Хмельницький національний університет
Факультет інформаційних технлогій
Кафедра комп'ютерної інженерії та інформаційних систем

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

Галузь знань 12 – Інформаційні технології

Спеціальність 123 –Комп'ютерна інженерія

на тему «Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання»

КвРКІ 16040.20.01.35 ПЗ

Виконав: студент 2 курсу, група КІ2м-20-1 _____ Ференс В.О.
Підпис Ініціали, прізвище

Керівник доктор техн. наук, професор _____ Бармак О.В.
Науковий ступінь, вчене звання Підпис Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КІІС, д.т.н., проф.
Т. О. Говорущенко
12 05 2022 р.

Хмельницький, 2022

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ

Освітній рівень МАГІСТР

Галузь знань 12 ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Спеціальність 123 КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Освітня програма ОСВІТНЬО-НАУКОВА ПРОГРАМА «КОМП'ЮТЕРНА ІНЖЕНЕРІЯ ТА ПРОГРАМУВАННЯ»

ЗАТВЕРДЖУЮ

Зав. кафедри Т.О.Говорущенко

" 01 " 09 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Ференс Володимир Олександрович
Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання

Керівник проекту (роботи) Бармак О.В.., д.т.н., проф.
Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 06.01.2021 р. № 1

2. Строк подання студентом проекту (роботи) на кафедру 06.05.2021 р.

3. Вихідні дані до проекту (роботи) Завдання на кваліфкаційне проектування

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Аналіз відомих методів захисту конфіленційності збереження даних
Проєктування та удосконалення наявних алгоритмів збереження конфіденційності
Метод виявлення проблем конфіденційності та збурення потоку даних;
Дослідження ефективності запропонованих рішень, реалізація.

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

6. Консультанти розділів кваліцікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормоконтроль	Лисенко С.М., професор кафедри КІС		
Антиплагіат	Нічепорук А.О., доцент кафедри КІС		

7. Дата видачі завдання «_06_»_____09_____2021 р.

КАЛЕНДАРНИЙ ПЛАН

№з/п	Назва етапів (розділів) кваліцікаційної роботи	Термін виконання етапів проекту (роботи)	Примітка
1	Вибір напряму дослідження та узгодження тематики ДРМ з керівником	05.09.2021	виконано
2	Ознайомлення з предметною областю; формулювання мети та задач дослідження; визначення об'єкта та предмета дослідження	05.10.2021	виконан
3	Робота над розділом 1 – аналіз відомих моделей, методів за темою; постановка задачі	05.11.2021	виконан
4	Робота над розділом 2 – розробка моделей для вирішення поставленої задачі	05.12.2021	виконан
5	Робота над науковою статтею та тезами	05.01.2022	викона
6	Робота над розділом 3 – розробка методів для вирішення поставленої задачі	15.02.2022	викона
7	Робота над розділом 4 – проектування та розробка системи конфіденційності для вирішення поставленої задачі, експериментальна частина	05.04.2022	викона
8	Оформлення пояснювальної записки згідно вимог	15.04.2022	викона
9	Попередній захист ДРМ	18.04.2022	викон
10	Захист ДРМ на засіданні ЕК	До 10.05.2022	

Студент _____ В.О. Ференс
Підпис Ініціали, прізвище

Керівник проекту (роботи) _____ О.В. Бармак
Підпис ініціали, прізвище

РЕФЕРАТ

Тема кваліфікаційної роботи: «Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання»

Автор роботи: Ференс Володимир Олександрович

Керівник роботи: Бармак Олександр Володимирович

Пояснювальна записка: 111 ст., 12 рис., 3 дод., 65 джерел.

ПЕРЕЛІК КЛЮЧОВИХ СЛІВ: збурення даних, алгоритм захисту конфіденційності, великі дані, потоки даних, швидкодія, випадкове генерування.

Об'єкт дослідження є процес збереження конфіденційності великих даних.

Предмет дослідження є методи і засоби для забезпечення конфіденційності великих даних на основі машинного навчання та створення методів на основі базових алгоритмів забезпечення конфіденційності.

Метою роботи є покращення конфіденційності зберігання великих даних.

Для розв'язання поставлених задач використовувалися початкові алгоритми захисту конфіденційності; шум Лапласа, синтетична генерація даних, методи збурення даних, криптогафічні методи, які впливають на якість інтелектуального аналізу для підтримки цілісності та збереження конфіденційності даних.

Наукова новизна одержаних результатів полягає в наступному:

- розроблено новий метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання.

Практичне значення одержаних результатів полягає у вдосконалених методах забезпечення конфіденційності для великомашстабних даних не тільки для великих корпорацій, які мають великі потужності пристроїв, але і з підтримкою стаціонарних машин із дотриманням їх працездатності під час обміну інформації між такими компаніями. Розроблені методи контролюють витоки інформації та запобігають використанню їх третіми особами, що зазвичай можуть бути зловмисниками, які бажають використати конфіденційну інформацію у власних цілях.

Теоретичні та практичні результати роботи впроваджено при виконанні студентських науково-дослідних робіт, які виконувались в Хмельницькому національному університеті.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

АЗК – алгоритм захисту конфіденційності

АЗРД – алгоритм збурення розподілених даних

АЗКдГН – алгоритм збереження конфіденційності для глибокого навчання

АЗДО – алгоритм збурення даних обличь

БАЗД – безпечний алгоритм збурення даних

МЗПД – метод збурення потоку даних

ПК – персональний комп’ютер

ЦП – центральний процесор

КС – комп’ютерні системи

IPFS – однорангова файльова система

GDRP - загальний регламент захисту даних

ВСТУП

Актуальність роботи. Основні сфери, як для прикладу енергетика, транспортування, банківська справа, охорона здоров’я та сільське господарство, трансформуються досягненнями в Інтернеті та пов'язаних з ними технологій, таких як Інтернет речей (англ. Internet Of Things).

Об'єднання великої кількості різних технологій, таких як Інтернет речей, хмарні обчислення, граничні обчислення та машинне навчання сприяли стрімкому та активному поширенню технологічного розвитку у різних сферах, таких як охорона здоров’я, енергетика, агропромисловість і також транспортування. Збільшення кількості загальнодоступних гаджетів сприяло швидкому зростанню Інтернету речей, ставши одним з основних джерел великих потоків даних.

Кіберпростір охоплює не тільки фізичну сферу, але і людську, величезна кількість інформації (даних) стають доступними для аналізу. З розвитком швидкості генерації даних аналітика останніх дає чудові результати – точність у реалізації основних ідей, що представлено великими об’ємами даних. Один із тих підходів, який привернув найбільшу кількість уваги – поглиблене навчання, яким забезпечується виняткова продуктивність з обсягами великих даних. З конфіденційністю даних осіб тісно пов’язані різні індустрії таких як банківська справа, охорона здоров’я, енергетика, агропромисловість і транспортна інфраструктура.

І як наслідок, помітна тенденція зростання наслідків через відкриття конфіденційних даних третім особам, атак на бази даних, тощо. До слова, корпорації та інші організації постійно прагнуть, забезпечення максимальної конфіденційності та щоб вся інформація зберігалась на локальних серверах компанії. Для запобігання порушень конфіденційності машинне навчання разом із аналітикою даних повинно запровадити всі можливі сценарії захисту конфіденційності для забезпечення неушкодженості конфіденційності користувачів. Існує багато підходів до збереження конфіденційності, однак, величезні розміри великої кількості даних і потоків даних роблять збереження конфіденційності важкою задачею. Головною

проблемою в числі існуючих підходів є їх неспроможність підтримки правильного балансу між конфіденційністю, корисністю та ефективністю при оперуванні з великою кількістю даних. Деякі ефективні підходи забезпечують хорошу конфіденційність, але не забезпечують достатньої швидкодії під час операцій із даними, у той час як інші – навпаки забезпечують хорошу швидкодію, але не забезпечують високий рівень конфіденційності.

Постійні вдосконалення у сфері аналітики даних і технік машинного навчання, таких як поглиблене навчання, довели, що дають високу точність, забезпечуючи надійні прогнози. Однак такі підходи часто покладаються на величезні обсяги даних, які, можливо, потрібно буде зібрати з різних джерел. Таким чином, численні розподілені організації беруть участь у процесі збору та аналізу даних, ризикуючи порушенням конфіденційності, передбачене стандартами чи протоколами, такими як загальний регламент захисту даних (GDRP) та акт про передачу та захист даних закладів охорони здоров'я (HIPAA). Для прикладу, моделі якщо тренуються на масиві бази даних, зазвичай, розкривають особисту інформацію, та є вразливими до атак, які ставлять за мету порушити конфіденційність даних. Окрім цього, відомо, що така область, як аналіз біометричних даних: це може бути як і данні із датчика аналізу відбитку пальців, що є чи не в кожному смартфоні, так і розпізнавання обличь та сканування сітківки ока. Ці області виконують ресурсозатратні та важкі завдання, які часто залучають сторонні сервера, а отже до яких можуть і отримати доступ кібер-зловмисники із своїми корисливими намірами. Тому, якщо біометрична інформація неконтрольовано доставляється на ненадійні та сумнівні сторонні сервери, то це можна вважати чималим витоком конфіденційності (тобто неконтрольований витік інформації), оскільки данні отримані із біометричних датчиків можна співвідносити із конфіденційними даними записів медперсоналу та інформації банківських операцій. Атаки на конфіденційність намагаються зробити все, для виявлення ідентичності осіб у вхідних даних, які використовуються для кожного екземпляру аналізу даних або створення машинного навчання моделі. Таким чином, порушення конфіденційності може значно вразити цілісність та безпеку

інформації, а також передати особисту інформацію в ненадійне середовище, до якого кібер-зловмисники звертаються зі зловмисним наміром.

Можливість ділитися інформацією, обмежуючи розкриття приватної інформації третім сторонам, має велике значення, оскільки обмін даними із зовнішніми сторонами є важливим для аналізу даних та машинного навчання. З розвитком розподілених хмарних середовищ машинного навчання, таких як Google і Amazon, все більше користувачів можуть стати вразливими до атак конфіденційності. Довіряючи цим середовищам, користувачі можуть передавати свої дані для навчання моделей і отримати доступ до них за допомогою білого чи чорного ящиків. Проте зловмисник може легко реалізувати алгоритми, які можуть завдати шкоди системі. Ці алгоритми можуть запам’ятовувати конфіденційну інформацію користувача як частину наведених моделей. Пізніше зловмисники можуть систематизувати збережену інформацію і таким чином отримати інформацію про користувачів і порушити їх конфіденційність. Атаки щодо конфіденційності, такі як висновки про членство, показують вразливість моделей поглибленого навчання, конфіденційних даних, навіть якщо вони випускаються як моделі чорного ящика. Аналогічним прикладом, який показує слабкість навчених моделей машинного навчання, є атаки інверсії моделей, які відновлюють зображення з системи розпізнавання обличчя.

Використання механізмів для збереження конфіденційності є дуже важливим для обмеження витоку конфіденційності при обміну даними. Прикладом можуть слугувати системи охорони здоров’я, котрі зберігають велику кількість даних про пацієнтів, для покращення клінічної аналітики та надання більш ефективних медичних послуг. За допомогою транзакцій клієнтів у роздрібній торгівлі створюється величезний обсяг даних, і вони є особливо необхідними для здійснення обслуговування клієнтів. Варто зауважити, що дані, котрі зберігаються в цих системах, містять приватну інформацію, яка потребує захисту. Якщо за даними не слідкувати належним чином, то можуть статися непередбачувані витоки конфіденційності. Іншим прикладом є додаток Facebook, який має незліченну базу програм, і використання його слабких місць може мати негативні наслідки для

конфіденційності користувачів. Проте вже випробувані механізми збереження конфіденційності могли б більше ефективно вирішити наявні проблеми.

Досить легко ідентифікувати користувачів у базі даних. Для цього було об’єднано кілька квазіідентифікаторів, таких як вік, поштовий індекс та стать. Перед публікацією видалення ідентифікаторів із бази є недостатнім для безпеки та захисту конфіденційності осіб. Не є новиною, що дані, які потрапляють до третіх осіб можуть нести за собою критичні наслідки. Контрольований випуск інформації – саме так можна визначити конфіденційність даних для їх аналізу і машинного навчання. Можливість вищесказаного неконтрольованого випуску інформації в методах аналізу даних є пропонованим інтелектуальним аналізом даних, безперечно, із всіма забезпеченнями конфіденційності за всіма стандартами. Застосування підходів до збереження конфіденційності аналізу даних і машинного навчання для створення цінної інсайтерської інформації без розкриття приватності юзерів є основною метою інтелектуального аналізу даних із збереженням конфіденційності.

Роблячи висновки щодо динаміки, згаданої вище, можна визначити три основні труднощі, які інтелектуальний аналіз даних із має подолати, щоб забезпечити надійні шляхи збереження конфіденційності для аналізу даних. Вони полягають у наступному.

Доступність великорозмірних статичних даних. Великорозмірний набір даних нерідко складається з досить великого обсягу екземплярів (наприклад, мільйонів кортежів); у деяких випадках масиви даних з великими розмірами також можуть вміщувати у собі величезну кількість атрибутів (наприклад, сотні чи тисячі). Для застосування конфіденційності у таких налаштуваннях, шляхи збереження мають використовувати ефективні та вже випробувані схеми. Однак, є складним завданням дотримання балансу між конфіденційністю та корисністю, при цьому підтримуючи ефективність.

Наявність потоків даних. Швидкість та динамічність генерування даних є основною складністю забезпечення безпеки для потоків даних. Доступний в умовах Інтернету речей, сценарій потоку даних досить швидко генерує дані, а також

надсилає запит із вимогою аналізу даних для отримання важливого уявлення про конкретну організацію. , що вищесказаний сценарій згенерований для збереження

Присутність великого розподілу сторін, що беруть участь у обміні даними. Здебільшого сучасні середовища аналізу даних і машинного навчання часто містять у собі багато розподілених сутностей. Об'єднане машинне навчання є прикладом останніх інновацій, які фокусуються на об'єднанні різних методів використання новітнього машинного навчання та аналітики у широкомасштабних налаштуваннях. Розподіл даних, представлений середовищем федеративного навчання. Нові сценарії збереження конфіденційності розглядають розподіл даних для обмеження непередбачуваних витоків даних та забезпечення безперечних рішень для збереження конфіденційності.

В роботі пропонується використання алгоритмів захисту конфіденційності, що були розроблені на основі базових методів.

Метою роботи є покращення конфіденційності зберігання великих даних.

Задачі дослідження формулюються в роботі наступним чином:

1. Виділити недоліки відомих методів та стратегій, систематизувати їх.
2. Розробити модель базової архітектури системи для покращення конфіденційності зберігання великих даних.
3. Метод збурення розподілених даних.
4. Алгоритми збурення даних із використанням локальної диференціальної конфіденційності для великих даних
5. Здійснити реалізацію азпропонованих рішень.
6. Провести експериментальні дослідження з розробленими засобами.

Об'єкт дослідження – процес збереження конфіденційності великих даних.

Предмет дослідження – методи і засоби для забезпечення конфіденційності великих даних на основі машинного навчання та створення методів на основі базових алгоритмів забезпечення конфіденційності.

Методи дослідження. Для досягнення поставлених задач використано основні положення:

1. Початкові алгоритми захисту конфіденційності.

2. Шум Лапласа.
3. Синтетична генерація даних.
4. Методи методи машинного навчання.
5. Методи збурення даних.

Наукова новизна одержаних результатів полягає в наступному:

- розроблено новий метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання.

Обґрунтованість і достовірність наукових положень, висновків і рекомендацій. Наукові положення, висновки і рекомендації кваліфікаційної роботи магістра обґрунтовані коректним та доцільним використанням математичного апарату, алгоритмами здійснення машинного навчання, успішною реалізацією, ефективним практичним впровадженням результатів, яке продемонструвало відповідність теоретичних розробок з реальними результатами застосування.

Практичне значення одержаних результатів. У результаті виконаного дослідження розроблено метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання.

Теоретичні та практичні результати роботи впроваджено при виконанні науково-дослідних робіт, які виконувались в Хмельницькому національному університеті.

Особистий внесок здобувача. Всі основні результати дослідження, які представлені до захисту кваліфікаційної роботи, одержані автором особисто. В роботах, опублікованих у співавторстві, автору належать основні ідеї, теоретична та практична розробка положень, відображених у характеристиці наукової новизни отриманих результатів, а саме: [1] – запропоновано використання алгоритму забезпечення конфіденційності який імплементовано в сферу охорони здоров’я для демонстрації результатів.

Апробація результатів дисертації. Основні положення та результати проведених у кваліфікаційній роботі досліджень доповідалися та обговорювалися на двох міжнародних наукових конференціях:

Публікації. За результатами проведених досліджень основні наукові

результати опубліковано у збірнику матеріалів конференції «Актуальні проблеми комп’ютерних наук «АПКН-2021» , 2021. С. 257-259.

Структура кваліфікаційної роботи. Кваліфікаційна робота складається з анотації, вступу, чотирьох розділів, висновків, списку використаних джерел з 65 найменувань на 7 сторінках та трьох додатків на 26 сторінках. Загальний обсяг роботи становить 111 сторінки, з них 84 сторінок основного тексту, 12 рисунків.

1 АНАЛІЗ ВІДОМИХ МЕТОДІВ І ЗАСОБІВ ПОКРАЩЕННЯ КОНФІДЕНЦІЙНОСТІ ДЛЯ ВЕЛИКИХ ДАНИХ

1.1 Огляд та поняття конфіденційності даних

Об'єднання великої кількості технологій, таких як Інтернет речей, хмарні обчислення, граничні обчислення та машинне навчання сприяли стрімкому поширенню технологічного розвитку у різних сферах, таких як охорона здоров’я, енергетика, агропромисловість і також транспортування. Збільшення кількості загальнодоступних гаджетів сприяло швидкому зростанню Інтернету речей, ставши одним з основних джерел великих потоків даних. Кіберфізичні системи, сенсорні системи (широкий спектр) та передові інструменти аналізу об’єднані для надання консолідованих послуг. Як наслідок, крім того, що накопичено звичайними способами, конкретна система може отримувати вигоду з кількох джерел. Отже, кіберпростір переповнений приватною інформацією, що стосується багатьох вразливих сфер. Для прикладу: охорона здоров’я та агроінженерія. Більше уваги вимагає застосування різноманітних механізмів для захисту конфіденційності та безпеки даних, враховуючи їхній вразливий характер. Є три тісно пов’язаних поняття, які вивчають режими технічних механізмів для захисту інформації: безпека, приватність та конфіденційність. Конфіденційність даних визначає ті речі, які суб’єкт може якими данні його збираються та використовуються. Дані, що були оприлюднені у відкритому доступі, мають відповідати законам про конфіденційність, щоб власника даних не можна було ідентифікувати, використовуючи його чи її особисту інформацію. Конфіденційна інформація зберігається у безпеці і у секреті наскільки це максимально можливо. Безумовно існують винятки при яких порушення конфіденційності є необхідністю. Для прикладу це може бути запит правоохоронних органів для виявлення загроз у кіберпросторі. Безпека даних запевняє користувачів про надійний захист за допомогою сучасних методів, таких як шифрування, і гарантує, що доступ отримують лише авторизовані юзери. У цій кваліфікаційній роботі досліджуються проблеми конфіденційності аналізу даних та машинного навчання.

Збільшення кількості даних, зібраних із кіберфізичного та людського світів було зумовлено останніми досягненнями у сфері комп’ютерних технологій. Велика кількість розумних кіберфізичних систем таких як розумна мережа, розумні транспортні засоби, інтелектуальні системи охорони, розумні системи охорони здоров’я та, безумовно, розумні будинки стають широко популярними завдяки величезному технологічному прогресу, особливо за останній десяток років. Перераховані системи часто взаємодіють із нашим навколишнім середовищем. Зазвичай вони збирають данні для аналізу. Для прикладу використовуються данні сенсорів, які сигналізують про зміни навколишнього середовища, а алгоритми штучного інтелекту приводять в дію автоматичні процеси для виконання дій, які потрібні саме в цей проміжок часу під час вищеназваних умов та показів датчиків. Окрім цього шалені досягнення останніх років у сфері Інтернет- технологій, таких як Інтернет речей значно змінили багато сфер, таких як важка промисловість, охорона здоров’я, енергетика та транспортування. Будь-які збори даних у великих масштабах можуть мати сенс, але якщо ці данні є практичними, та їх так чи інакше можна використовувати у прийнятті рішень та подальших діях. Досліджувати сподівані зв’язки даних і надання корисних даних аналітикам інформації дає можливість аналітика даних та машинне навчання. Збільшення об’ємів обсягів даних призвело до стрімкого розвитку технологій, на подобі як аналітика колосальних об’ємів даних і машинного навчання. Також були відкриті нові можливості. Основні технології, що пов’язані зі створенням та збором даних, постійно оновлюються, щоб підвищити гнучкість програмування та покращити зберігання даних. Удосконалення та покращення мережевої інфраструктури, для прикладу об'єднання програмно-конфігурованих мереж і віртуалізації функцій мережі, у взаємодії між граничними обчисленнями гарантують кращу якість обслуговування для складних додатків, на основі Інтернет речей, що сприяє розширеному аналізу даних і машинному навчанню.

Об'єднання великої кількості технологій, таких як Інтернет речей, хмарні обчислення, граничні обчислення та машинне навчання сприяли стрімкому поширенню технологічного розвитку у різних сферах, таких як охорона здоров’я,

енергетика, агропромисловість і також транспортування. Збільшення кількості загальнодоступних гаджетів сприяло швидкому зростанню Інтернету речей, ставши одним з основних джерел великих потоків даних. Кіберфізичні системи, сенсорні системи (широкий спектр) та передові інструменти аналізу об'єднані для надання консолідованих послуг. Як наслідок, крім того, що накопичено звичайними способами, конкретна система може отримувати вигоду з кількох джерел. Отже, кіберпростір переповнений приватною інформацією, що стосується багатьох вразливих сфер. Для прикладу, охорона здоров'я та агроінженерія. Більше уваги вимагає застосування різноманітних механізмів для захисту конфіденційності та безпеки даних, враховуючи їхній вразливий характер. Є три тісно пов'язаних поняття, які вивчають режими технічних механізмів для захисту інформації: безпека, приватність та конфіденційність. Конфіденційність даних визначає ті речі, які суб'єкт може якими данні його збираються та використовуються. Дані, що були оприлюднені у відкритому доступі, мають відповідати законам про конфіденційність, щоб власника даних не можна було ідентифікувати, використовуючи його чи її особисту інформацію. Конфіденційна інформація зберігається у безпеці і у секреті наскільки це максимально можливо. Безумовно існують винятки при яких порушення конфіденційності є необхідністю. Для прикладу – це може бути запит правоохоронних органів для виявлення загроз у кіберпросторі. Безпека даних запевняє користувачів про надійний захист за допомогою сучасних методів, таких як шифрування, і гарантує, що доступ отримують лише авторизовані юзери. У цій кваліфікаційній роботі досліджуються проблеми конфіденційності аналізу даних та машинного навчання.

Збільшення кількості даних, зібраних із кіберфізичного та людського світів було зумовлено останніми досягненнями у сфері комп'ютерних технологій. Велика кількість розумних кіберфізичних систем таких як розумна мережа, розумні транспортні засоби, інтелектуальні системи охорони, розумні системи охорони здоров'я та, безумовно, розумні будинки стають широко популярними завдяки величезному технологічному прогресу, особливо за останній десяток років. Перераховані системи часто взаємодіють із нашим навколишнім середовищем.

Зазвичай вони збирають данні для аналізу. Для прикладу, використовуються данні сенсорів, які сигналізують про зміни навколишнього середовища, а алгоритми штучного інтелекту приводять в дію автоматичні процеси для виконання дій, які потрібні саме в цей проміжок часу під час вищеназваних умов та показів датчиків. Окрім цього шалені досягнення останніх років у сфері Інтернет- технологій, таких як Інтернет речей значно змінили багато сфер, таких як важка промисловість, охорона здоров'я, енергетика та транспортування. Будь-які збори великих даних можуть мати сенс, але якщо ці данні є практичними, та їх так чи інакше можна використовувати у прийнятті рішень та подальших діях. Досліджувати сподівані зв'язки даних і надання корисних даних аналітикам інформації дає можливість аналітика даних та машинне навчання. Збільшення об'ємів обсягів даних призвело до стрімкого розвитку технологій, на подобі як аналітика колосальних об'ємів даних і машинного навчання. Також були відкриті нові можливості. Основні технології, що пов'язані зі створенням та збором даних, постійно оновлюються, щоб підвищити гнучкість програмування та покращити зберігання даних. Удосконалення та покращення мережевої інфраструктури, для прикладу об'єднання програмно-конфігурованих мереж і віртуалізації функцій мережі, у взаємодії між граничними обчисленнями гарантують кращу якість обслуговування для складних додатків, на основі Інтернет речей, що сприяє розширеному аналізу даних і машинному навчанню.

1.2. Відомі методи та засоби покращення конфіденційності для великих даних

Комплекси збору інформації, такі як Інтернет речей, часто поширюються та запроваджуються між етапами зборів та аналізом даних. У цьому процесі випливає різного роду інформація, що може призвести до порушення приватності та безпеки даних в цілому. Можливість ділитися інформацією, запобігаючи розголошенню персональної інформації, стає важливим аспектом конфіденційності інформації, і це одна з найважливіших технічних, правових, етичних і соціальних проблем. Різні

державні та комерційні організації збирають величезні обсяги даних користувачів, зокрема інформацію про стан здоров’я, фінансовий стан та особисті уподобання. Часто нехтують конфіденційністю соціальні мережі, системи у банківській сфері, охорони здоров’я та інші прикладні системи, які обробляють приватну інформацію. Це нехтування проявляється через те, що вони часто нехтують конфіденційністю через непряме використання приватних даних. Безліч інших галузей також використовують значну кількість приватної інформації для моделювання та прогнозування аномалій, які є пов’язані із людиною, такі як епідемії чи злочини. З вищесказаних слів можна зробити висновок, що саме збереження конфіденційності та безпека даних може стати вкрай складною проблемою та, безумовно, вимагає надійних рішень.

Машинне навчання – це область інформатики, що вивчає штучний інтелект, яка дозволяє комп’ютерам автоматизувати набуття нових знань. Головною умовою набуття знань є обов’язкове вдосконалення в режимі реального часу і дії відносно тієї чи іншої задачі на основі досвіду, отриманого із попереднього навчання. Глибинне навчання — це група алгоритмів та методів машинного навчання, являться частиною сімейства машинного навчання, які функціонують з представленнями даних під наглядом, напів наглядовим або неконтрольованим способом. Доведено, що основною здатністю алгоритмів глибинного навчання зображувати конкретну проблему у вигляді запровадження ієрархії понять є найбільш вдалою, ніж інші алгоритми машинного навчання. Алгоритми глибокого навчання показали достатньо велику кількість можливостей виняткових міркувань у реальних програмах. Ці алгоритми мають змогу використовувати величезні обсяги даних з мінімальним впливом предметної області. Також вони змогли обійти інші методи навчання.

Налічуються наступні три найбільшвикористовувані типи керованих моделей глибокого навчання: стандартні глибокі нейронні мережі, рекурентні нейронні мережі та згорткові нейронні мережі. Хоча вищезгадані мережі мають багато спільних властивостей, мають структурні відмінності та різну основну поведінку при навчанні. Нейромережі – це достатньо тісно пов'язана мережа

модулів обробки, що називаються нейронами. Слід взяти до уваги, що кожен нейрон виробляє послідовність реальних активацій. Рекурентні нейронні мережі формуються з великою кількістю нейронів, пов'язаних з петлями зворотного зв'язку, які підтримують контрольовану кількість попередніх часових станів. Вони вважаються найглибшими з трьох типів; вони працюють за принципом створення сегментів пам'яті довільних послідовностей вхідних шаблонів. Центральний блок обробки (процесор) більшою мірою нагадує топологію, яка має сильну схожість на сітку, що актуально у навчальній матриці властивостей.

Центральний процесор (ЦП) намагається у більшості випадків розпізнавати істотні ознаки зображень. Якщо точність навчання значно вища за точність тестування, то відбувається переобладнання. Переобладнання уникають якісні моделі. Процесом застосування будь-якої модифікації алгоритму навчання для зменшення помилки узагальнення називають регуляризацією. Коли певний відсоток нейронів у такий спосіб перемикається в кожну епоху (цикл навчання), то її може бути досягнуто за допомогою випадання, щоб уникнути переобладнання.

Техніка підготовки даних, що використовує наявні вхідні зображення в наборі навчальних даних і керує ними для створення багатьох трансформованих версій, використовуючи різні методи перетворення – це збільшення зображення. Для того, щоб зробити навчену модель більш узагальненою з високою надійністю, ця техніка дозволяє штучним нейронним зв’язкам вивчити більш широкий спектр вхідних даних.

Розмір пакету, кількість нейронів, функції активації, кількість епох і оптимізатор, змінюються на різних етапах навчання при налаштуванні гіперпараметрів вхідні дані для гіперпараметрів, щоб визначити найкращий приклад, який дає найвищі результати. Розмір пакету – це кількість навчальних прикладів, які будуть розповсюджені за один прохід вперед і назад. Нейрон(вузол) – основний компонент штучної нейронної мережі. Він є одним проходом, при якому весь набір даних вводиться вперед і назад через нейрон мережі і називається епохою. Оптимізатор (або алгоритм оптимізації) використовується для оновлення параметрів моделі.

Блокчейн — це розподілений реєстр записів даних, мережі яких переплітаються і не належать центральному органу. Концепція Блокчейну використовується в розділі, для того, щоб забезпечити можливість відстеження розподіленого машинного навчання для налаштувань промислового Інтернету речей. Дані транзакцій блокчейну є незмінними та загальнодоступними. Автоматично стає прозорою та стійкою до атак та програма, яка побудована на блокчейні. Етеріум — це блокчейн-платформа з відкритим вихідним кодом для децентралізованих додатків. Програми, які запускаються на віртуальній машині Етеріум (EVM), називаються «смарт-контрактами». Solidity і Vyper — дві з популярних мов, які використовуються для написання смарт-контрактів на Етеріумі.

IPFS — це однорангова файлова система, яка забезпечує високу пропускну здатність блочної моделі зберігання вмісту з гіперпосиланнями, що є унікальним хеш-значенням. Будь-яка зміна файлу знищить його початкове хеш-значення, що робить дані, збережені в IPFS, незмінними. IPFS утворює узагальнений спрямований ациклічний граф і поєднує розподілену хеш-таблицю, для підтримки створення версійних файлових систем і блокчейнів. Захист конфіденційності людей став проблемою з поширенням споживчих технологій, що підтримуються Інтернетом. У літературі представлені різні підходи. щоб вирішити цю проблему тоді як деякі підходи зосереджені на підвищенні обізнаності. Але деякі методи намагаються використовувати різні способи. Перш за все, великі обсяги великих даних створюють багато проблем з конфіденційністю, хоча питання, які є пов’язані з безпекою та конфіденційністю великих даних, не є абсолютно новими. Але це вимагає уваги через специфіку середовища та динаміку використовуваних пристроїв. Розвиток цих середовищ та різноманітність пристроїв ускладнює та постійно ускладнює безпеку та конфіденційність.

Серед кількох визначень конфіденційності, можна вважати найбільш прийнятним твердження для аналізу даних, що зберігає конфіденційність. Однак складність завдання полягає у тому, щоб визначити рівень інформації для належного аналізу конфіденційності. Її розглядають як таємниця приватної

інформації.

Отже, втрата конфіденційності є витоком приватної інформації. Для того, щоб виміряти витік конфіденційності використовують різні підходи, такі як ймовірність та інформаційний вміст. Хоча також можна виокремити шифрування, щоб повністю зупинити цей потік, адже воно обмежує можливість аналіз та обмін даними.

Таким чином, необхідно систематизувати всі можливі підходи та методи збереження конфіденційності, які дозволяють демобілізувати контрольовану інформацію, яку можна виміряти за допомогою різних засобів, таких як ймовірність.

Кількісно визначити конфіденційність конкретних методів конфіденційності за допомогою показників конфіденційності може бути складно. У літературі наведено велику кількість визначень показників конфіденційності. Відповідно до конкретних типів методів конфіденційності (наприклад, втручання в збурення багатовимірних даних). Таким чином можна показати показник конфіденційності, який є методом збурення даних основою якого є адаптивний шум, що базується на залежності від близькості між порушеним значенням і його вихідним значенням. Цей показник оцінює вірогідність [c] вихідної оцінки протягом інтервалу часу [a, b]. Наразі конфіденційність оцінюється за інтервалом [a, b] та вірогідністю [c]. Цей метод має проблему неприйняття. Запропоновано розподіл бази вихідних даних на рахунки. На основі ентропії даних існує більш загальний метод кількісної оцінки конфіденційності. Поняття «інформація, що пов’язана з даними» використовується для визначення рівня конфіденційності. Однак одним із основних недоліків цього підходу є те, що він не враховує фактори, які викликають ризики, що є напряму пов’язані із змагальними атаками. Для додаткового методу інтерференції даних на основі шуму з урахуванням близькості між початковими значеннями та вихідні значеннями. Одним із основних цього підходу є те, що він так чи інакше не враховує ризики, які є напряму пов’язані з атаками зловмисника на основі базових знань. У цьому сценарії для обміну даними конфіденційності існують різні атрибути. І передбачено, що в них присутні різні рівні захисту конфіденційності.

Вищесказане несе більш конфіденційну інформацію, ніж інші атрибути і також тісно пов’язано з ними. А показники конфіденційності, зображені раніше, були слідують цій думці. Між іншим, різниця присутня як і між збуреними та і між невиправленими атрибутами. Основна важкість – це оцінка початкових даних після використання збурення. Вища різниця дисперсії забезпечує відповідну її конфіденційність. А статистична різниця (названої дисперсії) передбачає більш високий рівень складності при аналізі вхідних даних, а саме їх оцінці. Покращення рівню конфіденційності (найслабшого атрибута) є основною метою збереження конфіденційності. Однак підходи до збереження конфіденційності, які використовують цей підхід часто не є ефективними. Така продуктивність зумовлена ітераційним характером пошуку оптимальної цілі для покращення конфіденційності атрибута, що володіє найменшою силою. Для подолання невизначеного рівня приватності була введена модель конфіденційності. Моделі конфіденційності включають: анонімність, багатоманітність, стислість і диференційну конфіденційність. Модель конфіденційності гарантує концептуальний підхід до виконання та дотримання суворих умов безпеки. Отже модель конфіденційності під час глибокого аналізу даних і машинного навчання забезпечує достатньо організований спосіб кількісної оцінки конфіденційності.

Є три технологічні підходи конфіденційності: контроль над розкриттям статистичних даних, інтелектуальний аналіз даних для збереження конфіденційності і технології, що підвищують конфіденційність. Шифрування на основі атрибутів, контроль доступу за допомогою аутентифікації, контроль доступу на основі часу та розташування, а також використання протоколів на основі обмежень – це деякі механізми, які використовуються для покращення конфіденційності систем у динамічних середовищах. Основні підходи для інтелектуального аналізу даних для збереження конфіденційності можна розмежувати за чотирма типами: криптографічні підходи, збурення даних, незбурювані підходи, штучне генерування даних.

Захищені багатосторонні обчислення гомоморфне шифрування і являються двома частопропонованими у різних напрямках криптографічними підходами до

інтелектуального аналізу даних, який забезепечує збереження конфіденційності. Між іншим, ці методи використовують великі криптографічні операції. Криптографічні підходи складні і це часто змушує вдаватися до несерйозних але водночас скурпульозних методів. Ними буде зроблено висновок про приватну інформацію лише шляхом обробки даних, які надаються для протокольного заключення. Не дивно, що існують деякі середовища, які містять лише зловмисні сторони. В цих середовищах нікому не можна довіряти. Отже, реалізація як і гомоморфного шифрування так і безпечних багатосторонніх обчислень (якщо розглядати для шкідливих середовищ) потребують ще більш ресурсозатратних криптографічних операцій. На сьогодні спостерігається надзвичайна складність допуску компромісу між конфіденційністю та відповідністю, яка дозволить підвищити певний ступінь точності за рахунок сподіваного витоку конфіденційності. Можемо зробити висновок, що криптографічні підходи для великомасштабних сценаріїв, які є поєднані із потоками величезних даних є нереальними. Зауважимо, що дане шифрування збільшує розмір даних під час кодування (для прикладу один біт перемножується на 16 біт), що володіє ненадійністю для збереження даних. Беручи до уваги всю складність, дослідники шукали ефективніші вирішення, що забезпечували б більшу податливість та продуктивність для великих наборів даних. У результаті, такі зусилля спрямували до розробки підходів до збурення даних, підходів без збурення та підходів до генерації синтетичних даних, які забезпечують доцільні рішення у порівнянні з криптографічними підходами.

Незбурювані підходи містять відбір вибірки, глобальне перекодування та локальне подавлення. Проте незбурювані підходи можуть бути невідповідними, щоб впоратися з можливостями більш оснащених супротивників. Коли незбурювані підходи застосовуються до великомасштабних позицій, таких як великі дані, сила збереження конфіденційності різко знижується.

Генерація синтетичних даних вважається ефективним підходом до збереження конфіденційності, оскільки аналітики не мають доступу до вихідних даних. Основною проблемою створення підходів синтетичних даних полягає в

тому, що існує певна можливість відповідності синтетичного запису інформації конкретного власника даних. Відповідно, це може спричинити витік конфіденційності.

Підходи до генерації синтетичних даних також можуть бути трудомісткими через використовувані ітераційні підходи.

Серед різноманіття підходів до інтелектуального аналізу даних часто надають перевагу збуренню (модифікації) даних через його простоту, ефективність та можливість налаштування компромісу між конфіденційністю. Ці властивості роблять збурення даних найкращою альтернативою, як, для прикладу, великі дані та потоки даних.

Коли справжні дані (вихідні) повністю змінюються перед тим, як вони повідомляються третій стороні, яка потенційно може бути зловмисником – збурення даних. Порівнюючи з гомоморфним шифруванням, або із аналогічним криптографічним підходом, змінені дані не піддаються жодним форматам перетворення, зважаючи на те, що просторова та часова важкість є меншою. Оскільки гомоморфні шифрування забезпечують достатньо високий рівень конфіденційності. Але слід враховувати, що під час збурення даних рівень витоку даних хоч і часто мінімальний, але присутній, тому слід передбачувати всі можливі виходи даних, щоб не було порушень цілісності даних чи їх конфіденційності. Початкове значення x згенеровується шляхом додавання до x випадкової величини r до x або шляхом застосування певного процесу рандомізації до x. На рисунку 1.1 показано збурення даних, котрі можна розділити на два класи: 1) збурення вхідних даних та 2) збурення вихідних даних.

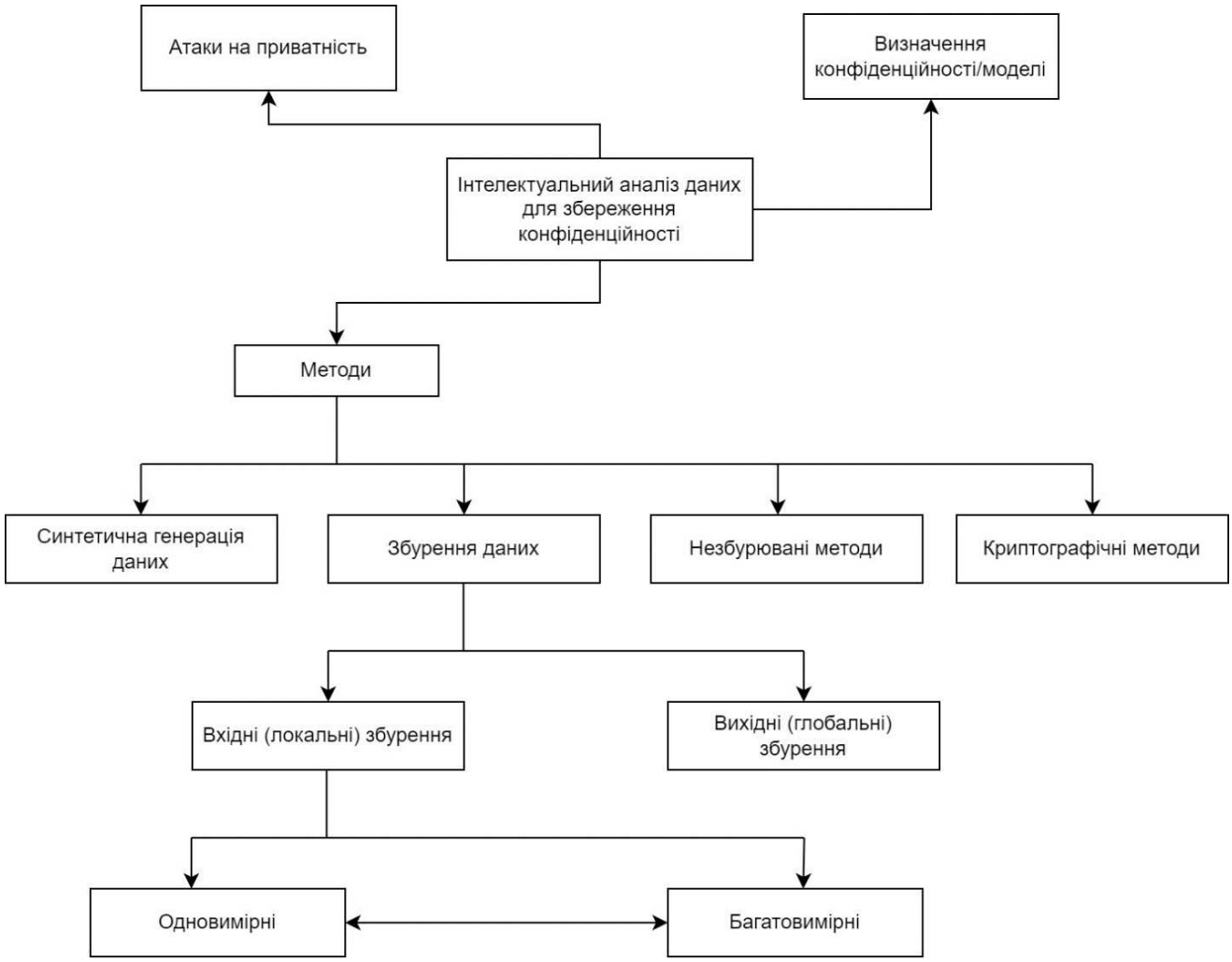


Рисунок 1.1 - Класифікація існуючих методів інтелектуального аналізу даних для збереження конфіденційності

Збурення вихідних даних називається глобальним збуренням даних, тоді як збурення вхідних даних -- локальним збуренням даних. На рисунку 1.1 показано, що збурення даних проводиться на даних, коли вони полишають власників даних. Довірена особа при збурені даних (переважно вихідних) застосовує збурення даних до вихідних даних, які отримуються в результаті виконання запитів на них. Не дивно, що регулярність застосування збурень як вхідних так і вихідних даних, що потребують конфіденційності є досить високою. По логіці речей збурення переважно варто застосовувати на збурені вхідних даних, саме для ненадійних середовищ, де треті особи мають доступ до інформації, або безпека середовища не є на високому рівні. Якщо порівнювати вхідне і вихідне збурення даних, то на

вхідних – присутнє застосування, якого неможливо уникнути вищого рівня випадковості, що дає забезпечення і посилення рівнів конфіденційності, ніж збурення власне вихідних даних.

На рисунку 1.1 показано, що вхідне збурення поділяють на одновимірне та багатовимірне. Прикладами одномірного вхідного збурення є адитивне обурення, мікроагрегація, заміна даних та повторна вибірка. Приклади підходів багатовимірного збурення -- це геометричне обурення, конденсація даних, випадкове обертання, випадкова проекція. Гібридне збурення – це злиття кількох форм збурення разом. Вихідне збурення виконується додаванням шуму та експонентного механізму. Випадковий шум під час адитивного збурення даних зазвичай додається таким чином, що ті статистичні характеристики, які виділені як основні зберігаються постійно. Низька корисність даних – істотна проблема при такому підході. Безумовно, рівень конфіденційності можуть знизити і методи, що розроблені у відповідь. На чітких правилах конфіденційності, які дозволяють публікацію наборів та мікронаборів даних грунтується мікроагрегація, яка дозводяє публікувати ці мікронабори даних із дотриманням всіх правил конфіденційності. Принцип дії полягає у поділі набору значень на кластери. які містять у собі k елементів. Зміна значень відбувається у кожному кластері на геометричне місце тих уявних точок, які перетинаються (центроїд). Коли поширені дані містять інформацію про метод захисту та його параметри, мікроагрегація однієї змінної вразлива до атак прозорості. Щодо багатовимірної мікроагрегації, її також було висунуто, але вона складна і, згідно з доведнням, NP-складна задача. Ще одним типом підходів є метод рандомізації. Приклаом може слугувати рандомізація відповідей опитаних, що зберігають конфіденційність респондентів. Часто забезпечують високий рівень конфіденційності через високий рівень рандомізації вхідних даних методи рандомізації. Відповідно, корисність може бути низькою. Округлення замінює вихідні значення набору даних на заокруглені значеннння. Тільки із безперервними даними працює округлення. Часто виконується у багатовимірному сценарії одновимірного режиму. У одновимірному режимі кожен атрибут збурюється окремо. Оскільки округлення працює у багатовимірному

сценарії, то він може бути неефективним для великих даних. Данна непродуктвність пояснюється наступним чином. Так при великих обсягах даних сила округлення зменшується, то це зменшує і силу обмеження розкриття кондоленційної інформації. Обмін між окремими записами відбувається таким чином при обміні інформацією, що це зберігає низький порядок частот. Питання компромісу між втратою інформації (незалежно від її розміру) чи розкриття її – одна із областей, що так чи інакше потребує додаткового захисту від аномалій, покращення та оцінки. Низький рівень можливості накладання методів на метод по відношенню до вихідних даних суттєво впливає на можливість використання у деяких сферах. Для прикладу інтелектуальний аналіз стану природнього середовища у сфері охорони природи та її ресурсів у світі. Тому можливість композиції – один із основних напрямків розвитку питань конфіденційності. Один із головних та важливих недоліків повторної вибірки є компроміс між розкриттям даних та складністю під час обчислень, що тягне за собою більш глобальнішу проблему вимогливості то кількості ресурсів пристрою, який виконує метод.

При скупченні даних вхідний набір ділиться на кілька груп таким чином, щоб різниця між записами в саме тій групі була мінімальною, і, власне, у кожній групі зберігається певний рівень статистичної інформації про різні записи. Після того формуються знезаражені дані за допомогою рівномірного випадкового розподілу. Це відбувається на основі окремих векторів, що генеруються за допомогою певного розкладання характеристичних коваріаційних матриць кожної однорідної групи. Конденсація має істотний недолік у тому, що вона може значно погіршити якість даних у міру їх зростання. Проте конденсація даних є гідним суперником серед алгоритмів збереження конфіденційності. Є випадки, що підхід конденсації не може забезпечити достатньої конфіденційності, особливо тоді, як просторова локальність підгруп недостатньо велика, і може бути характерною особливістю потоків даних. Натомість просторова локальність велика і якість набору даних різко знижується. З цього випливає, що знижується і точність. Випадкова матриця перекладу застосовується під час геометричного збурення. Збурення обертання, відстані та перекладу включені у метод. Існує три типи матричних методів

збурення даних: випадкового обертання, геометричних даних та випадкового збурення проекції. Перемноження матриці даних відбувається при випадковому обертанні. Властивості ортогональної матриці присутні при перемноженні, а застосування здебільшого відбувається до тих пір поки не буде досягнутий відповідний рівень конфіденційності і захисту від зловмисників. Випадкова матриця перекладу застосовується під час збурення геометричних даних. Це безпосередньо впливає на збільшення рівню збереження даних від зловмисників. Проєктування даних із простору (високорозмірного) у випадково вибраний (низькорозмірний) Випадкове збурення обертання й проєкційності, а також геометричне збурення даних можуть зберігати відстані між кортежами в наборі даних завдяки ізометричній природі перетворень. Як наслідок, вони забезпечують високу корисність при класифікації та кластеризації. Підхід базується на основі ескізіві створює збурені дані шляхом розгляду ескізів (r розмірів) вихідного запису x (d розмірів), де $r \leq d$ і $s_j = \sum_{i=1}^{d} x_i r_{ij}$. Де власне r_{ij} випливає за допомогою генератора псевдовипадкових чисел. Базуючись на основі ескізів, підхід досягає компромісу між корисністю та конфіденційністю, де зменшення r збільшує конфіденційність при зменшенні корисності. Відповідно до гібридного збурення кілька механізмів об'єднуються, щоб застосувати збурення до вхідних даних. Прикладом можуть слугувати мультиплікативні та адитивні властивості матриці, що часто використовуються водночас у гібридному збуренні, яке має схожі властивості із геометричним збуренням. Проте ці алгоритми мають досить високу обчислювальну складність і вимагають набагато більше часу, що робить їх недієздатними з великими наборами даних.

Також було виділено те, що сценарії збурення даних зазвичай вразливі то атак зловмисників пов’язані із відновленням даних. На основі незалежної аналітики, відомих атак на ввід та вивід даних та спектральна фільтрація не були пропущені. Сильне збурення даних забезпечує достатню стійкість до атак та аномалій. Атаки основою яких є незалежний компонентний аналіз часто використовують його для відновлення вхідних даних. Атаки, які мають відомі нам

введення та виведення інформації можна припустити, що зловмисник, який бажає отримати доступ до даних має у наявності або володіє інформацією про певну частину даних, та знає відповідність чи часткову відповідність між вхідними даними та вже відповідними їм збуреними даними. Зловмисник зазвичай застосовує це порівняння даних для подальшого синтезу решти інформації із вже наявної. Аналізуючи дані власних векторів, атаки, з урахуванням власного аналізу, намагаються відфільтрувати шум збурених даних. Фільтрація з сингулярним розклад матриці, спектральна фільтрація та фільтрація з використанням методу головних компонент формують три приклади атак , що базуються на основі власного аналізу. До розкриття інформації ведуть атаки на основі аналізу розподілу. У свою чергу, вони намагаються відновити функцію густини ймовірності вихідних даних. Вищеназвані атаки так чи інакше вимагають, щоб нові підходи до збурення даних. А найголовніша вимова для підходів – це стійкість.

Наразі машинне навчання стає у нагоді в багатьох сферах, таких як охорона здоров’я, відкриті банківські операції, які певним чином використовують конфіденційні приватні дані. Якщо належні процедури обробки не застосовуються до сценаріїв машинного навчання, то вразливі приватні дані можуть стати загальнодоступними. Диференціальна конфіденційність була слушним рішенням щодо обмеження витоку конфіденційності.

Глибинне навчання – одна з областей машинного навчання, яка привернула увагу завдяки високій точності, отриманій під час навчання на великих наборах даних. Це навчання злагоджено працює, коли є велика кількість даних, що може бути потенційною загрозою для конфіденційності. Зокрема, коли здійснюється у вразливих сферах, таких як охорона здоров’я та банківська справа.

Певні ключові функції намагаються зробити інверсні моделі атак. Зазвичай вони отримують на вході біометричні дані (наприклад відбиток пальця в системі розпізнавання відбитків пальців), що використовується для навчання моделі. Атака виключення моделі базується на витягуванні даних та їх параметрів, які є зв’язані із моделлю, що є навченою на приватних даних. Згодом зловмисник застосовує виділені параметри для навчання моделі, яка успішно копіює поведінку та

функціональність вихідної моделі.

Кілька спроб вирішення конфіденційності (витоку даних) в глибокому навчання шляхом активного машинного навчання. (АВТОР) розробив розподілений механізм багатостороннього навчання для нейронних мереж. Але цей механізм не використовував вихідні набори даних. Їх процес навчання проводився на алгоритмі оптимізації стохастичного градієнтного спуску. Алгоритм працював на приватній бібліотеці. А втрата конфіденційності розраховувалась за параметром моделі. Безумовно, існувало багато параметрів моделі. І ця особливість сприяє втраті приватності в цілому саме через те, що таких параметрів може бути тисячі.

Конфіденційність даних концентрується на перешкоджанні оцінці вхідних даних із знезаражених. В той час як утиліта концентрується на збереженні властивостей і інформації, специфічних для застосування. Зазначимо, що механізми збереження конфіденційності не сприяють продуктивності та швидкодії загалом. Тобто вони знижують корисність для покращення конфіденційності. А пошук компромісу між захистом конфіденційності та корисністю даних є важливим питанням. Вимоги конфіденційності і корисності часто є суперечливими: алгоритми, що сприяють збереженню конфіденційності за рахунок корисності. Спостерігається, що конфіденційність часто зберігається шляхом збурення даних. Також допускається зміна вихідних даних. Дослідження похибки збурення також зазначається поширеним способом вимірювання корисності методу, що зберігає конфіденційність. Цей вид обробки даних являє собою різницю між результатом запиту на збурених даних та результатом ідентичного запиту до вхідних даних. Науковці виявляли похибки різних типів (A, B, C та D) під час використання та досліджень різних методів збурення даних. Похибка типу A відбувається рівно тоді коли до зміни сумарних показників призводить саме обурення вхідного атрибута. Похибка B та C відповідно відбуваються тому, що перший є результатом того, що збурення змінює відносини між конфіденційними атрибутами, а другий – змінює відносини між неконфіденційними та конфіденційними атрибутами. І наостанок базовий розподіл даних було порушено процесом знезараження відноситься до похибки класу D. Зауважимо, що якщо

інструменти пошуку даних будуть працювати з збуреними даними менш точно ніж з вихідним. Але це за умови існування способу добування даних.

Існуючі алгоритми та підходи які допомагають збереженню конфіденційності також показують, що вони регулярно піддаються проблемам корисності чи конфіденційності, як показують дослідження. Але це розглядається тоді, коли дослідження концентруються на загальних додатках. Низьку корисність дає метод адитивного збурення з шумом. Це зумовлено випадковістю доданого шуму. Метод випадкової відповіді має інший підхід до збереження конфіденційності. Однак має аналогічний рівень випадковості. А метод багатовимірної мікроагрегації забезпечує низьку придатність до використання через високий рівень складності, що є обумовленою тим, що метод являється NP-складною задачею. Також існує конденсація даних. Дане рішення забезпечує збереження конфіденційності потоків даних, але їх якість погіршується в міру їх зростання. Відповідно до цього бажано застосовувати цей метод тільки для малої кількість даних. У інакшому випадку використання конденсації даних для великих об’ємів даних напряму знижує швидкодію обміну інформацією з ними та відповідно призводить до вкрай низької корисності та продуктивності. Також вивчалася велика кількість багатовимірних підходів. Обертальне та геометричне обертальне збурення є одними із них. Але вони не відзначалися високою обчислюваною складністю. Вони вимагають величезну кількість часу для виконання підходу, що робить його неприйнятним. Тому спостерігаються прямі ознаки того, що застосування вищевказаних методів не є раціональним рішенням для обробки даних великих розмірів. Це зумовлює необхідність структурованого підходу для забезпечення захисту інформації та збереження конфіденційності для потрібних даних та конкретних програм.

1.3. Постановка задачі дослідження

Згідно проведеного аналізу відомих методів та стратегій необхідно:

1. Виділити недоліки відомих методів та стратегій, систематизувати їх.

2. Розробити модель базової архітектури системи для покращення конфіденційності зберігання великих даних.

3. Метод збурення розподілених даних.

4. Алгоритми збурення даних із використанням локальної диференціальної конфіденційності для великих даних

5. Здійснити реалізацію азпропонованих рішень.

6. Провести експериментальні дослідження з розробленими засобами.

1.4. Висновки до першого розділу

Масштабні порушення конфіденційності великих даних є відкритою проблемою у світі не тільки у простих користувачів, які користуються ПК і наражають свої дані на небезпеку, нехтуючи базовими правилами, а і приватні особи, фірми та корпорації, у яких конфіденційність даних стоїть чи не на першому місці, для яких вивчення та удосконалення підходів конфіденційності є пріоритетним питанням. Регламентовано попередні умови, що використовуються у підходах, що застосовуються для вищевказаних цілей. Було досліджено деяку літературу та думки, які пов’язані з машинним навчанням, і у якій зроблено акцент на конфіденційності даних. Між слід підмітити, що є і хороші та надійні способи захисту даних, а є такі, що вкрай знижують продуктивність, що згубно впливає на швидкодію обміну даними під час тих чи інших операцій, що виконуються.

2. МОДЕЛЬ БАЗОВОЇ АРХІТЕКТУРИ СИСТЕМИ ДЛЯ ПОКРАЩЕННЯ КОНФІДЕНЦІЙНОСТІ ДЛЯ ВЕЛИКИХ ДАНИХ

2.1 Архітектура базового алгоритму зберігання конфіденційності великих даних та інтелектуальний аналіз

Питання збереження конфіденційності сьогодні є ключовим через те, що досить часто користувачі нехтують базовими правилами, а розробники алгоритмами, які могли би цю конфіденційність забезпечити. Для більш детального розгляду проблем безпеки та ефективності методів конфіденційності, які із високим ступенем достовірності можуть вплинути на стан інформації в цілому розглянемо два методи. Вони являються не просто методами, а двома ефективними та широкозастосовуваними алгоритмам, які є засновані безпосередньо на оптимальних геометричних перетвореннях. АЗК (алгоритм захисту конфіденційності) та АЗРД (алгоритм збурення розподілених даних) – їх назва. Перший – був протестоваий на ефективність, масштабованість і безпосередню стійкість. Також його розгляд був сфокусований на точному використанні більш як восьми наборів даних та до п’яти наборів класифікації цих даних включно. Перевищення швидкості виконання, впевненої стійкості до атак і великі точності при класифікації великомасштабних даних із збереженням конфіденційності показав масштабований алгоритм АЗК. В той же час похідний від АЗК АЗРД як зображено на рисунку 2.1 – розглядає питання як саме може бути розширений АЗК для роботи, включаючи більшість сфер та включаючи можливість роботи саме у розподіленій системі, що підтримує автоматизоване машинне навчання. У більшості випадків корпорації функціонують із великою кількість великих даних, які не відповідають ресурсам стардартного ПК, у яких часто недостатньо чи обчислювальної потужності ЦП чи мала кількість оперативної пам’яті чи взагалі імовірні проблеми із охолодженням, які чи користувачі чи навіть системні адміністратори не встигають лагодити завчасно. А якщо система не може витримати навантаження від простого захисту даних, то і функціонувати вона без аномалій та перевантажень не зможе. Як наслідок, можемо зробити висновок, що

без спрощення алгоритму чи послаблення обмежень у складності базових обчислень в алгоритмі застосовувати алгоритм нераціонально. Тому прийнято рішення розподілити затрати ресурсів на цей алгоритм у відповідному розподіленому середовищі. Безумовно, цей розподіл відбувається як і на пристрої із обмеженими ресурсами, частина із яких перелічені вище, так і на ПК чи серверах із дуже-високою продуктивністю роботи. Постійні дослідження та різні експерименти над спрощеним методом шифрування, показали, що він забезпечує не тільки беззаперечну стійкість до атак, але і чудову оптимізацію та відсутністю аномалій відносно роботи ПК чи сервера під час експериментів. Це доводиться тим, що під час виконання алгоритму у фоні системи працювали в нормі та ресурси витрачалися не тільки на захист конфіденційності даних. Слід взяти до уваги, що подальші експерименти під час федеративного навчання із збереженням повної та надійної конфіденційності даних показали, що спрощений метод є чи не найкращим рішенням для безпеки даних та забезпечення максимального збереження їх від рук третіх осіб та зловмисників, які можуть використати дані у власних інтересах незалежно від алгоритму. При цьому спостерігалися відсутні стрибки витрат ресурсів під час захисту даних під час розподіленого машинного навчання.

Рисунок 2.1 – Розподіл методів і назви основних алгоритмів

Збільшення кількості і доступності літератури допомагає робити революцію у великому списку сфер життя. Із цього списку можна відокремити актуальну на сьогодні охорону здоров'я та в загальному цю сферу. Винайдення нових технологій діагностики людини, на основі якого можуть бути виготовлені інноваційні препарати для покращення чи забезпечення життєдіяльності людського організму. Моніторинг ситуацій в регіонах та швидке виявлення, яке сприяє реакції та вирішенні ситуацій в суспільстві в режимі реального часу вірусології, а саме спалахів вірусів інфекцій та нових видів бактерій, які несуть небезпеку не тільки локальним містам чи країнам, але і всьому населенню всієї планети. У цьому розділі розглянуто конфіденційність статичних даних у особливо великих розмірах. Представлено два види збурення даних. Більш розширеним є другий алгоритм, який призначений для вирішення проблеми розподілу даних саме у розподіленому середовищі машинного навчання. Хоча у будь-якому випадку через диференційну конфіденційність, представлення у порівнянні із попередніми моделями, може призвести до низької корисності під час обробки великої кількості даних. Це якщо взяти до уваги із точки зору просунутої аналітики, зокрема і в системі охорони здоров'я. Низька корисність алгоритму у цій системі вкрай небажаною та критично важливою для її першочергового вирішення.

Перший підхід названий алгоритмом збереження конфіденційності (АЗК) для великих даних. Це є абсолютно нова модель конфіденційності, яка є гарантованою емпіричною властивістю. Цей підхід вартий уваги, бо він застосовує для подальшої випадковості перш за все випадкове відображення осі та багатовимірне обертання півплощини з наступним рандомним розширенням, а вже після – випадкову перестановку кортежів. Новим методом є рандомізоване розширення, який базується на збільшені позитивного чи негативного впливу до певного екземпляру даних. Як підмічено у вступі – проблема кількості ОЗУ не є новою та є важливою для ефективності рішень питань конфіденційності, тому зауважимо, що витрати пам'яті першого методу порівняно близькі до інших рішень. Це забезпечує стабільну роботу пристрою, швидкий обмін даними та безперечну стійкість до атак, яка гарно поєднана із збереженням всієї повноти доступу до інформації.

Підтримка найкращого емпіричного захисту конфіденційності за допомогою шляхів визначення глобально-оптимальних параметрів збурення з дотримкою принципу Ф – розподілу даних.

Але не все можливо передбачити, для прикладу він не розглядає доступні дані, які є розподілені у машинному навчанні. Як наслідок було вивчено нами питання, яке є тісно пов’язаним з розподілом, а аналізом даних. Все це для узагальнення та розробки оптимального рішення і для збереження конфіденційності для зазначеного розподіленого навчання.

Другим методом є простий розподілений алгоритм збереження конфіденційності та цілісності даних із використанням локального механізму збурення даних. Збурення даних підходу до розподіленого машинного навчання дозволяє збурювати великі набори даних, якими необхідно ділитися між розподіленими суб'єктами без витоку конфіденційності.

У розподілених сутностях таких, як пристроях edge/fog фактичне збурення даних проводиться локально з використанням глобальних параметрів збурення, згенерованих у центральному вузлі координації. У такий спосіб, цей підхід обмежує передачу вихідних даних через мережу, яка може бути атакована зловмисниками у будь-який момент обробки інформації.

Глобальна генерація параметрів обурення запевняє, що точність і стійкість до атак збурених даних стабільна. Спочатку було протестовано на шести наборах даних, отриманих із сховища даних. Отже, підхід до розподіленого машинного навчання забезпечує відмінну тоність класифікації, стійкість до атак та високу ефективність розподілення машинного навчання в умовах великих даних.

Алгоритм збереження конфіденційності (АЗК) для великих даних застосовує збурення даних із великою кількістю геометричних перетворень та регулярними обертаннями та з подальшим випадковим розширенням та випадковою розстановкою кортежів. За основою досліджень запропонованої моделі конфіденційності під назвою Ф – поділу визначено, що алгоритм є оптимальним для базового захисту конфіденційності. З погляду атак, що руйнують дані, пошкоджують сектори даних, які зберігаються і на постійних, і на змінних носіях.

АЗК який призначений для великих даних досягає шляхом перебору та вибору одних із найкращих можливих параметрів збурення даних. Основою для збурення слугують вхідні набори даних. Без адміністрування більшість алгоритмів просто не мають сенсу, тому, відповідно, АЗК забезпечує розподіл на користувачів, власників чи адміністраторів наборів даних відповідно до їх ролей у тій чи іншій сфері (рисунок 2.2), де ці дані застосовуються відповідно правил. Це забезпечують і додаткові релізи версій вихідного набору даних. Отже, вихідний набір буде всього-на-всього недоступним для користувачів, і не зможе бути доступним за жодних обставин та причин.

Рисунок 2.2 – Приклад розподіленої організаційної структури

На рисунку 2.2 зображено екосистема охорони здоров'я яка територіально розподілена між кількома місцями. Система охорони здоров'я може мати безліч розподілених філій, що сприяють та збирають безліч даних про охорону здоров'я, включаючи дані датчиків інтернету речей. Центральний орган у переважній більшості випадків координує роботу розподілених лікарень з погляду з погляду підтримки цілісності даних підтримки широкого спектра аналітики. Центральний орган (дослідний центр) також відповідає за обмін даними з третіми сторонами для підвищення інтелектуальності та якості обслуговування пацієнтів.

Другим пунктом у цьому розділі запропоновано систему збереження

конфіденційності, як підхід до розподіленого машинного навчання АЗРД. Цей алгоритм є розподіленим алгоритмом, що сприяє збереженню конфіденційності, який використовує збурення даних. Алгоритм підтримує збурення великих даних із підтримкою їх поширення між розподіленими об’єктами без порушення конфіденційності. Фактичне збурення відбувається в розподілених сутностях локальних пристроїв включаючи локальні правила та глобальні параметри збурень. Тому, АЗРД обмежує вихідні дані для передачі (до збурення) через мережу, через необхідність захисту інформації від третіх сторін, які можуть несанкціоновано отримувати доступ до даних, та діяти із ними у своїх, злочинних цілях для отримання власної вигоди. Генерація глобальних параметрів збурення інформації із високою імовірністю може гарантувати відсутність погіршення точності чи стійкості до атаки збурених даних. Як наслідок АЗРД успішно запобігає витоку конфіденційності.

АЗК збурює набір даних із використанням багатовимірних геометричних перетворень, випадковим перетасовуванням кортежів, Та із новим доповненням шуму, яке розглядається в розділі 4.

Важливим є використання метрики конфіденційності із багатьма колонками для оцінки запропонованого методу. На рисунку 2.2 схематично відображено основний потік та архітектуру запропонованого алгоритму збурення. Припускаючи, що всі атрибути однаково важливі, нормалізація застосовується до матриці даних вже на початку збурення. Тому чим вищий рівень конфіденційності збурених даних, тим важче оцінити початкові дані.

АЗК достатньо складний, у випадку, якщо він може продовжувати задовольняти тим самим вимогам конфіденційності, що використовується у моделі конфіденційності після неодноразового незалежного застосування алгоритму. Відповідно, щоб покращити компонування та випадковість збурень у алгоритмі, шум, який був отриманий із випадкового нормального розподілу (враховуючи середнє відхилення 0,3 відповідно середнього значення, що рівне 0) сумується із даними відповідно методу випадкового розширення. Тому введено шум, який так чи інакше посилює позитивність чи негативність певних значень, де нулі не

піддаються жодним змінам, що зображено на рисунку 2.3.

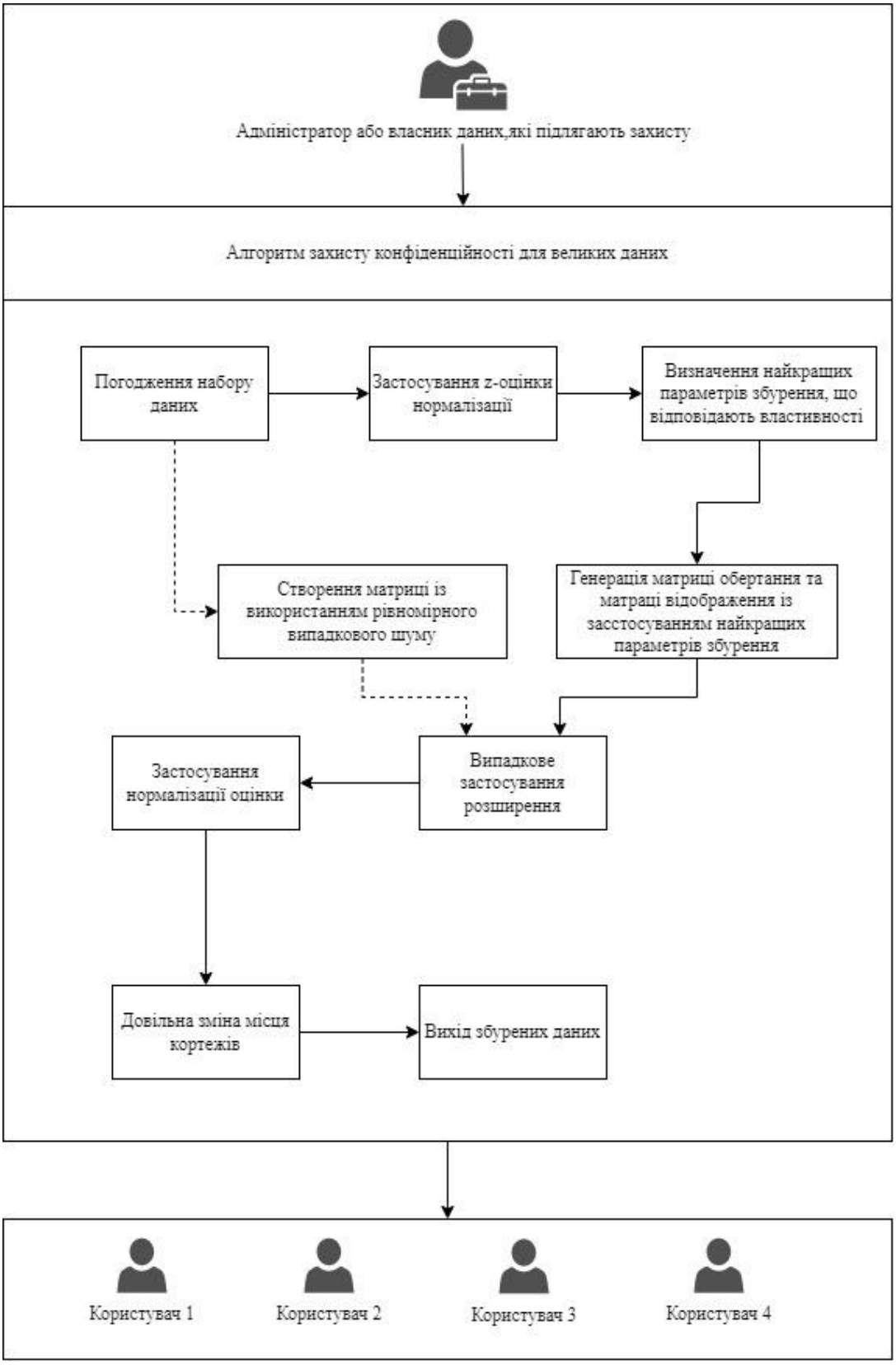


Рисунок 2.3 – Схематичне зображення потоків та архітектури АЗК для великих даних

Для розширення вказаної ідеї розглянуто дисперсію різниці між збуреними і незбуреними наборами даних. Атрибут, який розглянуто у цьому методі повертає мінімальну дисперсію для різниці, та розглядається як мінімальна гарантія збереження конфіденційності даних. Якщо X^p відображає збурений ряд даних атрибуту X, то рівень конфіденційності методу збурення варто визначати:

$$Var(P), \text{ при } P = (X^p - X). \tag{2.1}$$

В такому випадку має місце рівність:

$$Var(P) = Var(p_1, p_2, \dots, p_n) = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2. \tag{2.2}$$

Найкращі параметри збурення визначаються наступним чином. Під час кожної ітерації відбувається максимізація значення φ. Це власне і сприяє створеню значення ϕ. Як зазначено у рівняннях нижче. При цьому вісь відбиття змінюється від 1 до n (кількості атрибутів). Кут повороту при цьому збмінюється у діапазоні значень від 0° до 179° Таким чином повертається найбільше значення ϕ для мінімальної гарантії конфіденційності φ.

$$\phi = \max([[\varphi_j]_{j=1}^{179}]). \tag{2.3}$$

Наступним кроком є генерація матриці обертання за допомогою найкращих параметрів збурення. АЗК записує оптимальний кут повороту і вісь відбиття на Ф. Наступне рівняння слід використовувати відповідно оптимального кута.

$$RF_{\overline{ND}} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}_{(n+1)(n+1)} \quad (2.4)$$

Далі слід застосувати складене перетворення відображення, трансляції та обертання до нормалізованої матриці з використанням матриць оптимального відображення та оптимального обертання.

Обґрунтування та технічна новизна. Алгоритм збереження конфіденційності для великих даних застосовує геометричні перетворення із достатньо оптимальними та оптимізованими параметрами збурення при цьому збільшує випадковість за допомогою випадкових розширень та відповідних перестановок кортежів. АЗК визначає конфіденційність так, що результуючий набір даних має високий рівень різниці, який є оптимальним у порівнянні з вихідним набором даних. Це все відповідно нашого розподілу, що використовується в АЗК Оскільки відбувається мінімізація простору пошуку, та знаходження найкращої моделі можливого збурення даних, можна стверджувати, що вона є корисно для конкретного набора даних.

Оцінку продуктивності зосереджено на корисності, яку можна визначити як зручність використання або як ефективність використання збурення даних. Досліджено корисність АЗК з точки зору класифікації. Таким чином АЗК ставить за мету довести комфортність використання, яка заснована на тимчасовій складності, боротьбу із надмірними витратами пам’яті, масштабованості та упередженнях, що охоплюють основу оцінки. Вихідні набори даних були збурені за допомогою АЗК, геометричним та обертальним збуренням. Тому, це дало можливість порівняти результати за допомогою непараметричного статистичного тесту.

2.2. Метод збурення розподілених даних

Ризик викриття інформації методом збурення у стандартних ситуаціях вимірюється шляхом кількісної оцінки. Аналіз аспектів не є маловажливою частиною роботи, тому було проаналізовані такі із них: гарантії суворої конфіденційності даних, стійності до атак чи аномалій, які спричинені навантаженням та кількісного регулювання ентропії (переважно в сторону збільшення). На основі незалежних компонентних аналізів АЗК був протестований. Також важливо було зробити тест на відомі атаки введення-виведення, успішність якого є базовим критерієм для роботи у будь-яких, навіть самих агресивних та несприятливих середовищах для роботи. Стійкість АЗК до атак порівнювались із обертальним та геометричним збуренням за допомогою непараметричного статичного тесту, який є загальновідомим і не поступається надійності відносно аналогічних таких самих тестів. Особливістю АЗК – це збурення даних за найкращих параметрів та коефіцієнтів для досягнення оптимальної конфіденційності поділу. Це було доведено тестуванням наших параметрів, які виявляли опір реконструкції даних та їх елементів. Для покращення вимірів нагромадженості та забрудненості даних було досягнуто збільшенням ентропії. Але слід брати до уваги, що збільшення ентропії, яке допомагає досягати нам вищого її значення матиме за мету збільшення кількості відповідних записів. Збільшення ентропії – чудовий показник для визначення складності відновлення початкових вхідних даних. Як наслідок – використання методу у якому збільшено ентропію чудово допомогли перевірити дані на домішки та подальші оцінки конфіденційності, що забезпечуються АЗК для великих даних.

У цьому параграфі присутній базовий опис кроків, щодо розробки алгоритму збурення розподілених даних (АЗРД). Власне цей алгоритм є розширенням концепцій, які використовувались у АЗК.

Хоча наш АЗК розроблений так, щоб працювати із великими даними, але слід зауважити, що АЗК не підходить для розподіленого машинного навчання. Він не враховує розподіл даних, які є доступні у сценарії, коли ресурси пристроїв є обмеженими, як показано на рисунку 2.2. Перенос збурення даних на розподілені

гілки до того, як дані залишать локальну мережу – це основна мета АЗРД. Однак при цьому варто передбачити мінімальні втрати глобальної корисності. Для досягнення цієї мети наш дослідницький центр, та розподілені філії займаються проведенням обмінів параметрів збурення. Оскільки збурення даних здійснюється за допомогою глобальних та оптимальних параметрів, наш АЗРД дотримується правильного балансу між корисністю та конфіденційністю. Для пояснення використано слово вузол та сутність альтернатив. Тим не менш, виражаючись конкретно, вважається, що поняття "сутності" являється повноцінно наповненою організацією, тоді як поняття вузла представляю собою одну або більше обчислювальних / обробляючих приладів всередині цієї сутності.

На рисунку 2.2 зображено приклад сценарію в якому АЗРД було інтегровано до розподіленої системи охорони здоров’я для забезпечення збереження конфіденційності даних. Зображено різні рівні наборів. Найнижчий рівень представляє розподілений набір суб’єктів, у нашому випадку (лікарні) та їх взаємодія з АЗРД. На малюнку показано, що розподілені компоненти АЗРД інтегровані в кожну із розподілених лікарень для виконання збурень даних.

Запропонований алгоритм делегує обслуговування даних та збурення розподіленим суб’єктам, залишаючи центральному лише генерацію глобальних параметрів збурення. АЗРД є можливість використовувати для збурення даних охорони здоров’я. Таким чином, розподілені філії не повинні передавати конфіденційні дані третій особі, яка може бути потенційним зловмисником і використати дані у своїх зловмисних цілях. Проте, беручи до уваги чітку координацію запропонованого алгоритму при генерації параметрів (глобальних) дані можуть забезпечити достатньо високу корисність.

Федеративне навчання над збуреними даними. Федеративний модуль набуває чинності після збурення, а кожен із розподілених об’єктів буде використовувати локальні дані (збурені). Все це для навчання локальної моделі машинного навчання (для прикладу, глибока нейронна мережа) протягом певної кількості локальних епох. Після завершення яких – розподілені сутності займаються надисланням модель в репозиторій для створення глобального

представлення моделі. Центральний орган (сервер) передає далі агреговані параметри до розподілених сутностей для оновлення локальних моделей для їх узагальнення. Федеративна установка проведе достатню кількість місцевих епох і раундів федерації до моделей машинного навчання на основах вимог організації (для прикладу виробництва великих потоків даних), архітектури моделей машинного навчання та властивості набору даних.

У цьому розділі запропоновано ефективний алгоритм збереження конфіденційності (АЗК) задля збереження окремих одиниць у тих великих даних. АЗК застосовує збурення, ґрунтуючись новій моделі конфіденційності під назвою Ф – поділу. Доведено, що Ф – поділ гарантує емпіричну конфіденційності. Подальші введення механізмів випадкового шуму (переважно нових) мають назву випадкового розширення. Це зумовлено досягненнями під час систематичних підходів до параметрів не тільки збурення даних для випадкового зображення осі, але і з подальшим введенням нового механізму, який додає вищесказаний шум. Цей алгоритм переважає над тимчасовою ефективністю, а результати точності класифікації збуреного набору даних близькі до результатів класифікації вихідних даних. Експлуатаційну надійність та практичність АЗК проаналізовано з погляду тимчасової складності, масштабованості та споживання пам'яті. Наступне, що було проаналізовано – це конфіденційність АЗК та складових, таких як стійкість до атак, забезпечення конфіденційності та збільшення інформаційної ентропії.

Оскільки було проведено відповідні класифікації даних експериментів наш АЗК перевершив два споріднені методи такі як геометричне та обертальне збурення. Найвищий середній ранг, показаний в АЗК, показує підтвердження аргументів, які отримані в результаті попередніх тестів. Не все є так ідеально, тому, зауважимо, що данні точності класифікацій даних та їх наборів, які є збурені за допомогою АЗК поступається виключно точності відношення вихідному набору даних відносно вхідного. Аналіз та аналітика тимчасової складності алгоритму АЗК дорівнює $O(n^3 \times m)$, де число n – це кількість атрибутів, а m – позначає точну кількість кортежів. Часто набори даних зростають із різних причин, але зазвичай це відбувається через зростання кількості нових кортежів. Однак, зауважимо, що

при будь-яких умовах аналізу алгоритму лиш кількість атрибутів залишається постійною та суворо незмінною. Придатність алгоритму для збурення великих даних підтверджується цим.

Тестування АЗК дало свої результати. Проаналізувавши час, доведено, що алгоритм досягає завершення часу збурення набагато швидше, аніж обертальне та геометричне збурення. Вони ж не сходяться протягом максимального часу (100 одиниць часу) які були відведені для тесту на окремому ПК, що зайвий раз доводить, що новий метод без ніяких проблем може бути використаний для даних величезних кількостей та масштабів з вкрай низькими витратами енергії та потужності. А власне вимоги до АЗК є досить помірними. У парі із обертальним збуренням спостерігалася різниця у використовуваній пам’яті відповідно геометричного обурення у меншу сторону. У відповідності до непараметричного статичного тесту АЗК використовує менше пам’яті, ніж обертальне збурення. Однак, периферійні значення вказують на те, що різниця не є істотною. Набір даних, який містить собі велику кількість кортежів та значну кількість атрибутів може добре працювати із запропонованим методом.

Можна зробити висновок, що роблячи отримані дані непридатними для описового статистичного аналізу, збурення змінює описову статистику. У результаті зловмисник не може проводити атаки щодо виведення бази даних на основі описової статистики. АЗК не впливає на взаємозв'язок між атрибутами і має нижчі рівні зсувів типу B та C. Подальший аналіз зсувів показав, що поділ між вихідними та збуреними атрибутами схильний до впливу. Це унеможливлює пряме зіставлення між атрибутами.

Для цього способу оптимальними параметрами збурення набору даних є забезпечення конфіденційності Ф - поділу. Запропонований є рандомізованим розширенням, яке ще більше знижує унеможливлення відновлення вихідних даних. Експериментуючи зі стійкістю до атак, АЗК показав надійний захист від атак. АЗК застосовує зворотнє стандартизоване оцінювання нормалізації після перетворень, тому діапазони значень кінцевих атрибутів розташовані в межах значень вихідного набору даних. Насамперед, це може знизити ймовірність атак, оскільки

зловмисники не зможуть відрізнити вихідний набір даних від збуреного.

АЗК для великих даних протистоїть різним атакам. Незалежний компонентний аналіз для вводу-виводу демонструє рівень вразливості АЗК. Оскільки алгоритм виконує випадкові розширення та збурення відповідно до отриманого набору даних. Незначний вплив на конфіденційності лише незначно впливають на виконувані випадкові перемішування отриманих наборів даних. Отже, ті атаки, що допускають використання втрати конфіденційності через багаторазове звільнення не становлять величезної небезпеки. Цим забезпечується емпірична гарантія сумісності системи. АЗК не намагається забезпечити мінімальні збитки збереження інформації у загальному відповідно і атаки, які фокусуються на цих даних також не становлять небезпеки. Однак ми, як і будь-які інші науковці не можемо стверджувати, що наш АЗК для великих даних невразливий.

Зрештою, у збурених даних досліджено збільшення ентропії. Складнішим процес відновлення вихідних даних робить збільшення ентропії набору даних . АЗК показує високі результати у даному відношенні, збільшуючи ентропію. Для збереження конфіденційності великих даних АЗК забезпечує масштабоване та ефективне рішення, яке може бути ефективно використане.

Було розширено АЗК, щоб подати чіткий та ефективний розподілений механізм збереження конфіденційності. Названо механізм АЗРД для розподіленого машинного навчання. N – мірні геометричні перетворення з наступним цілком випадковим розширенням застосовує наш АЗРД, а Ф- поділ – як базова модель конфіденційності для видобування оптимальних параметрів збурення. Ф – поділ не виключає також і мінімальну конфіденційність для вхідних даних. Було протестовано ці методи на стійкість до атак за точністю класифікації та тимчасовою складністю. Затримки зв’язку на розподіленому збурені АЗРД та вплив кількості розподілених вузлів на класифікацію в процесі машинного навчання – теж були проаналізовані.

Відповідно до комплексної аналітики АЗРД виводить складність з часом $O(n^4)$ до центральної сутності, де n – це кількість атрибутів. Однак $O(n^4) = O(k)$, оскільки n залишається константою (k) для даного параметра. Тому, споживання

кількості часу є незалежною від кількості екземплярів (кортежів), які вводяться. Часова складність показана у наступній сутності $O(n^3m) = O(m)$, де n – число атрибутів, яке є постійною величиною для нашого набору даних, в такому випадку (m - позначає кількість кортежів). Нові параметри відносно не часто додаються з тією ж самою швидкістю, що і зростають дані. Це виключно при розгляді фіксованої установки. Із високою часткою імовірності вони залишаються незмінними. Тому, оптимальним для машинного навчання (яке є розподіленим) є те, що для даного сценарію власне кількість часу, що виділяється та витрачається розподіленим вузлом – ростиме лінійно. Звісно із забезпеченням конфіденційності. АЗРД за допомогою емпіричних даних показує, він в змозі забезпечити точність класифікації за централізованими алгоритмами. І відповідно забезпечити кращу продуктивність у порівнянні із геометричним та обертовим збуренням. АЗРД у порівнянні із АЗК та його централізованим підходом забезпечує трохи нижчу точність класифікації. Це зумовлено тим, що АЗРД надає підвищений рівень випадковості за різних умов. Адже всі кроки випадкових розширень саме окремо виконуються розподіленою структурою . Ці дії виконуються для покращення випадкової генерації даних. Емпіричні дані доводять, що на точність класифікації (а саме при розподіленому машинному навчанні) кількість клієнтів немає впливу на класифікацію, який можна помітити. Кількість часу потрібного для уподібнення моделей одна до одної росте в залежності від кількості клієнтів. Одна із цих особливостей робить АЗРД з допомогою федеративного навчання стає ідеальним рішенням для збереження конфіденційності.

АЗРД використовує Ф – поділ як базову модель конфіденційності. Використання саме цього поділу дозволяє оптимально збурювати дані для цього екземпляра. Не мало важливо те, що АЗРД є забезпеченим високим рівнем конфіденційності за рахунок зниження ймовірності атак реконструкції даних. Цю властивість підтримують геометричні перетворення даних та додавання випадкового шуму. Через те, що зворотна нормалізація кінцевого набору даних до вихідних значень та їх діапазонів атрибутів, зловмисники нездатні ні за яких обставин виділити чи відрізнити вхідні дані (оригінальні) від збурених наборів

даних. Стійкість до атак забезпечена краща і більш оптимізована у порівнянні із геометричним та обертовим збуренням. АЗРД цілком гідно забезпечує та гарантує кращу стійкість до атак у порівнянні із АЗК, який призначений для великих даних. АЗРД напряму впливає на підвищений рівень випадковості, оскільки поліпшення випадковості даних як випадкове розширення виконується чи не кожною розподіленою структурою окремо. Слід зауважити, що під час федеративного навчання ні в якому разі не відбувається поділ збуреними даними між розподіленими клієнтами чи сервером. У результаті, атак відновлення даних на збурені дані не буде, відповідно, стійке поняття конфіденційності порушених даних спричиняє високу конфіденційність підготовлених моделей. Простіше кажучи якщо корпорації чи будь-які заклади (наприклад охорони здоров'я) наш метод, то АЗРД без проблем може стати цілком зваженим та оптимальним рішенням, що допоможе зберегти і забезпечить повну конфіденційність даних та машинного навчання, що записують, обробляють та діляться великими обсягами даних, які власне і розгорнуті територіально розподілених системах.

2.3 Висновки до другого розділу

У цьому розділі описано новий метод збурення, який називається АЗК. Він слугує для того, щоб вирішити проблеми ефективності, конфіденційності, масштабованості та зручності використання. АЗК додає конфіденційності даним, використовуючи модель, яку запропоновано раніше і яку назвали Ф – розподілом, який забезпечує достатню гарантію конфіденційності завдяки ефективному вибору найкращих параметрів збурення для збурень великих даних АЗК. Асимптотична складність АЗК є лінійною. І слід взяти до уваги, що кількість атрибутів є значно більшою, ніж кількість атрибутів для даного набору даних. Зумовлено тим, що кількість атрибутів є часто контрастною для конкретного набора даних. І саме ця властивість дозволяє нашому АЗК працювати із великими наборами даних. Також цей алгоритм забезпечує низьке споживання часу на його виконання для збурення великих даних. Теж немаловажливим є те, що точності класифікацій перевершують

методи випадкового обертання та геометричного збурення і є вкрай близькими до точності класифікації вихідного методу. Додатковими перевагами алгоритму будуть можливість композиції та гарантія конфіденційності.

Сьогодні існує велике різноманіття систем, які є сучасними та прогресивними. Для приклада може слугувати банківська справа чи вищезгадана охорона здоров'я. Ці системи є часто обмежені. Зокрема це відбувається у аналітичному використанні належних механізмів для обміну даними із відповідним забезпеченням конфіденційності для аналітики. Оскільки наш алгоритм є композиційним без вагань розширено АЗК і представили алгоритм розподіленого збурення АЗРД. Він може забезпечити конфіденційність машинного навчання, яке є розподіленим. У запропонованій системі АЗРД весь контроль генерації даних та глобальних параметрів обурення належить центральному контролюючому органу, в той час як локальне збурення даних можуть бути проведені генерацією глобальних параметрів. Обчислювана складність АЗРД полягає у виконуваній центральній сутності, що відповідає $O(k)$ для кількості екземплярів, де k – постійна змінна.

3 ВИЯВЛЕННЯ ПРОБЛЕМ КОНФІДЕНЦІЙНОСТІ ДЛЯ ВЕЛИКИХ ДАНИХ

3.1 Безпечний алгоритм збурення даних із використанням локальної диференціальної конфіденційності для великих даних

Збереження конфіденційності з високим ступенем надійності та корисності визначено пріоритетним напрямком для аналізу та вивчення. Створення нових та ефективних механізмів – саме це потрібно випускати для забезпечення цієї пріоритетності. В даному розділі досліджено та представлено питання конфіденційності потоків даних, а наш метод названо методом збурення потоку даних (МЗПД). Забезпечення корисності представленим методом є вищою, ніж аналогічні а точності класифікації є дуже близькими до вхідних потоків даних. МЗПД забезпечує більш високу стійкість до атак що ставлять за мету не тільки викрасти дані, щоб використати у власних (можливо і зловмисних) цілях, але і до тих, що намагаються після атаки залишити пошкоджені дані для унеможливлення подальшого використання. Хоча наш метод займається забезпеченням лінійної складності для кількості екземплярів, він також забезпечує $O(n^3)$ складність числа атрибутів, що позначає значення (n). Тому, в цьому розділі наведено аналогічний алгоритм БАЗД (Безпечний алгоритм збурення даних із використанням локальної диференціальної конфіденційності), який засновано на одній із вже наявних інтерполяцій, що забезпечує чудову збалансованість між кількома факторами такими як: конфіденційність та корисність, масштабованість та високий рівень продуктивності, який фактично незалежний від кількості даних. Порівнюючи БАЗД з результатами порівняння емпіричних алгоритмів та існуючими алгоритмами збереження конфіденційності, можна побачити, що БАЗД перевершує їх за швидкістю виконання масштабованості та стійкості до атак. Особливу увагу надано точності виконання алгоритму. Гнучкість у виборі найкращих можливих параметрів на відміну від інших методів забезпечує БАЗД. Такі параметри як кількість шуму, що додаються, можуть бути адаптовані до набору даних та домену.

На сьогодні існує високе різноманіття пристроїв інтернету речей. Їх

доступність та відносно невисока вартість дозволила пришвидшити розвиток цієї сфери, а потоки все можливих датчиків стають найважливішими джерелами даних. Ризики порушення конфіденційності та недоторканості спричинені саме цією доступністю датчиків. Програми краудсерсингу дедалі частіше починаються задіюватися, а повсюдний збір даних у них зазвичай включає особисту інформацію включаючи звички, вподобання та способи життя користувачів, витоки якої до третіх осіб викликають побоювання із сторони звичайних користувачів та розробників щодо конфіденційності. Інформація, яку витягнуто із потоків даних не повинна бути пов'язаною із окремими особами. Але зауважено, що поділу цими потоками неможливо уникнути. У питаннях потоків даних виділено основні проблеми, серед яких найбільш важливими є ефективний збір та зберігання даних, їх обробка та захист конфіденційності безпека включаючи запобігання витоків інформації.

У цьому розділі розглянуто ці проблеми потоків даних і запропоновано два рішення. Перший – метод збурення потоку даних на основі алгоритму конденсації використовується для різних цілей. Його може бути використано для збурення високошвидкісних потоків даних, особливо це актуально для пристроїв інтернету речей та їх датчиків. У алгоритмі передбачено, що кількість даних може стрімко зростати і це не повинно заважати можливостям роботи з ними. Перший крок методу – це проведення формування власне однорідних груп (на основі відстаней між кортежами та повним врахуванням їх) і використання властивостей груп для створення матриці обертання, яка призначена для виконання збурення та виконання кожної групи. Доведено під час тестів на типових наборах даних ефективність запропонованого методу. А точності класифікації теж було доведено за допомогою різних алгоритмів класифікації. Запропонований метод є вкрай ефективним у класифікації потоків даних із передбаченим та необхідним збереженням конфіденційності інформації. На відомих атаках на введення чи виведення інформації та на незалежних компонентних аналізах було продемонстровано здатності алгоритму захисту конфіденційності. Показано перевагу запропонованого методу як точності класифікації і у конфіденційності

даних у порівнянні із випадковим збуренням та конденсацією даних. Очевидних переваг у алгоритмі достатньо. Відмінним рішенням для потоків даних, що вимагають моментальної чи оперативної обробки за умов обмеженої кількості ресурсів пристрою, її роблять характеристики ефективності точності та високої швидкодії.

Між іншим, МЗПД забезпечує лінійну складність для визначеного числа кортежів по формулі, що вказана вище для числа атрибутів (n). Число атрибутів для більшості потоків даних залишається постійними та незмінним. Враховуючи широкомасштабність розвитку компонентів інтернету речей та їх складових передбачено і часткові зміни їх зважаючи на нестабільні покази датчиків, які провокують наступні обчислення для забезпечення точних результатів. Наприклад, у потоках інформації із датчиків та екосистем інтернету речей (для прикладу, розумне місто чи сфера охорони здоров'я), які додають все нові і нові датчики для додавання новітніх функцій до вищеназваних систем. Отже, питання пошуку ефективності рішень є відкритим. Саме тому у цьому розділі запропоновано інший надійний алгоритм, який відрізняється ефективністю та забезпеченням лінійної складності як кількості екземплярів так і атрибутів. Алгоритм отримав назву БАЗД. Цей алгоритм використовує інтерполяцію, яка є полінеміальною та включає у себе поняття диференціальної конфіденційності. БАЗД – це система збурення на основі лінійності та поліномі Чебишева, що дозволяє працювати швидше, ніж аналогічні методи. Використано спільні набори даних, які взято для оцінки ефективності точності та стійності БАЗД до атак. Показано, що результати БАЗД чудово справляються з такими завданнями класифікації великих даних із дотриманням всх стандартів конфіденційності даних в інтелектуальних системах.

3.2 Метод збурення потоку даних

Призначення цього алгоритму - обробка потоків великих даних. У алгоритмі поєднано функції як і якості та ефективності, так і точності збурення, використовуючи властивості конденсації даних. Призначення алгоритму –

зберігання потоку збурених даних загального призначення в потоках даних, які надходять від інтернету речей та їх всеможливих датчиків. (Рисунок 3.1) Як показано на рисунку. Тобто дані, що є вихідними вже готові до зберігання із дотриманням всіх норм конфіденційності і є готові до довготривалого зберігання, пройшовши алгоритм МЗПД.

Рисунок 3.1 – Застосування МЗПД у потоках даних компонентів інтернету речей

На рисунку всі показники датчиків інтернету речей, які перетворюються у дані, групуються та генерують єдиний потік даних. Потік даних перш ніж зберігатися проходить обробку МЗПД.

МЗПД спочатку кластерує дані в кілька однорідних груп. Зауважено, що незалежно від того чи це угрупування чи кластеризація дані так чи інакше будуть використовуватися взаємозамінно. Обробка даних після алгоритмічних маніпуляцій, що направлені на збереження конфіденційності накладається на шматки даних, які здебільшого мають фіксований та незмінний розмір, що включають динамічну підтримку. Після цього відбувається процес генерації коваріаційної матриці для кожної групи включаючи застосування чіткої групової статистики. Далі – створення відповідних груп, які об’єднуються, а кортежі – довільно збурюються та випускаються.

У свою чергу після генерації коваріаційної матриці власні вектори їх визначаються шляхом розкладання $C(G_i)$. Колонки $P(G_i)$представляють власні вектори коваріаційної матриці від попередньої. Відповідні вектори утворюють

власну систему та роблять її ортогональною. Тому, властивості ортогональної матриці має матриця, яка є результуючою для власних векторів $P(G_i)$ конкретної коваріаційної матриці, у них стовпці та рядки є ортогональними.

Отже, $P(G_i)$ зберігає співвідношення:

$$P(G_i) \times P(G_i)^T = P(G_i)^T \times P(G_i) = I, \text{ при } P(G_i)^T \quad (3.1)$$

$P(G_i)^T -$ транспонована матриця $P(G_i)$ і де I – є тотожна матриця. А це прямо показує, що $P(G_i)$-матриця конкретної однорідної групи має всі властивості матриці обертання. Доведено, що отримана матриця є незмінною ортогональною матрицею і не змінила свій статус ортогональної відносно попереднього стану до обробки, не зважаючи на те, що порядок рядків або стовпців цієї матриці змінено. Тому ця матриця, що була переставлена по стовпцям (в результаті результуючого обертання), вестиме себе так, як матриця обертання. Тоді ця властивість була викликана для випадковості процесу збурення за допомогою випадкового перемішування стовпців.

$$C(G_i) = P(G_i) \times \Delta(G_i) \times P(G_i)^T. \quad (3.2)$$

На блок-схемі, що зображена на рисунку 3.2 показано, що після прийняття даних алгоритмом застосовано кластеризацію для однорідних груп. Вони у свою чергу зурюються. У випадку надто великого розміру статичного набору даних і коли він фізично не вміщується в доступну пам’ять, набір даних зазвичай розподіляють на кілька розділів для застосування одного і того ж самого набору кроків (алгоритму) до кожного розділу даних. Ці дії направлені на отримання остаточно збурених даних.

Рисунок 3.2– Блок-схема алгоритму збурення статичної бази даних

Алгоритм для статистичних наборів даних. Вище показано як відбувається збурення в статичному наборі даних, у якому весь набір даних подається в алгоритм лиш єдиний раз. Наприкінці обробки груп, повернені групи зазвичай об'єднуються. Кортежі у свою чергу випадково змінюються місцями та перемішуються для збільшення фактору випадковості у кінцевих наборах даних. Дані дії направлені виключно на підвищення конфіденційності даних. Значення так званих міток (прапорців) в алгоритмі являють собою конфігурації для двох можливих варіантів кластеризації чи угрупування. І тут існує відповідно два сценарії розвитку подій. У випадку першого – користувач який задав k як кількість груп чи кластерів (як вхідне значення), то прапорець буде встановлено в 1 і групування буде проводитися з використанням k – середньої кластеризації, як і було зафіксовано у алгоритмі. Однак при k' - вибрано як вхід для кількості кортежів у групі чи кластері, то прапорець буде встановлено на 0. В нашому випадку 0 за стандартами– стан за замовчуванням. Існують випадки коли розмір групи лише один кортеж. Тому варто передбачити і те, що лише відповідно до конфігурації попередньої групи, яка була найближчою буде здійснено поворот. Але це лише у випадку коли група більш ніж з одним кортежем. Основна мета цього кроку – посилення ефекту збурення. Перемноживши вектор і матрицю тотожності отримано один і той же вектор створюючи нульовий ефект збурення вихідного вектора. Це можна уникнути з допомогою корекції процесу збурення даних.

На блок-схемі МЗПД буферизує кількість кортежів даних під час кожного раунду збурення протягом t кількості ітерацій. Цей алгоритм також виробляє $t \times l$ кількість збурених кортежів даних наприкінці кожних t ітерацій. Цей процес збурення триває доти, поки потік даних не буде припинено.(Рисунок 3.3)

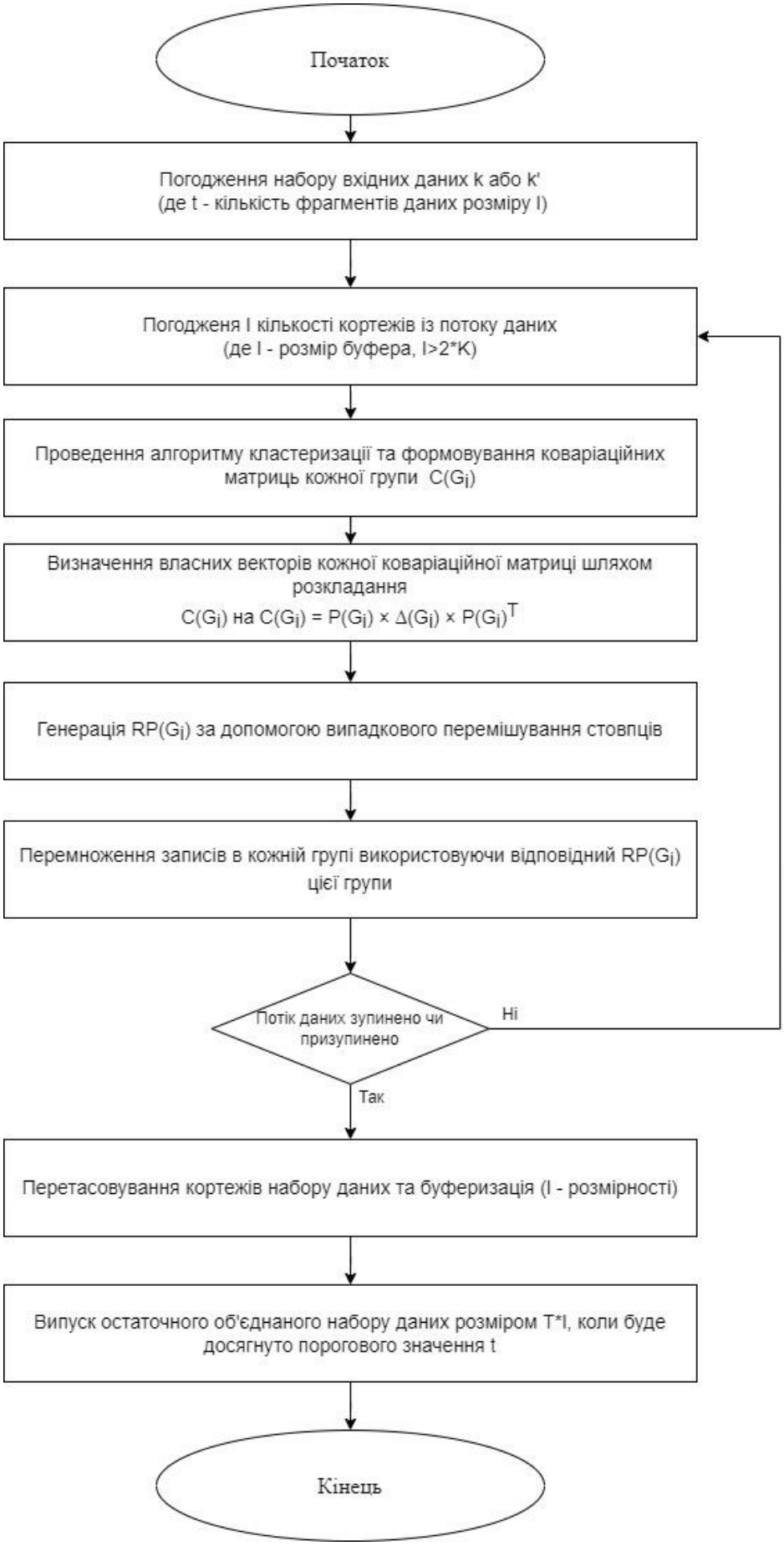


Рисунок 3.3 – Блок-схема алгоритму збурення потоку даних.

Алгоритм потоків даних. Розглянуто випадки коли у разі потоків даних приймається розмір буфера l та поріг (t) для вивільнення даних, а де (t) – кількість блоків даних із визначеним розміром l, які повинні бути вивільнені незалежно від того моменту, як потік даних буде тимчасово призупинено чи остаточно зупинено. Алгоритмом передбачено мінімальну розмірність групи, яка становить два кортежі. Іншими словами накладається наступне суворе обмеження $l < 2 \times k$. Необхідність обмеження заключається у уникненні можливості отримання будь-яких результатів із наявністю кількості груп де деякі з них матимуть єдиний кортеж, а деякі – жодного.

Потоки даних безкінечно й поступово зростають і це робить процес збереження конфіденційності складним. Як показано у алгоритмі, для зменшення складності динамічного режиму, буфер даних розміром l підтримується динамічно. Цей буфер даних розміром l через статичну природу кластеризації дозволяє генерувати кластери з кращою однорідністю. Під час виконання алгоритмом приймаються і буферизуються кількості кортежів. І у якості подальших дій проводиться збурення буферизованих даних. Тоді ці дані звільняються та відбувається процес об’єднання із поточними блоками даних. Зауважено, що буфер може прийняти наступний набір l –кортежів. Враховано теж наявність порогу t, що дозволяє алгоритму напряму уникнути можливості очікування вивільнених даних протягом необмеженої кількості часу.

Незважаючи на те, МЗПД забезпечує ефективне рішення щодо збереження конфіденційності для потоків даних, тимчасова складність цього методу для кількості атрибутів (n) становить $O(n^3)$. БАЗД рекомендовано використовувати для забезпечення більш ефективного рішення, адже він розроблений для збереження конфіденційності потоків і великих даних, що генеруються такими системами, як інтелектуальні кіберфізичні системи.

Однією з цілей було досягнення балансу між конфіденційністю та корисністю, оскільки вони можуть дещо негативно впливати один на одного. Прикладом може слугувати просторове розташування набору даних, яке

потенційно сприяє його корисності в аналізі даних, оскільки на результати, що були отримані за допомогою механізмів аналізу, таких як класифікація даних і кластеризація, часто впливає просторове розташування вхідних даних. Однак слід зауважити, що просторове розташування може бути порушено, коли механізми конфіденційності застосовують такий метод, як рандомізація. Хоч механізми обурення даних покращують конфіденційність, корисність, у свою чергу, може бути зниженою. І навпаки, збільшення корисності може негативно вплинути на конфіденційність.

Для вирішення цих труднощів, БАЗД обробляє дані трьома кроками: 1) визначає чутливість набору даних для калібрування; 2) проводить поліноміальну інтерполяцію з каліброваним шумом; 3) використовує апроксимовану функцію для створення збурених даних. Ці кроки гарантують, що БАЗД застосовує достатню рандомізацію для збереження конфіденційності просторового розташування вихідних даних. Слід зауважити, що БАЗД також використовує поліноміальну інтерполяцію, що супроводжується додаванням шуму і калібрується відповідно до інструкцій диференціальної конфіденційності.

На рисунку 3.4 показано інтеграцію БАЗД у потік даних загального призначення смарт кібер-фізичних систем. На ньому зображені дані, збурені за допомогою БАЗД, що надходять безпосередньо із СКФС. З цього випливає, що дані в модулі зберігання вже пройшли через збереження конфіденційності БАЗД і не містять жодних початкових даних.

Представлено систематичний потік кроків у випадковому генерування даних для отримання належної конфіденційності вхідного сигналу. Алгоритмом прийнято вхідний набір даних, запас конфіденційності, а розмір вікна та поріг (t) визначено як вхідні параметри. Розмір вікна визначає такий саме ту кількість екземплярів даних, що збурюватимуться за один цикл випадкової генерації (рандомізації). Підкреслено необхідність його для підтримки швидкості виконання аналізу чи модифікації після обробки (для прикладу розглянуто збурення даних, класифікація та кластеризація), процеси яких проводяться на рівні із потоком даних. Зазвичай параметр за замовчуванням порогу для статистичних наборів

даних -1. Для попередньоназваного набору даних $t = -1$ повністю ігнорує те, що певна кількість збурених вікон повинна бути вивільнена ще задовго до завершення всього набору даних. У випадку із потоками даних відмічено неабияку користність значень вікна та порогу (t). ООскільки ними підтримується буфер даних та може бути задане певне число для алгоритму значення (t) для вивільнення кожного із (t) вікон. Оскільки потоки даних зростають саме нескінченно у більшості випадків, то зроблено висновок, що підтримка значення (t) є вкрай важливою для потоків даних.

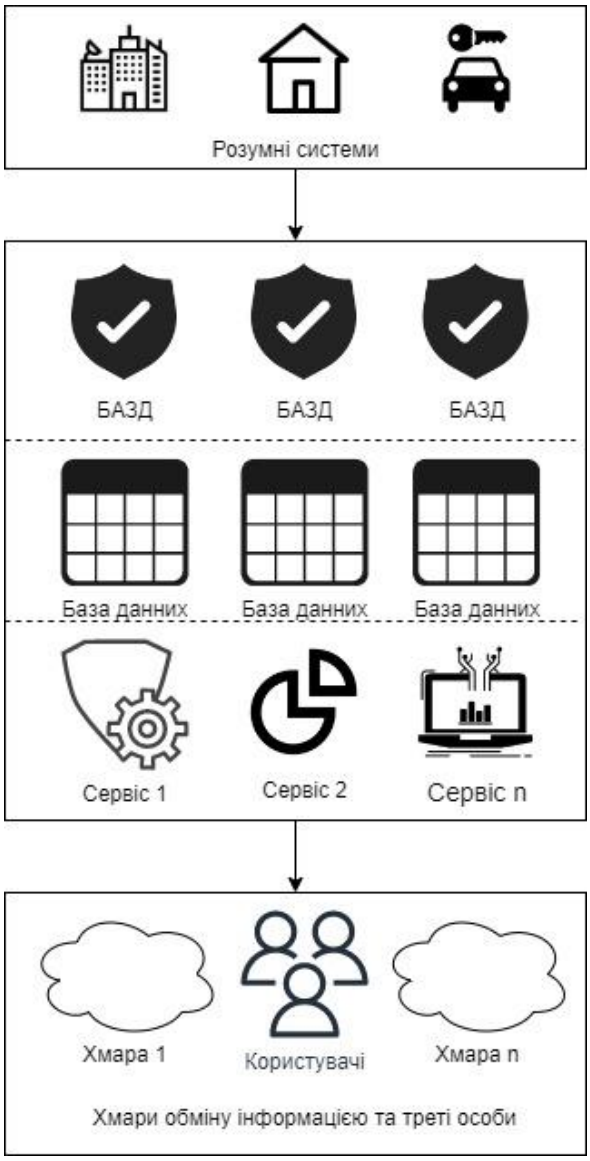


Рисунок 3.4 – Розташування БАЗД серед інтелектуальної системи

Припущено, що вихідні дані збурюються перед потраплянням на пристрої

зберігання. До збурених даних матимуть доступ або публічні, або приватні служби.

І алгоритм слідує за тим, щоб вивільнення даних відбулось у чітко визначені проміжки часу, які забезпечують ефективне та вчасне виконання алгоритму. Враховуючи всі відповідності до традиційної диференціальної конфіденційності, прийнятні значення шкали шуму Лапласа повинні знаходитися у інтервалі $(0, 9]$. Відповідно зниженню рівня конфіденційності та якості зберігання може призвести менша чутливість інтерполяції та збільшення значення шкали Лапласа більше 2. Користувачі власноруч підлаштовують значення і є відповідальні за них, і працюють вони відповідно своїх вимог. Запропоновано обрати значення 1. Це забезпечить баланс між конфіденційністю та корисністю. Передбачено різні сценарії алгоритму, тому у випадку, коли користувач вибере значення менше 1, то алгоритм забезпечить в свою чергу бездоганну конфіденційність інформації, що надана йому, але в той же час корисність та швидкодія будуть на низькому рівні. Зворотнім буде випадок, коли користувач вибере значення більше 1. Відповідно буде забезпечена більша корисність та швидкодія, але це згубно вплине на конфіденційність та її якість.

Використання БАЗД у кіберфізичній розумній системі охорони здоров’я. У системах охорони здоров’я існують багато проблем, не виключенням є проблема конфіденційності. І це чудовий виклик для кіберфізчних систем, щоб показати себе. Зокрема у операційних із залученням штучного інтелекту, у відділеннях травматології та різних видів терапії питання контролю процесів і також їх конфіденційності та безпеки стоїть чи не на першому місці для вирішення. Кількість інформації у системі охорони здоров’я дуже різноманітна. Вона включає в себе не тільки персональні дані пацієнтів, лікарів чи іншого персоналу будь то електрик чи охоронець, але і життєво-необхідні показники кисню у новонароджених, інформація про стан суглобів у літніх людей чи просто підтримка життя під час операції та всіх його складових чи показників. Також такі системи зазвичай містять актуальну інформацію із сфери вірусології та імунології, що завжди актуально та важливо. І власне це не весь список даних, які потребують

надійного захисту. БАЗД надає чудове та практичне рішення в плані витоку персональних даних та важливої інформації для підтримки життєдіяльності систем. На рисунку 3.5 зображено орієнтовний варіант використання БАЗД у розумній кіберфізичній системі охорони здоров'я.

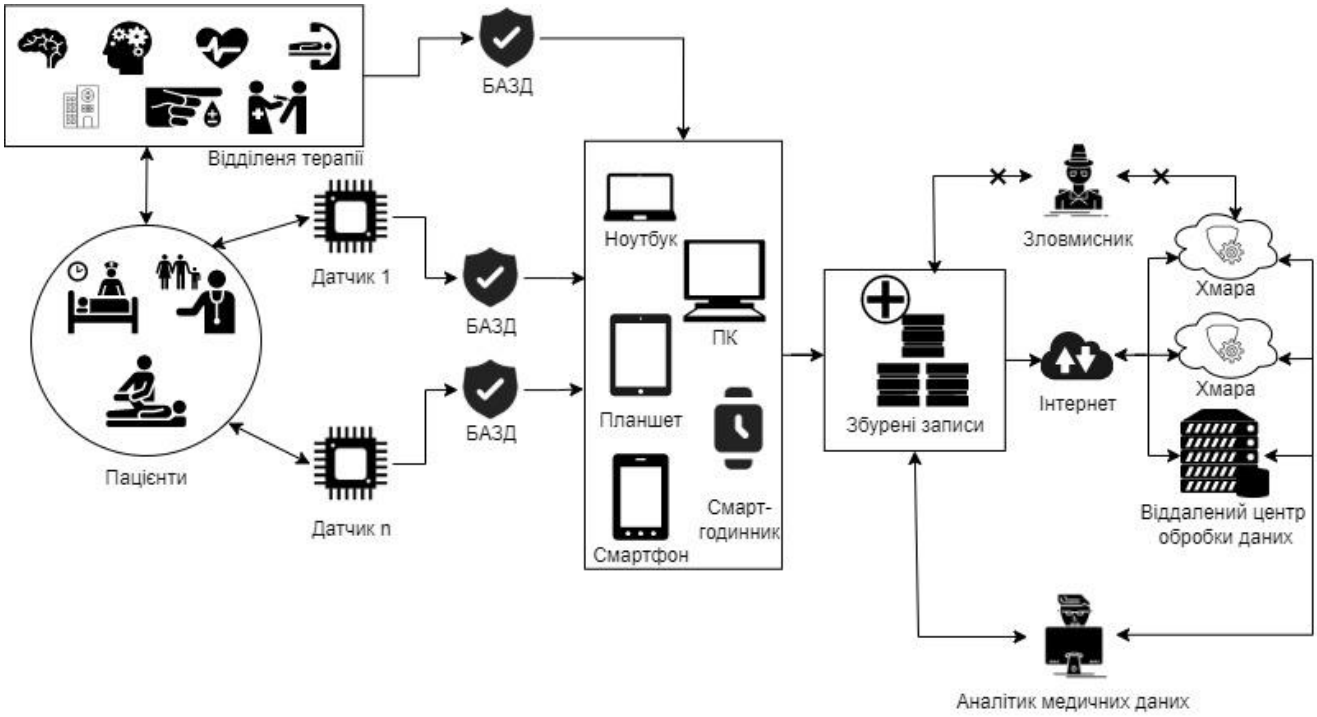


Рисунок 3.5 – Приклад використання БАЗД

Лікарям та пацієнтам доступна маса опцій для взаємодії із датчиками, які допомагають слідкувати за їх станом здоров'я, збирати інформацію про показники здоров'я та стан палати, температуру та навіть рівень вуглекислого газу у приміщені, який допоможе оцінити якість повітря та увімкнути вентиляцію, або ж просто відкрити кватирку. Лікарям доступні записи та можливість аналізувати маси фізичних різних параметрів, включаючи перевірку необхідних аналізів пацієнта та навіть наявності протипоказань до застосування анестезії. Таким чином всі дані передаються на центральний блок інформації, який виступає посередником між сервером та користувачем. Це може бути як і стаціонарний ПК чи ноутбук, так і смартфон. Інформація із датчиків кількості кисню в крові, виміри артеріального тиску збираються із пацієнтів чи не кілька разів на добу. Отже у даних умовах БАЗД

збурює всі чутливі входи, що пересилаються на центральний блок. Тому, центральні блоки ні в якому разі не отримують жодної чутливої інформації. У свій час, як зображено на рисунку, сховища даних (сервери) зберігатимуть лише збурені дані. Такий тип зберігання даних буде чудовим не тільки для локальних пристроїв та серверів, але і для віддалених центрів обробки, що можуть покривати цілі країни. Особи, які аналізують дані мають доступ виключно до збурених даних для проведення внутрішньої аналітики та збору статистики. Оскільки дані зберігаються у збуреному вигляді, атаки зловмисників на них навряд чи закінчать успіхом.

Було представлено два нових методи збурення даних. У МЗПД в основному поєднано два методи збурення даних. Метод конденсації даних може бути використаний як для статичних даних так і для обробки потокових даних. Метод обертального збурення може використовуватись виключно для обробки статичних даних. Відповідно до збереження відстані між кортежами, використовуючи метод обертального збурення, він регулярно забезпечує чудову точність класифікації та кластеризації даних. Проте цей алгоритм є ефективним з точки зору економіки ресурсів ПК та обробки. Метод конденсації даних з одного боку є досить швидким та ефективним у обробці даних в напрямку збереження конфіденційності. Однак, його штучна генерація даних часто знижує їх якість. Це відбувається лише в тому випадку, коли він налаштований на забезпечення високої конфіденційності даних. Але саме через ці налаштування для забезпечення такого високого рівня конфіденційності, що є із більш просторовою локальністю, він часто призводить до генерації із низькою точністю наборів даних. Найкращий метод для обробки даних як статистичних так і потокових – МЗПД, що поєднує у собі переваги методів конденсації та обертального збурення.

МЗПД показує зростаючу точність в той час як метод конденсації даних показує зниження точності і збільшення значення k'. Неможливо очікувати високого рівня класифікації від конденсації даних, це підкріплюється тим фактом, що при великих наборах даних (інакше кажучи коли більші розміри груп чи кластерів і просторова локалізація є більшою). Алгоритм конденсації припускає, що дані є рівномірно розподіленими і у них наявні невеликі просторові локалізації

для запобігання втрати точності. Оскільки розподіл скорочених даних є вкрай близьким до вихідних даних, це призводить то зниження конфіденційності. Отже, ефект від обертального збурення із результатами зібрані одним набором даних. Обертання ні в якому разі не спотворює відстань між кортежами, тому фактор точності в даному випадку не страждає. В результаті – МЗПД забезпечує кращу точність класифікації за більш високого рівня відхилення збурених даних, що є особливою властивістю для збурення. Це полягає у випадковому збурені даних, яке може бути як і плановим, так і на вимогу, що робить алгоритм гнучким і дружнім до користувача. Саме при збурені потоків даних результуюча кількість груп може стати необмеженою, але всі тонкощі захисту конфіденційності навіть найчутливіших даних буде забезпечено завдяки алгоритму збурення потоку даних.

Обчислювана складність МЗПД керована алгоритмом кластеризації. Так як кількість не тільки атрибутів, але і кортежів стрімко збільшується (до того ж одночасно), то відповідно їм зростатиме і часова складність алгоритму збурення. У звичайних умовах збільшення потоку даних є інкрементальним, де нові кортежі достатньо швидко додаються до набору даних, в той час як кількість атрибутів – постійна. Розглянуто нерівність m та n, як $m >>> n$, де кількість параметру n є меншою у порівнянні з надзвичайно великою кількістю m. Це говорить про те, що кількість кортежів мають більший вклад у тимчасову складність. Звернуто увагу на емпіричні дані також, витрати часу МЗПД – є досить низькі, а графіки показують що витрати часу є близькі до лінійних. Низькі часові затрати вкотре наголошують на придатність методу для швидких потоків даних тих, які можуть бути не тільки звичайної розмірності, але і тих, що можуть бути екстремальних розмірів здатні до збурення великих наборів даних.

Застосовано логічні міркування щодо вибору відповідних значень для k/k' із урахуванням відповідних значень результатів МЗПД на різних наборах даних. Для них існує відносна залежність між l та k/k'. Від швидкості конкретних потоків даних залежить вибір названих двох вихідних параметрів. Якщо врахувати те, що розглянуто, то чим швидше потік даних (для прикладу 8000 кортежів в момент

часу) тим меншим є значення l (для прикладу 4000). Отже значення k' повинно бути вище (щонайменше 400), або k має бути нижче (наприклад 10). Коли швидкість потоку даних є відносно нижчою 70 кортежів на момент часу, має місце застосування зворотного сценарію, де l може бути у свій час вищим (для прикладу 5000). І такі швидкості в той час, коли $k' = 110$ чи $k = 50$. Які би значення не вибрав користувач – значення для цих параметрів не впливатимуть значно на основну мету цього методу. Тому корисність та конфіденційність залишаться на попередньому рівні, і це скоріше вплине на ефективність методу та його швидкодію.

Не дивлячись на те, що МЗПД – ефективний алгоритм збурення потоку даних, він демонструє $O(\mathrm{n}^3)$ тимчасову складність для кількості атрибутів (n). Для динамічного середовища, яке поступово додає нові датчики інтернету речей (відповідно збільшуючи кількість атрибутів), МЗПД може не дійти до оптимального рішення. Інший механізм збереження конфіденційності, який називається БАЗД, розрахований для великих даних і потоків даних, що виконує збурення даних на основі інтерполяції та застосуванні диференційної приватнсті.

Під час інтерполяції БАЗД додає калібрований шум, щоб запровадити рандомізацію, і, відповідно, конфіденційність щодо збурених даних. Цей шум дозволяє виконувати процес інтерполяції з передбачуваною випадковою помилкою для мінімізації середньоквадратичної похибки. Також дотримується диференційна конфіденційність для додавання й введення шуму, що відповідає характерному бюджету конфіденційності. Він дозволяє користувачам БАЗД регулювати кількість шуму, тому його менші значення ε (зазвичай менше 1, але більше 0) додають більше шуму, для посилення фактора випадковості і генерації відповідних даних, натомість великі значення бюджету конфіденційності генерують значення із меншою випадковістю.

Здатність БАЗД зберігати форму вихідного розподілу даних після додавання шуму є очевидною перевагою і дозволяє забезпечити вищу корисність, ніж стандартний локальний диференційно приватний алгоритм. Ця характеристика

може мати свою важливість, а конфіденційність, яка забезпечується стандартним диференційно приватним механізмом, може бути трохи вищою, ніж у БАЗД.

Із статичними даними та із потоками даних БАЗД працює добре. З точки зору класифікації, стійкості атак та прикладу застосування було оцінено його. Брався до уваги також фактор тимчасової складності та масштабованість потоку даних. БАЗД у порівнянні із іншими алгоритмами збурення та конденсації. БАЗД є одним із найкращих вибором для збереження конфіденційності даних, які були отримані за допомогою розумних кібер-фізичних систем та інших технологій. Із точки зору класифікації він забезпечує високу корисність збурених даних завдяки властивості зберігати базові характеристики, такі як форма вихідного розподілу даних. Застосовуючи велику кількість шуму та невелике значення бюджету конфіденційності БАЗД у будь-якому випадку намагається зберегти форму вихідних даних. У надзвичайно несприятливих умовах БАЗД вестиме себе стабільно і зможе забезпечити вищу корисність у порівнянні з аналогічними збуреннями. Існує декілька кроків для підвищення конфіденційності збурення даних: масштабування чи нормалізація, апроксимація будь-яких масштабів включаючи шумну інтерполяцію, перемішування даних. Власне ці кроки допомагають БАЗД перевершити всі вищевказані у цьому параграфі методи збурення враховуючи всі норми конфіденційності.

Взято до уваги, що застосування методів збереження конфіденційності часто стикається із певними обмеженнями в особливості варто приділяти увагу конкретним адоптаціям. МЗПД та БАЗД розроблені лише для збурення числових даних. Однак, реальні сфери та способи застосування алгоритмів МЗПД та БАЗД обробляють великі дані з реальними значеннями. Наведено приклади використання датчиків температури, наближення, тиску та рівня вуглекислого газу та варіативні інфрачервоні датчики. Це дозволяє їм застосовуватися у різноманітних сферах, включаючи банківську, агроінженерію, соціальні мережі, погоду, менеджмент, маркетинг та розглянуту сферу охорони здоров’я.

3.3 Висновки до третього розділу

В даному розділі було представлено два алгоритми збурення потоку даних МЗПД та БАЗД. Стійкість до атак та точність МЗПД є вищою, аніж в аналогічних методах. Складність МЗПД – залежність від кластеризації, коли кількість атрибутів носять постійний характер. Алгоритм демонструє, коли кількість кортежів знаходиться у постійному стані найгіршу складність для виконання його. Кількість атрибутів дуже мала у порівнянні із потоками кортежів, тому МЗПД продемонстрував нижчі кількості часу, що був витрачений, піл час емпіричного аналізу. Це, власне, і забезпечує безперебійну роботу з зростаючими потоками даних і великими даними. У свою чергу БАЗД має переваги: вища точність класифікації, ефективність та масштабованість при збереженні більшого рівня конфіденційності та вищої стійкості до атак. Це означає, що другий алгоритм підходить більше для динамічних середовищ, що можуть бути представлені у розумних кіберфізичних середовищах. БАЗД – чудовий засіб для збереження конфіденційності для будь-яких сфер. Розглянуто алгоритм у сфері охорони здоров’я, тому що вона є чи не основною передовою сферою у світі кіберфізичних середовищ. Алгоритм може ефективно обробляти потоки даних, що генеруються датчиками, які контролюють пацієнта чи навіть сім’ю, що знаходиться на догляді у сімейного лікаря із передбаченим шифруванням та подальшою обробкою у хмарі перед подачею їх для аналізу вище поставленому спеціалісту із аналітики охорони здоров’я.

4 ЕФЕКТИВНІСТЬ ЗАСТОСУВАННЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

4.1 Ефективність застосування

АЗКдГН на згорткових глибоких нейронних мережах показує високу точність (91%-96%) з чудовою якістю моделі навіть за низьких бюджетах конфіденційності (ε = 0,5). Інший алгоритм збурення даних обличь (АЗДО), що зберігає конфіденційність, використовує локальну диференціальну конфіденційність. Використовуючи диференційну конфіденційність АЗДО зберігає лише збурені дані на сторонніх серверах для запуску стандартного алгоритму розпізнавання власної поверхні.

У цьому розділі описуються та досліджуються проблеми конфіденційності машинного навчання та охарактеризовуються два окремих алгоритми до глибокого навчання і збурення даних облич, що зберігають конфіденційність. Запропонований алгоритм збереження конфіденційності для глибокого навчання (АЗКдГН) дозволяє власнику даних залучити рівень рандомізації до того, як дані покинуть пристрої власників і досягнуть потенційно ненадійної служби машинного навчання. Це може функціонувати шляхом поділу згорткової нейронної мережі на три прошарки: 1) згортковий модуль; 2) модуль рандомізації; 3) повністю підключений модуль. Модуль рандомізації використовує протокол розподілу міток, що покращує випадковість, і дозволяє АЗКдГН підтримувати високу корисність у порівнянні з наявними протоколами.

Порівнюючи з традиційними підходами машинного навчання, глибоке навчання демонструє відмінний успіх у вирішенні складних питань, таких як розпізнавання мовлення, класифікація зображень, обробка мови. Моделі глибокого навчання часто ґрунтуються на конфіденційних даних краудсорсингу, таких як фінансові записи, записи про стан здоров’я, особисті зображення. Коли моделі глибокого навчання зазначаються на масивних базах даних, що містять конфіденційні дані, можуть виставляюти напоказ приватну інформацію.

Все більше користувачів можуть стати вразливими до атак при розвитку

розподіленого хмарного машинного навчання у таких середовищах величезних корпорацій, як Amazon і Google. Маючи довіру до цих середовищ, користувачі можуть передавати свої дані й отримати доступ до них за допомогою білого чи чорного ящиків. Проте слід зауважити, що зловмисник може легко втілити у реальність шкідливі алгоритми та дати їм хід під виглядом частин процесу навчання. У той же час, зловмисні алгоритми можуть запам'ятовувати конфіденційну інформацію користувача. В інших ситуації зловмисники можуть використовувати запам'ятовані відомості, і таким чином отримувати інформацію про користувачів, порушуючи їх конфіденційність.

Вразливість моделей глибокого навчання зазначають атаки щодо конфіденційності, такі як висновки про членство, навіть якщо вони прилеглі як моделі чорного ящика. Ще одним прикладом, який показує слабкість моделей машинного навчання, можуть слугувати атаки інверсії моделей, що відновлюють зображення з алгоритму збурення даних облич. Важливо, щоб машинне навчання як використовувало практичні механізми збереження конфіденційності, щоб обмежити витік конфіденційності. Варте уваги те, що ці підходи для збереження конфіденційності для машинного навчання можна було використовувати для додатків на основі IoT, таких як розумна охорона здоров'я, IIoT та Industry 4.0.

Розглянуто питання конфіденційності глибокого навчання. Для розробки розподіленого механізму збереження конфіденційності з використанням диференціальної конфіденційності залучено розподілений механізм збереження. Для забезпечення контролю та обмеження витоку конфіденційної інформації під час глибокого навчання диференціальна конфіденційність дає надійну та стабільну основу, що гарантує відповідний рівень конфіденційності. Існує достатня кількість наявних еталонних підходів, що грунтуються на єдиній глобально-диференційальній конфіденційності, що використовує довіреного куратора для отримання диференційальної конфіденційності за допомогою каліброваного шуму. Необхідність довіреного куратора по конфіденційності робить методи даного виду забезпечення збереженості даних непридатними для практичного та стабільного. Алгоритм в такому сценарії працює на сервері однієї із великих центральних

корпорацій. В такому сценарії обробляє вхідні дані для завантаження на сервер. Такий підхід може згубно впливати на конфіденційність через те, що зловмисники зазвичай здійснюють атаки, що є орієнтовані на сервер. Більш того розглянуто, що алгоритми глибокого навчання якими би вони не були легкими для систем мають тенденцію бути вкрай складними та вимогливими до обчислювальної потужності техніки користувачів чи потужностей корпорацій. Як наслідок, алгоритми, які тісно пов'язані з глобальним диференційним захистом конфіденційності слід використовувати на високопродуктивних комп'ютерах, а власники інформації, що не є достатньо захищеною не можуть використовувати її чи обробляти у незахищених середовищах. Крім що використання периферійних методів для калібрування шуму (Лапласкіна та Гауса для моделей глибокого навчання) може погіршувати результати (забезпечувати меншу точність) або сприяти вкрай високому рівню витоку конфіденційності. Ці проблеми зумовлюють складність даних методів. Під час виконання методу локальної диференціальної конфіденційності данні власників піддаються збуренню перед тим, як їх відправити в обробку. Такі дії дозволяють уникнути необхідності в довіреній третій стороні, що гарантує максимальну конфіденційність. Локальний підхід до збурення даних призводить до змін розробки розподілених алгоритмів для збереження конфіденційності для маси сценаріїв (для прикладу основного Інтернету речей).

Пропонований підхід для збереження конфіденційності для глибокого навчання (АЗКдГН) є розподіленим механізом локальної диференційної конфіденційності, що включає в себе новий протокол для обмеженя витоку даних моделей конволюційних нейронних мереж, що випускаються як моделі «чорної скриньки». Основним методом АЗКдГН є використання властивостей випадкової відповіді. Даних метод є відносно популярним та відповідає вимогам локальної диференційної конфіденційності. ЇЇ налаштування та багаторівнева архітектура АЗКдГН допомагають для забезпечення поставленої мети. Зберігання конфіденційності між кількома сторонами, що зазвичай є неможливим при використанні існуючих методів глобальної диференційної конфіденційності для глибокого навчання. Дані проблеми нівелює АЗКдГН. Оскільки, метод глобальної

диференціальної конфіденційності до АЗКдГН повністю задовольняє потребу в контролі бюджету конфіденційності до процесу збурення. Налаштування точності відбувається незалежно. Узагальнивши АЗКдГН значно зменшує вплив бюджету конфіденційності на точність, і саме це призводить до стрімкого збільшення рівня захисту та кращої точності у порівнянні із іншими існуючими рішеннями. Для прикладу, якщо порівняти глобальну диференційну конфіденційність для глибокого навчання АЗКдГН забезпечує достатню точність (близько 88%) в екстримальниз випадках бюджету конфіденційності (наприклад із точністю 0.5), що забезпечує мінімальний витік. Чітко показано, що комп'ютер загального призначення достатньо ефективний для надійного виконання необхідних обчислень власника даних. Відповідно даним твердженням та порівнянням із існуючими методами, АЗКдГН може бути більш практичним, гнучким, безпечним та надіним інструментом для обмеження витоків конфіденційності інформації із моделей глибокого навчання.

Під час розробки підходів надійних методів розпізнавання осіб варто врахоувувати проблеми конфіденційності також.У галузі обробки зображень розпізнавання облич не тільки не займає останнє місце, але і містить безліч додатків у галузі обробки зображень та аналізу. Стрімкий розвиток відповідних технологій дозволяє ефективно та точно інтегрувати системи розпізнавання осіб навіть в повсякденні системи розумного будинку, які тісно пов'язані із смартфонами у яких технології розпізнавання облич вже не є новими. Ще одним прикладом може слугувати сфера розпізнавання облич для виявлення правопорушників, які вже є в базі на момент їх ідентифікації під час процесу розпізнавання облич в момент проходу турнікету у метро чи каси у супермаркеті. Компаніями інвестовано велику кількість коштів у системи розпізнавання. Однак, питання безпеки розпізнаних даних є відкритим. Широке використання біометраї є серйозною загрозою для конфіденційності людей та їх приватного життя. Це зумовлено тим, що у моменти (за виключенням моменту розблокування смартфону) проходження біометричної ідентифікації методом розпізавання облич людям не повідомляється цей факт. Тому, ідентифікація особи (її розпізнавання за допомогою обличчя) із збереженням

всіх норм конфіденційності нам необхідно ідентифікувати людину за зображенням не розкриваючи нікому вагомих даних біометрії та особливостей власника. Як наслідок варто розглянути дві основні сторони зображення людини. Перша, має містити зображення, а інша – базу даних зображень. Ретельне шифрування даних дозволить стороні 1 дізнатися результат, не дізнвашись про виконання адгоритму ідентифікації користувача. В цей же час сторона 2 не знатиме процесів і будь-яких вхідних зображень, бо доступ до них буде повністю перекритий. У будь-якому випадку, враховуючи високу обчислювану складність сторін та необхідність довіри до них можуть виникати певні проблеми взаємодії. У цьому розділі запропоновано збурення даних, яке є менш складним з обчислюваних точок зору (і відповідно з точки зору навантаження на системи), але яке спричиняє певний рівень втрати корисності.

У більшості існуючих методів, які допомагають збереженню конфіденційності заснування їх відбувається на базі гомоморфного шифрування. А це у свою чергу теж неабияк впливає на продуктивність. Такі методи зумовлюють наявність проблем з корисністю для великих даних та сценаріїв для їх обробки для мільйонів осіб. Виділено наступний переліч вирішень питань, що пов’язані із конфіденційністю існуючих методів для розпізнавання осіб.

1: Біометрія особи, що ідентифікується ні в якому випадку не повинна бути пов'язана з іншими конфіденційними даними.

2: Метод має бути масштабованим та дружнім до ресурсів (як зовнішніх, так і внутрішніх).

3: Біометрія особи не повинна бути доступна нікому ні за яких причин і не слід виконувати перетворення в багатьох порядках (тобто слід використовувати виключно одностороннє перетворення).

4: Біометричні дані особи однієї і тієї ж людини з двох різних додатків не повинні бути пов'язані між собою.

5: Біометрія повинна мати опцію, яка може її відкликати, що актуально для додатків у яких можливий витік даних. (це направлено для запобігання будь-якому зловмисному використанню)

Наступним методом, що запропоновано, є метод, що контролює витік конфіденційності власне при розпізнаванні осіб та задовольняє вищеперераховані питання краще, ніж аналогічні методи збереження конфіденційності. Запропоновано підхід, що використовує збурення для зберігання даних, та зберігає їх в такій же формі. Принцип метода – це диференціальна конфіденційність. Йому назва АЗДО (алгоритм збурення даних облич) із збереженням даних у безпеці. Для збурення до вихідних даних зображення (щоб обмежити потенційний витік конфіденційності через залучення ненадійних та недовірених сторонніх серверів та користувачів) використовується локальна диференціальна конфіденційність. Щоб уникнути необхідності у контакті із третіми особами, які можуть бути потенційними зловмисниками застосовано випадкову генерацію відносно даних, які зазвичай використовуються для начальних або тестувальних цілей. Відповідно низькій складності АЗДО може бути легко реалізований на пристроях, які обмежені за ресурсами, що допускають можливість збурення на вході. Контроль рівня конфіденційності за допомогою бюджету конфіденційності є додатковою перевагою запропонованого методу. Позначення бюджету конфіденційності застосовується для позначення його рівня. Тому, чим показник нижчий, тим нижчим і є рівень конфіденційності. АЗДО використовує диференціальну конфіденційність, яка є локальною. Зниження допускається лиш на 6 пунктів. Для прикладу від 84% до 78% за бюджету ε=8, де $0 < \varepsilon \leq 9$ розглядається як прийнятний рівень конфіденційності. За рахунок додавання шуму у показники бюджету АЗДО може регулювати і компроміс між точністю та рівнем конфіденційності.

4.2 Реалізація алгоритму захисту конфіденційності для глибокого навчання

У цьому розділі йдеться про диференційовано приватний механізм, що часто використовується в АЗКдГН. Цей алгоритм збереження конфіденційності для глибокого навчання можна систематизувати, як похідний диференційно приватний алгоритм, який заснований на техніці випадкової відповіді. Він використовує дві

властивості диференціала конфіденційності: композицію і інваріантність постобробки при застосуванні диференційної конфіденційності до згорткової нейронної мережі. АЗКдГН використовує регуляризацію, а саме збільшення зображення та налаштування гіперпараметрів, щоб оптимізувати свою продуктивність. АЗКдГН протестувано на згорткових нейронних мережах за допомогою прикладного програмного інтерфейсу нейронних мереж Python Keras, який злагоджено працює на основі механізму потоку даних TensorFlow, який розробив Google. Keras надає високорівневий прикладний програмний інтерфейс нейронної мережі, перш за все розроблений для швидкого дослідження.

Структура згорткової нейронної мережі розподілена на два основних модулі, та виводимо проміжний модуль випадковості у згорткові нейронні мережі. У даних мережах вхідні ознаки зазвичай піддаються скороченням розмірності із використанням набору згорткових та об'єднувальних шарів. Вихідні дані останніх – згладжуються в єдиний одновимірний масив, а потім подаються у повністю підключену штучну нейронну мережу, тому ця частина названа згортковим модулем. АЗКдГН слідує після згорткового модуля, за допомогою якого використано модуль згортки тільки для генерації 1-U сплощеного вихідного сигналу, який відповідає певному вихідному зображенню. Названий сигнал і є одномірним вектором стовпців плаваючих значень. До того ж АЗКдГН перед випадковою генерацією перетворює вхідні значення на двійкові. Вхідні данні зазвичай можуть мати різні діапазони. Перетворення як і малих дробів так і великих значень у двійкові може зазвичай включати велику кількість бітів. Це призводить до непостійного рівня складності (непослідовного рівня складності) алгоритму. Для уникнення застосовано нормалізацію до значень одномірних векторів, що походять із шару згладжування. Згортковий модуль використовують лише для генерування одновимірного виходу, який відповідає певному входу зображення. Цей вихід є просто одновимірним вектором стовпця чисел із рухомою комою та дійсними значеннями.

Визначення межі бітового зразка встановлює діапазон конкретного нормалізованого вхідного значення L-score. Преш за все варто дати оцінку верхій

та нижній межі конкретного входу. Існують три основні сегменти двійкового рядка. Перший біт — знак входу (1 для негативного і 0 для позитивного). Дві інші частини призначені для цілого числа та тільки деякої частини вхідного числа. Кількість бітів для цілого числа залежить від максимального значення цілого числа, яке необхідно відобразити. Завдяки нормалізації оцінки, кількість бітів, необхідних для представлення цілого числа, обмежена. Кількість бітів для дробу залежить від точності (наскільки близьке десяткове значення двійкового дробу до значення вхідного дробу). Щоб досягти максимальної точності для дробу варто використовувати більшу кількість бітів.

Визначивши довжину компонентів, всі без виключення входи можуть бути зіставлені. Двійкове представлення співставлене із цілим значенням. Створюється за допомогою n та m – номерів двійкових цифер цілих чисел, та цілих частин дроба відповідно. Значення x являє собою кількість двійкових цифр цілого числа, де $x \in R$, а $g(i)$ являє собою $i-$ ий біт двійкового рядка. Де молодший біт записується як $k = -m$ Знаковий біт дорівнює 1 для від'ємних значень, і відповідно 0 для додатних. Знаковий біт присвоюється старшому біту двійкового рядка.

$$g(i) = (\lfloor 2^{-k}|x| \rfloor \ mod\ 2)_{k=-m}^{n}, \text{ де } i = k + m \tag{4.1}$$

Далі об'єднано всі двійкові рядки в один довгий двійковий рядок для уникнення втрати конфіденційності через одну із властивостей вищезгаданої диференційальної конфіденційності. При застосуванні факторів випадковості на кожному двійковому рядку, що відповідає кожному одновимірному $1 - D$ вектору окремо –це складе бюджети конфіденційності окремо для всіх етапів випадкової генерації. У випадку якщо r двійкових рядків було випадково згенеровано, то результуючою буде втрата конфіденційності остаточної випадковості $r \times \varepsilon$. Оскільки АЗКдГН проводить рандомізацію на певному об'єднаному двійковому рядку відразу, існує імовірність загрози втрати конфіденційності при вихідному значенні ε.

Так як рандомізація відбувається після згорткового модуля, модуль згортки та модуль АЗКдГН просувають до кінцевих даних власника. Модуль диференційно приватних штучних нейронних мереж є ненадійним провідником і може виконуватися на хмарному комп'ютері або на будь-якому високопродуктивному комп'ютерному сервері. Випуск моделі передбачає не тільки модуль диференційно приватних штучних нейронних мереж, який може використовуватися для тестування. У наведених параметрах згортковий модуль не розрахований для функцій, що залишають мінімальне обчислювальне навантаження на конкретного власника даних.

Конфігурація локальної диференціальної конфіденційності полягає в алгоритмі збереження конфіденційності для глибокого навчання. У розподіленні налаштувань згортковий модуль працює на боці власників даних. Рандомізовані дані накопичуються на сервері, наприклад, у загальнодоступній хмарі, де диференційована приватність генерується на повністю підключеному мережевому модулі.

Рисунок 4.1 Складовий розподіл схеми згорткової нейронної мережі

На рисунку 4.1 зображено мережевий модуль, який переміщує згортковий модуль до власника даних, створючи провідну конфіденційність навіть до рандомізації, є одновимірним вектором зі зменшеним розміром. Крім того, в контексті великих даних, де надзвичайна кількість власників даних мають зв’язок із сервером, поширення моделі згорткової нейронної мережі (рисунок 4.2).може забезпечити додаткову гнучкість та ефективність обробки даних

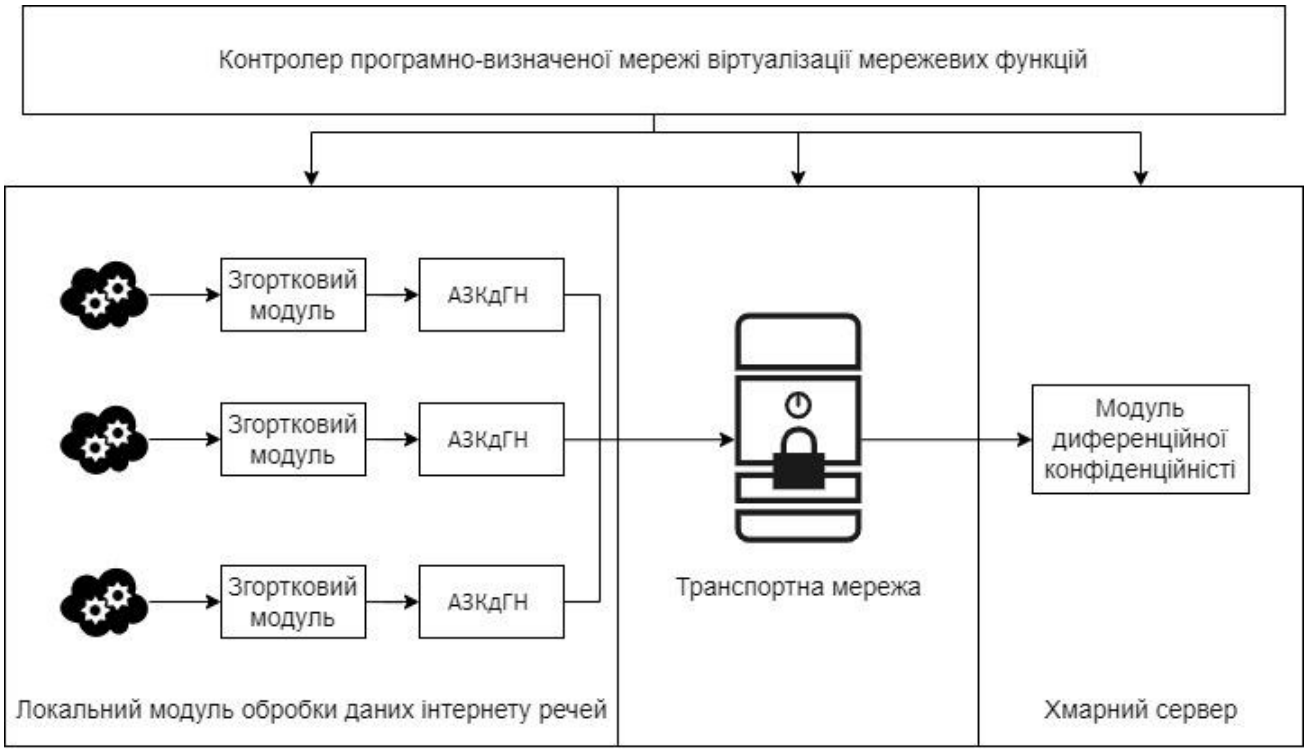


Рисунок 4.2 – АЗКдГН інтеграція у програмно визначену мережу

Алгоритм захисту конфіденційності для глибокого навчання та віртуалізацію мережевих функцій зображено на рисунку 4.2 у взаємодії між межами та хмарами. Як показано на малюнку АЗКдГН забезпечує рішення, яке легко налаштовувати, що дозволяє ефективну інтеграцію модуля випадковості в рівень віртуалізації мереж, для збереження конфіденційності машинного навчання в розподіленому середовищі програмно визначених мереж та віртуалізованих мережевих функцій у взаємодії між межами та хмарами.

Однак, отримано можливість перенести чи не всю архітектуру диференційної конфіденційності ЗНМ на один носій інформації, який обслуговується довіреним

куратором. В цьому випадку є можливість застосувати данні архітектури у наборах даних довіреного куратора, де модель буде випущена з усією архітектурою ЗНМ у якій міститься згортковий модуль (ненавчений) з АЗКдГН та із модулем диференційовано приватної моделі штучних нейронних мереж.

Порівнюючи результати моделі із врахуваннями параметрів моделі визначено, що подано коректний параметр ε для моделі із забезпеченням хорошої точності. Імовірне накопичення великого та неприйнтного значення ε під кінець виконання генерації моделі. При $\varepsilon = 2$ та при $\delta = 10^{-5}$ методом була забезпечена чудова точність. Однак, адитивна межа значення δ може бути ненадійною у випадку використання методу для набагато більних наборів даних, що містять інформацію. АЗКдГН забезпечує кращу точність для вкрай низького бюджету конфіденційності. Варто мати на увазі, що відсутність будь-яких адитивних зв'язків гарантує, що АЗКдГН має низькй рівень витоку конфіденційності, коли відбувається надання йому наборів великих даних. У методах, що згадані в попередньому розділі, які є засновані на глобальній диференційній конфіденційності, наявність довіреної сторони неминуча. Однак, у реальних сценаріях роботи наявність цієї сторони часто залишається під питанням. В такому випадку АЗКдГН може бути набагато кращим рішенням. Можливість працювати із надійними та ненадійними сторонами (кураторами) це показує.

У цьому параграфі буде розглянуто балові кроки, зроблені АЗДО (алгоритмом збурення даних облич), який є похідним від АЗКдГН, який використовує метод диференційної конфіденційності для розпізнавання облич. АЗДО ставить за мету збереження конфіденційності вхідних зображень використовуючи рандомізацію, як показано на рисунку 4.3. В загальному, будь-який пристрій введення, який використовується для отримання зображень облич матиме можливість застосовувати АЗДО для того, щоб задовольнити мету рандомізації безпосередньо перед відправкою на будь-які пристрої, що застосовуються для зберігання інформації чи сервери.

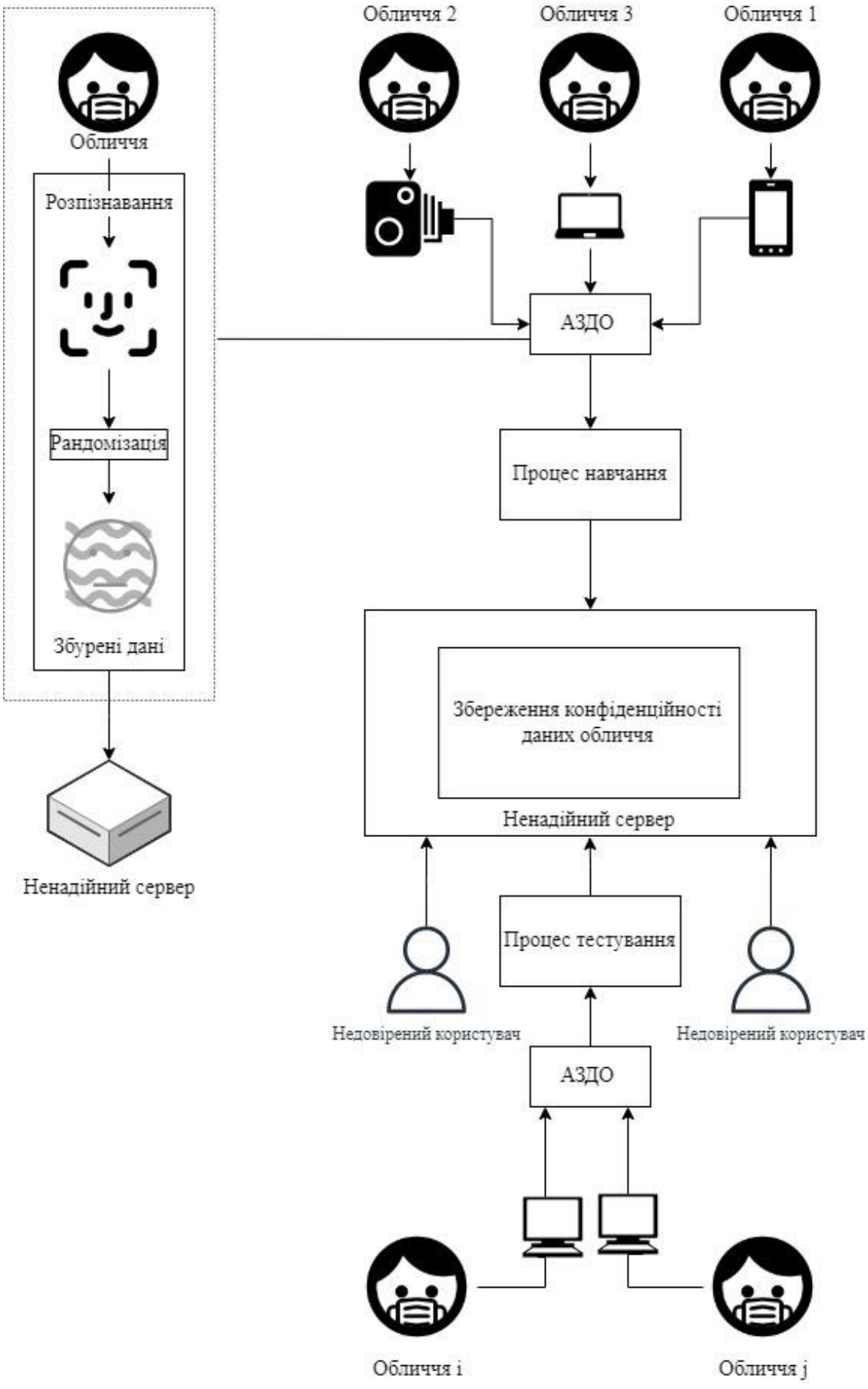


Рисунок 4.3 – Схематичне зображення АЗДО

На рисунку 4.3 показано розміщення АЗДО в системі розпізнавання обличчя. Алгоритм випадково розподіляє як навчальні, так і тестові зображення, щоб ненадійні сторонні сервери не передавали чутливі дані, які вимагають дотримання всіх норм конфіденційності ненадійним користувачам, оскільки вони можуть бути потенційними зловмисниками. На малюнку виноски в лівій частині показано основний потік рандомізації всередині АЗДО, який застосовує шум Лапласа до власних граней. На рисунку вказано, що існують 3 базових кроки до забезпечення конфіденційності даних облич. Перш за все вони приймають та погоджують оригінальні дані обличчя. Другим кроком є генерування власних граней, а останнім – додавання шуму Лапласа для рандомізації зображень. Тому у запропонованій моделі використання алгоритму чітко видно, що ненадійний сервер буде містити лиш версію даних, що зберігає конфіденційність моделі розпізаного обличчя.

Погодивши зображення, що були введені АЗДО нормалізує зображення, щоб воно відповідало попередньо визначеній роздільній здатності, яка є для АЗДО вхідною попередньо визначено розмір 56×74. Але алгоритм не обмежується цими значеннями і не заганяти у рамки користувача. Тому, виходячи з розмірів вхідного зображення та обчислювальної потужності граничних пристроїв, користувачами може бути як і зменшено, але у більшості випадків збільшено допустиме значення розширення вхідного зображення. За допомогою коваріаційної матриці відбувається обчислення головних компонент (під час розгляду власних векторів). Оскільки нормовані вектори, що застосовуються для аналізу основних компонентів обмежені від 0 до 1, то збільшення чутливості за допустимі рамки призведе до значного рівня шуму, що різко зменшить швидкість під час використання локальної диференційної конфіденційності для механізму застосування шуму. Тому слід обирати чутливість як максимальну різницю між двома індексами, що дорівнює 1. І лише у цьому випадку шум Лапласа дорівнюватиме $1/\varepsilon$. Для створення конфіденційних версій зображень відбувається процес збурення кожного індексу. Після визначення параметрів АЗДО додає шум Лапласа до кожного індексу, що бере участь у аналізі основних компонентів.

Під час проведення навчання моделі розпізнавання облич, АЗДО зберігає

конфіденційність. Оскільки основним завданням розпізнавання облич є класифікація зображень, то кожне обличчя уособлює клас. Для отримання хорошої точності модель класифікації повинна мати достатньо чітке представлення зображення. Отже, кількість зображень на одне обличчя є цінним параметром, який безпосередньо впливає на точність, де більша величина, безумовно, сприятиме вищій точності. Таким чином, кількість зображень на одне обличчя дозволяє алгоритму витягувати власні грані, які забезпечують краще представлення вхідних зображень. Кількість вибраних компонентів методу головних компонент не перевищує дозволений поріг.

Процес використання АЗДО показано на рисунку 4.3, де кожне введене зображення буде під впливом рандомізації АЗДО перед навчанням чи тестуванням. Генерація та рандомізація власної поверхні відбуваються в межах локальної границі. Усі пристрої введення пов’язані із сторонніми серверами тільки через АЗДО, а база даних розпізнавання обличчя зберігає лише порушені зображення. Оскільки модель розпізнавання обличчя зосереджується лише на використанні збурених зображень, навчена модель не буде пропускати приватну інформацію. Будь-який ненадійний доступ до сервера відхилить можливість втрати цінних біометричних даних. Оскільки АЗДО перешкоджає тестуванню даних, є мінімальний витік конфіденційності з даних тестування введених зображень.

Так як додаткові обчислення виконуються базуючись на диференційно приватному алгоритмі, вони не послаблюють гарантію конфіденційності. Результати додаткових обчислень все одно будуть диференціальною конфіденційністю. Ця властивість диференціальної конфіденційності називається інваріантністю обробки. АЗДО також використовує цю конфіденційність й успадковує інваріантність після обробки. Властивість інваріантності постобробки гарантує, що навчена модель збурених даних також задовольняє ту саму конфіденційність, яку встановлює АЗДО. Таким чином, вказаний метод гарантує мінімальний рівень витоку конфіденційності сторонніх ненадійних серверів.

4.3 Висновки до четвертого розділу

Розроблено два алгоритми згідно машинного навчання, що зберігають конфіденційність. Перший — новий локальний диференційно приватний механізм для навчання глибокої нейронної мережі з високою конфіденційністю та точністю. Показує відмінну точність навіть при екстремально низьких бюджетах конфіденційності порівняно з наявними диференційовано приватними підходами. Завдяки широкому простору функцій, створеному АЗКдГН протягом процесу рандомізації, він створює вищу точність для набору даних, у порівнянні з базовою моделлю згорткової нейронної мережі без будь-якої конфіденційності. Наявні диференційовано приватні механізми реалізовуються, здійснюючи глобальну диференційну конфіденційність, тому їм важлива наявність надійного куратора. Налаштування ненадійного куратора підходу забезпечує вищий рівень конфіденційності, залишаючи при цьому низький рівень обчислювального навантаження на власників даних. Оскільки вивід згорткового модуля є вихідним результатом зі зменшеним одновимірним розміром, переміщення згорткового модуля до власників даних створює додаткову конфіденційність навіть без застосування рандомізації. Розподіл структури згортково нейронної мережі між власниками даних і серверами є більш гнучким стосовно обробки даних у середовищі великих даних. Вказаний розподіл допомагає АЗКдГН без перешкод адаптуватися до інновацій, таких як об’єднання програмно-конфігурованої мережі і віртуалізації функцій мережі у взаємодії між межами і хмарами. Коли багато власників даних спілкуються з одним сервером, його завдання полягає у створенні диференційовано приватної моделі штучної нейронної мережі. Використання раніше згаданого методу у налаштуваннях ненадійного куратора має можливість приватно обмінюватися конфіденційними даними та воднчас обмежувати витік конфіденційності в структурі розподіленого машинного навчання. Таким чином, вибір вхідних параметрів диференційно приватного компонента АЗКдГН не залежить від процесів налаштування,наприклад, регуляризації, збільшення зображення та налаштування гіперпараметрів. Це дозволяє без перешкод

налаштовувати модуль з більш високим рівнем точності та надзвичайним конфіденційності, що в результаті спричиняє неабиякий обсяг відмінного балансу між конфіденційністю та корисністю.

Як другий підхід до машинного навчання має назву алгоритм збурення даних облич (АЗДО), що зберігає конфіденційність за допомогою даних на турбацію. АЗДО застовує властивості диференційної конфіденційності, що, віповідно, має можливість забезпечити достатньо високий рівень конфіденційності для технологій розпізнавання обличчяя. АЗДО не потребує довіреної сторони і використовує локальний підхід, при застоуванні рандомізації до того, як зображення опинятьчя на ненадійному сервері. АЗДО передає лише рандомізовані дані, які не вимагають захищеного каналу. АЗДО – це кваліфікований і легкий підхід, який можна без зайвих зусиль можна інтегрувати в будь-який пристрій із обмеженими ресурсами. Так як тестування й розпізнавання зображень обличчя здійснюються виключно базуючись на рандомізованих даних, АЗДО не втрачає ефективність протягом розпізнавання облич. Диференційно приватні поняття дозволяють користувачам налаштовувати параметри конфіденційності згідно з вимогами домену. Підсумовуючи вище зазнечене, можна зробити висновок, що АЗДО — це найсучасніший підхід до розпізнавання облич, що зберігає конфіденційність.

ВИСНОВКИ

У роботі виконано практичні та теоретичні дослідження, за результатами яких, розроблено алгоритми захисту конфіденційності для великих даних із використанням збурення даних та машинного навчання для покращення ефективності виявлення та знешкодження зловмисних дій, яким може піддаватися чутлива інформація у найрізноманітніших сферах.

При цьому отримано такі основні результати:

1. Виділено недоліки відомих методів та стратегій, а також систематизація їх у подальших розробках.

2. Розроблено модель базової архітектури системи для покращення конфіденційності зберігання великих даних із подальшим її вдосконаленням та нарощуванням методів, як і наявних, так і новостворених на основі наявних методів. Спроєктована система виконана таким чином, що її компоненти можуть обмінюватися результатами обробки даних.

3. Удосконалено початковий алгоритм захисту конфіденційності до методу збурення розподілених даних, який може забезпечити конфіденційність під час машинного навчання. У даній системі АЗРД весь контроль генерації даних та глобальних параметрів належить центральному контролюючому органу, в той час як локальне збурення провозиться безпосередньо генерацією глобальних параметрів.

4. Розроблено метод збурення потоку даних, який продемонстрував безперебійну роботу із зростаючими потоками даних. Також його було удосконалено до безпечного алгоритму захисту даних, який слугує чудовим засобом для збереження конфіденційності у різноманітних сферах.

5. Здійснено реалізацію рішень та імплементацію їх у систему охорони здоров’я для демонстрації роботи методу забезпечення конфіденційності для великомасштабної аналітики на основі машинного навчання.

За темою кваліфікаційної роботи магістра опубліковано одна стаття у збірнику наукових праць Всеукраїнської наукової конференції та дві публікаці у журналах міжнародних наукових конференцій.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1.Ференс В.О, Бармак О.В. Дослідження аспектів проектування та оптимізації взаємодії компонентів інтернету речей за стандартом Nb-IoT. *Збірник наукових праць XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021».* Хмельницький, 2021. с. 15-16.

2. V. Ferens, O. Pasichnyk, Key Trends of Smart Home Development. *Current Trends in Young Scientists' Research V All Ukrainian Scientific and Practical Conference,* April 12, 2018. Zhytomyr: ZSTU, 2018. P. 74-76

3. V. Ferens, O. Pasichnyk, Key Trends of Smart Home Development. *Current Trends in Young Scientists' Research VI All Ukrainian Scientific and Practical Conference* (April 11, 2019). Zhytomyr: ZSTU, 2018. P. 110-113.

4. Wawryn, K., Widuliński P. Detection of anomalies in compiled computer program files inspired by immune mechanisms using a template method. *Journal of Computer Virology and Hacking Techniques.* 2020. URL: https://doi.org/10.1007/s11416-020-00364-w

5. Zeng J., Tang W. Negative Selection Algorithm Based Unknown Malware Detection Model. In: Gong M., Linqiang P., Tao S., Tang K., Zhang X. (eds) Bio-Inspired Computing - Theories and Applications. BIC-TA 2015. *Communications in Computer and Information Science*, vol. 562. Springer, Berlin, Heidelberg. URL: https://doi.org/10.1007/978-3-662-49014-3_53

6. Корченко А. О. Методи ідентифікації аномальних станів для систем виявлення вторгнень: *автореф. дис. ... д-ра техн. наук*: 05.13.21, Київ, 2019, 40 с.

7. Лукова-Чуйко Н. В. Методологічні основи забезпечення функціональної стійкості розподілених інформаційних систем до кібернетичних загроз: *автореф. дис. ... д-ра техн. наук:* 05.13.06, Київ, 2018, 40 с.

8. Virusbulletin. URL: https://www.virusbulletin.com/testing/ (дата звернення 20.04.2021).

9. Av-test. URL: https://www.av-test.org/en/statistics/malware/ (дата звернення

20.04.2021).

10. Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications Vol. 60*, January 2016. P. 19-31.

11. Bernadette J. Stolz, Jared Tanner, Heather A. Harrington, Vidit Nanda Geometric anomaly detection in data. *Proceedings of the National Academy of Sciences* Aug 2020, 117 (33) 19664-19669; DOI: 10.1073/pnas.2001741117

12. Xiang Yu, Hui Lu, Xianfei Yang, Ying Chen, Haifeng Song, Jianhua Li, Wei Shi An adaptive method based on contextual anomaly detection in Internet of Things through wireless sensor networks. *International Journal of Distributed Sensor Networks* 2020. Vol. 16(5) DOI: 10.1177/1550147720920478

13. Goldstein M., Uchida S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. 2016. *PLOS ONE* 11(4): e0152173. URL: https://doi.org/10.1371/journal.pone.0152173

14. Hayes, M.A., Capretz, M.A. Contextual anomaly detection framework for big sensor data. *Journal of Big Data*. 2015. URL: https://doi.org/10.1186/s40537-014-0011y

15. Liu, L., Hu, M.; Kang, C., Li, X. Unsupervised Anomaly Detection for Network *Data Streams in Industrial Control Systems.* Information *2020,* 11*, 105.* URL: https://doi.org/10.3390/info11020105

16. Xiaodan Xu, Huawen Liu, Minghai Yao. Recent Progress of Anomaly Detection. Complexity, vol. 2019, Article ID 2686378, 11 pages, 2019. URL: https://doi.org/10.1155/2019/2686378

17. Jianwen Huang, Zhen Chai and Hailong Zhu. *Detecting anomalies in data center physical infrastructures using statistical approaches. Journal of Physics: Conference Series, Volume 1176, Issue 2. Jianwen Huang* et *al* 2019 *J. Phys.: Conf. Ser.* 1176 022056

18. Fisch, A., Grose, D., Eckley, I.A., Fearnhead, P., & Bardwell, L. (2020). anomaly: Detection of Anomalous Structure in Time Series Data. *arXiv: Applications*.arXiv:2010.09353

19. Lu, X., Wang, S., Kang, F., Liu, S., Li, H., Xu, X. and Cui, L. (2019). *An*

anomaly detection method to improve the intelligent level of smart articles based on multiple group correlation probability models. International Journal of Crowd Science, Vol. 3, № 3. P. 333-347. URL: https://doi.org/10.1108/IJCS-09-2019-0024

20. Solarz A., Bilicki M., Gromadzki M., Pollo A. , Durkalec A., Wypych M. Automated novelty detection in the WISE survey with one-class support vector machines. *Published online*: 05 October 2017. DOI: 10.1051/0004-6361/201730968

21. Mukrimah Nawir, Amiza Amir, Naimah Yaakob, Ong Bi Lynn. Effective and efficient network anomaly detection system using machine learning algorithm. *Bulletin of Electrical Engineering and Informatics* Vol.8, No.1, March 2019, pp. 46~51ISSN: 2302-9285, DOI: 10.11591/eei.v8i1.1387

22. Anta A., Hadjistasi T., Nicolaou N. et al. Tractable low-delay atomic memory. *Distrib. Comput*. 34, 33–58 (2021). URL: https://doi.org/10.1007/s00446-020-00379-y

23. Ouyang L., Huang Y., Wei H., Lu J. Achieving Probabilistic Atomicity With Well-Bounded Staleness and Low Read Latency in Distributed Datastores. *in IEEE Transactions on Parallel and Distributed Systems.* vol. 32, № 4, pp. 815-829, 1 April 2021, doi: 10.1109/TPDS.2020.3034328.

24. Lakshman A. and Malik P. Cassandra: A decentralized structured storage system, *SIGOPS Operating Syst. Rev.*, vol. 44, no. 2, pp. 35-40, Apr. 2010.

25. Ganesh, C., Patra, A. Optimal extension protocols for byzantine broadcast and agreement. *Distrib. Comput*. 34, 59–77 (2021). https://doi.org/10.1007/s00446-020-00384-1

26. Huang, Z., Radunovic, B., Vojnovic, M. et al. Communication complexity of approximate maximum matching in the message-passing model. *Distrib. Comput*. 33, 515–531 (2020). https://doi.org/10.1007/s00446-020-00371-6

27. Czumaj, A., Konrad, C. Detecting cliques in CONGEST networks. *Distrib. Comput*. 33, 533–543 (2020). https://doi.org/10.1007/s00446-019-00368-w

28. Abboud, A., Censor-Hillel, K., Khoury, S. et al. Fooling views: a new lower bound technique for distributed computations under congestion. *Distrib. Comput*. 33, 545–559 (2020). https://doi.org/10.1007/s00446-020-00373-4

29. Di Luna, G.A., Flocchini, P., Izumi, T. et al. Fault-tolerant simulation of

population protocols. *Distrib. Comput*. 33, 561–578 (2020). https://doi.org/10.1007/s00446-020-00377-0

30. Ellen, F., Gelashvili, R., Shavit, N. et al. A complexity-based classification for multiprocessor synchronization. *Distrib. Comput*. 33, 125–144 (2020). https://doi.org/10.1007/s00446-019-00361-3

31. Chatterjee, S., Pandurangan, G. & Robinson, P. The complexity of leader election in diameter-two networks. *Distrib. Comput.* 33, 189–205 (2020). https://doi.org/10.1007/s00446-019-00354-2

32. Busch, C., Herlihy, M., Popovic, M. *et al.* Time-communication impossibility results for distributed transactional memory. *Distrib. Comput.* 31, 471–487 (2018). https://doi.org/10.1007/s00446-017-0318-y

33. Boczkowski, L., Korman, A. & Natale, E. Minimizing message size in stochastic communication patterns: fast self-stabilizing protocols with 3 bits. *Distrib. Comput.* 32, 173–191 (2019). https://doi.org/10.1007/s00446-018-0330-x

34. Min B., Varadharajan V. (2014) Feature-Distributed Malware Attack: Risk and Defence. In: Kutyłowski M., Vaidya J. (eds) Computer Security - ESORICS 2014. ESORICS 2014. Lecture Notes in Computer Science, vol 8713. Springer, Cham. https://doi.org/10.1007/978-3-319-11212-1_26

35. Nath H. V., Mehtre B.M. (2014) Static Malware Analysis Using Machine Learning Methods. In: Martínez Pérez G., Thampi S.M., Ko R., Shu L. (eds) Recent Trends in Computer Networks and Distributed Systems Security. *SNDS 2014. Communications in Computer and Information Science,* vol 420. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-54525-2_39

36. Merayo, M. G., Hierons, R. M., Núñez, M. Passive testing with asynchronous communications and timestamps. *Distrib. Comput.* 31, 327–342 (2018). https://doi.org/10.1007/s00446-017-0308-0

37. Michail, O. Terminating distributed construction of shapes and patterns in a fair solution of automata. *Distrib. Comput.* 31, 343–365 (2018). https://doi.org/10.1007/s00446-017-0309-z

38. Angluin, D., Aspnes, J., Eisenstat, D.: Fast computation by population

protocols with a leader. *Distrib. Comput. 21,* 183–199 (2008)

39. Aspnes, J., Ruppert, E.: An introduction to population protocols. In: Garbinato, B., Miranda, H., Rodrigues, L. (eds.) *Middleware for Network Eccentric and Mobile Applications* pp. 97–120. Springer, Berlin (2009)

40. Attiya, H., Welch, J.: Distributed Computing: Fundamentals, Simulations, and Advanced Topics, vol. 19. Wiley, New York (2004)

41. Chen, H.-L., Doty, D., Soloveichik, D.: Deterministic function computation with chemical reaction networks. *Nat. Comput*. 13, 517–534 (2014)

42. Censor-Hillel, K., Parter, M. & Schwartzman, G. Derandomizing local distributed algorithms under bandwidth restrictions. *Distrib. Comput.* **33,** 349–366 (2020). https://doi.org/10.1007/s00446-020-00376-1

43. Malwarebytes Endpoint Security. URL: https://ru.malware bytes. com/business/endpoint security (дата звернення: 3.04.2022).

44. Barenboim, L., Elkin, M., Gavoille, C.: A fast network-decomposition algorithm and its applications to constant-time distributed computation. *SIROCCO*, pp. 209–223 (2015)

45. Benjamini, I., Gurel-Gurevich, O., Peled, R.: On k-wise independent distributions and Boolean functions. *arXiv preprint arXiv:1201.3261* (2012)

46. Antoniadis, K., Blanchard, P., Guerraoui, R. *et al.* The entropy of a distributed computation random number generation from memory interleaving. *Distrib. Comput.* **31,** 389–417 (2018). https://doi.org/10.1007/s00446-017-0311-5

47. Alistarh, D., Sauerwald, T., Vojnovic, M.: Lock-free algorithms under stochastic schedulers. *In: Proceedings of the 2015 ACM Symposium on Principles of Distributed Computing, PODC 2015*, Donostia-San Sebastián, Spain, July 21–23, 2015, pp. 251–260 (2015). doi:10.1145/2767386.2767430

48. Barker, E., Kelsley, J. Recommendation for random bit generator (rbg) constructions. SP 800-90C (2012)

49. Zhou, H., Bruck, J.: Generalizing the Blum-Elias method for generating random bits from markov chains. In: *Proceedings of IEEE International Symposium on Information Theory (ISIT)* (2010)

50. Savenko O. Interoperability of distributed multiple system for malware detection based on components levels of safety. *Проблеми інформаційних технологій*. 2018. № 24. С. 78-92.

51. Corporate Endpoint Protection Products Group Test: Socially-Engineered Malware Q2. URL: http://www.nsslabs.com/ research/ endpoint-security/ anti-malware/q2-2010-endpoint-protection-product-group-test.html (дата звернення: 3.04.2022).

52. ESET Endpoint Security. URL: http://www.eset.com/ (дата звернення: 3.04.2022).

53. Symantec Endpoint Protection. URL: https://www.anti-malware.ru/reviews/ Symantec_Endpoint_Protection (дата звернення: 3.04.2022).

54. Branitskiy A., Kotenko I. Hybridization of computational intelligence methods for attack detection in computer networks. Journal of Computational Science. 2017. №23. P. 145–156.

55. Bitdefender. URL: http://www.bitdefender.com/ (дата звернення 20.04.2021).

56. Securelist. FAQ: Disabling the new Hlux / Kelihos Botnet. URL: https://securelist.com/blog/research/32634/faq-disabling-the-new-hluxkelihos-botnet-13/ (дата звернення: 3.04.2022).

57. COMSS1. URL: http://www.comss.ru/page.php?id=2758 (дата звернення: 3.04.2022).

58. Enterprise End Point Protection Comparative Analysis - Socially Engineered Malware: Report Overview. URL: www.nsslabs.com/reports (дата звернення: 3.04.2022).

59. Bakotech. URL: https://bakotech.ua/ (дата звернення: 3.04.2022).

60. ClamAV. URL: https://www.clamav.net/ (дата звернення: 3.04.2022).

61. Wireshark.org URL: http://www.wireshark.org/docs/dfref. (дата звернення 20.04.2021)

62. Kaspersky Lab. URL: http://www.kaspersky.ru (Viewed on April 2, 2019).

63. Avast!. URL: https://www.avast.com/index (дата звернення 20.04.2021).

64. AVG. URL: http://www.avg.com (дата звернення 20.04.2021).

65. Avira. URL: http://www.avira.com (дата звернення 20.04.2021).

ДОДАТОК А

(обов'язковий)

КОПІЯ ТЕЗ ДОПОВІДІ

УДК 004.7.056.5

Ференс В. О., Бармак О. В.

Хмельницький національний університет

ОСОБЛИВОСТІ ВИКОРИСТАННЯ ПРОТОКОЛУ NB-IOT ДЛЯ ПРОЕКТУВАННЯ ТА ОПТИМІЗАЦІЇ ВЗАЄМОДІЇ КОМПОНЕНТІВ ІНТЕРНЕТУ РЕЧЕЙ

Розглянуто та описано взаємодію компонентів Інтернету речей за протоколом NB-IoT, що є сучасною технологією доступу до передачі даних за еволюцією (LTE), для широкополосних мереж малої потужності (LPWAN). Основною метою NB-IoT є підтримка масового використання комунікацій машинного типу (mMTC) та забезпечення низької потужності, недорогої та низької швидкості передачі даних при комунікації. NB-IoT базується на LTE технології з деякими змінами, що відповідають вимогам mMTC. Метою роботи є огляд змін у конструкції, які внесені у стандарт NB-IoT та перспективи застосувань.

The interaction of the Internet components of Things according to the NB-IoT protocol - a modern technology for access to data transmission by evolution (LTE) for low-bandwidth networks (LPWAN) – has been considered and described. The main aim of NB-IoT is to support massive machine-type communication (mMTC) and enable low-power, low-cost, and low-data-rate communication. NB-IoT is based on LTE design with some changes to meet the mMTC requirements. The aim of this survey is to provide a comprehensive overview of the design changes brought in the NB-IoT standardization along with the application prospects.

Протокол Nb-IoT – це специфікація стандарту стільникового зв'язку, яка розроблена для обслуговування пристроїв, що генерують невеликий обсяг даних. Технологія підходить для підключення до цифрових мереж зв'язку широкого спектру автономних пристроїв. Екологія – моніторинг стану навколишнього середовища: рівень забруднення повітря, температура тощо. Енергетика – контроль і оптимізація водних, газових, енергетичних ресурсів. Smart city – освітлення, системи розумного сміття, паркування, розумні зупинки, аналіз стану люків і каналізаційних систем. Охорона і безпека – сигналізаційні системи: відкриття/закриття дверей, контроль присутності в приміщенні тощо. Житлове та комунальне господарство ЖКГ – автоматизований збір даних по газу, воді, теплу, своєчасне визначення аварійних ситуацій на всіх рівнях системи. Сільське господарство – віддалений аналіз стану ґрунту, розумне зрошення. Широко застосовується для контролю персоналу у різноманітних сферах, зокрема медицині.

Останнім часом Інтернет речей використовується як компонент в експлуатаційних технологіях атомних електростанцій моніторингових радіаційних мережах. У цих галузях потрібна надійна та ефективна мережева інфраструктура

для оптимальної роботи. На відміну від людського трафіку, профіль комунікації IoT повністю відрізняється від того, який є у смартфонах у стільникових мережах четвертого покоління (4G).

Рисунок 1 – Галузі застосування протоколу Nb-IoT

Своєю фізичною структурою та архітектурою мережа Nb-IoT практично все успадкувала від LTE, тому побудова інфраструктури для Інтернету речей не вимагає нічого, окрім оновлення ПЗ на наявних базових станціях. За рахунок простоти системи оператори можуть надавати низькі тарифи для клієнтів Інтернету речей.

Ефективні та прибуткові комунікаційні системи Інтернету речей повинні мати низьку вартість, розмір, вагу та потужність. Зі збільшенням залежності від технології NB-IoT у житті людей, велику кількість терміналів NB-IoT потрібно підключити до мережі, щоб задовольнити різні потреби користувачів.

Технологія NB-IoT має низку переваг по швидкості, масовості виробництва і, як наслідок, по дешевизні. Можливість використовувати для NB-IoT освоєні ліцензовані діапазони частот для операторів мобільних мереж 4G, а також вже розгорнуте мережеве обладнання призведе до інтенсивного розвитку сектора IoT в структурі бізнесу мобільних операторів.

Швидке зростання кількості масових пристроїв Інтернету речей додало проблему для дизайну 5G через його контрастні вимоги з двох основних позицій. По-перше, вузька пропускна здатність може обмежувати продуктивність передачі IoT на застосунок, збільшувати накладні витрати на зв'язок, необхідні перед передачею даних. Тому важливо досягати постійних удосконалень щодо ефективного використання цих ресурсів для підтримки подальшого зростання компонентів Інтернету речей. По-друге, протоколи резервування доступу, призначені для мереж 5G, стикаються з величезною кількістю підключень для

застосунків IoT, оскільки, як очікується, щільність пристрою буде більшою, ніж здатність методів реалізувати процедури резервування доступу, і може суттєво вплинути на наскрізну затримку. Ці питання будуть критичним, особливо для сценаріїв швидкого доступу. Запити доступу та сигналізація, пов'язана з кожним періодичним процесом передачі IoT, є тягарем як для пристрою IoT, так і для мережевих ресурсів. Ці проблеми далі передаються у трафік висхідної лінії зв'язку і залишаються відкритими.

На наведеному нижче зображенні (рисунок 2) видно дві гілки розвитку мереж: на одній підвищується швидкість Інтернету, а на іншій - знижується енергоспоживання клієнтського обладнання. Nb-IoT відноситься саме до другої гілки: вводяться нові обмеження для приймачів. Смуга зв'язку істотно звужується, це як раз відображено в назві самої мережі: NB - Narrow Band, тобто вузька смуга. Обмеження призводять до зниження швидкості передачі даних, яка вимірюється у кілобайтах в секунду або навіть кілобітах.

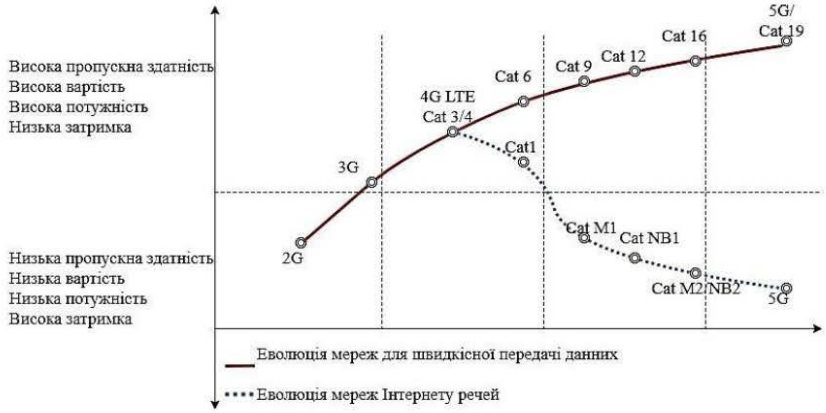


Рисунок 2 – Розвиток NB-IoT

Через великі перспективи використання інтернету речей (IoT) і постійно зростаючий попит на нього, стимулює процес оптимізації, оскільки якщо ця технологія стане використовуватись набагато ширше, навантаження на мережу може бути колосальним.

Перелік посилань

1. Abbas A. M. et al. NB-IoT optimization for smart meters networks of smart cities: Case study //Alexandria Engineering Journal. – 2020. – Т. 59. – №. 6. – С. 4267-4281.
2. Raj J. S. et al. QoS optimization of energy efficient routing in IoT wireless sensor networks //Journal of ISMAC. – 2019.
3. NB IoT (Narrowband Internet of Things) [Електронний ресурс]. Режим доступу: bit.ly/3uE3wUC

ДОДАТОК Б

(обов'язковий)

КОПІЯ ТЕЗ ПУБЛІКАЦІЙ В ЖУРНАЛІ

1. V. Ferens, O. Pasichnyk, Key Trends of Smart Home Development. Current Trends in Young Scientists' Research V All Ukrainian Scientific and Practical Conference (April 12, 2018) – Zhytomyr: ZSTU, 2018. P. 74-76

by detecting motion, by turning on the flood light, starting the alarm or even recording. The detection can be done by measuring the speed or the vector of the objects in the field of view [1].

Raspberry Pi 2 (a small single-board computer), Raspberry Pi Camera Module v2 and Passive Infrared sensor (PIR) were used for the creation of the motion detection system. This computer is the main module for image capturing and processing. The camera plugs directly into Raspberry Pi and delivers 5 MP image resolution; 1080p, 720p, 480p video recording [2]. A highly effective, low-cost technology for detecting unauthorized activity by monitoring changes in infrared heat patterns in the area under observation. The PIR sensor detects the change in temperature and triggers when an unauthorized object enters the perimeter. It can turn an alarm or security recording on. The great advantage of the PIR detection systems is a small memory bank that receives the amount of infrared energy typically focused on its surface when there is no activity in the area [3].

Current system operates as follows: PIR sensor executes motion detection, camera executes image capture, and Raspberry Pi executes data processing. System analysis showed that PIR sensor has a number of disadvantages, such as necessity of special conditions for the arrangement, range of motion detection and probability of false alarms. So, it was decided to remove it from the system and to implement camera-based algorithm for recognition of alive objects.

REFERENCES

1. Advantages of installing motion detecting camera [Електронний ресурс] // [сайт] URL: http://www.cctv-camera.com.sg/articles/5-advantages-of-installing-motion-detecting-camera.html (дата звернення: 30.03.2018).

2. The Raspberry Pi Foundation [Електронний ресурс] // [сайт] URL: https://www.raspberrypi.org (дата звернення: 30.03.2018).

3. Security and self-defence store [Електронний ресурс] // [сайт] URL: http://www.securityandselfdefensestore.com/resources/47-security-cameras/151-pir-detection.html (дата звернення: 30.03.2018).

UDC 004.8

V. Ferens, BSc Student
O. Pasichnyk, PhD in Education, As. Prof., language advisor
Khmelnitsky National University

KEY TRENDS OF SMART HOME DEVELOPMENT

In today's world technology is advancing with a fast pace, radically changing our lives. A lot of companies which produce various household appliances and gadgets implement some kind of artificial intelligence technologies in them. Moreover, these devices can be linked into a single network. This enables them to "communicate" with each other and with the host, thus forming a system of "smart house".

Of course, such systems are a far cry from real smart homes from science fiction shows, where systems of artificial intelligence behave like a human, fully in charge of the house and are able to create Virtual Reality. But this is only the beginning.

There's no doubt that technology advances faster than we can even keep up these days, and the smart home sector is one of the fastest-growing. At CES 2018, hundreds of companies showed off a variety of new smart home solutions and gadgets, from the useful and innovative to the repetitive and uneventful. Here we offer an outline of key technologies in the sphere, which are believed to shape the nearest trends of "smart home" further development.

1. Voice assistants in everything.

Smart home tech allows you to automate your home, from lighting and security to entertainment and cooking. More and more of these devices are starting to get voice-activated upgrades baked right into the hardware. Last year, we saw countless devices and gadgets that work with voice assistants, but you needed a dedicated smart speaker to take advantage.

Now, one can basically skip the speaker entirely. But if your home has already got an Echo or a Google Home, having more devices that respond to person's voice commands could lead to over-saturation and frustration. The breadth of options to come in 2018 will definitely create more ways to build the perfect smart home, but it will also require smarter planning on the part of the consumer.

2. **Smart home security suites that suit all needs**.

There are numerous smart home security devices available on the market: motion sensors, security cameras, smart locks, video doorbells, etc. However, if you are buying different devices from different brands, syncing them all up can prove difficult, and juggling multiple apps is a hassle.

Various smart home companies are launching security solutions that make it easy to keep an eye on things from one access point whether you want to stick with a single brand or mix and match. What is really compelling about all of the new tech is that it can be customized to individual homes and preferences for a completely DIY experience rather than an out of the box solution that won't always fit every home.

This year, they will finally be releasing the Ring Alarm Security Kit (originally the Ring Protect) as well as motion sensors, night-lights, outdoor lighting, smoke and water leak detectors, and new wireless, standing smart cameras for a holistic way to keep an eye on your home all in one app.

GlobalLogic, a proven design company, has a wealth of experience in automating smart home projects for diverse customers from around the world, from startups to technological giants. Our practice proves that while mobile app interfaces are a great tool, it's easier and more natural for a user to communicate at home through voice.

3. **Conversation with a device** that has artificial intelligence, understands a person and performs respective commands, brings a completely unusual impression. According to a study by Gartner, by the end of 2018, 30% of our technology interaction will take place in a conversation format with smart machines. Leaders in technology products and services must now invest in improved voice interfaces, which are still

limited. The future comes in the form of a home personal assistant, and in the end, the language itself will become a versatile interface for home use.

An excellent example is Aleksa from Amazon: Among other things, Aleksa organizes the voice control with smart home appliances. Over the years of Siri's upgrade, Google Now and Cortana have also become technologically advanced personal assistants at the head of the smart home-ecosystems.

Voice-guided intellectual functions of the house develop gradually. So, Mark Zuckerberg created a system called Jarvis for his own home. The name of this device is similar to a computer with artificial intelligence belonging to Tony Stark in the movie "Iron Man": his name is J.A.R.V.I.S. stands for Just A Rather Very Intelligent System.

4. **A wonderful time for progress ... But there are also difficulties**.

Artificial intelligence is a key element of human interaction with a machine. However, the implementation of even simple functions, such as turning on and off the light, is a complicated task, since commands can be formulated very differently: "Turn off the lights in the bedroom" or "Keep the lamp in the baby".

Teaching home assistants to understand different linguistic senses is an important step towards positive user experience. Voice biometrics and intelligent decision-making will help you to come here. For the phrase "Play my favorite song," the home assistant must be able to distinguish the voice of a person and choose music according to their preferences.

To implement the potential of intelligent home systems, market leaders need to collaborate and form unexpected partnerships. This will ensure the emergence of high-quality new devices and services interconnected, which will make the life of the user even easier. Currently, the focus of the experts is the global benefits of smart homes, and for a long time such systems should move from the "bad mother" category to the "must have" category.

In the end, security and reliability are key parameters that should remain at the height of the evolution of technology for housing. For example, given the fact that the voice command can activate the security-related function, it is necessary to ensure the exact delimitation of a certain person's live voice and audio recordings in order not to hit the catchers. And soon we can all have Iron Man technology or a computer to manage our home, car, and home appliances.

Undoubtedly, artificial intelligence and systems of "smart house" are our future. And this is not only convenient, but it also preserves the planet's condition and saves our time.

REFERENCES

1. A smart home can finally become your home Available at: https://www.globallogic.com/ua/gl_news/the-smart-home-finally-comes-home/

2. 4 smart home trends to watch in 2018 Available at: http://smarthome.reviewed.com/features/4-smart-home-trends-to-watch-in-2018

2. V. Ferens, O. Pasichnyk, Key Trends of Smart Home Development. Current Trends in Young Scientists' Research VI All Ukrainian Scientific and Practical Conference (April 11, 2019) – Zhytomyr: ZSTU, 2018. P. 110-113

Synthesis under the influence of ultraviolet radiation. It lies in the fact that the oxygen-containing gas is passed through a cooled and transparent ultraviolet radiation (for example, quartz) reactor irradiated by a source of ultraviolet radiation having a corresponding spectrum. The gas, as a rule, is embodied in the form of pure oxygen.

All methods have a number of advantages and disadvantages. For example, ultraviolet light syntheses are simpler to implement, but far less productive than all other methods. Therefore it is not used in industrial devices.

For non-industrial scale, it would be advisable to use the quiet method as its performance is rather high, and the complexity and cost of realization is moderate. The following is a block diagram of the apparatus for generating ozone.

Fig.1:

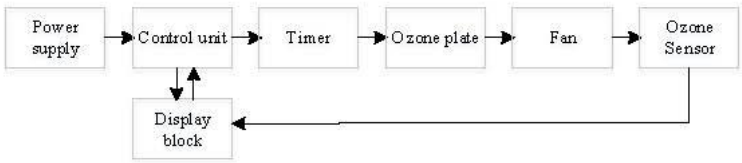


Fig.1: An apparatus for generating ozone

V. Ferens, Student
O. Pasichnyk, PhD in Educ., As. Prof., language advisor
Khmelnitsky National University

KEY TRENDS OF SMART HOME DEVELOPMENT

In today's world technology is advancing with a fast pace, radically changing our lives. A lot of companies which produce various household appliances and gadgets implement some kind of artificial intelligence technologies in them. Moreover, these devices can be linked into a single network. This enables them to "communicate" with each other and with the host, thus forming a system of "smart house". Of course, such systems are a far cry from real smart homes from science fiction shows, where systems of artificial intelligence behave like a human, fully in charge of the house and are able to create Virtual Reality. But this is only the beginning. In the past the concept of "smart home" often meant a system of remote control of light, heating and washing machine. Today this concept has become much wider. Nowadays there are "smart" materials, programs for smartphones that allow you to control appliances at a distance. They can do cleaning for you and many more. We have identified three main technologies that can change our houses and become their integral part in the future.

1. Drowning In To-Do Lists

For my smart home podcast series, I've been interviewing my friends to find out what tools they use to manage their list of to-do's. "I keep them in a Google doc," one friend told me. "I keep it multiple Google Docs," said another friend. "Each one is dated, and I when I think I'm no longer serious about following a list, I simply create

another one with a new date." One guy used Evernote. Best of all was a friend of mine who explained how his to-do lists are memorialized with stickies on his bedroom wall, much to the chagrin of his wife.

While the tools were all different, the one thing that everyone seemed to have in common was a general feeling of failure when it came to crossing enough things off their list and an abiding belief that there was too much to do in too little time. Everyone seemed to be searching for a magic elixir that would save them more time

I often wonder how technology can improve our lives. One area in particular that fascinates me is identifying tasks that technology can handle so that they don't need to appear on my to-do list, and just as importantly, so that they won't occupy space in my mind. I read once about the dressing habits of people like Steve Jobs and Mark Zuckerberg who seemingly wear the same outfit everyday. Upon closer inspection, it turns out that both men have multiple identical pants and identical shirts. The reason for wearing the same outfit every day? If you always wear the same thing, then there is no decision to make. You can then turn to more important decisions and lead a more productive life.

How, you might ask, are to-do lists and clothes connected to the smart home? I've explored how technology like the smart thermostat or smart lighting could save me money if they only turned on when I was in a room in need of heat or air conditioning or light. That's interesting, but what's infinitely more exciting to me is if the smart home could offload my decisions and work by completing tasks independently of me. Fewer decisions that I need to make means more time for me to focus on the things that really matter.

2. A Smart Home Driven By Artificial Intelligence

In many industries, when you interview an ambitious leader, he or she will talk with you about how they will reinvent factory-built housing or the fitness space or retail. However, in some, people will talk about how they are part of an ecosystem and how their success is in large part predicated on the success of other companies in the ecosystem. In the case of the smart home, almost all of the players I interviewed talked about a future where the holy grail was a home driven by Artificial Intelligence.

Think of Artificial Intelligence as computing power that is able to perform particularly complex tasks that would otherwise require a human brain to perform. A motion sensor might trigger a light to turn on. But if a home had Artificial Intelligence, it might consider the time of day, the person walking around the home, and where she was walking in deciding which light to turn on and how long to keep it on for. Not every person I spoke to used the words "Artificial Intelligence." A hot phrase you'll hear again and again from experts is that a house needs to be "aware" or "contextually aware" before you can bring Artificial Intelligence into the home.

Let's imagine the universe of things a house can be aware of: it can be aware of the presence of the people who live in the house (along with their personas); it can be aware of what they're doing; it can even be aware of what every device in the house is doing. If you want the house to think like a human, the house needs to be able to analyze the data a human would analyze before making a decision.

3. Your Home As Your Personal Caretaker

How would it work for a smart home to free me of some of my decision-making? How could it lighten the load for me, literally and figuratively? Let's imagine a day

together. You wake up in the morning and your alarm goes off. It's not a buzzer. You want to discover new music on Spotify and this song is on your suggested Discover Weekly list. What's really interesting, though, is not the song. It's the fact that you didn't have to set the alarm the night before.

That's because there is some level of intelligence in the cloud that's watching over you and trying to simplify your life. It knows that today you have a spin class because it checked your workout goals, which then checked availability for a class at SoulCycle, which then purchased the class, which then put it on your calendar. The system was smart enough to calculate travel time and set the alarm appropriately.

You stagger out of bed and walk down the stairs to the kitchen. The coffee just finished brewing. You have your smart home to thank for that. Your yogurt and granola is ready in the exact proportions you want inside the refrigerator. The refrigerator knew earlier in the week that you were running low on breakfast foods and placed an order online. You're in a rush, so you walk out the door and leave for the gym.

There's no time to set the alarm or draw the blinds (which is something you do when you leave the house so that people can't look in while you're away). You don't think to turn off the music or the lights or lower the heat, as you won't need to heat the house to 72 degrees while you're away. It's not that you forget to do all of those things. You just don't have to think about them, because the house knows that you left. It knows to lock the door behind you, to turn off the coffee maker, to pull the blinds, to reduce the heat, to shut off the music, and to turn off the lights.

Today is shopping today. Really, every day is shopping day. The sensors in your drawers measure the toilet paper that is left, and the sensors in the closet monitor cleaning supplies and laundry detergent. You're running low on a few things. The online order is placed. When it arrives, the cameras at your front door will recognize the FedEx truck and coordinate with the lock to pop open your front door. The delivery man's picture will be taken and a gentle voice will come on over your speakers, asking him to set down the packages just inside the house. Cameras will be watching him from beginning to end, and the door will close on its own behind him when he leaves. Your home's robot then proceeds to unpack the items and place them where they belong.

4. A wonderful time for progress ... But there are also difficulties.

Artificial intelligence is a key element of human interaction with a machine. However, the implementation of even simple functions, such as turning on and off the light, is a complicated task, since commands can be formulated very differently: "Turn off the lights in the bedroom" or "Keep the lamp in the baby".

Teaching home assistants to understand different linguistic senses is an important step towards positive user experience. Voice biometrics and intelligent decision-making will help you to come here. For the phrase "Play my favorite song," the home assistant must be able to distinguish the voice of a person and choose music according to their preferences.

To implement the potential of intelligent home systems, market leaders need to collaborate and form unexpected partnerships. This will ensure the emergence of high-quality new devices and services interconnected, which will make the life of the user even easier. Currently, the focus of the experts is the global benefits of smart homes, and for a long time such systems should move from the "bad mother" category to the "must have" category.

In the end, security and reliability are key parameters that should remain at the height of the evolution of technology for housing. For example, given the fact that the voice command can activate the security-related function, it is necessary to ensure the exact delimitation of a certain person's live voice and audio recordings in order not to hit the catchers. And soon we can all have Iron Man technology or a computer to manage our home, car, and home appliances.

Undoubtedly, artificial intelligence and systems of "smart house" are our future. And this is not only convenient, but it also preserves the planet's condition and saves our time.

REFERENCES

1. A smart home can finally become your home Available at: https://www.globallogic.com/ua/gl_news/the-smart-home-finally-comes-home/

2. 4 smart home trends to watch in 2018 Available at: http://smarthome.reviewed.com/features/4-smart-home-trends-to-watch-in-2018

3. How Homes Powered By Artificial Intelligence Will Know & Care For You https://www.forbes.com/sites/andrewweinreich/2018/02/08/the-future-of-the-smart-home-how-homes-powered-by-artificial-intelligence-will-know-care-for-you/

R. Gavrilyuk, Student
O. Pasichnyk, PhD in Educ., As. Prof., language advisor
Khmelnitsky National University

METHODS OF TEXT RECOGNITION

In the course of everyday activities, government structures, business, and academic institutions educational institutions use a large number of paper documents, most of which are handwritten. A large amount of data and knowledge is contained in printed or handwritten documents that are archived. The need is growing digitization of paper documents in order to further process their content computerized computer systems.

Text recognition can be divided into several areas that are sufficient significantly differ in their methods of solving. The text can be printed either manuscript. Any of them can be extra structured. For example, Formulas can contain different levels of records, such as superscripts, sublines, special marks for mathematical actions, etc.

To date, there are a number of methods that solve the problem print text recognition, but there are still no systems capable of to recognize any handwritten text [1]. Existing systems can suffice not qualitatively recognize specific handwriting. Therefore, the task of development is relevant handwriting recognition method that will allow you to process it handwritten documents.

Before the text is recognized, there is always a previous one processing the input image. The first step is to improve the quality image. At this stage, increase the contrast and sharpness of the image, as well as filtration from noise. The next step is segmentation [2], by which the structure of the text is determined. Segmentation

ДОДАТОК В

ПРЕЗЕНТАЦІЯ ДОПОВІДІ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Кафедра комп'ютерної інженерії та системного програмування

Ференс Володимир Олександрович

Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання

Науковий керівник – д.т.н., проф. Бармак О.В.

Хмельницький - 2022

1

Мета і задачі дослідження

Метою роботи є покращення конфіденційності зберігання великих даних.

Об'єкт дослідження – процес збереження конфіденційності великих даних

Предмет дослідження – методи і засоби для забезпечення конфіденційності великих даних на основі машинного навчання та створення методів на основі базових алгоритмів забезпечення конфіденційності.

2

Мета і задачі дослідження

Поставлена мета досягається розв'язанням наступних задач:

- дослідження та розгляд понять конфіденційності великих даних;
- проаналізувати відомі методи для покращення конфіденційності великих даних;
- виділити основні недоліки методів і стратегій захисту даних та систематизувати їх
- розробити алгоритми захисту конфіденційності на основі відомих методів, метод збурення розподілених даних, алгоритми збурення даних із використанням локальної диференціальної конфіденційності для великих даних
- здійснити реалізацію азпропонованих рішень;
- провести експериментальні дослідження з розробленими засобами;
- схематично імплементувати алгоритми і зобразити їх роботу.

3

Наукова новизна отриманих результатів

Розроблено новий метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання

Практичне значення отриманих результатів

Практичне значення одержаних результатів полягає у вдосконалених методах забезпечення конфіденційності для великомашстабних даних не тільки для великих корпорацій, які мають великі потужності пристроїв, але і з підтримкою стаціонарних машин із дотриманням їх працездатності під час обміну інформації між такими компаніями.

4

Відомі методи, на яких ґрунтуються методи захисту конфіденційності для великих даних

- методи збурення даних
- криптогафічні методи
- шум Лапласа

5

Актуальність дослідження

- Дослідження вкрай актуальне у теперішній час. Це зумовлено помітною тенденцією зростання наслідків через відкриття конфіденційних даних третім особам, атак на бази даних, тощо.
- Завдяки програмно-технічній базі для запобігання порушень конфіденційності машинне навчання разом із аналітикою даних відбувається запровадження всі можливі сценарії захисту конфіденційності для забезпечення неушкодженості конфіденційності користувачів.
- Забезпечення конфіденційності даних слугують основним напрямом у ІТ-боротьбі із країною-агресором.

6

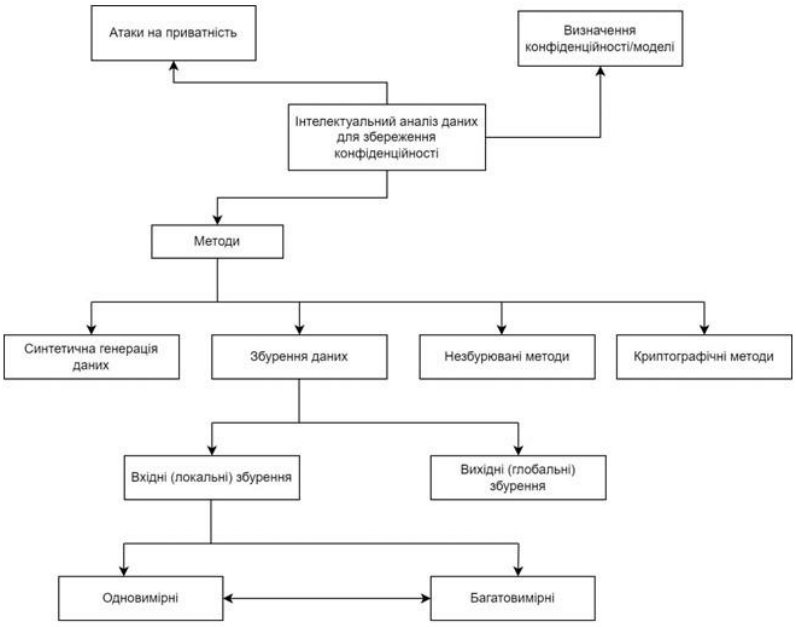
МЕТОД ЗАБЕЗПЕЧЕННЯ КОНФІДЕНЦІЙНОСТІ ЗБЕРЕЖЕННЯ ДАНИХ

Етапи методу

1. Аналіз існуючих методів
2. Здійснення збору даних, зокрема тих, що підлягають захисту
3. Застосування безпечного алгоритму збурення даних
4. Збурення потоку даних
5. Модуль віртуалізації функцій
6. Аналітика та подальша обробка даних чи їх збереження

7

Аналіз існуючих методів



Класифікація існуючих методів інтелектуального аналізу даних для збереження конфіденційності

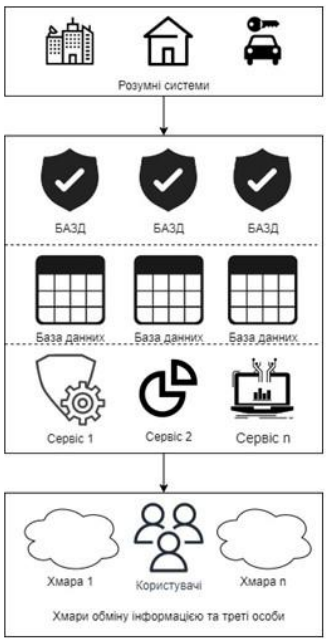
8

Здійснення збору даних, зокрема тих, що підлягають захисту

Застосування модуля забезпечення конфіденційності із подальшою підготовкою до обробки

9

Застосування безпечного алгоритму захисту даних



Розташування БАЗД серед інтелектуальної системи

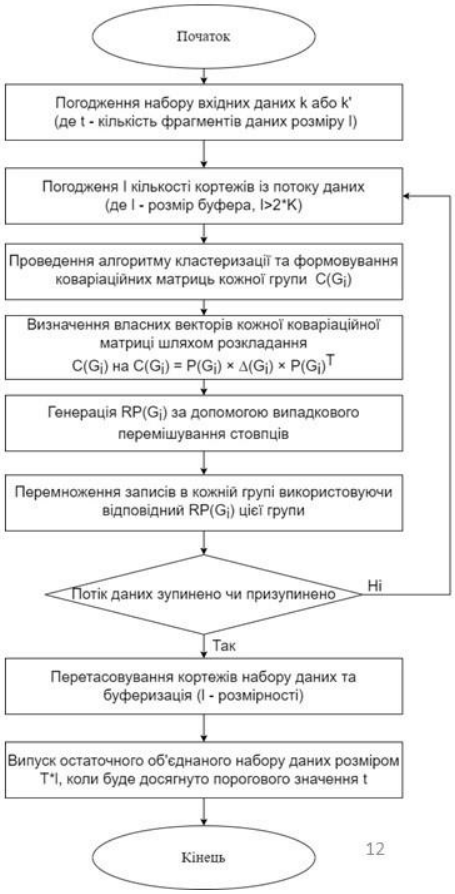
10

Територіальне розподілення організаційної структури системи охорони здоров'я

Приклад розподіленої організаційної структури 11

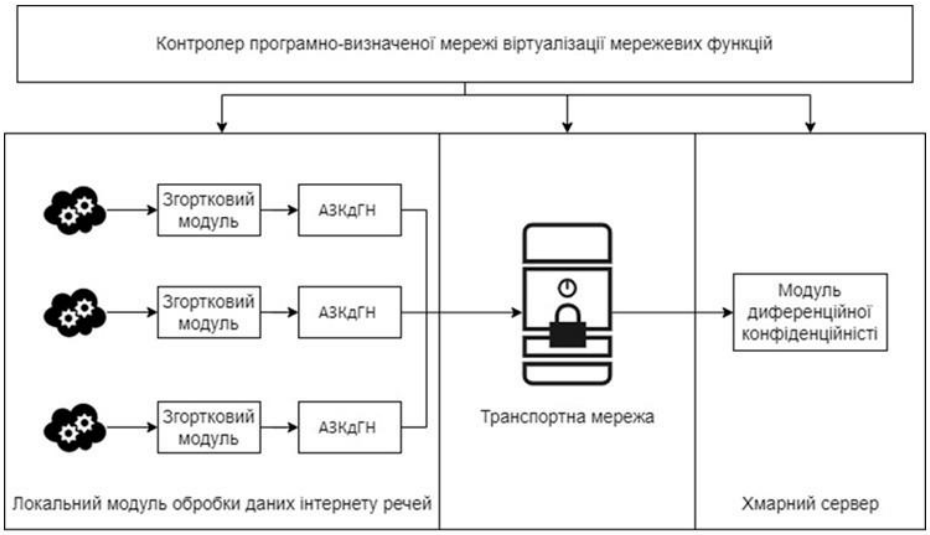
Збурення потоку даних

На блок-схемі МЗПД буферизує кількість кортежів даних під час кожного етапу збурення протягом t кількості ітерацій. Цей алгоритм також виробляє $t \times l$ кількість збурених кортежів даних наприкінці кожних t ітерацій. Цей процес збурення триває доти, поки потік даних не буде припинено.



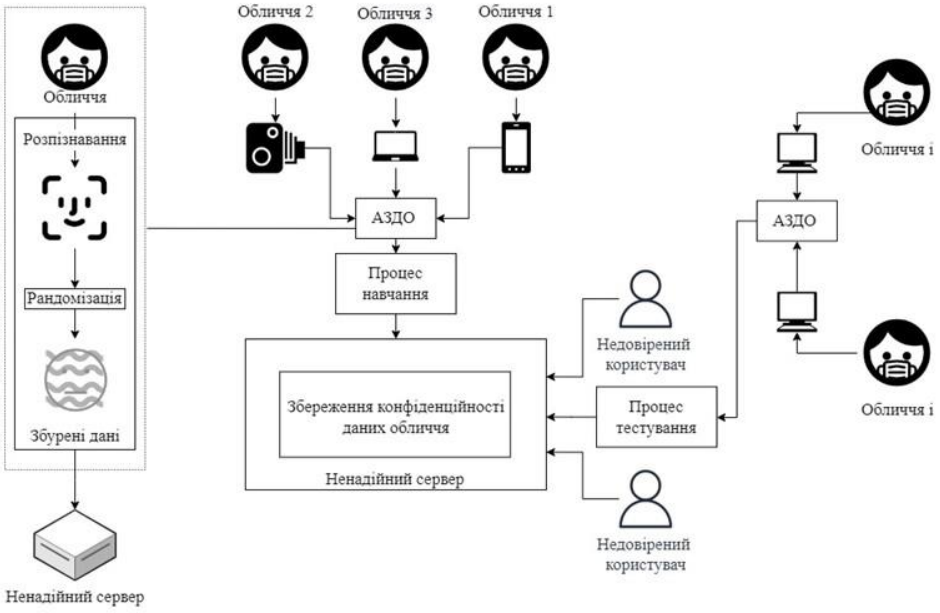
Блок-схема алгоритму збурення потоку даних 12

Мережевий модуль віртуалізації функцій



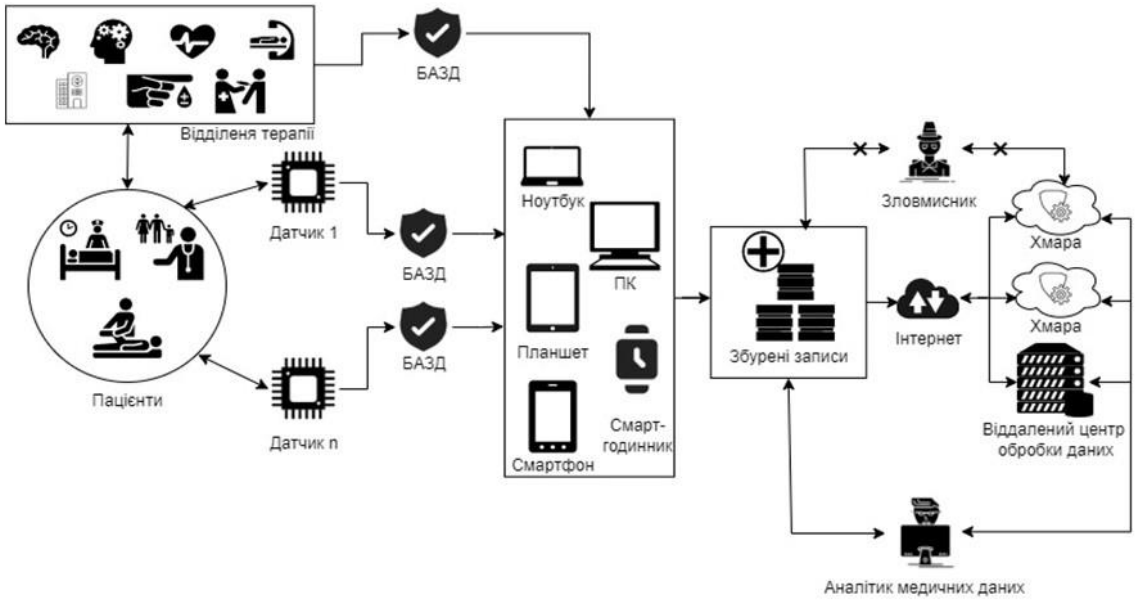
Інтеграція у програмно визначену мережу

Мережевий модуль алгоритму захисту даних обличчя



Алгоритму захисту даних обличчя

Використання безпечного алгоритму захисту даних у розумній кіберфізичній системі охорони здоров’я



Імплементація алгоритму захисту даних у систему охорони здоров’я 15

Публікації за матеріалами кваліфікаційної роботи

- За темою кваліфікаційної роботи опубліковано тези на тему «Key Trends of Smart Home Development» на Міжнародній науково-практичній конференції студентів, магістрів, аспірантів «Current Trends in Young Scientists’ Research V All Ukrainian Scientific and Practical Conference» та «Current Trends in Young Scientists’ Research VI All Ukrainian Scientific and Practical Conference» за 2018-2019 роки [1], а також опубліковано тези у Всеукраїнській науково-практичній конференції Актуальні Проблеми Комп’ютерних Наук (АПКН-2021) [2]. Було взято участь у Всеукраїнській науково-практичній конференції Актуальні Проблеми Комп’ютерних Наук.

18

Висновки

- У роботі виконано практичні та теоретичні дослідження, за результатами яких, розроблено алгоритми захисту конфіденційності для великих даних із використанням збурення даних та машинного навчання для покращення ефективності виявлення та знешкодження зловмисних дій, яким може піддаватися чутлива інформація у найрізноманітніших сферах.
- Виділено недоліки відомих методів та стратегій, а також систематизація їх у подальших розробках.

17

Висновки

- Розроблено модель базової архітектури системи для покращення конфіденційності зберігання великих даних із подальшим її вдосконаленням та нарощуванням методів, як і наявних, так і новостворених на основі наявних методів. Спроєктована система виконана таким чином, що її компоненти можуть обмінюватися результатами обробки даних.
- Удосконалено початковий алгоритм захисту конфіденційності до методу збурення розподілених даних, який може забезпечити конфіденційність під час машинного навчання. У даній системі АЗРД весь контроль генерації даних та глобальних параметрів належить центральному контролюючому органу, в той час як локальне збурення провозиться безпосередньо генерацією глобальних параметрів.

18

Висновки

- Розроблено метод збурення потоку даних, який продемонстрував безперебійну роботу із зростаючими потоками даних. Також його було удосконалено до безпечного алгоритму захисту даних, який слугує чудовим засобом для збереження конфіденційності у різноманітних сферах.
- Здійснено реалізацію рішень та імплементацію їх у систему охорони здоров'я для демонстрації роботи методу забезпечення конфіденційності для великомасштабної аналітики на основі машинного навчання.

19

Доповідь закінчено
Дякую за увагу!

20

UNICHECK

Ім'я користувача:
Кафедра KI

ID перевірки:
1011145466

Дата перевірки:
11.05.2022 17:14:38 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
11.05.2022 17:17:11 EEST

ID користувача:
100005591

Назва документа: Ференс_Метод забезпечення конфіденційності збереження даних для великомасштабної ...

Кількість сторінок: 84 Кількість слів: 18438 Кількість символів: 146666 Розмір файлу: 967.52 KB ID файлу: 101104221

1.52% Схожість

Найбільша схожість: 1.08% з джерелом з Бібліотеки (ID файлу: 1007657363)

0.41% Джерела з Інтернету	30	Сторінка 86
1.4% Джерела з Бібліотеки	95	Сторінка 86

0.04% Цитат

Цитати	1	Сторінка 87

Не знайдено жодних посилань

0% Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	2

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 1.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 10%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 103409 Название: Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання Добавлено в БД: 2022-05-11 Авторы: Ференс В.О. Руководители: Бармак О.В. Консультанты: Опоненты:	Символы	Лексемы	Символы	Лексемы
	137937	1033	822 (1%)	11 (1%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

РЕЦЕНЗІЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

Дипломник студент групи КІ2м-20-1 Ференс В.О.

Тема Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання

Спеціальність 123 – Комп'ютерна інженерія

Обсяг кваліфікаційної роботи:

Кількість листів креслень 0; кількість сторінок записки 91

1.Короткий зміст роботи та прийнятих рішень Робота присвячена актуальній темі в області забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання і складається з наступних розділів: вступ, аналіз предметної області та постановка задачі, архітектура розподіленої системи, метод головних компонент, ефективність запропонованих рішень, реалізація систем та експерименти, висновки, додатки.

2. Висновок про відповідність КР поставленому завданню Кваліфікаційна робота виконана у відповідності з виданим завданням із дотриманням всіх встановлених вимог.

3. Характеристика виконання кожного розділу роботи, ступінь використання останніх досягнень науки і техніки і передових методів роботи: В першому розділі здійснено детальний аналіз предметної області, розглянуто методи захисту конфіденційності, їх переваги та недоліки, особливості і вимоги при розробці, використання методів захисту конфіденційності в наукових задачах з предметної області з комп'ютерної інженерії, згідно проведеного аналізу було сформульовано актуальність роботи і визначено вимоги для створюваних методів. В другому розділі згідно досліджених джерел було розроблено алгоритм захисту конфіденційності та похідний від нього алгоритм збурення розподілених даних, їх апаратну реалізацію, детально розглянуто та описано всі компоненти алгоритму та їх взаємодія, а також крім удосконалення алгоритму було проведено перший непараметричний статичний тест. В третьому розділі було удосконалено метод збурення потоку даних та приведено приклад його використання у сфері охорони здоров'я у вигляді покращеного безпечного алгоритму збурення даних. У четвертому розділі подана методика алгоритму збереження конфіденційності для глибокого навчання, яку було вдосконалено до алгоритму захисту даних обличчя, якого було досягнуто згідно розподіленого механізму локальної диференційної конфіденційності, що включає в себе новий протокол для обмеження витоку даних, розроблений алгоритм та результати експериментальних досліджень з розробленими алгоритмами та проаналізовано результати.

4. Позитивні сторони роботи До позитивних сторін роботи слід віднести актуальність напрямку дослідження, отримані наукові і практичні результати з предметної області з комп'ютерної інженерії, розроблені методи захисту конфіденційності, які актуальні не лише для великих корпорацій, але і для невеликих компаній зважаючи на низьку ресурсузатратність та ефективність рішень, реалізацію запропонованих рішень, експериментальні дослідження з розробленою системою та .експериментальні дослідження

5. Негативні сторони роботи недостатньо деталізовано представлення тестування системи

6. Оцінка графічного оформлення та пояснювальної записки роботи Матеріали кваліфікаційної роботи є структурованими у чіткій та логічній формі та відображають послідовність виконання поставлених завдань.

7. Відгук про роботу в цілому Зміст представленої роботи в повній мірі розкриває обрану тему. Дослідження, проведені в матеріалах є достатньо аргументованими.

8. Інші зауваження

9. Оцінка дипломної роботи Робота заслуговує оцінки «відмінно», а її автор – присвоєння кваліфікації «магістра з комп'ютерної інженерії».

Рецензент (прізвище, ім'я, по-батькові, посада, місце роботи) Кльоц Юрій Павлович, кандидат технічних наук, доцент, завідувач кафедри кібербезпеки

"12" Травня 2022 р. _____
(підпис)

Завідувачу кафедри КІІС
д-р.техн.наук, проф. Говорущенко Т. О.

Ференса Володимира Олександровича
ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КІ2м-20-1

ЗАЯВА

З правилами чинного Положення «Про дотримання академічної доброчесності в Хмельницькому національному університеті» від 26.09.2020 (зі змінами від 26.11.2020), згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування заходів дисциплінарної та академічної відповідальності, ознайомлений (а). Про використання програмно-технічних засобів для перевірки кваліфікаційних робіт здобувачів вищої освіти на плагіатоповіщений (а) та надаю свою згоду на обробку та збереження університетом моєї роботи в інституційному репозитарії університету.

Також надаю університету право на передачу моєї роботи для обробки та збереження в базах даних програмно-технічних засобів (Unicheck та Anti-Plagiarism) та використання роботи для виявлення плагіату в інших роботах, які перевіряються програмно-технічними засобами та користувачами, що мають доступ до цих програмно-технічних засобів, виключно в обмежених цілях для виявлення плагіату в текстах робіт.

Робота для перевірки університетом надається в друкованому та електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

11.05.2022р
дата

підпис

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА СИСТЕМНОГО ПРОГРАМУВАННЯ

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: Метод забезпечення конфіденційності збереження даних для великомасштабної аналітики на основі машинного навчання

Автор: Ференс Володимир Олександрович

Спеціальність: 123 – Комп'ютерна інженерія та програмування

Освітня програма: освітньо-наукова

Науковий керівник: Бармак О.В., д.т.н, професор

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є незначними, законними і не є плагіатом, оскільки:

1) окремі виявлені збіги є загальновживаними фразами або виразами, про що свідчить посилання системи на збіг з 95 джерелами з бібліотек та 30 джерелами з мережі Інтернет;
2) всі зафіксовані системою ознаки модифікації тексту відносяться до комбінування латинських символів зі україномовними скороченнями індексів в формулах, що не є модифікацією тексту.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 1.52% і адресується до 26 першоджерел, а найбільша схожість становить 0.12% що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ О.В. Бармак

Гарант ОП _____ О. С. Савенко

Завідувач кафедри КІІС _____ Т. О. Говорущенко