польгуль, та А. А. Яровий, "Метод подолання різнорідності даних для виявлення шахрайства при інсталюванні мобільних додатків", *Вісник СНУ ім. В. Даля – Сєвєродонецьк: СНУ ім. В. Даля*, № 7 (248), с.60-69, 2018.
[10] Т. Д. Польгуль, та А. А. Яровий, "Аналіз різнорідних даних в інтелектуальних системах виявлення шахрайства", *Вісник Вінницького політехнічного інституту*, № 2, с. 78-90, 2019.
[11] Т. Д. Польгуль, "Інформаційна технологія побудови інтелектуальних систем виявлення шахрайства при інсталюванні мобільних додатків", *Інформаційні технології та комп'ютерна інженерія*, № 1, с. 4-16, 2019.
[12] T. Polhul, "Development of an intelligent system for detecting mobile app install fraud", in *Proc. IRES 156th International Conference*, Bangkok, Thailand, 2019, pp. 25-29.
[13] А. А. Яровий, та Т. Д. Польгуль, "Комп'ютерна програма "Програмний модуль збору даних інформаційної технології виявлення шахрайства при інсталюванні програмних додатків", *Свідоцтво про реєстрацію авторського права на твір № 76348 від 26.01.2018 р.*, К.: Міністерство економічного розвитку і торгівлі України, 2018.
[14] А. А. Яровий, та Т. Д. Польгуль, "Комп'ютерна програма "Програмний модуль визначення схожості користувачів інформаційної технології виявлення шахрайства при інсталюванні програмних додатків", *Свідоцтво про реєстрацію авторського права на твір № 76347 від 26.01.2018 р.*, К.: Міністерство економічного розвитку і торгівлі України, 2018.

[15] Т. Д. Польгуль, та А. А. Яровий, “Визначення шахрайських операцій при встановленні мобільних додатків з використанням інтелектуального аналізу даних”, *Сучасні тенденції розвитку системного програмування. Тези доповідей*, Київ, 2016, с. 55-56. Режим доступу: http://ccs.nau.edu.ua/wp-content/uploads/2017/12/%D0%A1%D0%A2%D0%A0%D0%A1%D0%9F_2016_07.pdf

[16] А. Яровий, Т. Польгуль, та Л. Крилик, “Розробка методу виявлення шахрайства при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних”, *XIV Міжнародна конференція Контроль і управління в складних системах (КУСС-2018). Тези доповідей*, Вінниця, 2018, с. 35.

[17] А. А. Яровий, та Т. Д. Польгуль, “Подолання різнорідності вхідних даних при виявленні шахрайства при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних”, на *П’ятій міжнародній науково-технічній конференції студентів, магістрів, аспірантів «Інформатика, управління та штучний інтелект»*, Національний технічний університет «Харківський політехнічний інститут», Харків, 2018, с. 109.

[18] Т. Д. Польгуль, та А. А. Яровий, “Визначення шахрайських операцій при інсталяції мобільних додатків з використанням інтелектуального аналізу даних”, на *XLVI науково-технічній конференції підрозділів ВНТУ*, Вінниця, 2017. Режим доступу: http://ir.lib.vntu.edu.ua/bitstream/handle/123456789/17200/2158.pdf?sequence=3

[19] Т. Д. Польгуль, “Моделювання процесу виявлення шахрайства при інсталюванні мобільних додатків”, на *XLVIII науково-технічній конференції підрозділів ВНТУ*, Вінниця, 2019. Режим доступу: https://conferences.vntu.edu.ua/ index.php/all-fitki/all-fitki-2019/paper/view/6863

[20] Т. Д. Польгуль, “Порівняльний аналіз Apache Spark та Apache Flink для роботи з Big Data”, на *XLV науково-технічній конференції підрозділів ВНТУ*, Вінниця, 2016. Режим доступу: https://ir.lib.vntu.edu.ua/handle/123456789/11619

АНОТАЦІЯ

Польгуль Т. Д. **Інформаційна технологія виявлення шахрайства при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних. – Кваліфікаційна наукова праця на правах рукопису.**

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Вінницький національний технічний університет, Вінниця, 2020.

Робота присвячена розробленню інформаційної технології виявлення шахрайства при інсталюванні мобільних додатків з використанням інтелектуального аналізу даних.

Вперше запропоновано метод подолання різнорідності вхідних даних, що являє собою сукупність процедур вибору ознак, зниження розмірності та нормалізації даних, відмінність якого полягає у новій моделі процесу подолання різнорідності даних шляхом шкалювання за інформативністю, що дозволяє всю

множину різнорідних даних про користувачів звести до вектору уніфікованих ознак без зменшення діагностичної цінності інформації.

Удосконалено модель класифікації користувачів на основі глибинних нейронних мереж у частині зниження розмірності та нормалізації даних згідно запропонованого методу подолання різнорідності даних, яка є основою для створення узагальненого портрету шахрая з метою спрощення процесів їх виявлення.

Вперше розроблено узагальнений метод виявлення шахрайства при інсталюванні мобільних додатків, відмінність якого полягає у використанні запропонованої моделі класифікації користувачів та методу подолання різнорідності вхідних даних, що дозволяє визначити класи користувачів та підвищити точність виявлення шахрайства при інсталюванні мобільних додатків.

Ключові слова: інформаційна технологія, інтелектуальний аналіз даних, виявлення шахрайства, виявлення аномалій в даних, коефіцієнти подібності, модель класифікації, машинне навчання, глибинні нейронні мережі, метод подолання різнорідності, матриця невідповідності, інсталювання мобільних додатків.

АННОТАЦИЯ

Польгуль Т. Д. **Информационная технология обнаружения мошенничества при инсталлировании мобильных приложений с использованием интеллектуального анализа данных. – Квалификационная научная работа на правах рукописи.**

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.06 «Информационные технологии». – Винницкий национальный технический университет, Винница, 2020.

Работа посвящена разработке информационной технологии обнаружения мошенничества при инсталлировании мобильных приложений с использованием интеллектуального анализа данных.

Впервые предложен метод преодоления разнородности входящих данных, который представляет собой совокупность процедур выбора признаков, снижения размерности и нормализации данных, отличие которого заключается в новой модели процесса преодоления разнородности данных путем шкалирования по информативности, что позволяет все множество разнородных данных о пользователях свести к вектору унифицированных признаков без уменьшения диагностической ценности информации.

Усовершенствована модель классификации пользователей на основе глубинных нейронных сетей в части снижения размерности и нормализации данных согласно предложенного метода преодоления разнородности данных, которая является основой для создания обобщенного портрета мошенника с целью упрощения процессов их обнаружения.

Впервые разработан обобщенный метод выявления мошенничества при инсталлирования мобильных приложений, отличие которого заключается в использовании предложенной модели классификации пользователей и метода преодоления разнородности входящих данных, что позволяет определить классы

пользователей и повысить точность обнаружения мошенничества при инсталлирования мобильных приложений.

Ключевые слова: информационная технология, интеллектуальный анализ данных, обнаружение мошенничества, обнаружение аномалий в данных, коэффициенты сходства, модель классификации, машинное обучение, глубинные нейронные сети, метод преодоления разнородности, матрица несоответствия, инсталлирование мобильных приложений.

ABSTRACT

Polhul T. D. **Information technology for fraud detection during mobile applications installation using data mining. – Qualification research paper, manuscript copyright.**

Thesis for the degree of a candidate of technical sciences in specialty 05.13.06 «Information technology». – Vinnytsia National Technical University, Vinnytsia, 2020.

The dissertation research is dedicated to developing the informational technology to detect fraud during mobile application installation using data mining.

The purpose of the dissertation research is to increase the accuracy and speed of fraud detection processes during mobile application installation.

The scientific novelty of the qualification research paper is:

1. For the first time, a method of overcoming the heterogeneity of input data is proposed, which is a set of procedures for feature selecting, dimensionality reduction and data normalization, the difference of which lies in the new model of overcoming the heterogeneity of data by scaling information, which allows the whole set of heterogeneous user data to be reduced to a vector, reducing the diagnostic value of information.

2. The model of users' classification based on deep neural networks in terms of dimensionality reduction and data normalization has been improved according to the proposed method of overcoming heterogeneity of data, which is the basis for creating general fraudsters fingerprint in order to simplify their detection processes.

3. For the first time, a generalized method for fraud detection in during installation of mobile applications has been developed, the difference being the use of the proposed users' classification model and the method of overcoming the heterogeneity of the input data, which allows defining users’ classes and increasing the reliability of fraud detection during mobile app installs.

The practical value of the results obtained in the qualification research paper is as follows:

– classification of heterogeneous data was carried out, which made it possible to simplify the process of analysis of data which is heterogeneous by metrics, dimensions and data templates and to automate it;

– an algorithm for detecting anomalies in data, algorithms for the process of overcoming the heterogeneity of input data, algorithm for fraud detection during mobile applications installations, an algorithm for generalized fraudster fingerprint formation, and algorithm for minimizing the time of fraud detection based on the parallelization of computing processes, which are the basis of information technology, which has increased the accuracy and speed of detection of fraudsters, were developed;

– information technology for fraud detection during mobile applications installation has been proposed for the first time, that uses the following: a generalized method for fraud detection during mobile applications installation, a method for overcoming the heterogeneity and classification model of user input data, which, unlike existing Antifraud technologies, allowed to improve accuracy of users' classification to 99,14 %, in particular, detecting fraudulent users to 82,95 %;

– "Mobile App Install Fraud Detection System" software for fraud detection during mobile application installation has been developed.

The results of the qualification research paper were implemented at Garuda AI B.V. (Netherlands) – information technology; LLC «WinInteractive» – algorithms for overcoming data heterogeneity, model for the process of overcoming data heterogeneity; LLC «4HighTech» – a model of the process of overcoming the heterogeneity of the input data, the method of fraud detection when installing mobile applications; PE «Litsoft» – algorithm for overcoming heterogeneity of data, a generalized method of detecting fraud when installing mobile applications, a method for creating a generalized fraudster's fingerprint; to the educational process of the Computer Science Department of Vinnytsia National Technical University – a generalized method for fraud detection during mobile applications installation; and to the educational process of the Department of Informatics, Software Engineering and Economic Cybernetics of Kherson State University – information technology for fraud detection during mobile applications installation using data mining.

Keywords: information technology, data mining, fraud detection, anomaly detection, similarity coefficients, classification model, machine learning, deep neural networks, method for overcoming heterogeneity of data, confusion matrix, mobile applications installation.

Підписано до друку 24.02.2020 р. Формат 21х29.7 1/4.
Наклад 100 прим. Зам. № 2020-038.
Віддруковано в комп'ютерному інформаційно-видавничому центрі
Вінницького національного технічного університету.
м. Вінниця, вул. Хмельницьке шосе, 95. Тел.: 65-18-06
Суб’єкт видавничої справи
серія ДК №3516 від 01.07.2009р.