DOI 10.31891/2307-5732-2023-329-6-274-280
УДК 004.056, 004.8

ОНАЙ МИКОЛА
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0000-0002-4938-8355
e-mail: onay@pzks.fpm.kpi.ua
СЕВЕРІН АНДРІЙ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0009-0009-1366-8054
e-mail: severinandrey97@gmail.com

МЕТОДИ ЗБЕРЕЖЕННЯ ПРИВАТНОСТІ В МАШИННОМУ НАВЧАННІ

В роботі наведено результати аналізу атак на системи машинного навчання, а також методів протидії для збереження приватності приватних наборів даних: анонімізація, федеративне навчання, гомоморфне шифрування, безпечні багатосторонні обчислення та диференційна приватність.

Ключові слова: машинне навчання із збереженням конфіденційності, федеративне навчання, гомоморфне шифрування, безпечні багатосторонні обчислення, диференційна приватність.

ONAI MYKOLA
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
SEVERIN ANDRII
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

METHODS FOR PRIVACY-PRESERVING IN MACHINE LEARNING

Data security and confidentiality are the biggest problems today, as much of the information is stored electronically and transmitted through a variety of devices (smartphones, computers) that have become widespread in public life. This is confirmed by the strengthening of legislation aimed at ensuring data protection. In particular, in 2016, the European Union adopted the General Data Protection Regulation (GDPR), and the California Consumer Privacy Act (CCPA) was adopted in California in 2018. The legal acts mentioned above encourage companies, developers, and researchers to develop information systems that are confidential by design (implementing the "Privacy by Design" approach). Data privacy is also in the central place during the construction of data analysis and artificial intelligence systems. First, such systems penetrate deeper into many areas of human activity every year, such as e-commerce (product recommendations, online assistants), human resource management (candidate resume analysis), social sphere (anti-spam, removal of offensive content), medicine, the gaming industry, and even politics. Also, an important element for building a reliable and accurate system is the availability of sufficient data for training. However, not all collected data can be used for training in decisions using artificial intelligence, as the data may contain a significant amount of private information: secret (eg, financial, military data), confidential (identifying data: passport data, registration number) taxpayer) or sensitive (medical data containing patient diagnoses). Under such conditions, finding a dataset to build an artificial intelligence system is an important task.

The paper presents the results of the analysis of attacks on machine learning systems, as well as countermeasures to preserve the privacy of private data sets: anonymization, federated learning, homomorphic encryption, secure multilateral computing, and differential privacy.

Keywords: privacy-preserving machine learning, federated learning, homomorphic encryption, secure multiparty computation, differential privacy.

Постановка проблеми

У багатьох сферах людського життя досить стрімко впроваджуються системи, що використовують методи машинного навчання. Для побудови більшості таких систем необхідно використовувати набори даних для навчання та тестування побудованих моделей, що ґрунтуються на реальних даних. Це потрібно, щоб забезпечити необхідну точність систем, що використовують методи машинного навчання. Однак, здебільшого у такі набори даних, що ґрунтуються на реальних даних, містять принаймні частину приватних даних. У зв'язку з цим актуальною задачею є захист приватних наборів даних у системах з використанням штучного інтелекту.

Мета роботи: аналіз атак на системи машинного навчання й аналіз методів, що дозволяють їм протидіяти й забезпечити конфіденційність приватних наборів даних.

Об'єкт дослідження: процеси шифрування, дешифрування та генерації інформації для побудови загальнодоступних систем аналізу даних і штучного інтелекту.

Предмет дослідження: методи, способи та моделі захисту наборів даних з використанням нейронних мереж.

Аналіз останніх джерел

Розроблення й аналіз методів захисту приватних наборів даних у системах з використанням штучного інтелекту є областю досліджень багатьох науковців й провідних IT компаній (Google, Apple, Microsoft). Це проявляється, як в значній кількості наукових статей на цю тему за останні кілька років; так і в проведенні різноманітних воркшопів та конференцій для розгляду останніх досліджень (наприклад, на постійній основі є щорічний воркшоп PPML, що присвячений лише питання збереження приватності даних у машинному навчанні). У роботах [1-2] наведені основні виклики в цій галузі, а також розглянуті основні методи що дозволяють їм протидіяти. З іншого боку, у роботах [3-9] основний акцент зроблено на докладному розгляді певного методу або його модифікації. У зв'язку з цим, актуальною задачею є комплексний порівняльний аналіз методів збереження приватності.

Виклад основного матеріалу

Перш ніж розглядати методи захисту приватних наборів даних, варто сформулювати основні загрози приватності даних. Такими загрозами є наступні типи атак на системи машинного навчання [1]:

- отруєння (poisoning) – атаки, що порушують цілісність навчальних наборів даних (шляхом впливу на навчальних набори даних; наприклад, їх зміни);
- атаки на логічний висновок (inference attacks) – атаки, які дозволяють зробити висновок про особисту інформацію окремого учасника на основі навчальних даних, проміжної, навченої або агрегованої моделі машинного навчання;
- атаки ухилення/дослідницькі (evasion/exploratory) – атаки, які змушують навчену модель видавати неправильні результати.

З точки зору захисту приватних наборів даних, основними загрозами є атаки на логічний висновок. Типові приклади таких атак [1]:

- атаки логічного висновку (inference attack), у яких зловмисник може зробити висновок, чи використовувався конкретний профіль пацієнта для навчання класифікатора, пов'язаного із захворюванням;
- атаки на інверсію моделі (model inversion attacks), які можуть використовувати доступ «чорної скриньки» (без відомостей про деталі реалізації) до моделей передбачення, щоб оцінити аспекти чиєїсь геномної інформації;
- глибокий витік із градієнтів (deep leakage from gradients), який отримує приватні навчальні дані зі спільних градієнтів у випадках застосування машинного навчання для завдань комп'ютерного зору та обробки природної мови..

Важливо знати про ці загрози та вживати заходів для їх запобігання, як-от використання методів збереження приватності, впровадження надійних заходів безпеки та регулярний моніторинг і оцінка продуктивності моделей машинного навчання, щоб переконатися, що вони не роблять упереджених або несправедливих прогнозів. Основними способами забезпечення захисту приватних наборів даних є [2]:

- генерація синтетичних наборів даних (synthetic data generation);
- обробка приватних наборів даних (анонімізація даних, диференційна приватність, гомоморфне шифрування);
- федеративне навчання.

Розглянемо їх докладніше.

Синтетичні дані – штучні дані, які згенеровані за певним алгоритмом на відміну від оригінальних даних, які ґрунтуються на реальній інформації. Ідея таких даних полягає в тому, що вони складаються з нових точок даних, що не є простою модифікацією існуючого набору даних. Однак, варто зауважити, що не всі синтетичні дані є анонімними. Синтетичні дані використовують для навчання моделі з наміром перенести результати навчання на реальні дані. За умови успішної побудови генератора синтетичних даних, можна отримати такі переваги:

- швидко і дешево можна згенерувати стільки даних, скільки необхідно;
- згенеровані дані можуть мати ідеально точні мітки, включаючи мітки, які може бути складно отримати на реальних даних;
- синтетичні дані можна використати як заміну певних реальних сегментів даних, які містять, приватну інформацію.

Генератор синтетичних даних можна побудувати використовуючи різні підходи (зокрема, доменну рандомізацію чи комбінування реальних даних), шляхом побудови генеративних моделей серед яких [3]: прихована модель Маркова, Баєсівська мережа, агентне моделювання, варіаційний автокодувальник, генеративні конкуруючі нейронні мережі [10].

Іншим способом захисту приватних наборів даних є попередня обробка набору даних, основними методами якої є: анонімізація даних, диференційна приватність (differential privacy) та гомоморфне шифрування. Розглянемо ці методи докладніше.

Анонімізація даних – це процес захисту приватної або конфіденційної інформації шляхом видалення або зміни ідентифікаторів, які з'єднують особу із збереженими даними. Іноді анонімізацію даних також називають «обфускація», «маскування» або «деідентифікація» даних. Анонімізація даних здійснюється багатьма галузями, які працюють з конфіденційною інформацією, зокрема, такими як охорона здоров'я, фінансова та цифрова індустрії. Це сприяє цілісності обміну даними. Наприклад, коли лікарні потрібно передати дані до медичної дослідницької лабораторії, вона захищає конфіденційні дані й зберігає анонімність своїх пацієнтів. Це можна зробити шляхом видалення імен, номерів соціального страхування, дати народження та адрес своїх пацієнтів з загальнодоступного списку, залишаючи лише важливі компоненти, які необхідні для медичних досліджень, такі як вік, хвороби, зріст, вага, стать, раса тощо.

Анонімність даних може забезпечуватись різними способами. Серед них базовими є [11-12]: придушення атрибутів і/або записів, перестановка даних, підміна даних (частковим випадком якої є псевдонімізація), маскування символів, дисперсія чисел і дат, узагальнення.

Придушення атрибутів і/або записів полягає у їх видаленні з набору даних, якщо вони містять ідентифікуючу інформацію, але є несуттєвими для подальшого аналізу або не можуть бути анонімізовані іншими способами. Перестановка даних полягає у перестановці окремих значень атрибутів, в результаті

чого, вони залишаться присутніми в модифікованому наборі даних, але не відповідатимуть оригінальним засобам. Підміна даних – це процес заміни даних зі стовпця-атрибуту випадковими значеннями зі списку фальшивих, але схожих на вигляд даних (напрриклад, номери кредитних карток можуть бути замінені випадковим рядком з 16 чисел). Частковим випадком підміни даних є псевдонімізація – заміна ідентифікаційних даних кодованими даними (складеними за певним принципом). Маскування символів – це заміна символів у значенні атрибута даних, наприклад, використовуючи сталий символ, такий як «*» або «x». Маскування символів зазвичай виконують частково, тобто застосовується лише до деяких символів в атрибуті. Дисперсія дат і чисел полягає у зміні значень атрибута на певний значення з інтервалу. Дати і числа часто використовуються в якості параметрів пошукового запиту до набору даних, а також вони є важливими елементами медичної та фінансової статистики. Оскільки додавання/віднімання певного заданого однакового значення до атрибуту легко можна декодувати, то необхідно додавати/віднімати випадкове значення з певного довірчого інтервалу, що дозволить зберегти розподіл даних за цим атрибутом. Узагальнення – заміна специфічних даних більш загальною, але ще корисною інформацією. За рахунок цього відбувається зниження точності даних. Цей спосіб зазвичай полягає у використанні діапазонів (перетворення віку людини у віковий діапазон) або менш точної інформації (перетворення точного місця розташування у менш точне). Цей прийом іноді також називають перекодуванням даних. Його доцільно використовувати для значень, що можуть бути узагальнені й ще будуть корисні в перетвореному вигляді.

Розглянемо базові способи анонімізації на прикладі структурованих даних. Нехай задано набір даних, що представлений у таблиці 1.

Таблиця 1

Оригінальний набір вхідних даних

Користувач	Вік	E-mail	Поштовий код	Адреса	Кількість покупок
Олег	21	oleh@gmai.com	02222	вулиця Вишнева 1	3
Руслан	34	ruslan@i.ua	02217	вулиця Крайня 12б	8
Валентин	18	valentyn@outlook.com	01103	вулиця Вишнева 10	2
Степан	27	stepan@gmai.com	03061	вулиця Франка 7	5

Таким чином, застосувавши розглянуті вище способи придушення атрибутів (атрибут Адреса), перестановки даних (атрибут Користувач, маскування символів (атрибути Поштовий код та E-mail) і дисперсії чисел (атрибут Вік), можна отримати анонімізований набір, що представлений у таблиці 2.

Таблиця 2

Набір даних, до якого застосовано способи анонімізації

Користувач	Вік	E-mail	Поштовий код	Кількість покупок
Степан	23	****@gmai.com	02xxx	3
Валентин	32	******@i.ua	02xxx	8
Олег	15	********@outlook.com	01xxx	2
Руслан	26	******@gmai.com	03xxx	5

Якщо ж застосувати способи розглянуті вище способи підміни даних (атрибут Користувач), узагальнення даних (атрибути Вік та Адреса), псевдонімізацію (атрибут Користувач) та маскування символів (атрибут E-mail), можна отримати анонімізований набір, що представлений у таблиці 3.

На основі розглянутих прикладів, можна зробити висновок, що головна перевага анонімізації полягає в тому, що вона дозволяє приховати чутливі аспекти даних. Однак, навіть за умови очищення набору даних від ідентифікуючої інформації, анонімізовані дані можуть бути розшифровані і розкриті за допомогою методів деанонімізації (також називають повторна ідентифікація). Оскільки дані зазвичай зберігаються в декількох джерел (частина з яких доступна для широкого загалу, наприклад, через державні реєстри даних), методи деанонімізації можуть перехресно посилатися на джерела та виявляти особисту інформацію в анонімізованому наборі даних. У зв'язку з цим, критики вважають, що анонімізація надає помилкове відчуття безпеки.

Таблиця 3

Оригінальний набір вхідних даних

Користувач	Вік	E-mail	Поштовий код	Адреса	Кількість покупок
User3	20-30	****@gmai.com	02222	вулиця Вишнева	3
User1	> 30	******@i.ua	02217	вулиця Крайня	8
User4	< 20	********@outlook.com	01103	вулиця Вишнева	2
User2	20-30	******@gmai.com	03061	вулиця Франка	5

Диференційна приватність – метод захисту даних, який захищає конфіденційність користувача шляхом додавання випадкового шуму до даних. Мета цього методу – забезпечення жорстких статистичних гарантій того, що зловмисник не зможе зробити висновок про приватні дані, на основі результатів даних, що

отримані за допомогою рандомізованого алгоритму. Іншими словами, абсолютна різниця ймовірності p того, що результатом запиту до оригінального набору даних D_1 є значення s, і ймовірності p_2 того, що результатом запиту до модифікованого набору даних D_2 є те саме значення, повинна бути менша певного значення (тобто така різниця ймовірностей має бути в межа похибки). Математично метод диференційної приватності можна сформулювати наступним чином [4]:

Нехай $\varepsilon < 0$, A – рандомізований алгоритм, який приймає на вхід приватний набір даних D_1, а $E(A)$ – область значень цього алгоритму. Алгоритм A є ε-диференційно приватним, якщо для всіх записів наборів даних D_1 і D_2, які відрізняються єдиним записом, для всіх підмножин $S \subseteq E(A)$ виконується наступна нерівність $p\left\{A(D_1) \in S\right\} \le e^{\varepsilon} \cdot p\left\{A(D_2) \in S\right\}$, де p – ймовірність отримана з випадковості рандомізованого алгоритму. Зазначеного результату можна досягти шляхом додавання випадкового шуму, зокрема шуму від розподілу Лапласа або Гауса. Варто зауважити, що чим більше шуму додається до вхідних даних, тим менше цінності вони представляють, наслідком чого є. менша точність їх аналізу. Цю тезу чудово ілюструє приклад, зображений на рис. 1. Застосувавши до зображення значну кількість шуму, можна досягти високого рівня захисту конфіденційності даних, при цьому екземпляр даних майже повністю втрачає свої властивості, що унеможливить здійснення подальшого аналізу цих даних. Зважаючи на це, під час застосування методу диференційної приватності необхідно знайти баланс між захистом приватності й збереженням характерних особливостей даних.

Рис. 1. Залежність цінності даних, від кількості доданого шуму

Гомоморфне шифрування – це форма шифрування, яка дозволяє виконувати обчислення над зашифрованим текстом, розшифрований результат яких буде таким самим, як і результат операцій над відкритим текстом [5]. Основною концепцією цього методу захисту приватних даних є гомоморфізм. Математично це можна сформулювати наступним чином: функція f: між двома групами G і H є гомоморфною, якщо: $f(xy) = f(x)\,f(y)$ для будь-яких x, y, що належать G. Найпростішими прикладами таких функцій є c $(x + y) = cx + cy$, $|xy|=|x||y|$, $(xy)^c=x^cy^c$. Можна сформулювати гомоморфне шифрування наступним чином:

Нехай $a \cdot b \equiv 1$, тоді $E(x) = x^b \bmod n$, $D(x) = y^a \bmod n$, $E(x_1) \cdot E(x_2) = x_1^b \cdot x_2^b \bmod n = (x_1 \cdot x_2)^b \bmod n = E(x_1 \cdot x_2)$, де x_1, x_2 – звичайний текст.

Комбінуючи дві операції (наприклад, додавання і множення), можна побудувати будь-яку довільну функцію. Сучасні схеми шифрування, які підтримують гомоморфні обчислення в необмеженій кількості, називаються повністю гомоморфними схемами шифрування. Перша згадка про гомоморфне шифрування була в 1978 році, але повністю гомоморфна криптосистема була побудована Крейгом Гентрі лише в 2009 році. З тих пір кількість досліджень в цій області значно зросла. У 2015 році Microsoft розробила бібліотеку SE + (Simple Encrypted Arithmetic Library) з відкритим кодом, написану на C++, яка реалізує різні форми гомоморфного шифрування, включаючи BFV, CKKS [2], що використовується в системах штучного інтелекту. Однак обчислювальна потужність і пам'ять все ще залишаються серйозними технічними перешкодами для цього, оскільки обробка зашифрованих даних потребує набагато більше ресурсів – фактично в мільйон разів більше, ніж потрібно для обробки незашифрованих даних або відкритого тексту. Завдяки прогресу в технології за останні 10 років гомоморфне шифрування тепер не тільки можливо, але й все частіше використовується на практиці. Тому певні операції можна виконувати безпосередньо над зашифрованими текстами, щоб при дешифруванні отримати ту саму відповідь, що й при виконанні операцій над вихідними повідомленнями. Це головна перевага цього методу. Це впливає на зашифровані дані на AI-рішеннях, а також оригінальних. Основним недоліком є кількість зашифрованої інформації порівняно з вихідними даними. Таким чином, на 1 МБ інформації обсяг зашифрованих даних може досягати 10 ГБ.

Одним із підходів до інформаційної безпеки є протокол конфіденційних багатосторонніх обчислень (MPC), який забезпечує конфіденційність розподілених даних. Цей підхід є гомоморфним [9]. Нехай набір користувачів (учасників) $P_1,\ldots, P_n$, кожен з яких має певні особисті дані x_i і прагне обчислити загальну функцію $y = f(x_1,\ldots,x_n)$, яка є результатом цих даних. У той же час зв'язок між суб'єктами має бути безпечним, а дані – коректними. Залежно від вхідних даних і кількості учасників конфіденційний протокол може бути побудований різними способами. Розглянемо приклад такого протоколу. Припустимо, потрібно знайти суму $\sum_{i=1}^{n} x_i$ значень цих учасників, кожне зі значень яких не перевищує модуля m ($x_i < m$). Потім

вводиться випадкове число $r < m$, яке відоме тільки першому учаснику. Кожен учасник по черзі обчислює значення $\tilde{x}_i = x + \tilde{x}_{i-1}$ де $\tilde{x}_0 = r$, і передає його наступному. Коли n-й учасник завершує обчислення, загальну суму $y = \sum_{i=1}^{n} x_i = \tilde{x}_n - r$ можна знайти без розголошення особистої інформації зашифрованих суб'єктів. Цей підхід має низку переваг: він готовий до комерційного використання, усуває компроміс між безпекою даних і корисністю (які є характерними для анонімізації та диференціальної конфіденційності), а результат є точним і сторони не бачать приватних даних інших учасників. Однак, він є обчислювально (потрібно генерувати випадкові числа, виконувати операції над даними кожного учасника) і комунікаційно (передача інформації між учасниками) витратним.

Стандартні підходи до створення систем штучного інтелекту вимагають централізації навчальних даних на одній машині. Однак у 2016 році дослідники з Google запропонували децентралізований архітектурний підхід – федеративне навчання (Federated Learning) [6, 13]. Ідея федеративного навчання полягає в навчанні алгоритму штучного інтелекту на багатьох кінцевих пристроях або серверах, які містять локальні набори даних, які залишаються на пристрої під час навчання (ними не обмінюються між пристроями). Цей підхід відрізняється від традиційних централізованих методів машинного навчання, коли всі зразки даних завантажуються на один сервер, а також від більш класичних децентралізованих підходів, які припускають, що локальні зразки даних рівномірно розподіляються між пристроями. Розглянемо приклад такого архітектурного підходу до навчання, який зображено на рис. 2.

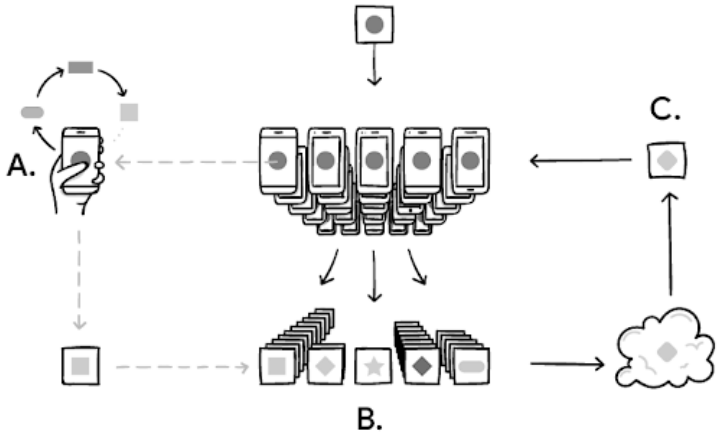

Рис. 2. Архітектурний підхід федеративного навчання (Federated Learning) [13]

Припустимо, необхідно вирішити задачу класифікації різних геометричних фігур (коло, квадрат, ромб, овал тощо). За умови застосування федеративного навчання, поточна версія моделі зберігається на сервері, наприклад, у хмарі. Пристрій користувача завантажує початкову версію моделі на свій пристрій (наприклад, телефон або планшет) і вдосконалює її, вивчаючи локальний набір даних (блок А на малюнку 1.3), а потім узагальнює зміни у вагових коефіцієнтах моделі та надсилає їх у хмару як невелике оновлення. Як тільки це оновлення, надіслане в зашифрованому вигляді, надходить на сервер, воно негайно усереднюється з оновленнями від інших користувачів (блок В на рис. 2), і вага загальної моделі покращується (блок С на рис. 1.3). Потім цю процедуру повторюють. Однак усі дані, які використовувалися для навчання, залишаються на пристрої користувача, а оновлення не зберігаються в хмарі.

Алгоритм усереднення оновлень на основі задачі обчислення скінченних сум, який охоплює не тільки лінійну та логістичну регресію, але й більш складні алгоритми, включаючи нейронні мережі: $\min_{w \in R^d} f(w)$, де $f(w) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$. Для проблеми машинного навчання $f_i(w) = l\left(w; x_i; y_i\right)$ це втрата точності прогнозу для прикладу $\{x_i; y_i\}$, зробленого моделлю з параметрами w. Припустимо, що існує K користувачів, які беруть участь у моделюванні за допомогою локальних наборів даних. Позначимо $\{P_k\}_{k=1}^{K}$ розділ індексу точок даних $\{1, \ldots, n\}$, оскільки P_k є набором точок даних, що зберігається в користувачеві k і визначає $n_k = |P_k|$. Потім сформулюємо визначення функції втрат для загальної моделі як лінійної комбінації функцій втрат [6, 13]: $f(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \overset{def}{=} \sum_{k=1}^{K} \frac{n_k}{n} \cdot \frac{1}{n_k} \sum_{i \in P_k} f_i(w)$.

Федеративне навчання дає змогу створювати кращі моделі з меншою затримкою та меншим споживанням енергії, забезпечуючи приватність. Також цей підхід, крім надання оновлення спільної моделі, дозволяє використовувати вдосконалену модель на кінцевому пристрої з мінімальною затримкою. Ідеальними завданнями для федеративного освіти є завдання, які характеризуються наступним:

- навчання на реальних даних у режимі реального часу з мобільних пристроїв має явну перевагу

перед навчанням на надійних даних, які зазвичай доступні в центрі обробки даних;
- дані чутливі до конфіденційності або мають великі розміри (порівняно з розміром моделі), тому їх бажано не зберігати в центрі обробки даних виключно з метою навчання моделей;
- для контрольованих завдань, мітки даних для яких можна отримати під час взаємодії з користувачем.

Федеративне навчання не може вирішити всі проблеми машинного навчання (наприклад, навчитися розпізнавати різні породи собак шляхом навчання на ретельно маркованих прикладах), а для багатьох інших моделей необхідні дані про навчання вже й так зберігаються в хмарних сховищах (наприклад, фільтр спаму для Gmail). Зараз федеративне навчання застосовується для інтелектуальної клавіатури Google Gboard на мобільних телефонах з операційною системою Android [13]: коли клавіатура пропонує запит, телефон користувача локально зберігає інформацію про поточний контекст та про те, чи обрав користувач пропозицію. Потім процеси федеративного навчання, які використовують історію на пристрої, пропонують покращити наступну ітерацію моделі пропозицій користувацьких запитів. Іншим прикладом застосування цього підходу є сфера охорони здоров'я [14]: французький стартап Owkin, де були розроблені моделі навчання біомедичних засобів, заснованих на алгоритмах збору неоднорідних даних (фармацевтичних компаній і медичних установ) шляхом застосування федеративного навчання з використанням технологій високої відстежуваності (технологія distributed ledgers, однією з форм якої є blockchain).

Проведемо порівняльний аналіз методів захисту приватних наборів даних, використовуючи наступні п'ять критерії: складність, практичність, потреба у великій кількості даних для використання методу, надійність (чи значний рівень приховування даних), точність системи штучного інтелекту (на модифікованих даних); Результати порівняння наведені в таблиці 4.

Таблиця 4

Порівняння методів захисту приватних наборів даних

Метод	Складність	Практичність	Потреба у великій кількості даних	Надійність	Висока точність системи
Генерування синтетичних даних	+	+	+	+	+/–
Анонімізація даних	–	+	–	–	+
Диференційна приватність	+/–	+	+	+/–	+/–
Гомоморфне шифрування	+	–	–	+	+
Федеративне навчання	–	+/–	+	+	+

З проведеного аналізу можна зробити висновок, що генерація синтетичних наборів даних є практичним і надійним методом, але досить складним і вимагає великої кількості вхідних даних (умов, прикладів) для формування більш точної системи штучного інтелекту. Анонімізація даних досить проста, практична і не потребує великих масивів даних, але цей метод недостатньо надійний. Диференціальна конфіденційність – це практичний метод, який вимагає великих наборів даних, і залежно від кількості використовуваного шуму може бути як дуже надійним, так і неточним, і ненадійним, але дуже точним. Гомоморфне шифрування є надійним і може бути використане для побудови високоточних систем, але цей метод є складним, зокрема обчислювальним витраним, і може застосовуватися до обмеженого класу завдань. Федеративне навчання є надійним і точним методом без розповсюдження локальних даних навчання, але його можна використовувати, якщо є принаймні кілька незалежних користувачів, які мають достатньо даних навчання.

Висновки

Забезпечення захисту приватних наборів даних для систем машинного навчання є актуальною задачею, зважаючи на швидкість впровадження таких систем у більшість сфер життя. Розглянуто основні типи атак на системи машинного навчання, а також методи протидії атакам, що загрожують витоку приватних даних. Проаналізовано переваги та недоліки методів захисту приватних наборів даних. Зокрема, були розглянуті методи генерування синтетичних даних, анонімізацію даних, диференційну приватність, гомоморфне шифрування та федеративне навчання. Розглянуті методи дозволяють вирішити деякі проблеми приватності даних, але кожен з них має як переваги, так і недоліки, які треба враховувати при вирішенні задачі.

Отже, спираючись на проведений аналіз можна зробити висновок, що розроблення альтернативних і модифікація існуючих методів захисту інформації, які дозволять мінімізувати розглянуті недоліки, є актуальними напрямками подальших досліджень.

References

1. Xu R. Privacy-preserving machine learning: Methods, challenges and directions / R. Xu, N. Baracaldo, J. Joshi. // arXiv preprint arXiv:2108.04417. — 2021 — DOI: 10.48550/arXiv.2108.04417.

2. Lauter K. Faculty Summit 2017: Private AI [Electronic resource] / Kristin Lauter // Microsoft Research. — 2017. — Access mode: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/07/Private_AI_Kristin_Lauter.pdf.

3. Nikolenko S. I. Synthetic Data for Deep Learning [Electronic resource] / Sergey I. Nikolenko. — 2019. — Access mode: https://arxiv.org/pdf/1909.11512.pdf.

4. Dwork C. The Algorithmic Foundations of Differential Privacy [Text] / C. Dwork, A. Roth. // Foundations and Trends® in Theoretical Computer Science. — 2014. — Vol. 9, №3-4. — С. 211–407. — DOI 10.1561/0400000042.

5. Minelli M. Fully homomorphic encryption for machine learning [Text] / Michele Minelli., 2018. — 157 p.

6. Communication-efficient learning of deep networks from decentralized data. [Text] / [H. Brendan McMahan, E. Moore, D. Ramage and others]. — 2016.

7. Konečný J. Federated Optimization: Distributed Optimization Beyond the Datacenter [Electronic resource] / J. Konečný, B. McMahan, D. Ramage. — 2015. — Access mode: https://arxiv.org/pdf/1511.03575.pdf.

8. Abadi M. Learning to Protect Communications with Adversarial Neural Cryptography [Electronic resource] / M. Abadi, D. G. Andersen. — 2016. — Access mode: https://arxiv.org/abs/1610.06918.

9. Lindell Y. Secure Multiparty Computation (MPC) [Electronic resource] / Yehuda Lindell — Access mode: https://eprint.iacr.org/2020/300.pdf.

10. Generative Adversarial Nets [Text] / [I. J. Goodfellow, J. Pouget-Abadie, M. Mirza and others]. // Advances in neural information processing systems. — 2014. — Vol. 2 — P. 2672–2680.

11. Data Anonymization Techniques [Electronic resource]. — 2019. — Access mode: https://www.solarwindsmsp.com/blog/data-anonymization-overview.

12. Guide to basic data anonymisation techniques [Electronic resource] // Personal Data Protection Commission Singapore (PDPC). — 2018. — Access mode: https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf.

13. Brendan McMahan H. Federated Learning: Collaborative Machine Learning without Centralized Training Data [Electronic resource] / H. Brendan McMahan, D. Ramage. — 2017. — Access mode: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

14. Kuchler H. Pharma groups combine to promote drug discovery with AI [Electronic resource] / Hannah Kuchler // Financial Times. — 2019. — Access mode: https://www.ft.com/content/ef7be832-86d0-11e9-a028-86cea8523dc2.