МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ
КАФЕДРА УПРАВЛІННЯ КІБЕРБЕЗПЕКОЮ ТА ЗАХИСТОМ ІНФОРМАЦІЇ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: “МЕТОДИ ТА ЗАСОБИ ЗАХИСТУ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ В СОЦІАЛЬНИХ МЕРЕЖАХ З ВИКОРИСТАННЯМ ЗАСОБІВ ГЕНЕРАТИВНОГО ІНТЕЛЕКТУ”

на здобуття освітнього ступеня магістра
зі спеціальності 125 Кібербезпека та захист інформації
освітньо-професійної програми Управління інформаційною та кібернетичною безпекою

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Микита КАРПЕЧЕНКОВ
(підпис) *Ім’я, ПРІЗВИЩЕ здобувача*

Виконав: Здобувач вищої освіти гр. УБДМ-61
Микита КАРПЕЧЕНКОВ
Керівник:
д.е.н, доцент Тетяна КАПЕЛЮШНА
Рецензент:
д.т.н, професор Галина ГАЙДУР

Київ 2026

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
Навчально-науковий інститут кібербезпеки та захисту інформації

Кафедра Управління кібербезпекою та захистом інформації
Ступінь вищої освіти магістр
Спеціальність 125 Кібербезпека та захист інформації
Освітньо-професійна програма Управління інформаційною та кібернетичною безпекою

ЗАТВЕРДЖУЮ
Завідувач кафедрою УКБЗІ
_____ Світлана ЛЕГОМІНОВА
"____" _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Студенту Карпеченкову Микиті Павловичу
(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: "Методи та засоби захисту персональної інформації в соціальних мережах з використанням засобів генеративного інтелекту"

керівник кваліфікаційної роботи Тетяна КАПЕЛЮШНА, д.е.н., доцент
(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій від "30" жовтня 2025 р. № 467.

2. Строк подання кваліфікаційної роботи "20" грудня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: нормативно-правове забезпечення з питань кібербезпеки та захисту інформації (ЗУ "Про захист персональних даних", ЗУ "Про інформацію", регламент GDPR), інформація щодо політик безпеки досліджуваних соціальних мереж, статті українських та зарубіжних вчених. Перелік питань, які потрібно розробити:

1. Розглянути теоретико - методичні засади захисту персональної інформації в соціальних мережах.

2. Проаналізувати методи захисту персональної інформації в соцмережах та виявити прогалини у захисті контенту користувачів соцмереж

3. Сформувати архітектуру моделі, формалізувати та сформувати навчальний набір даних для захисту інформації користувача соцмереж з використанням засобів генеративного інтелекту (на основі DP-GAN-HD)

4. Оцінити запропоновану модель та проаналізувати стійкість до витоку даних з подальшою розробкою рекомендацій щодо впровадження в інфраструктуру соцмереж.

5. Перелік ілюстративного матеріалу; *презентація*

6. Дата видачі завдання "02" жовтня 2025 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назви етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Визначення об’єкту, предмету, мети та завдань дослідження.	10.10.2025	
2.	Збір та аналіз літератури.	23.10.2025	
3.	Аналіз основних характеристик інформаційного середовища користувачів та середовища загроз.	27.10.2025	
4.	Дослідження нормативно - правової бази у сфері захисту персональних даних та аналіз із міжнародними стандартами	10.11.2025	
5.	Дослідження можливих сфер впливу, генерації потенційних загроз внаслідок імплементації генеративного інтелекту	15.11.2025	
6.	Формулювання висновків за результатами дослідження.	22.11.2025	
7.	Оформлення роботи.	04.12.2025	
8.	Оформлення презентації.	14.12.2025	
9.	Отримання рецензії на роботу.	18.12.2025	
10.	Захист в ЕК.	19.01.2026	

Здобувач вищої освіти _____ Микита КАРПЕЧЕНКОВ
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Тетяна КАПЕЛЮШНА
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ КІБЕРБЕЗПЕКИ ТА ЗАХИСТУ ІНФОРМАЦІЇ

ПОДАННЯ
ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня магістра

Направляється здобувач Карпеченков М.П. до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю 125 Кібербезпека та захист інформації
(код, найменування спеціальності)
Освітньо-професійної програми Управління інформаційною та кібернетичною безпекою
(назва)
на тему: “Методи та засоби захисту персональної інформації в соціальних мережах з використанням засобів генеративного інтелекту ”
Кваліфікаційна робота і рецензія додаються.

Директор ННІКБЗІ _____ Євгенія ІВАНЧЕНКО
(підпис) *(Ім’я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувачем **КАРПЕЧЕНКОВИМ Микитою** у процесі виконання роботи досліджено особливості використання соціальних мереж як джерела даних та потенційного вектора загроз, проаналізовано ризики витоку інформації, потенційних атак щодо порушення безпеки інформації, проведено комплексний аналіз теоретичних засад функціонування технологій генеративного інтелекту та принципів застосування в цифровому середовищі. Систематизовано існуючі моделі генеративного інтелекту, визначено їх функціональні можливості, обмеження та безпекові аспекти використання в інформаційних системах підприємств.

КАРПЕЧЕНКОВ Микита продемонстрував високу теоретичну і практичну підготовку, володіння науково-дослідницькими методами, вміння самостійно здійснювати пошук шляхів вирішення проблеми дослідження.

Означене вище дозволяє оцінити кваліфікаційну роботу здобувача **КАРПЕЧЕНКОВА Микити** на оцінку “відмінно” та присвоїти йому кваліфікацію “Магістр з кібербезпеки та захисту інформації за освітньо-професійною програмою “Управління інформаційною та кібернетичною безпекою””.

Керівник кваліфікаційної роботи_____Тетяна КАПЕЛЮШНА
“____“_____2025 року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувач Карпеченков М.П. допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою
Управління кібербезпекою та захистом
інформації _____Світлана ЛЕГОМІНОВА

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну магістерську роботу

здобувача вищої освіти - Карпеченкова Микити Павловича
на тему: “Методи та засоби захисту персональної інформації в соціальних мережах з використанням засобів генеративного інтелекту”

Актуальність. Активне використання соціальних мереж та цифрових сервісів обумовлюють нові ризики для безпеки персональних даних користувачів. З платформ збираються великі обсяги даних, які містять чутливу інформацію, що робить цифрові профілі користувачів вразливими до несанкціонованого доступу, профілювання та маніпуляцій. За даних умов традиційні механізми захисту не повною мірою забезпечують належний рівень конфіденційності даних та стійкості до нових типів загроз. Тому актуалізується питання поєднання технічних засобів захисту з використанням методів захисту інформації користувачів соціальних мереж на основі генеративного інтелекту, як дієвого інструменту у системі кіберзахисту.

Позитивні сторони

1. Здобувачем ґрунтовно проаналізовано теоретико-методичні засади загроз соціальної інженерії, нормативно - правового забезпечення питань захисту інформації, позитивних та негативних сторін використання генеративного інтелекту в кібербезпеці у контексті захисту інформації в соціальних мережах.
2. Проведено порівняльний аналіз сучасних підходів до захисту даних у соціальних мережах та обґрунтовано наявність прогалин у захисті користувацького контенту.
3. Позитивно значиться практична частина роботи - запропонована архітектура моделі захисту персональної інформації на основі DP-GAN-HD, що поєднує механізми генеративного інтелекту та диференційної приватності. Окрім того, проведено оцінювання стійкості моделі до витоку даних.
4. Кваліфікаційна робота оформлена відповідно до вимог. Виклад матеріалу логічний та структурований згідно завдань, наведено логічні висновки.

Недоліки

Доцільно було б доповнити дослідження порівнянням запропонованої моделі з альтернативними підходами (federated learning, secure multi-party computation або ін.). Однак, вищезгадані зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи.

Висновок: Кваліфікаційна робота виконана на високому науково-методичному рівні, заслуговує позитивної оцінки, а здобувач - Карпеченков Микита Павлович заслуговує присвоєння кваліфікації “Магістр кібербезпеки за освітньо-професійною програмою “Управління інформаційною та кібернетичною безпекою”.

Рецензент: завідувач кафедри
кібербезпеки та захисту
інформації,
д.т.н, професор _____ Галина ГАЙДУР
підпис

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра; 86 стор., 19 рис., 15 табл., 62 джерела.

Метою роботи є теоретичне обґрунтування та практична розробка підходу до захисту персональної інформації користувачів соціальних мереж із використанням засобів генеративного інтелекту (на основі моделі PD-GAN-HD).

Об'єктом дослідження є процес забезпечення захисту персональної інформації користувачів у соціальних мережах.

Предметом дослідження є прогарамні засоби захисту персональної інформації користувачів у соціальних мережах, що базуються на використанні алгоритмів штучного інтелекту (генеративно-змагальної мережі даних - GAN).

Методи дослідження. Для вирішення поставлених завдань у роботі використано комплекс загальнонаукових та спеціальних методів; системний аналіз (для визначення функціональних вимог до моделі захисту персональної інформації користувачів соціальних мереж, ідентифікації основних компонентів архітектури та їх взаємозв'язків) та синтез (для вивчення структури загроз), порівняльний аналіз (при дослідженні політик соціальних мереж), узагальнення (для формування висновків та рекомендацій), метод архітектурного проєктування нейромереж (для побудови генеративно – змагальної мережі з урахуванням обмежень диференційної приватності (DP-SGD, gradient clipping, noise injection)), метод формальної специфікації (для формалізації вхідних, вихідних та внутрішніх параметрів моделі, визначення обмежень чутливості даних та допустимих рівнів приватності).

Короткий зміст роботи. Як результат у роботі проведено юридичний та технічний аналіз аспектів роботи соціальних мереж, нормативно - правове порівняння законодавства України та міжнародних стандартів з питання захисту персональних даних. Досліджено питання генеративного інтелекту та розроблені рекомендації щодо його експлуатації з додаванням алгоритмів модерації технологічними гігантами.

Галузь застосування одержаних результатів полягає у комплексній систематизації сучасних загроз персональним даним у соціальних мережах та розробці багаторівневої моделі рекомендацій, що поєднує поведінкові, технічні та організаційні аспекти захисту персональних даних з використанням генеративного інтелекту.

КЛЮЧОВІ СЛОВА; ІНФОРМАЦІЙНА БЕЗПЕКА, СОЦІАЛЬНІ МЕРЕЖІ, ГЕНЕРАТИВНИЙ ІНТЕЛЕКТ, ШТУЧНИЙ ІНТЕЛЕКТ, ШІ.

ABSTRACT

Text part of the qualification work for obtaining a master's degree; 86 pages, 19 figures, 15 tables, 62 sources.

The purpose of the work is the development of an approach and recommendations for ensuring the protection of users' personal information in social networks based on the use of software tools using artificial intelligence algorithms (generative-adversarial networks (GAN)) in order to increase resistance to information leaks

Object of research is the process of ensuring the protection of users' personal information in social networks.

Subject of research are software tools for protecting users' personal information in social networks based on the use of artificial intelligence algorithms (generative-adversarial data network - GAN).

Research methods. To solve the tasks set in the work, a complex of general scientific and special methods was used: system analysis (to determine the functional requirements for the model of protection of personal information of social network users, to identify the main components of the architecture and their relationships) and synthesis (to study the structure of threats), comparative analysis (when studying social network policies), generalization (to form conclusions and recommendations), the method of architectural design of neural networks (to build a generative-adversarial network taking into account the limitations of differential privacy (DP-SGD, gradient clipping. Noise injection)), the method of formal specification (to formalize the input, output and internal parameters of the model, to determine the limitations of data sensitivity and permissible levels of privacy).

Brief content of research. As a result, the work provides a legal and technical analysis of aspects of social networks, a regulatory and legal comparison of Ukrainian legislation and international standards on the protection of personal data.

Field of research consists in the comprehensive systematisation of modern threats to personal data in social networks and the development of a multi-level model of recommendations that combines behavioural, technical, and organisational aspects of personal data protection using generative intelligence.

KEYWORDS; INFORMATION SECURITY, SOCIAL NETWORKS, GENERATIVE INTELLIGENCE, ARTIFICIAL INTELLIGENCE, AI.

ЗМІСТ

ВСТУП

З розвитком генеративного інтелекту та технологій зв'язку (інтернет, мобільний зв'язок, IP-телефонія), виникнення глобальних платформ для соціальної взаємодії було очікуваним етапом інформаційного середовища. Соціальні мережі, шлях розвитку яких починався у вигляді примітивних засобів для спілкування та обміну контентом, з часом перетворилися на складну екосистему, яку інтегровано в суспільство та, якою користуються як інструментом для оприлюднення здебільшого персональної інформації.

Водночас, унаслідок такої інтеграції виника нова загроза у галузі захисту персональних даних, що стоїть у центрі дослідження. Користувачі, публікуючи свої дані (фотографії, дати народження, геолокаційні дані) та генеруючи метадані (у випадку з месенджерами, як-от Telegram, Viber, WhatsApp), власноруч створюють ризики. Ці дані, централізовано зібрані на серверах корпорацій, перетворилися на цінний актив сучасної економіки. Однак, на відміну від традиційних ресурсів, такий актив є частиною та власністю особистості, а його несанкціоноване використання, втрата чи витік можуть мати критичні наслідки для власників інформації: від шахрайства та крадіжки ідентичності до психологічного тиску та маніпуляції громадською думкою.

Для подолання проблеми забезпечення конфіденційності в умовах стрімкого розвитку шахрайських методів та адаптивних алгоритмів виникає нагальна потреба в імплементації адаптивних механізмів захисту, зокрема інтеграції генеративного інтелекту у захисний контур. Це зумовлює актуальність дослідження архітектур, заснованих на змагальному та диференційному принципі приватності.

Метою магістерської роботи розробка підходу та рекомендацій щодо забезпечення захисту персональної інформації користувачів у соціальних мережах на основі застосування програмних засобів, що використовують алгоритми

штучного інтелекту (генеративно-змагальні мережі (GAN)), з метою підвищення стійкості до витоків інформації

Для досягнення поставленої мети необхідно вирішити такі завдання;

1. Проаналізувати теоретико - методологічні засади захисту персональної інформації в соціальних мережах.
2. Ідентифікувати, систематизувати та дослідити основні методи захисту інформації.
3. Дослідити нормативно-правову базу у сфері захисту персональних даних, зокрема положення Закону України "Про захист персональних даних" та європейського Загального регламенту про захист даних (GDPR).
4. Дослідити модель генеративного інтелекту DP-GAN-HD на можливість імплементації в інфраструктуру медійних корпорацій та встановити перспективність впровадження.
5. Сформувати рекомендації щодо впровадження DP-GAN-HD в інфраструктуру соціальних мереж.

Об'єктом дослідження є процес забезпечення захисту персональної інформації користувачів у соціальних мережах.

Предметом дослідження є методи, засоби, технології генеративного інтелекту (DP-GAN-HD) та нормативно-правові засади захисту персональних даних у соціальних мережах.

Методи дослідження. Для вирішення поставлених завдань у роботі використано комплекс загальнонаукових та спеціальних методів; системний аналіз та синтез (для вивчення структури загроз), порівняльний аналіз (при дослідженні політик соціальних мереж), узагальнення (для формування висновків та рекомендацій); експериментальне моделювання (для оцінки ефективності алгоритму DP-GAN-HD, аналізу збіжності процесу навчання та верифікації балансу між рівнем приватності та корисністю синтезованих даних).

Наукова новизна отриманих результатів полягає у комплексній систематизації сучасних загроз персональним даним у соціальних мережах та розробці рекомендацій щодо впровадження генеративного інтелекту в захисний контур організації

Практичне значення одержаних результатів. Результати дослідження можуть бути використані організаціями для підвищення рівня цифрової безпеки інформації користувачів, фахівцями з кібербезпеки та розробниками соціальних платформ для вдосконалення механізмів захисту, а також у навчальному процесі при підготовці курсів з інформаційної безпеки та цифрової грамотності.

Апробація результатів. Результати дослідження апробовано на всеукраїнських конференціях: науково-практичній конференції *«Актуальні проблеми кібербезпеки»* матеріали всеукр. науково-практ. інтернет конф., м. Київ, 29 жов. 2025 р. 152-155, URL: https://duikt.edu.ua/uploads/p_2779_58326207.pdf; Всеукраїнській науково-практичній конференції *«Актуальні проблеми кібербезпеки»* матеріали всеукр. науково-практ. інтернет конф., м. Київ, 25 жов. 2024 р. 82-85 URL: https://duikt.edu.ua/uploads/p_2661_48963150.pdf, за результатами яких опубліковано тези.

Структура та обсяг роботи. Магістерська робота складається зі вступу, трьох розділів, загальних висновків та списку використаних джерел.

РОЗДІЛ 1
ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ЗАСАДИ ЗАХИСТУ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ В СОЦІАЛЬНИХ МЕРЕЖАХ

1.1 Цифровий профіль користувача та ризики безпеки персональних даних за використанням ШІ

Головним елементом сучасних соціальних комунікацій є персональна інформація. Якщо раніше така інформація існувала переважно у вигляді статичних записів у державних чи корпоративних базах даних, то з появою соціальних мереж цей актив перетворився на динамічний та легкодоступний ресурс. З кожною дією користувача, яку можливо зафіксувати генерується новий фрагмент даних з сукупності яких формується унікальний цифровий відбиток, через що соціальні мережі перетворилися із простих засобів спілкування на систему фікусування та аналізу людської поведінки.

Законом України “Про захист персональних даних” [1] (далі - ЗУ) відносини між суб’єктом, володільцем та розпорядником персональних даних встановлено наступним чином; у випадку, якщо у розпорядженні суб’єкта персональних даних присутня будь-яка інформація, необхідна для реєстрації на умовному веб-ресурсі А, такому суб’єктові необхідно надати своє погодження на обробку таких даних за умови зберігання чутливої інформації, що міститься у таких даних, у порядку, встановленому володільцем персональних даних та у відповідності до законодавства, після чого такі дані у визначеному вигляді надходять до розпорядника персональних даних. Сторонам взаємодії у ЗУ надані наступні визначення:

- Інформація - будь-які відомості та/або дані, які можуть бути збережені на матеріальних носіях або відображені в електронному вигляді.

- Володілець персональних даних - фізична або юридична особа, яка визначає мету обробки персональних даних, встановлює склад цих даних та процедури їх обробки, якщо інше не визначено законом. Тобто, володільцем називають таку сторону відносин, яка отримує згоду на обробку та зберігає надану інформацію у відповідності до встановленої нею-ж політики.
- Суб'єкт персональних даних - це саме людина, яка надає своє погодження або відхиляє згоду на обробку своїх персональних даних.
- Розпорядник персональних даних - фізична чи юридична особа, якій володільцем персональних даних або законом надано право обробляти ці дані від імені володільця;

Також, важливо розуміти різницю між володільцем та розпорядником даних, оскільки роз'яснення закону є неточним, і різниця між ними погано розкрита.

Отже, володільцем є безпосередньо фізична або юридична особа, якій суб'єкт персональних даних надає дозвіл на їх обробку; завданням розпорядника є зберігання, оброблення та, згідно із ЗУ "про захист персональних даних", їх передачу[1-2].

Згідно із Законом України "Про інформацію», інформація про фізичну особу, що підлягає захисту, визначається як конфіденційна інформація. Ця інформація охоплює будь-які відомості чи сукупність відомостей, що ідентифікують особу та є предметом правового регулювання [1-2].

Слід зазначити, що хоча ЗУ "Про інформацію" встановлює загальні категорії, більш детальна та пряма регламентація захисту саме персональних даних (як частини конфіденційної інформації про особу) здійснюється законом України "Про захист персональних даних". Згідно зі статтею 11 ЗУ "Про інформацію", до конфіденційної інформації про фізичну особу належать відомості про національність, освіту, сімейний стан, релігійні переконання, стан здоров'я, адреса проживання, дата і місце народження, її професійна, майнова приналежність та інші

відомості, які відповідно до закону не підлягають розголошенню. Головна вимога ЗУ “Про інформацію” полягає в тому, що збирання, зберігання, використання та поширення конфіденційної інформації про особу без її попередньої згоди забороняється, окрім випадків, прямо визначених законом (наприклад, в інтересах національної безпеки, економічного добробуту чи прав людини). За ЗУ “Про захист персональних даних” інформація, що підлягає захисту, поділяється на дві основні категорії; загальні та чутливі персональні дані (табл. 1.1)[1-3].

Таблиця 1.1

Категорії інформації, що підлягають захисту

Персональні дані	Ідентифікаційні дані Контактні дані Дані про громадянство та освіту Дані про професійну діяльність Фотографічне зображення (біометричні дані)
Чутливі персональні дані	Расове або етнічне походження Політичні, релігійні чи світоглядні переконання Членство у політичних партіях та/або професійних спілках Стан здоров’я та статеве життя Біометричні дані (крім фотографічних) - наприклад, відбитки пальців, сканування сітківки Генетичні дані Дані щодо притягнення до адміністративної чи кримінальної відповідальності

У соціальних мережах передбачено автоматичне видалення EXIF-даних (Exchangeable Image File Format) з завантажених зображень, налаштування чого необхідно контролювати з метою недопущення витоку небажаних даних, наприклад, про місцеположення та захисту приватності (щоб не розкривати точні GPS-координати місця зйомки, серійний номер камери тощо), а також для стандартизації файлів. Однак, це створює хибне відчуття безпеки. Хоча сторонні особи не можуть легко завантажувати фото і переглядати його метадані, сама платформа аналізує цю інформацію в момент завантаження. До подальшої

реєстрації даних підпадає геолокація, час, модель пристрою та інші доступні мета-дані, після чого зафіксована інформація реєструється до цифрового профілю користувача. [3]

Платформа також повністю перевіряє всю наявну інформацію, добровільно надану користувачем; хто ще позначений на фото, яка реакція інших користувачів на зображення, а також аналізує додані зображення за допомогою технологій комп'ютернох ідентифікації для встановлення об'єктів, облич, брендів та сцен. Інформація, що надається суб’єктом на добровільних підставах, можна поділити на такі категорії[4-5];

1. Explicitly Provided Data; дані, які надаються користувачем явно;

- ім'я, прізвище, дата народження, стать, номер телефону, адреса електронної пошти.
- місце проживання, освіта, місце роботи, сімейний стан, політичні та релігійні погляди.
- текст, фотографії, відео, історії, коментарі, особисті повідомлення в месенджерах.
- список друзів, підписки на сторінки та групи, позначки інших людей на фотографіях.

Користувач розуміє, що ця інформація стає доступною (принаймні для соціальної платформи або веб-застосунку), однак часто не до кінця усвідомлюється користувачем можливі способи її подальшого використання.

2. Observed Behavioral Data; дані, що збираються шляхом спостереження за поведінкою

Observed Behavioral Data є глибинним шаром даних. Компанії-власники соціальних мереж збирають інформацію про користувачів, та на підставі такої інформації виробляють аналітичні дані за результатами яких фіксують кожну

взаємодію користувача з такими платформами та їх контентом. Користувач не передає ці дані активно, вони є побічним продуктом його діяльності. Сюди входять;

- “вподобайки”, поширення (shares), кліки на посилання, час, проведений за переглядом певного допису чи відео.
- час та частота входів у мережу, типи пристроїв (мобільний телефон, ПК), операційна система, IP-адреса, дані геолокації (GPS, Wi-Fi точки доступу, стільникові вежі).
- вся інформація про те, які профілі, групи чи теми користувач шукав або переглядав.
- яка реклама була показана, чи клікнув на неї користувач, чи здійснив покупку після переходу.

Ці відомості для прошарку спостереження, зазначені за поведінкою є головними тригерами для аналізу профілів та розробки таргітованої реклами в соціальних мережах, оскільки вони дозволяють створювати точні поведінкові та прогнозтичні моделі. Схематично добровільно надані користувачем дані зображено в рис. 1.1.

3. Inferred or Predicted Data; результуючі або прогнозовані дані.

Цей прошарок самий складний з точки розуміння користувачем і є найменш прозорим. На основі перших двох шарів на базі алгоритмів машинного навчання та штучного інтелекту, платформа створює нову інформацію - прогнози та припущення про користувача. Ці дані не надаються і не спостерігаються безпосередньо, а генеруються самою системою. До цього прошарку включають прогнозовані інтереси, які на основі попередніх даних можуть бути випадково згенерованими; наприклад, визначення соціального статусу [5]. Порівняльний аналіз трьох шарів даних подано в таблиці 1.2.

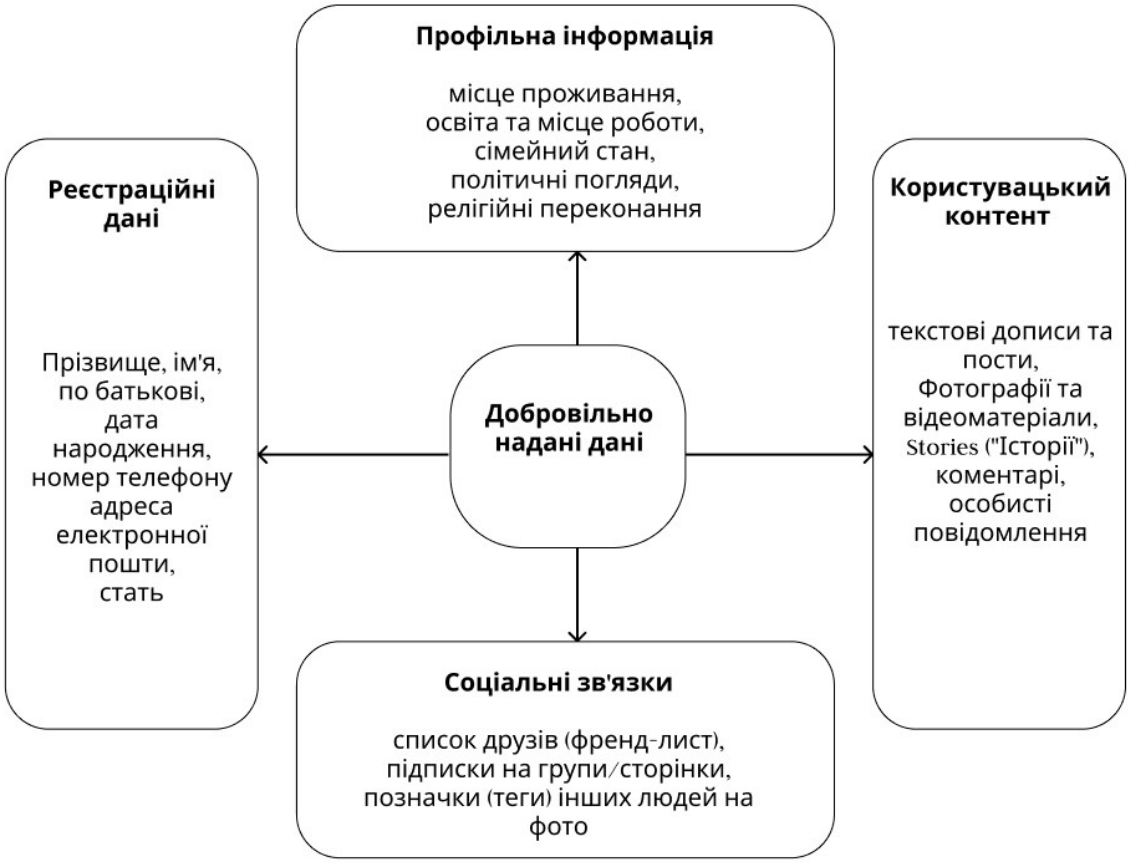

Рис. 1.1 Категоризація добровільно наданих суб’єктом інформаційних даних у цифровому профілі

Таблиця 1.2

Шари даних для аналізу профілю в соціальних мережах

Назва слою	**Властивості**	**Прозорість шару**
Explicitly Provided Data (дані, які надаються користувачем явно)	прізвище, дата народження, стать, номер телефону, адреса електронної пошти, місце проживання, освіта, місце роботи, сімейний стан, політичні та релігійні погляди, текстові дописи, фотографії, відео, "історії" (Stories), коментарі, особисті повідомлення в месенджерах та ін	Вища, оскільки сам користувач виконує дії, вказані в стовпчику властивості

Продовження табл. 1.2

Назва слою	**Властивості**	**Прозорість шару**
Observed Behavioral Data (дані, що збираються шляхом спостереження за поведінкою)	Дані про залученість, метадані активності, історія пошуку та переглядів, дані про взаємодію з рекламою	Середня, оскільки наведений прошарок є побічним продуктом
Inferred or Predicted Data (результуючі за попередньо нашарованими даними)	Платформа створює нову інформацію - прогнози та припущення про користувача.	Низька

Детальний цифровий профіль користувача формується шляхом об'єднання інформації з усіх трьох рівнів спостереження та збору даних. Саме завдяки цим цифровим слідам формуються бази даних, які використовуються корпораціями, що є власниками персональних даних, для подальшого коригування своїх бізнес-моделей та стратегій. Ця система використовує дані користувачів як актив на ринку цільової реклами, концепція, також відома як капіталізм спостереження. В обмін на дозвіл безкоштовно користуватися платформою користувачі надають свою увагу та особисті дані. Веб-додаток або інформаційна платформа обробляє дані та перетворює їх на прогнозні продукти для рекламодавців. Графічне зображення накладання даних для формування результуючих даних зображено в рис. 1.2

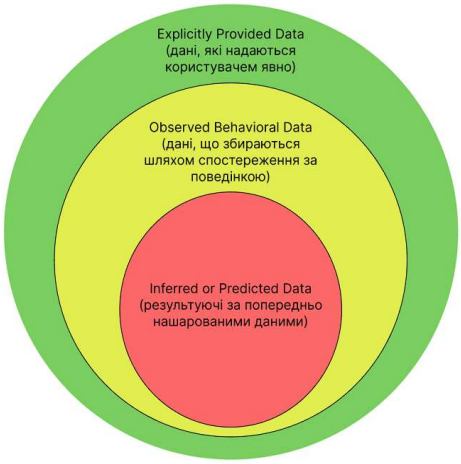

Рис. 1.2 Шари даних для створення цифрового портрету користувача соцмережі

Продукт корисний для надсилання повідомлення певній особі у потрібний час відповідно до її психологічного стану та інтересів разом з її передбачуваними намірами. Такий підхід створює системні ризики. Скандал з Cambridge Analytica [5] продемонстрував, що дані, зібрані з мільйонів профілів у Facebook, використовувалися для розробки детальних психометричних профілів виборців.

Ці профілі стали основою політичних повідомлень, які були розроблені та поширені з метою психологічного впливу на невизначених або незадоволених виборців, щоб змінити їхню виборчу поведінку. Алгоритми машинного навчання також стикаються з можливістю алгоритмічної дискримінації під час процесу навчання. Існує ймовірність збереження та неправильного тлумачення стереотипів з історичних даних, що базуються на певних групах населення. Алгоритми, що спостерігають за цим, можуть робити помилкові висновки про ці групи населення з точки зору фінансової спроможності або відсутності належної кваліфікації для роботи з високим рівнем доходу. Системи можуть припинити розміщувати оголошення про кредити або роботу для цих груп населення, які підпадають під такі стереотипи. Система залишається невидимою для користувачів і власників даних, отже, бар'єр, який вона створює, посилює соціальну нерівність у суспільстві. Сам

факт того, що будь-яка діяльність постійно реєструється, аналізується та оцінюється, викликає недовіру до соціальної мережі з точки зору людей [4-5].

Інша проблема штучного інтелекту полягає в тому, що він може бути неоднозначним. З одного боку, завдяки безперервному навчанню штучний інтелект може бути вічно пильним у захисті інформації або зробити її нечитабельною для інших моделей чи людей. З іншого боку, він також може стати джерелом інцидентів через забруднення навчальних наборів даних та витік даних через непродумані висновки. Та сама потужність великих мовних моделей та інших генеративних систем надає повний набір інструментів для моделювання атак, автоматизації, прогнозування вразливостей, коли дані доступні, та навчання системи для оперативного реагування на інциденти. Однак ті самі технології, якщо вони потрапляють до рук зловмисників, створюють або трансформують існуючі загрози до більш продвинутого рівня [10]. Надалі наведено декілька прикладів щодо можливого зловживання генеративним інтелектом;

- Генерація фішингових листів, оскільки генеративна модель при правильному навчанні здатне до повного відтворення оригінального листа;
- Повторення дизайну веб-сайтів;
- Використання Deep-Fake для отримання довіри;

У випадку прийняття генеративного ШІ до підрозділів кібербезпеки в організаціях, переваги наступні[6-10]:

- Генеративний інтелект , маючи доступ до баз даних, системних логів SIEM систем та до мережі, здатен до аналізу трафіку, зменшення відсотку помилкових спрацювань, виявлення нападів і загроз нульового дня та визначення пріорітетів до загроз. Наведені функції корисні для новостворених підрозділів, яким необхідний інструмент для фіксації та консультацій.

- Генеративний інтелект при формуванні коректних запитів здатен фільтрувати масиви даних для усунення помилкових спрацювань, виявлення аномалій і закономірностей та надання рекомендацій щодо подальшого плану дій.
- Використання генеративного інтелекту також може бути корисним для оновлення вже наявних ландшафтів загроз в конкретній організації завдяки автоматичному пошуку вразливих ділянків коду, аналізу мережевого трафіку з подальшою генерацією детального звіту щодо поводіння загроз або окремих сценаріїв
- Генеративний інтелект здатен до прискорення прийняття рішень аналітиків в стадії реагування на інциденти.

Крім цього, моделі генеративного інтелекту дають змогу проводити реалістичне тестування атак за допомогою великих мовних моделей та змагальних мереж, які імітують поведінку зловмисника. Це дозволяє створювати підроблені цілі або приманки для зовнішніх загроз, розробляти сценарії фішингових атак або відпрацьовувати інциденти без ризику для реальної інфраструктури. ГІ аналізує код, виявляє вразливості та пропонує можливі виправлення. Це пришвидшує процес захисту, зменшує вікно вразливості та знижує ризик людської помилки при реалізації патчів.

Ризики, пов’язані з використанням генеративного інтелекту в організації умовно можна поділити на ризики;

- пов’язані з використанням в організації інструментів ШІ
- як результат його використання фізичними особами, включаючи зловмисників. Схематично подано на рис. 1.3.

Ризики, пов’язані внаслідок використання генеративного штучного інтелекту організацією виникають внаслідок впровадження та експлуатації LLM, GAN-моделей для синтезу даних або встановлення інструментів автоматизації у робочий процес [10].

- коли співробітники використовують публічні хмарні сервіси (наприклад, загальнодоступні чат-боти) для аналізу, резюмування або генерації коду, вони можуть ненавмисно завантажити в систему комерційну таємницю, персональні дані клієнтів або захищений інтелектуальний продукт. Ці дані можуть бути використані для навчання публічної моделі, що фактично робить їх доступними для третіх сторін.

- Якщо організація навчає власні генеративні моделі на несанітизованих або скомпрометованих даних, це може призвести до того, що модель почне генерувати шкідливий, упереджений або помилковий результат. Це особливо небезпечно для моделей, які використовуються у фінансовому чи медичному секторах.

- Генеративний інтелект створює контент, навчаючись на величезних масивах даних. Використання організацією згенерованих зображень, текстів або коду може призвести до судових позовів, якщо кінцевий продукт містить значну схожість із захищеним матеріалом, який був у тренувальному наборі моделі.

- У випадку, коли генеративний інтелект (особливо складні LLM) приймає бізнес-рішення, його непрозорість ускладнює аудит та пояснення прийнятого рішення. Це створює юридичні та регуляторні ризики, особливо у сферах, де вимагається повна прозорість процесу (наприклад, скоринг, фінансовий аналіз).

Ризики використання генеративного інтелекту фізичними особами, включаючи зловмисників виникають в результаті відсутності нагляду за периметром організації і спрямовані на співробітників, клієнтів або цифрову інфраструктуру цільової організації [9-11] .

- Зловмисники використовують генеративний інтелект для створення бездоганних фішингових електронних листів, які не містять граматичних помилок

і можуть імітувати стиль спілкування конкретного керівника чи співробітника. Це різко підвищує ймовірність успішної атаки.

- Зловмисники використовують GAN-моделі для створення реалістичних аудіо- та відеофейків, що імітують голос та обличчя керівництва (наприклад, для "CEO Fraud"). Мета - ввести в оману або отримати несанкціонований доступ до важливих вузлів підприємства.

- Генеративний інтелект використовується для прискорення та автоматизації процесу розробки нового шкідливого програмного забезпечення або для написання високо ефективних змагальних промптів, що експлуатують вразливості в системах організації. В наслідок цього частота атак може суттєво збільшитись.

- Широке використання діпфейків і ботів, що генерують текст, руйнує довіру до цифрового контенту. Співробітники організації можуть стати жертвами зовнішньої дезінформації або витрачати значний час на верифікацію автентичності кожного повідомлення чи документа.

- Зловмисники можуть використовувати генеративний інтелект для обробки великих обсягів публічно доступних даних (Oversharing) для деанонімізації ключових співробітників або створення детальних психологічних профілів, які пізніше використовуються для прицільних атак соціальної інженерії. Структуровану інформацію наведено в таблиці 1.3.

Таблиця 1.3

Ризики інформаційній безпеці в організації при використанні ШІ

Тип загрози	**Назва ризику**	**Сутність ризику**	**Об'єкт Ураження**
І. Використання генеративного інтелекту в організації	Витік конфіденційної інформації	Ненавмисна передача чутливих даних у публічні хмарні ГІ - сервіси для їх навчання	конфіденційність, комерційна таємниця.

Продовження табл. 1.3

Тип загрози	Назва ризику	Сутність ризику	Об'єкт Ураження
I. Використання генеративного інтелекту в організації	Зараження навчальних даних	Свідоме або ненавмисне включення скомпрометованих зразків у внутрішні навчальні набори, що призводить до генерації внутрішньою моделлю шкідливого результату	Цілісність, якість рішень.
	Порушення авторського права	Використання згенерованого ГІ контенту (тексту, зображень, коду), який може мати значну схожість із захищеним	Правова відповідність, інтелектуальна власність.
	Автоматизована розробка шкідливого коду	Використання генеративного інтелекту для прискорення генерації нових варіацій шкідливого ПЗ (поліморфний код) або автоматичного написання експлойтів для використання вразливостей.	Інфраструктура, безпека мережі.
	Деанонімізація та посилення профілювання	Обробка великих обсягів публічних даних генеративним інтелектом для створення профілів	Приватність, психологічна стійкість.
Використання генеративного інтелекту зловмисниками та фізичними особами	Посилення фішингових атак	Використання генеративного інтелекту для створення бездоганних електронних листів, які імітують стиль спілкування керівника або колеги.	Співробітники, фінанси.
	Генерація діпфейків для шахрайства	Створення реалістичних аудіо- та відеофейків з використанням голосу та обличчя керівництва організації.	Довіра, фінанси, репутація.

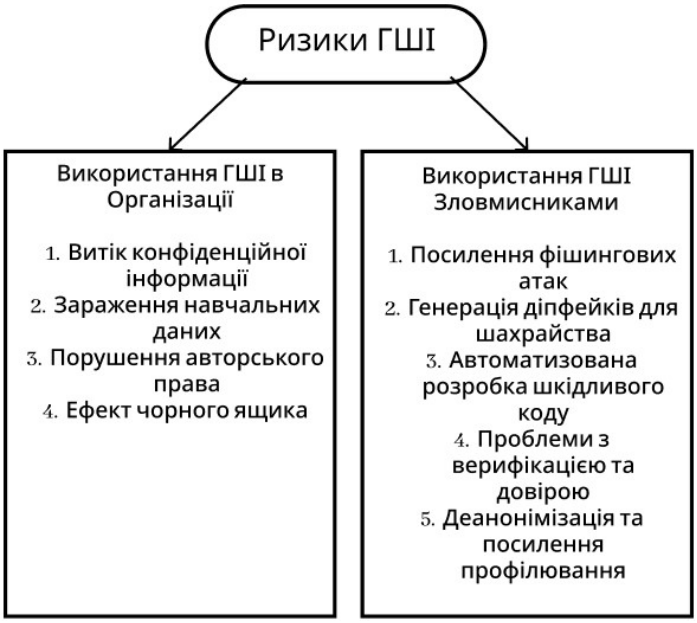

Рис. 1.3 Ризики генеративного інтелекту для інформаційній безпеці організації

Головним завданням мінімізації ризиків застарілих технологій безпеки є впровадження архітектури нульової довіри, заснованої на методах виявлення аномалій. Це помітно підвищує рівень виявлення загроз і реагування на інциденти, зменшуючи шанси на успішну кібератаку. Водночас політика DLP повинна застосовуватися як на кінцевих точках, так і на периметрі мережі. Цифрові активи повинні бути захищені шляхом послідовного застосування шифрування, жорсткого контролю доступу, а також регулярного аудиту та оцінки ризиків з метою запобігання витоку несанкціонованої інформації. Кодекси безпеки повинні бути вбудовані безпосередньо на етапі розробки моделі машинного навчання. Ігнорування етичних стандартів гарантує упередженість даних, що знову ж таки призводить до ризику розробки дискримінаційних алгоритмів. У цьому світлі етичні принципи повинні бути впроваджені як невід'ємна частина системи контролю якості в операційних процесах[11].

1.2 Нормативно-правове забезпечення регулювання питань захисту персональних даних

Приватність та конфіденційність персональних даних є головним правом людини в інформаційному полі, яке є необхідністю для повноцінного життя у вільному суспільстві. В сучасності дані стали «валютою» цифрової епохи, і набуває чинності питання контролю над розпорядниками, володільцями та об'єктами персональних даних.

Якщо раніше основна загроза полягала у надмірному державному нагляді, то з появою цифрових технологій ситуація ускладнилася в результаті появи корпорацій, які формують бізнес - модель навколо персональних даних та цифрових двійниках. Приклади навколо нас; соціальні мережі відстежують наші дії, пошукові системи формують профілі користувачів, онлайн-магазини прогнозують покупки на основі історії замовлень.

Такі зміни змусили держави шукати нові підходи до регулювання. Звідси й поява нормативних актів на кшталт Загального регламенту захисту даних ЄС (GDPR).

Фундамент сучасних норм конфіденційності заклали ще у 1948 році із прийняттям Загальної декларації прав людини. Тоді стаття 12 проголосила недоторканність особистого та сімейного життя. Згодом цю тезу підсилив Міжнародний пакт 1966 року, проте довгий час ці документи залишалися здебільшого декларативними - реальних важелів впливу на потоки інформації вони не давали [11].

Ситуація почала змінюватися у 1980-х. Спершу ОЕСР опублікувала рекомендації щодо справедливої обробки даних, а вже за рік Рада Європи ухвалила Конвенцію 108. Це був переломний момент; з'явився перший юридично обов'язковий міжнародний договір, до якого Україна приєдналася у 2010 році [12].

Ключова зміна сталася у 2018 році з набуттям чинності Загального регламенту захисту даних (GDPR). Цей документ змінив правила гри не лише для Європи. Завдяки принципу екстериторіальності, будь-яка компанія світу, що працює з даними європейців, змушена грати за цими правилами. GDPR не просто оновив старі норми, а жорстко кристалізував сім ключових принципів, які тепер є золотим стандартом захисту.

В основі всього лежить законність, справедливість та прозорість. Це означає кінець ери дрібного шрифту та заплутаних умов. Користувач має чітко розуміти; хто, як і навіщо бере його дані. Якщо українське законодавство поки що дозволяє певні розмиті формулювання, то GDPR вимагає конкретики. Згода має бути активною дією - жодних «галочок за замовчуванням».

Не менш важливим є обмеження мети. Дані збираються під конкретну задачу. Не можна взяти телефон для доставки піци, а потім використати його для розсилки реклами. Це запобігає «розповзанню» функціонала. Хоча в Україні це правило теж існує на папері, на практиці контроль за цільовим використанням значно слабший.

Потрібно брати лише той мінімум, без якого неможливо надати послугу. Якщо сервісу достатньо емейлу, вимагати дату народження - порушення [4].

Дані також повинні відповідати реальності - це принцип точності. Застаріла або помилкова інформація має виправлятися або видалятися. Це тісно переплітається з правом людини на виправлення (rectification), яке гарантують як європейські, так і українські норми.

Принцип обмеження терміну зберігання вимагає видаляти або анонімізувати дані, як тільки мета їх збору досягнута. Це питання безпеки; менше даних в архіві - менші ризики при можливому зламі. На відміну від чітких вимог GDPR, українські норми тут часто відсилають до розмитих галузевих стандартів.

Окремий блок - це цілісність та конфіденційність (безпека). Шифрування, контроль доступу, псевдонімізація - це база. Але GDPR пішов далі; він запровадив

жорсткий дедлайн у 72 години для повідомлення регулятора про витік даних. Це змушує компанії не замовчувати інциденти[7-9].

Особливого значення у забезпеченні внутрішньої відповідності набуває інститут інспектора з питань захисту даних (Data Protection Officer, DPO), закріплений у статтях 37-39 GDPR. Призначення DPO є обов'язковим для державних установ та приватних компаній, які проводять великий або системний обробіток персональних чи чутливих даних. DPO повинен мати експертні знання у сфері захисту даних і працювати незалежно; володілець і процесор не мають права давати йому вказівки або накладати санкції за виконання його функцій. До його обов'язків входять консультування працівників, контроль за дотриманням GDPR, рекомендації щодо проведення DPIA та взаємодія з наглядовим органом. В Україні подібного інституту з чітко визначеною незалежністю та функціональними повноваженнями поки немає, що призводить до формального виконання вимог і можливих конфліктів інтересів.

Принцип підзвітності (accountability). Вже недостатньо просто дотримуватись закону. Це перехід до проактивної позиції; документування процесів, оцінка ризиків, впровадження приватності на етапі розробки та призначення DPO. В українському правовому полі механізми підтвердження відповідності поки що значно слабші.

Законодавство не лише обмежує бізнес, а й озброює користувача правами.

Базовим є право на доступ (ст. 15 GDPR, ст. 8 Закону України). Людина може запитати у компанії: «Що ви про мене знаєте?», отримати копію даних та дізнатися, кому їх передавали. Якщо інформація виявиться помилковою, вступає в дію право на виправлення[12].

Найбільш дискусійним питанням стало право на видалення («право на забуття»). Воно дозволяє вимагати стирання даних, якщо вони більше не потрібні або якщо людина відкликала згоду. І хоча український закон дозволяє подати

вмотивовану вимогу про знищення даних, реалізувати це на практиці у нас значно складніше, ніж у юрисдикції ЄС.

Контролем за дотриманням вимог щодо захисту персональних даних в Україні займається Уповноважений верховної ради з прав людини (ст. 23 закону України). Йому належить розгляд скарг, проведення перевірок та видача приписів. Однак санкції, передбачені кодексом України про адміністративні правопорушення (ст. 188 п.39), залишаються порівняно низькими, що не забезпечує достатньої стримуючої сили. Для порівняння, GDPR (ст. 83) передбачає штрафи до 20 млн євро або 4 % від річного обороту компанії.

Відповідно до статей 44-50 GDPR, такі передачі можливі лише у випадках, коли держава, до якої здійснюється передача, забезпечує «адекватний рівень захисту», що підтверджується рішенням Європейської комісії. Для України отримання такого рішення є одним із пріоритетних завдань приєднання до європейського цифрового ринку. Національне законодавство потребує відповідної реформи, що передбачає підвищення незалежності наглядового органу, запровадження принципу підзвітності та застосування більш суворих санкцій відповідно до європейських стандартів[12-14].

Таким чином, положення GDPR не тільки формулюють права суб'єктів даних, але й надають засоби для забезпечення таких прав. Українське законодавство відтворює загальні принципи, але відстає в деталях і практичному застосуванні. Уніфікація національного законодавства з GDPR зможе забезпечити захист прав українських громадян, а також подальшу відповідність Європейському Союзу в цифровій сфері.

Є також суттєві відмінності оброблення «спеціальних категорій» персональних даних (чутливих даних). Стаття 9 GDPR забороняє обробку даних про расове чи етнічне походження, політичні переконання, релігійні та філософські погляди, членство у профспілках, генетичну, біометричну інформацію, стан

здоров’я, сексуальне життя чи орієнтацію, окрім кількох чітко визначених винятків. Українське законодавство передбачає обробку подібних даних лише за умови «однозначної згоди», що є менш жорсткою вимогою у порівнянні з «явною» (explicit) згодою, яку вимагає GDPR. Така різниця ослаблює захист чутливої інформації. Не менш важливим є питання інституцій. GDPR передбачає створення незалежних органів захисту даних (DPA) з широким спектром повноважень - від проведення перевірок і аудитів до накладення виправних заходів та штрафів, їхня діяльність координується Європейською радою з захисту даних (EDPB). В Україні аналогічні функції виконує Уповноважений Верховної Ради з прав людини, проте через брак коштів, персоналу і занадто широкі повноваження цей орган не може ефективно виконувати свої обов’язки. Принцип підзвітності, закріплений у GDPR, реалізується через конкретні юридичні інструменти, зокрема Data Protection by Design and by Default (стаття 25) та оцінку впливу на захист даних (DPIA, стаття 35). Ці механізми зобов’язують впроваджувати захисні заходи ще на етапі проєктування систем і бізнес-процесів, що дозволяє запобігати проблемам заздалегідь. В Україні ж поки що реагують на проблеми за реактивною моделлю, реагуючи на інциденти лише після їх виникнення. Варто також згадати про передачі даних між країнами. Українське правове регулювання доступу державних органів до персональних даних потребує суттєвих змін задля забезпечення пропорційності, незалежного контролю та судового захисту. Без цих реформ будь-яка гармонізація комерційного законодавства залишатиметься лише частковою. Отже, реформування сфери захисту персональних даних в Україні потребує комплексних змін у різних сферах, і тут недостатньо просто скопіювати європейські норми - треба переосмислити базові засади правової держави та цифрових прав людини[2-3].

Оновлення законодавчої бази у сфері захисту персональних даних в Україні вимагає всебічного, міждисциплінарного підходу і перегляд базових принципів функціонування правової основи в сфері цифровіації. Подальший аналіз слід

спрямувати на розгляд конкретних інститутів і механізмів, які утворюють фактичну структуру правозастосування в цій галузі, адже саме вони формують різницю між декларативним проголошенням прав і їх реальним ефективним втіленням.

Однією з ділянок є розрізнення юридичних статусів і відповідальності володільця та процесора(розпорядник) персональних даних. У статті 4 GDPR чітко визначено; володілець - фізична чи юридична особа, яка самостійно або спільно з іншими встановлює цілі та засоби обробки даних; процесор(розпорядник) - суб'єкт, який виконує обробку від імені володільця.

Володілець несе головну відповідальність за дотримання принципів GDPR; він зобов'язаний мати законну підставу для обробки, гарантувати реалізацію прав суб'єктів даних і підтверджувати підзвітність. Процесор діє лише за інструкціями володільця, проте регламент накладає на нього прямі обов'язки; впровадження адекватних заходів безпеки, негайне повідомлення про інциденти розкриття даних та заборону залучати субпроцесорів без письмової згоди володільця. Відносини між володільцем і процесором регулюються юридично зобов'язуючим договором - Data Processing Agreement (DPA), який визначає предмет, мету, обсяг, категорії даних та обов'язки обох сторін.

В українському законодавстві існують терміни «володілець» і «розпорядник» персональних даних, які концептуально співпадають із володільцем і процесором. Однак Закон України «Про захист персональних даних» не передбачає настільки ж суворих і деталізованих вимог до угод між ними, що створює прогалини у ланцюгу відповідальності і ускладнює ефективний моніторинг обробки даних[12-15].

GDPR зосереджується на визначенні цілей та загальних принципів, тоді як власник може на власний розсуд застосовувати конкретні технологічні рішення, залежно від таких критеріїв, як «сучасний стан техніки», вартість впровадження, контекст та ризики. Така модель дозволяє регулюванню залишатися актуальним протягом тривалого часу та бути гнучким. Водночас в українській практиці часто

виникає необхідність дотримуватися державних стандартів ДСТУ або сертифікованих засобів захисту, що може суттєво гальмувати розвиток інновацій і призвести до лише формальної відповідності без фактичного забезпечення безпеки даних.

Впровадження вимог GDPR є ресурсоємним процесом: юридичні аудити, реінжиніринг бізнес-процесів, нові інформаційні технології, навчання персоналу та призначення DPO. Це вже є серйозною перешкодою для малих та середніх підприємств. Але поширення штучного інтелекту створило нові виклики, які зараз активно обговорюються Європейською Комісією.

Принцип прозорості стикається з проблемою «чорного ящика», коли навіть розробники не можуть чітко пояснити причини, через які нейронна мережа дійшла певного висновку. Принцип обмеження мети є надто неоднозначним, оскільки модель, навчена виконувати конкретне завдання, може бути легко перенаправлена на інші, непередбачені цілі. Принцип мінімізації даних суперечить природі сучасних ML-моделей, які потребують величезних обсягів інформації для досягнення високої точності[15-18].

Механізми захисту передбачені GDPR, зокрема статтею 22, яка гарантує суб'єктам даних право не підлягати автоматизованим рішенням, включаючи профілювання, у випадках, коли їх наслідки мають юридичні або інші значні наслідки. Цього захисту недостатньо, і в результаті Європейський Союз видав окремий нормативний акт - Закон про ШІ, який доповнює GDPR і встановлює правила використання систем ШІ залежно від їхнього профілю ризику.

Другим фактором є стрімкий розвиток Інтернету речей (IoT). Мільярди підключених пристроїв постійно записують інформацію про звички, здоров'я, пересування і навіть розмови користувачів проти їхньої волі. Створення мережі постійного спостереження становить загрозу для приватного життя. Забезпечте

належну інформацію та згоду, коли йдеться про пристрої, які можуть не мати екрану або взаємодія з якими є обмеженою.

Більш активне застосування біометричних даних у процесах ідентифікації та автентифікації підвищує безпеку, але одночасно створює ще один серйозний ризик - формування баз даних ідентифікаторів людей, витік яких може мати негативні наслідки для людей.

У цьому контексті слід підкреслити, що побудова ефективної системи захисту персональних даних є для України довготривалим процесом. Він базується на політичній волі до проведення далекосяжних інституційних реформ, на інвестиціях у такий розвиток з боку бізнес-спільноти, а також на підвищенні рівня цифрової грамотності та формуванні культури конфіденційності в суспільстві.

Національна стратегія України щодо цифрової трансформації держави передбачає широке застосування ШІ та IoT, тому питання полягає не лише у простій адаптації GDPR у його вигляді, а і в формуванні правової та інституційної основи, достатньо гнучкої і стійкої, щоб адекватно реагувати на розвинені технології.

Підсумовуючи, варто підкреслити, що шлях України до створення дієвої системи захисту персональних даних - це тривалий процес. Він потребує політичної волі для проведення суттєвих інституційних реформ, готовності бізнес - спільноти інвестувати, а також підвищення цифрової грамотності та формування культури приватності в суспільстві.

Отже, У межах українського законодавства питання захисту даних переважно регулюється загальними нормами та базовими принципами недоторканності приватного життя. Хоча на нормативному рівні існує вимога забезпечувати конфіденційність, цілісність і доступність персональних даних, практична реалізація цих принципів суттєво залежить від технічної спроможності суб’єктів обробки та рівня контролю з боку держави. Держава висуває вимоги щодо попередження несанкціонованого доступу, зберігання інформації на території

України у випадках роботи з критичними реєстрами, а також можливості доступу до даних правоохоронних органів за процедурою.

Європейський Союз, на відміну від України, формує жорстку систему регулювання, що проявляється у вимогах щодо територіальності зберігання, контролю над передачею даних до третіх країн та формування обов'язку демонстрації відповідності принципам GDPR. Переважно це пов'язано з прагненням ЄС зберігати контроль над інформацією задля уникнення можливості отримання цієї інформації іноземними державами або корпораціями для профілювання користувачів, впливу на їхню поведінку або створення систем масового спостереження.

Міжнародні вимоги щодо розміщення даних визначаються балансом між державними інтересами, корпоративною відповідальністю та прагненням користувачів до приватності[18-19].

Усі системи регулювання, незалежно від моделі, спираються на концепцію чутливості даних, яка є базою для оцінки ризиків і вибору засобів захисту. Чутливість визначається не лише формальним типом інформації, а й контекстом її отримання, можливістю однозначної ідентифікації особи, потенційною шкодою, яку може спричинити витік, ступенем приватності змісту, обсягом даних, тривалістю зберігання, колом суб'єктів, що мають доступ, та географією передачі. Так само важливим є рівень захищеності від несанкціонованого доступу; застосування шифрування, наявність технічних і організаційних заходів, контроль доступу та архітектура системи в цілому. Дані стають особливо чутливими, коли вони дозволяють відтворити детальний цифровий профіль особи, визначити її звички, місцезнаходження, соціальні зв'язки, медичний стан чи політичні уподобання, а також коли обробка відбувається у нестабільному середовищі, зокрема в умовах війни або з використанням платформ, що розміщують інформацію за межами юрисдикцій з високим рівнем правового захисту.

Таким чином, міжнародні вимоги до розміщення інформації, національні особливості регулювання та параметри оцінки чутливості утворюють комплексну структуру, яка визначає фактичний рівень безпеки даних користувачів. Україна поступово рухається до гармонізації з європейськими вимогами, однак на практиці зберігається суттєвий дисбаланс між формальною нормативною базою та реальним рівнем контролю. ЄС, своєю чергою, зосереджується на створенні суворого правового поля, орієнтованого на мінімізацію ризиків, обмеження необґрунтованого збору даних та забезпечення прозорості. Наочно оцінку стану впровадження GDPR в Україні наведено в таблиці 1.4.

Таблиця 1.4

Порівняння правил та контролю щодо захисту персональних даних

Положення	**Правила**		**Контроль**		**Параметри оцінки чутливих даних**
	Нормативно-правові документи (Україна)	**GDPR**	**Нормативно-правові документи (Україна)**	**GDPR**	
Принципи обробки	Формальні	Чіткі	слабкий контроль;	аудит	Тип даних, можливість ідентифікації
Мета обробки	Гнучкі	Жорсткі	розмитий	ретельна перевірка	Обсяг, контекст збору
Підстави обробки	Основні	6 підстав	за згодою майже формальний	потребує глибокого аналізу	Контекст, шкода
Згода	Однозначні	Явна	за підміною згоди	жорсткий	Тривалість, обсяг
Категорії даних	Слабке розмежування	Жорстке	широкий спектр дозволів	заборони	Тип, шкода
Оцінка ризиків (DPIA)	Немає	Обов’язкова	Не потрібна	Жорсткі вимоги	Всі існуючі параметри
Повідомлення про витік	Невизначено	72 години	часто приховують	штрафи	Потенційна шкода
Передача даних за кордон	Слабкі	Суворі	майже без контролю	SCC/BCR	Географія, доступ

Продовження табл. 1.4

Положення	Правила		Контроль		Параметри оцінки чутливих даних
	Нормативно-правові документи (Україна)	GDPR	Нормативно-правові документи (Україна)	GDPR	
DPO	Не потребується	Потребується	відсутній	обов'язковий	Аналіз всіх параметрів
Privacy by Design	Немає	Є	відсутність вимог	аудит	Механізми доступу, шифрування
Права суб'єктів	Формальні	Фіксовані	ігнорування запитів	штрафи	Географія передачі
Безпека даних	Загальна	Чіткі	сертифікати	аудит	Механізм доступу, чутливість контенту
Право на забуття	Важко реалізуються	Обов'язкові та просто реалізуються	повільне видалення	видалення	Залежить від виду інформаційного активу та терміну його зберігання

GDPR має перелік законних підстав для обробки персональних даних, зазначений у статті 6 регламенту, який виключає будь-яку обробку, що не підпадає під шість чітко визначених випадків. Окрім згоди суб'єкта даних, обробка допускається у разі виконання договору, виконання юридичного зобов'язання, захисту життєво важливих інтересів особи, реалізації завдань, що виконуються в інтересах суспільства, або при використанні офіційних повноважень. Особливу увагу привертає підстава «законного інтересу», яка дозволяє здійснювати обробку для досягнення легітимних цілей володільця чи третьої сторони, за умови, що права та свободи суб'єкта даних не переважають над цими інтересами. Застосування цієї підстави передбачає триетапний аналіз; визначення законного інтересу, аргументація необхідності обробки та балансування інтересів. В українському законодавстві передбачені загальні підстави, наприклад «укладення та виконання правочину» або «необхідність виконання обов'язку, встановленого законом».

Відсутність чіткої категорії «законного інтересу» та відповідного тесту на баланс створює правову неоднозначність і змушує операторів надмірно покладатися на згоду, що не завжди є виправданим рішенням[14-16].

1.3 Основні підходи та методи щодо захисту інформації в соціальних мережах

Тріада основних принципів, таких як конфіденційність, цілісність і доступність, є основою для побудови системи інформаційної безпеки [13]. У разі дотримання державних нормативних вимог до цієї тріади додається принцип спостережності. Конфіденційність означає, що конфіденційні дані не можуть бути доступними нікому, крім уповноважених осіб. Цілісність дозволяє уникнути несанкціонованих змін інформації, забезпечуючи її надійність і повноту. Доступність, з іншого боку, передбачає вільний доступ уповноважених користувачів до інформації в потрібний час. Відповідальність означає, що дії персоналу реєструються і фіксуються в автоматизованій системі. Недотримання хоча б одного з основних принципів призводить до негативних наслідків як для фізичних, так і для юридичних осіб. Таким чином, всі наведені нижче методи захисту будуть розглядатися як такі, що забезпечують ці три основи безпеки [13].

Система аутентифікації «логін-пароль» продемонструвала свою слабкість, і багатофакторна аутентифікація, в більшості випадків двофакторна, 2FA, стала стандартом сучасної безпеки. Перевірка ідентичності повинна здійснюватися за допомогою двох різних факторів: чогось, що знає користувач - пароля, і чогось, що він має - одноразового коду з мобільного додатку, або чогось, що є його біометричною особливістю - відбитка пальця.

Із урахуванням розвитку методів та атак на користувацькі системи, виникають і нові загрози. Простір ризиків змінюється під впливом штучного інтелекту, що вимагає переходу до проактивних стратегій захисту.

У відповідь на розвиток ландшафту загроз модернізуються і методи захисту. концепція управління цифровим слідом - сукупність всієї інформації, залишеної користувачем у мережі. Ефективне управління передбачає періодичний аудит власної онлайн-присутності з метою зменшення обсягу даних, доступних для зловмисників. Крім того, безпека акаунту пов'язана із безпекою пристрою, що вимагає дотримання базових правил кібербезпеки; своєчасного оновлення ПЗ, використання антивірусних програм та обережності при завантаженні файлів.

Соціальні мережі надають своїм користувачам гнучкі налаштування конфіденційності, щоб вони могли точно вказати, хто бачить кожну публікацію і скільки інформації надходить до сторонніх додатків.

Однак побудувати надто складну систему безпеки неможливо через низку обмежувальних факторів; людських, економічних та технологічних. Серед найпоширеніших загроз можна згадати соціальну інженерію, особливо фішинг. Зловмисники обманюють психологію користувачів, змушуючи їх добровільно розкривати конфіденційні дані на певних підроблених веб-сайтах. Єдиним протиотрутою є те, що користувачі повинні бути невпинно пильними та критичними і вміти розпізнавати підроблені повідомлення. Поряд з цим, культура поводження з інформацією базується на принципі мінімізації даних - свідомому обмеженні публікації конфіденційної інформації.

Протокол HTTPS захищає канали зв'язку між клієнтом і сервером, створюючи зашифроване з'єднання, яке мінімізує перехоплення даних. Повідомлення шифруються безпосередньо на серверах платформи за допомогою алгоритмів шифруванняAES і KALYNA (за національною стандартизацією), що забезпечує шифрування даних у стані спокою. Особлива увага приділяється захисту комунікацій між користувачами, де все більшого поширення набуває наскрізне шифрування як провідний стандарт. Основні підходи та методи щодо захисту інформації в соціальних мережах подано в таблиці нижче (табл. 1.5).

Таблиця 1.5

Основні підходи та методи щодо захисту інформації в соціальних мережах та на пристрої

Підхід	**Принцип**	**Характеристика**	**Мета**
Модель безпеки (Тріада CIA + O)	Конфіденційність(C)	Доступ до чутливих даних надається виключно авторизованим суб'єктам.	Запобігання несанкціонованому ознайомленню з даними.
	Цілісність(I)	Достовірність та повнота інформації.	Захист даних від несанкціонованих модифікацій (змін).
	Доступність(A)	Можливість безперешкодного доступу до інформації для авторизованих користувачів у потрібний час.	Гарантування безперебійної роботи системи.
	Спостережність (норми)(O)	Фіксація та реєстрація дій персоналу в межах автоматизованої системи.	Забезпечення аудиту та відстеження подій.
Методи та механізми захисту	HTTPS протокол	Створення зашифрованого з'єднання між клієнтом та сервером.	Захист каналів зв’язку (“в процесі переміщення”).
	Алгоритми шифрування (AES, KALINA)	Шифрування безпосередньо повідомлень (data-at-rest) на серверах платформ.	Захист даних.
	Наскрізне шифрування (E2EE)	Доступ до змісту повідомлень лише відправнику та одержувачу.	Захист комунікацій між користувачами.

Продовження табл. 1.5

Підхід	Принцип	Характеристика	Мета
Методи та механізми захисту	Багатофакторна автентифікація (MFA, 2FA)	Вимагає підтвердження особистості за допомогою двох різних факторів.	Захист доступу до облікових записів.
	Розмежування доступу	Інструменти для гнучкого налаштування конфіденційності (визначення аудиторії для публікацій).	Контроль обсягу інформації, доступної стороннім.
	Управління цифровим слідом	Періодичний аудит власної онлайн-присутності.	Зменшення обсягу даних, доступних для зловмисників.
	Базові правила кібербезпеки	Своєчасне оновлення ПЗ, використання антивірусних програм, обережність при завантаженні файлів.	Забезпечення безпеки пристрою.
Ризик-орієнтований (у контексті використання соцмереж)	Фішинг	Маніпулювання психологією користувачів.	Порушення Конфіденційності через людський фактор.

Висновки до розділу 1

На основі проведеного дослідження встановлено, що в умовах сучасних соціальних комунікацій персональна інформація перетворилась із статичного запису в реєстрі, альбомі, документі на динамічний, багатошаровий актив. Його формування в соціальних мережах відбувається через три ключові категорії; Explicitly Provided Data (Явно надані користувачем реєстраційні та контентні дані), Observed Behavioral Data (Поведінкові дані, зібрані шляхом спостереження за

активністю, кліками та метаданими) та Inferred or Predicted Data (Прогнозовані дані), які генеруються самою платформою за допомогою алгоритмів машинного навчання, створюючи детального «цифрового двійника» користувача.

У цьому контексті, Генеративний Інтелект,з одного боку, є джерелом нових загроз; технології, як-от Великі Мовні Моделі (LLM) та Deepfake, використовуються зловмисниками для створення фішингових атак підвищеної якості, автоматизації генерації шкідливого програмного забезпечення та маніпулювання даними. Існує також високий ризик непрямих атак та ненавмисного витоку конфіденційної інформації через використання загальнодоступних моделей із закритим кодом.

З іншого боку, генеративний виступає інструментом захисту, дозволяючи реалізувати концепцію «Захист через обфускацію» (Security by Obfuscation). Це досягається через специфічні механізми; Image Cloaking (використання GAN для внесення змагальних збурень, що отруюють бази даних розпізнавання облич), обфускацію стилю (використання LLM для перефразування текстів і протидії стилометрії) та генерацію синтетичного шуму

Однак, ефективне застосування цих захисних інструментів залежить від надійної правової бази.

РОЗДІЛ 2
АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ЗАХИСТУ ІНФОРМАЦІЇ В СОЦІАЛЬНИХ МЕРЕЖАХ

2.1 Огляд існуючих методів та засобів захисту інформації в соціальних мережах

Аналізуючи архітектуру сучасних соціальних платформ, доцільно виокремити чотири ключові категорії (рис 2.1):

– платформи - універсали
– професійні мережі
– медіа - хостинги
– месенджери.

Ці платформи оперують мільярдами одиниць інформації, в деяких випадках спільно, що приваблює зловмисників. Як наслідок, вектори атак розвинулись від брутфорсу інфраструктури до соціальної інженерії[20].

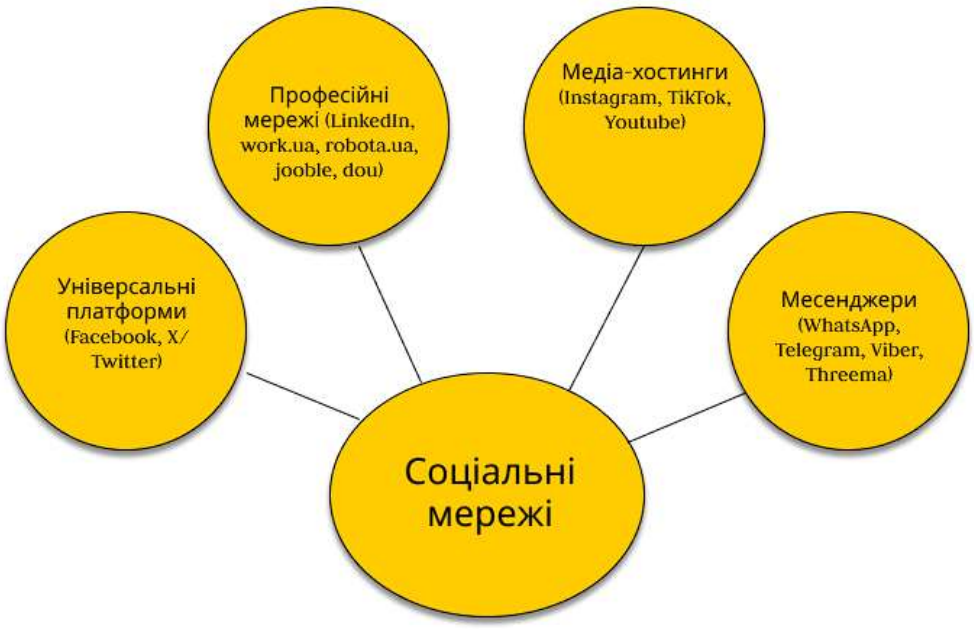

Рис 2.1 Соціальні мережі за функціональним призначенням

Кожна з цих платформ є одночасно джерелом і об'єктом загроз. Існуючі загрози діють за допомогою складного профілювання, зіставлення цифрових відбитків пальців та порівняння фрагментів профілів. Більше того, алгоритми зіставляють фрагменти публічних даних із зовнішніми базами даних, що призводить до деанонімізації з метою ідентифікації анонімних осіб. Загрози для персональних даних через генеративний штучний інтелект є нетиповим ризиком, оскільки останній не стосується виявлення логічних помилок у коді або слабких паролів. Генеративні моделі LLM і GAN створюють ризики для довіри, автентичності та конфіденційності інформації[21-22].

1. Експлуатація персонального стилю та ідентичності

- Сталі методи захисту не враховують стиль мовлення, через що генеративний інтелект може використати як вектор атаки.

- ГІ здатен аналізувати лексику, синтаксис та пунктуацію користувача, через що з’являється загроза копіювання стилістики мовлення.

- Використання GAN - моделей для генерації аудіо, відео та зображень, голос або почерк користувача створює загрозу підміни ідентичності.

2. Створення прогнозного продукту

- ГІ використовується для автоматичного виведення чутливої інформації із раніше "безпечних" або анонімізованих наборів даних.

- Генеративні моделі можуть автоматично адаптувати фішингові повідомлення або дезінформацію під психологічний профіль конкретного користувача, використовуючи його профільну інформацію.

3. Технологічні вразливості генеративних моделей

- Зловмисник може використовувати спеціальні запити до LLM, щоб змусити модель видалити конкретні приватні дані або конфіденційні деталі, які були використані для її навчання.

- ГІ може генерувати спеціальні "шумові" дані, які є непомітними для людини, але здатні цілеспрямовано обдурити або зіпсувати іншу модель ШІ (наприклад, модель розпізнавання обличчя чи модель модерації контенту).

Юридичні механізми захисту персональних даних у соціальних мережах формуються внутрішніми політиками приватності корпораціями. Саме корпоративні політики визначають, який обсяг даних збирається, з якою метою він обробляється, в яких ситуаціях надається третім сторонам й у яких межах користувач може контролювати їх використання. Незважаючи на формальну відповідність вимогам GDPR у Європейському Союзі та Закону України «Про захист персональних даних», фактичні практики обробки даних суттєво відрізняються між платформами та відображають бізнес-моделі відповідних компаній.

Політика конфіденційності Facebook та Instagram передбачає міжсервісне об’єднання даних і створення консолідованих профілів, які включають поведінкові характеристики, активність користувача, дані з пристроїв, місцеперебування, історію взаємодій, а також інформацію, отриману від зовнішніх партнерів. Юридично це оформлено як “законний інтерес платформи” та “покращення роботи сервісу”, однак на практиці така модель створює n - ступінь прозорості користувача для корпоративних та маркетингових алгоритмів[23]. Політика також надає Meta право обробляти дані для прогнозних цілей, що розширює обсяг inferred data і збільшує ризик надмірного профілювання. Хоча платформа формально надає механізми контролю за приватністю, їх функціональність є фрагментованою та не забезпечує повноцінного обмеження використання поведінкових атрибутів [9].

Зі свого боку, Telegram заявляє про мінімалістичний підхід до збору персональних даних, рекламуючи відсутність реклами та обмежену взаємодію з урядовими органами. Дійсно, згідно з політикою, сервіс зберігає лише номер телефону, IP-адресу та метадані, необхідні для роботи платформи. Telegram

офіційно залишає за собою право розкривати ці дані за рішенням суду, а відсутність налаштування наскрізного шифрування за замовчуванням - воно активується лише в «секретних чатах» - створює значний ризик доступу до комунікацій за допомогою засобів, на які користувач погодився, навіть не підозрюючи про небезпеку. Така правова структура забезпечує декларативну конфіденційність, але на практиці надає платформі широкі можливості для обробки метаданих, які є дуже чутливими в контексті ідентифікації користувачів[24].

Політика Linkdin, будучи професійно орієнтованою, включає велику кількість інформації про профіль у порівнянні з іншими соціальними мережами. Сервіс обробляє інформацію про освіту, досвід, місце роботи, корпоративні зв'язки, рекомендації, а також дані про поведінкову активність, що дозволяє йому прогнозувати зміни в кар'єрі, оцінювати професійні компетенції та створювати автоматизовані рекомендації. Незважаючи на відповідність GDPR, характер даних, що збираються на LinkedIn, є дуже чутливим, а їх обробка створює основу для цільового фішингу, корпоративного шпигунства та атак на відділи кадрів[25].

Користувачеві формально надається можливість обмежити видимість деяких даних, однак інші атрибути - зокрема інформація про професійну діяльність - залишаються публічними за природою функціоналу платформи.

Політика X включає положення про можливість збору біометричних атрибутів, зображень, метаданих пристрою, історії працевлаштування та освіти з метою «ідентифікації та безпеки». Формулювання політики дозволяє сервісу збирати дані в значно ширшому обсязі, ніж необхідно для функціонування платформи, а також агрегувати їх з джерел третіх сторін. Окремою юридичною вразливістю є дозволений скрепінг публічних профілів, що дає зовнішнім акторам можливість масово збирати персональні дані без фактичного порушення умов користування. Через це X часто виступає джерелом масових баз, які згодом

використовуються для фішингу, деанонімізації та моделювання соціальних графів[26].

Надалі було проведено порівняльний аналіз політик безпеки персональної інформації вищезазначених мереж (табл. 2.1).

Таблиця 2.1

Аналіз політик безпеки персональних даних популярних соціальних мереж

Платформа	**Основний фокус**	**Ключові розбіжності та уразливості**
Facebook / Instagram	Політика чітко дозволяє об'єднання даних про користувача, зібраних у всіх продуктах Meta (Facebook, Instagram, WhatsApp, Messenger) та від зовнішніх партнерів. Це створює максимально детальний, єдиний профіль.	Надмірний збір Observed Data (поведінкові дані) для прогностичних цілей. Наскрізне шифрування є непослідовним.
X	У нових редакціях політики X прямо зазначено можливість збору біометричних даних та історії працевлаштування/освіти з метою безпеки або ідентифікації.	Вразливість до скрепінгу (Scraping) - масового збору публічних даних, що використовується для деанонімізації та створення зовнішніх баз.
LinkedIn	Політика дозволяє використання даних для рекомендацій та аналізу ринку праці. Дані можуть використовуватися для прогнозування кар'єрного зростання або ймовірності зміни роботи (що є формою Inferred Data).	Висока чутливість профільної інформації (посада, компанія) робить користувачів ідеальною мішенню для цільового фішингу та корпоративного шпигунства.
Telegram	Наголос на відсутності реклами та використанні наскрізного шифрування (End-to-End). Заявлена мінімальна співпраця з урядами, хоча політикою передбачена можливість розкриття IP-адрес та номерів телефонів за рішенням суду.	Шифрування не є типовим (необхідна активація "Секретного чату"). Звичайні чати та, що важливіше, метадані є вразливими.

Масштабні інциденти витоку даних у соціальних мережах є доказом вразливості систем зберігання та обробки персональної інформації компаній, через що виникає необхідність перегляду систем безпеки. Інцидент з Facebook торкнувся понад 87 мільйонів користувачів через використання API соціальною мережою сторонніх додатків.

Виявлено, що ці дані застосовувалися для цільової політичної маніпуляції, що створило загрозу через експлуатацію персональних даних. Попри резонанс навколо Cambridge Analytica вразливість платформи до збору даних зберігалася, в результаті якого інцидент повторився. У 2021 році стався витік понад 533 мільйонів записів контактної інформації; номерів телефонів, повних імен та ідентифікаторів профілів, в результаті чого інформація була використана для формування бази даних для Spear Phishing - атак. Паралельно з проблемами технічних вразливостей, інцидент з LinkedIn охопив понад 700 мільйонів записів, підкреслив можливість масового скрепінгу (Scraping) публічно доступної інформації. Надалі, виходячи із попередньо-отриманих даних, графічно зображено відношення соціальних мереж між безпекою та вразливістю (рис. 2.2).

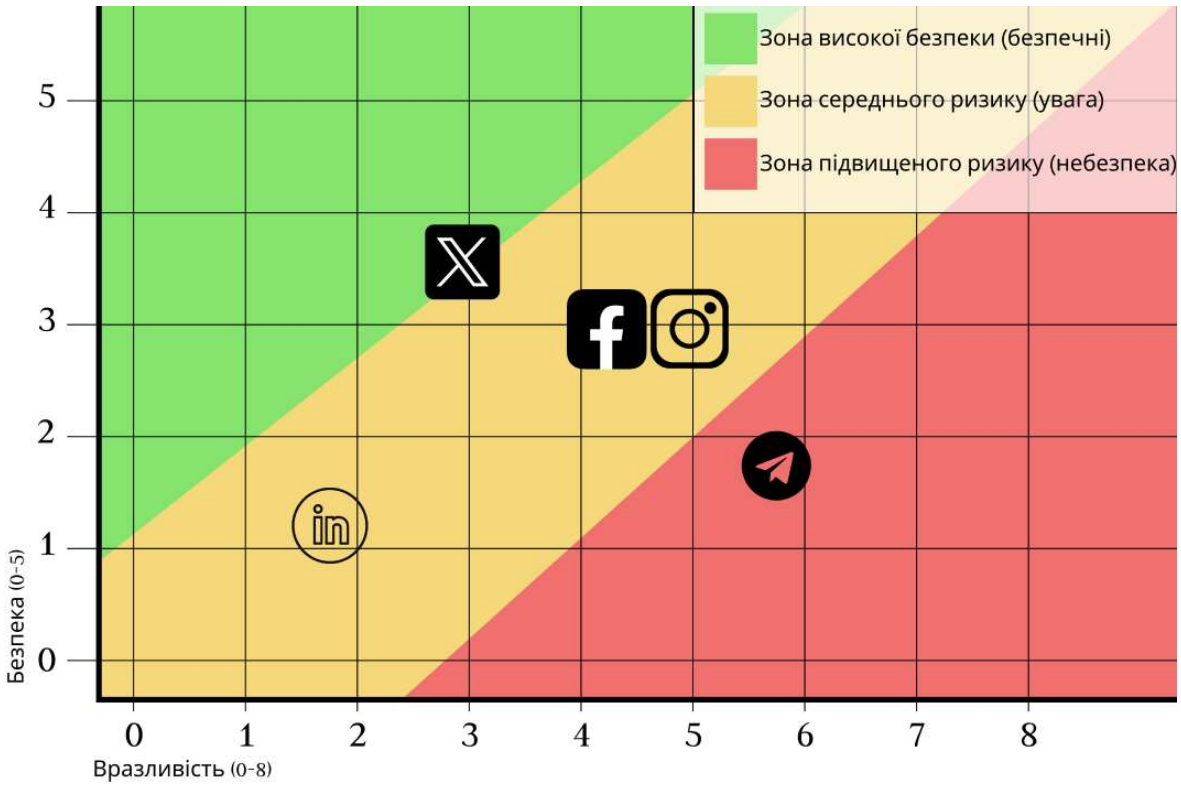

Рис. 2.2. Надійність соціальних мереж відносно параметрів безпека/вразливість

Вертикальна вісь (0 - 5) - рівень безпеки персональних даних (наявність шифрування, контроль доступу, обмеження збору даних).

Горизонтальна вісь (0 - 8) - рівень вразливості (масштаб збору даних, ризики скрепінгу, фішингу, витоку метаданих).

Оцінювання виконано експертно-аналітичним методом на основі аналізу політик безпеки та задекларованих практик обробки персональних даних (дані табл. 2.1).

Отже, Telegram характеризується високим рівнем вразливості та низьким рівнем загальної безпеки, що зумовлено агрегацією великих обсягів користувацьких даних, відкритістю публічних каналів і непослідовним застосуванням наскрізного шифрування.

X демонструє відносно найкращу позицію, однак залишається схильним до масового скрепінгу даних, аналізу відкритих профілів і метаданих.

LinkedIn перебуває у зоні середнього ризику: платформа забезпечує базові механізми захисту, проте висока чутливість професійних та ідентифікаційних даних користувачів значно підвищує ймовірність цільових атак, зокрема spear-phishing і BEC.

Продукти Meta (Instagram\Facebook) займає дещо кращий баланс між безпекою та вразливістю відносно найкращу позицію завдяки використанню криптографічних механізмів і обмеженню доступу до змісту повідомлень. Водночас вразливості метаданих і модель централізованої інфраструктури не дозволяють вважати платформу повністю безпечною.

2.2 Захист від кіберзагроз у соціальних мережах

З огляду на вразливі місця, вже визначені в попередньому розділі, та розвиток генеративної інтелектуальної системи, не можна переоцінити необхідність оцінки ефективності існуючих інструментів захисту. Поточна ситуація щодо захисту персональних даних на платформах соціальних медіа базується на інтеграції технологічних інструментів платформи, законодавства та налаштувань, ініційованих користувачами. Стратегії захисту включають інструменти, що залежать від технологій, законодавства та користувачів.

Ефективність захисту персональних даних у соціальних мережах досягається завдяки координації алгоритмічних, організаційних та поведінкових підходів. Цей підхід принципово відрізняється від загальноприйнятої парадигми безпеки, яка зосереджується на захисті інфраструктури від внутрішніх та зовнішніх загроз. Сучасні системи, з іншого боку, працюють на потоках даних, де персональні дані розглядаються як ресурс і потенційна загроза. Отже, системи захисту намагаються зменшити ризики, пов'язані з взаємодією між користувачем і платформою, обчислювальними процесами та діями третіх сторін, які становлять потенційну загрозу, включаючи злочинців[27].

Технічний аспект сучасних безпечних систем включає наскрізне шифрування, багатофакторну автентифікацію, контроль доступу до API та обмеження на обробку метаданих. Проте ефективність роботи цих систем значно варіюється залежно від політики надання послуг та економічної моделі, якої вони дотримуються. Послуги, які орієнтовані на отримання доходу від реклами, використовують складний аналіз поведінки, що зумовлює необхідність захисту даних. Наприклад, послуги, які надають пріоритет конфіденційності, встановлюють обмеження на обробку атрибутів, але слабким місцем є дані, що містяться на рівні метаданих, які є їхнім основним інструментом для профілювання[28].

Захист також доповнюється правовими рамками, які встановлюють параметри прийнятного збору та використання даних, регулюючи відносини між платформою, державою та користувачем. Юридичні вимоги щодо прозорості обробки даних, строків зберігання та можливості відкликати згоду на обробку даних, відповідно до вимог GDPR, були введені правовими системами України. Проте в межах параметрів правового контролю ефективність політики конфіденційності також не є послідовною, оскільки платформи часто рекламують широкі можливості свободи для користувача, а служби управління конфіденційністю часто є розрізненими або менш ефективними (табл. 2.2).

Подібна суперечність створює так звану «асиметрію знань», коли користувач не здатен повністю оцінити обсяг даних, які він фактично передає системі[28-30].

Поєднуючи технічні, юридичні та поведінкові механізми формується змішана модель безпеки, у межах якої жоден із компонентів не здатен забезпечити повноцінний захист автономно (рис. 2.3). Технічні засоби мінімізують ризики несанкціонованого доступу, юридичні обмеження забезпечують прозорість та підзвітність процесів обробки даних, а індивідуальні дії користувача зменшують ймовірність маніпуляцій та вторгнення через соціальну інженерію. Одночасно з цим генеративний штучний інтелект суттєво ускладнює реалізацію традиційних

механізмів захисту, адже створює нові типи атак, що поєднують технологічні й поведінкові вектори, розмиваючи межу між технічною експлуатацією та маніпулятивним впливом[28].

Таблиця 2.2

Рівень ефективності методів захисту персональної інформації в соцмережах

Методи захисту	**Рівень ефективності**	**Пояснення**
Шифрування даних (End-to-End)	Високий	Забезпечує захист від стороннього доступу до даних під час передачі та зберігання, але ефективність залежить від правильності впровадження.
Багатофакторна автентифікація (MFA)	Середній/Високий	Знижує ризик компрометації акаунта, але користувачі часто не активують MFA або використовують слабкі методи підтвердження.
Контроль доступу до API та обмеження метаданих	Середній	Зменшує ризики витоку даних через сторонні додатки, проте платформи з рекламною моделлю часто залишають відкриті точки збору даних.
Комплаєнс (GDPR, українські закони)	Середній	Забезпечують прозорість і підзвітність обробки даних, але на практиці дотримання обмежене.
дії користувача (налаштування приватності, обережність)	Низький/Середній	Ефективність сильно залежить від обізнаності та дисципліни користувача; помилки або недооцінка ризиків підвищують вразливість.
Поведінковий аналіз і профілювання платформами	Низький	Використання платформами для збору поведінкових даних створює додаткові точки ризику, важко контролюється користувачем.
Комплексний модель захисту (технічні + юридичні + поведінкові механізми)	Високий	Забезпечує комплексний підхід, проте жоден компонент сам по собі не гарантує повний захист.

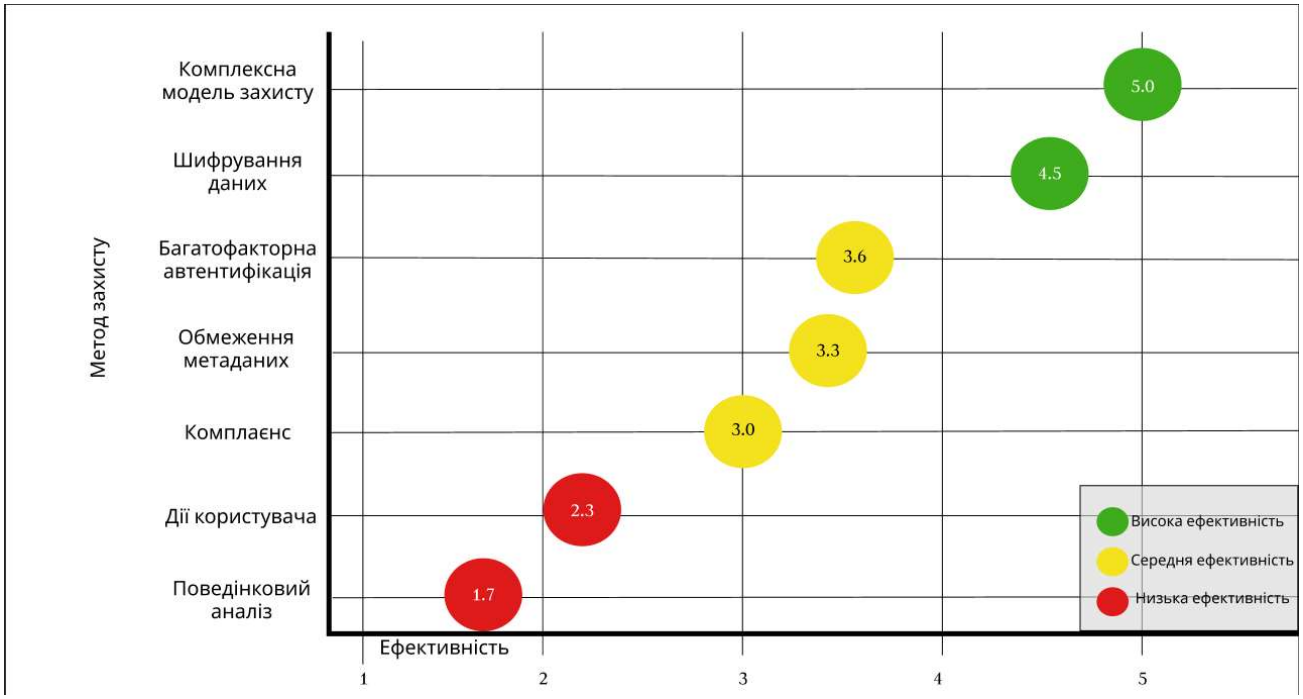

Рис. 2.3 Оцінка ефективності методів захисту

Горизонтальна вісь (0 - 5) - рівень ефективності механізмів захисту.

Оцінювання виконано експертно-аналітичним методом на основі аналізу аналітичних та галузевих звітів[22-23].

Отже, поведінковий аналіз та дії користувача характеризуються найнижчим рівнем ефективності (1.7–2.3 бали), що підтверджує критичну вразливість людського фактору та недостатність реактивних методів як єдиної лінії оборони.

Комплаєнс, обмеження метаданих та багатофакторна автентифікація перебувають у зоні середньої ефективності (3.0–3.6 бала). Ці методи створюють необхідні бар'єри та забезпечують базову гігієну безпеки, проте їх ізольоване використання залишає систему вразливою до складних цільових атак.

Шифрування даних демонструє високу ефективність (4.5 бала), забезпечуючи надійний криптографічний захист інформації навіть у разі перехоплення.

Комплексна модель захисту займає найкращу позицію з максимальною оцінкою (5.0 балів), доводячи, що лише синергія технічних, криптографічних та організаційних заходів дозволяє досягти високого рівня стійкості до загроз та мінімізувати ризики компрометації наскільки це можливо.

2.3 Виявлення прогалин у захисті контенту користувачів соцмереж

Технічна архітектура безпеки сучасних соціальних мереж формується на основі комплексу рішень, які мають на меті захист персональної інформації та мінімізацію ризиків несанкціонованого доступу до даних користувачів. Незважаючи на ефективність E2EE у захисті самого змісту комунікацій, ця технологія не охоплює IP-адреси, час комунікацій, список контактів, характеристики пристроїв і геолокаційні дані[31]. Саме ці метадані стають джерелом інформації для побудови поведінкових профілів користувачів, що дозволяє зловмисникам прогнозувати активність та взаємодії, розробляти таргетовані атаки та маніпуляції без доступу до зашифрованого контенту. Додатково виникає ризик сторонніх каналів атаки, коли дані аналізуються на клієнтському пристрої до шифрування або після дешифрування, що створює потенційні можливості для компрометації[31-33].

Багатофакторна автентифікація (MFA) впроваджується для підвищення рівня безпеки облікових записів, додаючи додатковий етап перевірки через біометрію, апаратні токени або одноразові паролі. Проте ефективність MFA знижується під впливом генеративних технологій, які дозволяють створювати синтетичні біометричні дані, включаючи голосові зразки та зображення обличчя, що використовуються для обходу систем перевірки. Такі методи застосовуються для проведення цільових атак, наприклад, шахрайства від імені керівників організацій («CEO Fraud») або соціальної інженерії нового покоління[34].

Спроби платформ обмежити доступ сторонніх додатків через API та регулювати обробку метаданих часто виявляються неефективними, оскільки генеративний штучний інтелект здатний агрегувати дані з різних відкритих джерел та формувати детальні профілі користувачів. Цей процес дозволяє створювати великі масиви інформації, які застосовуються для таргетованого фішингу, деанонімізації та маніпуляцій із соціальними графами користувачів. На практиці це забезпечує функціонал, аналогічний масштабним скандалам на кшталт Cambridge Analytica, проте з більшою точністю та швидкістю обробки даних. [34-38]

Для протидії автоматизованим атакам соціальні мережі застосовують системи виявлення аномалій, фільтри контенту, CAPTCHA та механізми обмеження частоти запитів (rate-limiting). Однак сучасні генеративні моделі дозволяють обходити ці захисні механізми. Зокрема, алгоритми створюють змагальні приклади (adversarial examples), які викривляють дані у спосіб, що залишається непомітним для людини, але блокують автоматичну класифікацію та модерацію контенту. Також генеративні алгоритми здатні імітувати поведінкові патерни користувачів, такі як швидкість набору тексту, рух курсору або типові патерни взаємодії з інтерфейсом, що дозволяє автоматично проходити перевірку CAPTCHA та обходити бот-фільтри[35-38].

Методи шифрування, MFA та обмеження доступу до API надають певний рівень захисту, проте вони виявляються недостатніми у протидії адаптивним атакам з використанням генеративного ШІ. Наприклад, навіть при реалізації E2EE та складної MFA, синтетичні дані та поведінкові профілі дозволяють створювати персоналізовані фішингові повідомлення, які важко розпізнати за класичними ознаками. З огляду на це, сучасні технічні механізми виявляються ефективними лише частково та у поєднанні з іншими шарами безпеки, такими як юридичні регламенти та поведінкові практики користувачів [38].

Системи виявлення аномалій та фільтри контенту, які орієнтовані на традиційні атаки, не можуть повністю врахувати складність і швидкість генеративних алгоритмів. Використання посиленого навчання та адаптивних моделей дозволяє ШІ створювати контент, що обходить алгоритмічні обмеження та автоматично підлаштовується під поведінку користувача. Це означає, що існуючі технічні засоби, хоча і знижують ймовірність стандартних загроз, виявляються уразливими для інноваційних атак, де поєднуються технологічні та поведінкові вектори [39]. Порівняльна характеристика інструментів безпеки цифрового профілю в соцмережах подана в таблиці 2.3.

Внаслідок цього з’являється потреба у розробці інтегрованих, гібридних моделей безпеки, які поєднують алгоритмічні рішення, організаційні заходи та поведінкові практики користувачів. Такий підхід дозволяє мінімізувати ризики витоку інформації, маніпуляцій із контентом та профілювання, проте навіть у цьому випадку неможливо гарантувати повний захист даних, оскільки генеративні технології постійно еволюціонують та адаптуються до існуючих бар’єрів.

Таблиця 2.3

Слабкі місця інструментів безпеки цифрового профілю в соцмережах проти генеративного ШІ

Технічний механізм	**Призначення**	**Слабкі місця проти генеративного ШІ**	**Рівень ефективності**
End-to-End Encryption (E2EE)	Захист змісту повідомлень між відправником та отримувачем	Не захищає метадані, аналіз яких дозволяє побудувати поведінкові профілі користувачів	Частково
Багатофакторна автентифікація (MFA)	Додає додатковий рівень перевірки через біометрію, токени або одноразові паролі	Генеративні моделі можуть створювати синтетичні біометричні дані, голосові зразки та зображення обличчя, обходячи MFA;	Частково
Обмеження доступу через API / контроль метаданих	Захист від сторонніх додатків та агрегування даних	Генеративний ШІ може збирати та агрегувати дані з різних відкритих джерел, формуючи детальні профілі для таргетованого фішингу та деанонімізації	Низький
Системи виявлення аномалій, фільтри контенту, CAPTCHA, rate-limiting	Протидія автоматизованим атакам та ботам	Генеративні моделі обходять CAPTCHA та бот-фільтри, створюють adversarial examples та імітують поведінкові патерни користувачів	Частково
Інтегровані гібридні моделі безпеки	Поєднання алгоритмічних рішень, організаційних заходів та поведінкових практик	Все ще не гарантують повного захисту через постійний розвиток генеративного ШІ	Середній

Висновки до розділу 2

Дослідження архітектури сучасних платформ соціальних мереж, включаючи універсальні платформи, професійні мережі, хостинг-сервіси та месенджери, підтверджує вразливість цих платформ. Методи атак перейшли від атак грубої сили на інфраструктуру до профілювання та розвинутих методів соціальної інженерії. Ця складність значно посилюється появою генеративного штучного інтелекту (ГШІ), де виникають ризики збільшення частоти атак за рахунок швидкості та підвищення якості методів атак. Генеративний інтелект використовує індивідуальні мовленнєві патерни, створює підроблені біометричні дані для підміни особи (особливо з використанням підходів генеративно-змагальної мережі) та використовується для створення прогностичних продуктів шляхом автоматизованого вилучення конфіденційних даних з анонімізованих наборів даних, а також для персоналізації фішингових повідомлень за допомогою психологічного профілювання користувача. Такі ризики значно посилюються власними вразливостями генеративних моделей ШІ, які потенційно використовуються зловмисником для вилучення конфіденційних даних або для створення «шумових» даних, які навмисно вводять в оману або обманюють інші системи ШІ. Правові бази, хоча й гармонізовані на рівні GDPR та Закону України «Про захист персональних даних», є складними та створюють асиметрію. Правові положення, починаючи від корпоративних політик, що підтримують агрегацію даних між сервісами (Meta), збору біометричних атрибутів (X) або проголошеної конфіденційності Telegram, скомпрометовані на рівні вразливостей метаданих та нездатності використовувати стандартне наскрізне шифрування. Табличний аналіз показує, що поточні технологічні захисні заходи, такі як використання наскрізного шифрування (E2EE) та багатофакторної автентифікації (MFA), забезпечують недостатню стійкість до адаптивних атак генеративної соціальної інженерії та інфільтрації, оскільки E2EE

не може забезпечити стійкість до профілювання метаданих, MFA є небезпечною при використанні синтетичних біометричних даних, а системи виявлення аномалій потенційно можуть бути обдурені за допомогою прикладів зловмисників. Це, серед інших міркувань, породжує необхідність впровадження інтегрованої гібридної моделі безпеки, яка скасовує розрив між технологічними та правовими нормами, а також практиками поведінки користувачів. Однак навіть ця інтегрована стратегія не забезпечує комплексної безпеки, яка захищає користувачів, оскільки підхід продовжує розвиватися та ефективно адаптуватися за допомогою генеративного інтелекту. Вторинна прогалина, серед інших міркувань, полягає в тому, що сучасні механізми безпеки не можуть належним чином захистити нову загрозу гібридного типу, яка вже поєднує технологічну експлуатацію за допомогою генеративного інтелекту зі зловживанням впливом, що вимагає впровадження нових адаптивних алгоритмічних підходів.

РОЗДІЛ 3
ЗАХИСТ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ КОРИСТУВАЧА СОЦМЕРЕЖ З ВИКОРИСТАННЯМ ЗАСОБІВ ГЕНЕРАТИВНОГО ІНТЕЛЕКТУ (НА ОСНОВІ МОДЕЛІ DP-GAN-HD)

3.1 Архітектура моделі

Інтеграція інструментів генеративного інтелекту в інфраструктуру соціальних мереж надає можливість щодо підвищення рівня конфіденційності, завдяки якій зменшується вірогідність несанкціонованого доступу до персональних даних користувачів. Водночас зберігаються основні функції моніторингу поведінки, керування контенту та тестування алгоритмів зі сторони розробників. Для таких цілей необхідно застосовувати модель, здатну поєднати високу якість генерації фальсифікованого контенту, доведені криптографічні гарантії приватності та стійкість до атак на реконструкцію інформації. Модель DP-GAN-HD є основним кандидатом для впровадження в архітектуру соціальних мереж, оскільки об’єднує архітектуру змагального інтелекту з механізмом генератора диференційної приватності та модулем високопіанової конфіденційності[40].

В DP-GAN-HD механізми приватності інтегровані безпосередньо в процес навчання моделі. Шум, який додається до градієнтів дискримінатора, унеможливлює витік чутливої інформації через параметри мережі, в той самий час паралельно алгоритми обрізання градієнтів гарантує що жоден окремий запис користувача не вплине на кінцеву конфігурацію моделі більше, ніж дозволено обраним рівнем ε-диференційної приватності. Завдяки цьому досягаються консенсус щодо ризику неправильного навчання моделі в соціальних мережах, внаслідок якої буде розкрито реальну інформацію профілю, фотографії чи поведінкові шаблони користувача.

Якість генерації DP-GAN-HD забезпечується завдяки використанню мультигенераторної структури генерації. Різні генератори працюють над окремими масивами даних, паралельно зі створенням даних дискримінатор та генератор змагаються між собою для встановлення рівня валідності матеріалу через що різноманітність синтетичних записів зростає та дає можливість моделювати складні, багатовимірні та мультимодальні структури. Це робить модель придатною для синтезу соціальних графів, текстових повідомлень, мультимедійних характеристик та поведінкових патернів [40-43].

Такі синтетичні дані можуть використовуватися для навчання алгоритмів рекомендацій, аналізу динаміки мережі, виявлення аномалій, антиспам - систем і розробки нових функцій платформи - без необхідності залучати реальні персональні дані користувачів.

Порівняння зі схожими моделями вказує, що DP-GAN-HD здатен забезпечити найбільш збалансований та практично придатний підхід серед схожих за принципом дії моделей. Архітектура, побудована на механізмі PATE демонструє забезпечення досить високе забезпечення теоретичної приватності але не здатен до стабільного генеративного процесу та масштабування. Диференційно-приватні варіаційні автоенкодери забезпечують хорошу реконструкцію табличних даних, але не здатні відтворювати складну структуру соціальних мереж, яка потребує багатовимірного моделювання. Дифузійні моделі здатні до генерації якісного контенту зображень та потребують великого об’єму ресурсів та енергії під час інтеграції механізмів приватності та втрачають стабільність. Інструменти створення синтетичних даних SDV застосовують диференційну приватність модульно, через що наскрізний захист приватних даних спрацьовує в n-частині випадків генерації. Аналогічно, моделі на базі GraphGAN намагаються моделювати структуру графів, однак без вбудованих механізмів диференційної приватності

можуть ненавмисно відтворювати унікальні підграфи, що створює високі ризики витоку [42-45].

Уся інформація, наведена в таблиці 3.1 демонструє, що DP-GAN-HD забезпечує оптимальне співвідношення між якістю, стабільністю і можливістю впровадження у реальних умовах. Вона створює передумови для переходу соціальних мереж до архітектури приватності за замовчуванням, у якій первинні дані використовуються мінімально, а більшість обчислювальних процесів відбувається на синтетичних, повністю деперсоналізованих вибірках.

Таблиця 3.1

Порівняння моделей ШІ: безпековий огляд

Модель	**Рівень приватності**	**Якість синтетичних даних**	**Стійкість до атак**	**Масштабованість**	**Придатність соцмереж**
DP-GAN-HD	Високий, інтегрована ε-DP	Висока, мультимодальна	Висока, захист від прямих і непрямих атак	Висока	Дуже висока
PATE-GAN	Дуже високий	Середня	Висока	Низька	Обмежена
DP-VAE	Високий	Низька-середня	Середня	Середня	Низька
DP-Diffusion	Середній-високий	Дуже висока	Середня	Низька, висока вартість обчислень	Середня
GraphGAN + DP	Середній	Середня	Низька-середня	Середня	Середня
SDV + DP	Середній	Висока або середня	Середня	Висока	Середня

Запропонована архітектура DP-GAN-HD залучає механізми диференційної приватності (рис.3.1) зі змагальними алгоритмами (GA) для вдосконалення генеративної складової[40]. Запропонована система містить дискримінатор та сукупність генераторів. Блок дискримінатора відповідає за гарантування захисту приватних даних шляхом застосування диференційної приватності, тоді як група генераторів прагне створювати якісні штучні дані, котрі імітують розподіл справжніх даних.

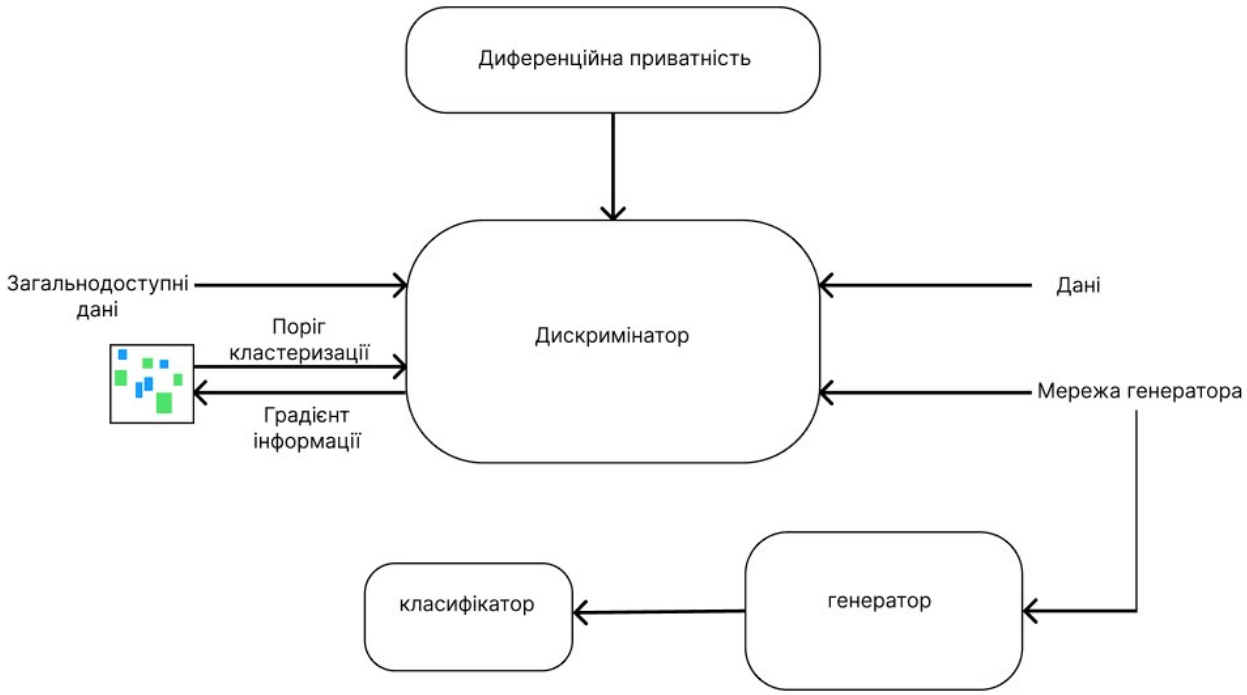

Рис 3.1 Концептуальна модель DP-GAN-HD

Робота моделі базується на класичному змагальному навчанні (GAN), де генератор намагається створити штучні дані, а завданням дискримінатора є розрізнення справжніх та згенерованих даних.

Для уникнення витоку конфіденційної інформації через градієнтні відомості дискримінатора, у процес навчання дискримінатора вбудовано механізм диференційної приватності. Цей механізм передбачає реалізацію двох основних кроків: обрізання градієнта для обмеження чутливості та введення гауссового шуму до обрізаного градієнта для формування диференційної приватності. Звичайне використання сталого порогу зрізу градієнта може спричинити надмірне зменшення градієнта або недостатній нагляд за викидами, негативно позначаючись на збіжності моделі. Для подолання цієї проблеми у моделі DP-GAN-HD впроваджено адаптивний спосіб відсікання градієнта, що базується на алгоритмі кластеризації DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Якість

навчання генератора обмежена через шумове збурення, що його вносить дискримінатор для забезпечення DP. Задля подолання цієї складності та підвищення якості створених даних запропоновано застосування кластеру генераторів. Однак застосування шуму DP має побічний ефект у вигляді погіршення якості навчання генератора.

Через додаткові випадкові коливання градієнтів генератор може повільніше пристосовуватися до розподілу реальних даних, що може викликати недостатньо точне відтворення статистичних закономірностей. Для подолання цієї проблеми у DP-GAN-HD використовується кластерний метод до генераторів. Замість одного генератора застосовується група генераторів, кожен з яких має окремі ваги та параметри. Кожен генератор у кластері навчається на приватних сигналах від дискримінатора, а на етапах відбору оцінюється його ефективність. Генератори з високими оцінками залишаються у кластері, тоді як менш влучні виключаються. Такий добір дозволяє зосередити обчислювальні ресурси на найбільш перспективних генераторах, збільшуючи стабільність навчання та якість синтетичних даних [42-46].

Подальше збільшення об’єму масиву згенерованих даних забезпечується через механізми кросовера та мутації параметрів обраних генераторів. На етапі кросовера вагові параметри батьківських генераторів комбінуються для генерації нових варіантів комбінацій. Мутації з додаванням гауссового шуму до параметрів ваг ще більше розширюють пошук, що особливо корисне для моделювання складних, гетерогенних даних соціальних мереж для аналізу текстових постів, структурних графів зв’язків та поведінкових патернів.

Також адаптивне відсікання градієнта у дискримінаторі може бути не зовсім ефективним у задачах, де дані мають неоднорідну структуру та різну чутливість. Застосування адаптивного методу на основі алгоритму DBSCAN дозволяє кластеризувати градієнти та визначати оптимальні пороги обрізання для кожного

кластера. Це зменшує ризик надмірного скорочення градієнта або недостатнього контролю над викидами, що, своєю чергою, покращує збіжність моделі та стабільність навчання.

На стадії відбору генератори з вищими оцінками зберігаються, а гірші відсікаються. На етапах кросовера та мутації параметри ваги обраних батьківських генераторів рекомбінуються та збурюються гауссовим шумом для забезпечення можливості дослідження ширшого простору параметрів.

Диференційна приватність забезпечується механізмом гауссового збурення шуму та відсіканням градієнта у дискримінаторі, через що модель DP-GAN-HD дотримується диференційних обмежень конфіденційності.

Генератор, отримуючи лише унікальний сигнал від дискримінатора, не може додатково посилити ризик витоку даних, навіть якщо подальше опрацювання створених синтетичних записів передбачає додаткові перетворення чи дослідження. Навіть після багаторазових ітерацій навчання та зведення параметрів у групі генераторів, модель підтримує дотримання граничних умов диференційної приватності. Генератор отримує лише диференційно приватний зворотний зв'язок від дискримінатора, не маючи прямого доступу до вихідних даних. Згідно з властивістю постобробки диференційної приватності, дані, створені генератором, не вносять додаткової втрати конфіденційності та не споживають додаткових [42-46].

3.2 Формалізація та формування навчального набору даних із профілю користувача соцмереж

Випадковим чином було відібрано 80 користувачів найбільш вразливої соціальної мережі Telegram (рис. 2.2). Для автоматизованого збору відомостей використовувався API-інтерфейс Telegram (MTProto) із застосуванням спеціалізованих клієнтських бібліотек Telethon та Pyrogram. Це дозволило

реалізувати асинхронний збір об'єктів профілю у форматі JSON із подальшою десеріалізацією для формування первинної структури навчального набору.

Для подальшого формування навчального матеріалу для моделі генеративного інтелекту було вилучено наступні категорії атрибутів:

- Текстовий контент: тексти постів, коментарі та описи профілів;
- Мультимедійні ознаки: метадані медіа-контенту (зображення, відео, аудіо);
- Динамічні показники: частота активності, середній інтервал між діями, реакції на контент та часові мітки;
- Системні метадані: геолокація (на рівні провайдера/регіону) та тип пристрою.

Процес збору супроводжувався первинною фільтрацією (data cleaning), що включала видалення прямих ідентифікаторів (номерів телефонів, посилань на особисті профілі) ще до етапу формалізації. Такий підхід забезпечує підхід до DP-GAN-HD лише семантично значущої інформації, позбавленої явних ознак персоналізації.

Оскільки серед цих типів інформації наявна висока частка чутливих відомостей, потрібно визначити, які саме атрибути можуть бути задіяні після приватизації, а які мусять бути відхилені або узагальнені до статистичних показників. Формальне сортування відомостей дає змогу виконати точне розмежування між інформацією, що може зашкодити приватності, та даними, які можна безпечно залучати в генеративному процесі.

Наступним етапом є формування уніфікованих вимог перетворення для кожного виду даних. Для письмових даних уживається повний цикл попереднього опрацювання; токенізація, лематизація, вилучення стоп-слів, усунення нечастих унікальних слів, які можуть бути ідентифікаторами, та векторизація за допомогою ембеддингів. Замість зберігання вихідних текстів формуються репрезентації

вищого рівня, що містять лише узагальнені семантичні ознаки без відтворення первинних висловлювань користувачів[47].

Дані про поведінку формалізуються засобами агрегування та статистичного вирівнювання, до яких додається гаусівський шум відповідно до параметрів диференційної приватності. Структурні відомості соціальних мереж замінюються графовими дескрипторами (ступінь вузла, коефіцієнт згрупування, центральність PageRank, спектральні показники графа), що відображають макробудову взаємодій без певних сполучень. Метадані усереднюються до рівня країни чи області, часові мітки змінюються на категорії, а дані про засоби збираються до типу платформи. Медіа-дані опрацьовуються через видобуток ознак та, у разі потреби, додавання шуму [47].

Після завершення перетворення кожен користувач або одиниця активності представляється у вигляді формалізованого приватного запису (record), що включає семантичні, поведінкові, структурні та метадані, але у суворо приватизованій формі. Записи стандартизуються до єдиної структури й нормалізуються для забезпечення коректної роботи дискримінатора в DP-GAN-HD.

Етичні принципи формування навчального набору наступні;

Прозорість та обґрунтованість збору даних – використання лише анонімізованих або узгоджених із користувачами даних.

Конфіденційність та приватизація – усі персональні ідентифікатори видаляються; застосовується диференційна приватність.

Недискримінаційність – виключення даних, що можуть спричинити дискримінацію певних груп.

Обмеження на використання чутливих категорій – дані про релігію, політичні переконання, сексуальну орієнтацію піддаються суворій анонімізації.

Безпечна агрегація та публікація – синтетичні або статистично узагальнені дані не розкривають приватну інформацію окремих користувачів.

Формування навчальної вибірки (табл. 3.2) завершується її розподілом між модулями DP-GAN-HD. Дискримінатор отримує приватні записи у складі стабільних батчів і навчається відрізняти реальні приватизовані структури від синтетичних. Генератори працюють із вибірками підмножин даних для моделювання різних модальностей, що підвищує різноманітність результатів[47-54].

Таблиця 3.2

Навчальний набір даних із профілю користувача соцмереж

Категорія даних	**Приклади**	**Формалізація**	**Приватизація (DP)**	**Результат**
Контентні дані	Тексти постів, коментарів, описи профілів	Токенізація, лематизація, видалення стоп-слів, усунення унікальних ідентифікаторів, векторизація	Векторизація без збереження тексту; семантичні ознаки без відтворення фраз	Високорівнева репрезентація тексту без конкретних персональних даних
Поведінкові дані	Частота активності, середній інтервал між діями, реакції на контент	Агрегування статистик; середня активність, медіана, дисперсія, часові проміжки	Додавання гаусового шуму за диференційною приватністю	Агреговані статистики поведінки користувача, приватні і без часових міток
Метадані	Геолокація, пристрій, часові мітки	Узагальнення геолокації до рівня країни/регіону, округлення або категоризація часу	Повна приватизація та агрегування	Статистично безпечні метадані, що не дозволяють реконструювати маршрути чи звички

Продовження табл. 3.2

Категорія даних	**Приклади**	**Формалізація**	**Приватизація (DP)**	**Результат**
Агреговані мультимедійні характеристики	Відео, зображення, аудіо	Витяг ознак; колірні гістограми, векторні ембеддинги	При необхідності додавання шуму	Узагальнені мультимедійні ознаки для генеративного навчання
Навчальний запис	Один користувач / одиниця активності	Об'єднання всіх оброблених типів даних у єдину структуру	Використання приватизованих даних для всіх компонентів	Приватний запис для подачі в DP-GAN-HD

Гістограма розподілу втрат генератора (рис. 3.2) демонструє частоту генерації контенту (батчі), з якою були досягнути певні значення втрат (loss) протягом усього процесу навчання. Вісь X - Низький loss означає, що генератор здатен обманути дискримінатор - його згенеровані дані схожі на реальні. Вісь Y - Кількість ітерацій, для яких генератор отримав певне значення loss. Порівняно з дискримінатором, генератор має більший розкид значень loss, що нормально для GAN; дискримінатор і генератор постійно «змагаються».

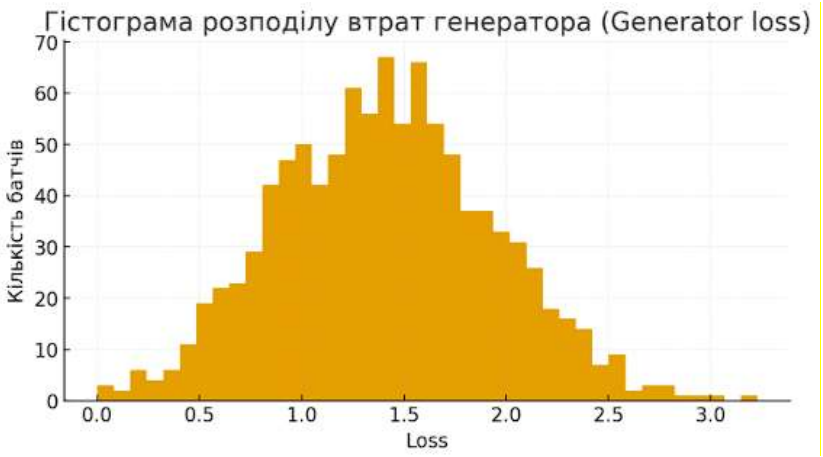

Рис 3.2 Гістограма втрат генератора

Гістограма розподілу втрат генератора дає змогу простежити, яким чином DP-шум впливає на стабільність оновлення ваг. Здебільшого розподіл генераторних втрат має ширший діапазон та більш виражений «хвіст» на великих значеннях. Це пов’язано з тим, що генератор отримує лише приватизований зворотний зв’язок, який уже містить додаткове шумове збурення. Наявність широкого варіативного поля втрат є типовою ознакою моделей, у яких дискримінатор навчається з використанням диференційної приватності, оскільки величина шуму безпосередньо впливає на інформативність градієнтів. Аналіз такої гістограми дозволяє визначити оптимальний баланс між рівнем приватності та якістю генерованих даних. Наприклад, занадто важкий хвіст розподілу може свідчити про необхідність зменшення параметра шуму або корекції параметрів DBSCAN-кластеризації для адаптивного відсікання градієнтів. Вісь X - Чим нижчий loss, тим краще дискримінатор відрізняє реальні дані від згенерованих.

Гістограма розподілу втрат дискримінатора (рис. 3.3) є оберненою залежністю з генератором, тобто він демонструє ефективність розрізнення справжніх та синтетичних даних. Вісь Y - скільки разів навчання дискримінатора отримувало певне значення loss.

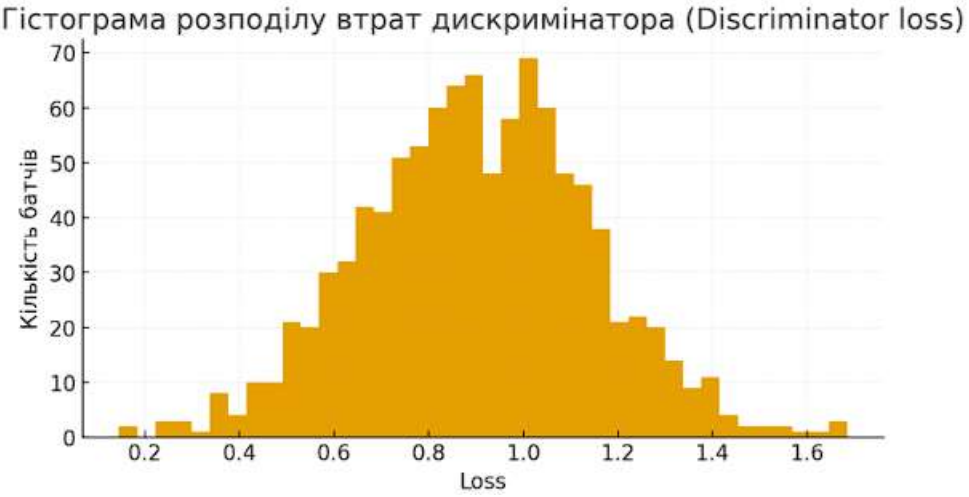

Рис. 3.3 - Гістограма втрат дискримінатора

Натомість гістограма втрат дискримінатора зазвичай демонструє компактніший розподіл, оскільки дискримінатор безпосередньо працює з приватизованими, але структурно стабільними даними. Низька варіативність втрат свідчить про те, що модель успішно навчається розрізняти синтетичні та реальні записи навіть за умов обмеженого доступу до інформації. Проте надмірно низькі значення втрат можуть вказувати на ризик непередбачуваного перенавчання, коли дискримінатор надто швидко адаптується до зашумлених патернів, що знижує ефективність змагального навчання та ускладнює процес роботи генератора.

Аби забезпечити безпечне використання сформованого навчального набору відомостей у практичних ситуаціях, впроваджуються особливі засоби для оприлюднення анонімізованих та синтетичних даних у соціальних мережах. Такі інструменти надають змогу[54-58];

- оприлюднювати відомості без загрози розкриття особистої інформації певних користувачів;
- керувати рівнем диференційної приватності (ε-параметр), що гарантує оборону від зворотної ідентифікації;
- автоматично вилучати потенційно чутливі ознаки та забезпечувати дотримання етичних норм;
- перетворювати дані на синтетичні чи статистично зібрані записи, які зберігають корисні відомості для аналітики та моделювання, проте не містять конкретних персональних даних.

Отже, поєднання технічного опрацювання, етичних засад та застосування засобів безпечного оприлюднення гарантує високоякісний, приватно-захищений навчальний набір, готовий до вправного застосування у генеративних моделях, як-от DP-GAN-HD, і водночас знижує небезпеки порушення конфіденційності користувачів.

3.3 Оцінка моделі та аналіз стійкості до загроз витоку даних користувача соцмережі

Надалі проведемо аналіз стійкості інформації з використанням DP-GAN-HD, вводимо наступні метрики [33; 52].

Таблиця 3.3

Метрики конфіденційності

Метрика	Інтерпретація	Рекомендоване порогове значення	Результат
Теоретична сила DP (ε)	Менше → сильніша приватність	0.1-1 - сильна приватність; 1-10 - збалансовано; >10 - слабка приватність (рекомендація для соцмереж; ε ≤ 1-2 для чутливих атрибутів)	1/N = 1/80 = 1.25, входить в рекомендований діапазон ≤
Додатковий параметр DP (δ)	Повинно бути малим	δ ≪ 1/N (N - розмір навчального матеріалу)	10^{-5} << 1.25 – Умова дотримана
Стійкість до атак (MIA AUROC / success rate)	Ближче до 0.5 - безпечніше	AUROC ≤ 0.6 вважається прийнятним; ≤0.55 - хороший рівень	Якщо Loss Discriminator = 0.69 то AUROC = 0.5. завдяки шуму DP рівень = 0.53 – 0.55
Міра успіху реконструкції error (MSE / similarity)	Більший MSE / нижча схожість → краще	Залежить від даних; наприклад SSIM ≤ 0.4 для зображень або normalized MSE > 0.2	$MSE\frac{1}{n} = \sum_{i=1}^{n}\left(x - x^{i}\right)^{2} = \text{MSE} > 0.2 - 0.3$
Параметр чутливості даних при зміні (Influence / sensitivity (L2))	Менше → краще	Ціль; sensitivity після clip ≤ заданого S (визначається під час DP-SGD)	Стале значення відсіку градієнтів DP-SGD = 1.0 – 1.5.
Практичний ризик (Re-identification rate (OSINT)	Нижче → краще	<1% бажано для публічної синтетики; для внутрішнього використання <5%	Отримані збурені дані відправляються в додаток пошуку облич. 1 збіг = +1.25%
Збереження корисності (Downstream utility ratio)	Ближче до 1 → краще	≥0.9 - відмінно; 0.7-0.9 - прийнятно; <0.7 - проблеми	Якість реальних даних/якість синтетичних даних. 0.75-0.82.

Для забезпечення стабільного функціонування DP-GAN-HD у виробничому середовищі соціальної мережі необхідно створити окремий контур моніторингу та аудиту, який дозволить своєчасно виявляти аномалії, деградацію моделі, дрейф даних та надмірне споживання ресурсів. Основна увага має бути приділена трьом напрямам: контролю якості згенерованих даних, контролю приватності та контролю обчислювального навантаження.

Для відстеження коректності роботи генераторів та уникнення зміщень у розподілах рекомендується впровадити регулярний Data Quality Monitoring, який включає перевірку стабільності семантичних ознак, ентропії текстів, частотності рідкісних патернів та індексів схожості між синтетичними та реальними даними.

По-друге, необхідно впровадити Privacy Budget Monitoring. Кожна операція, яка взаємодіє з реальними даними, повинна автоматично реєструвати обсяг витраченого ε-бюджету. Цей механізм має включати журналювання, сповіщення та блокування викликів у разі наближення до встановленого порогу.

Важливим компонентом також є контроль обчислювального завантаження генераторів. Навчання DP-GAN-HD потребує значних GPU-ресурсів, тому варто застосовувати динамічне масштабування на основі відстеження часу навчання, кількості одночасних запитів та частоти оновлення ваг. Оптимальною є гібридна модель, де постійно працюють базові GPU-екземпляри, а додаткові запускаються у пікові періоди за допомогою авто-скейлінгу контейнерів (наприклад, на Kubernetes). [57-60]. Окрему увагу слід приділити аудиту доступу до моделей та даних. Усі дії операторів, аналітиків та адміністраторів щодо навчання моделі, зміни параметрів приватності, вивантаження підмножин даних та експорт ваг повинні логуватися у незмінювані журнали (immutable audit logs). Такі журнали дозволяють розслідувати потенційні порушення, виявляти внутрішніх зловмисників та дотримуватися комплаєнс-вимог.

Впровадження перелічених підсистем (табл. 3.4) дозволяє забезпечити прогнозовану роботу DP-GAN-HD, кероване масштабування, відповідність вимогам приватності та мінімізує операційні ризики у виробничій інфраструктурі соціальних мереж.

Таблиця 3.4

Метрики моніторингу DP-GAN-HD та їх допустимі порогові значення

Метрика	**Опис**	**Спосіб вимірювання**	**Допустиме порогове значення**	**Фактичне значення**	**Інтерпретація порушення порогу**
Loss_Generator	Відображає якість роботи генератора та здатність моделі створювати реалістичні синтетичні дані.	Середнє значення втрати за епоху	≤ 2.0	1.65	Модель генерує некоректні або нестабільні дані; можливі проблеми з приватністю через погану узагальненість.
Loss_Discriminator	Показує, наскільки дискримінатор правильно розрізняє реальні та синтетичні дані.	Значення втрат під час навчання.	0.4-1.5	0.69	Дискримінатор або занадто сильний (ризик overfitting), або занадто слабкий (занадто легка генерація).
Gradient Penalty	Контролює стабільність тренування та забезпечує коректність роботи SN/GP-модуля.	Норма градієнта	0.5-1.0	0.82	Занадто великі значення - нестабільність; занадто малі - ризик деградації моделі.
Privacy Budget (ε)	Основний показник диференційної приватності, що регулює рівень шуму.	Обчислюється у процесі DP-обліку	$1 \leq \varepsilon \leq 8$	1.25	ε надто високий: слабкий захист, можливий витік даних; ε надто низький:

Продовження табл. 3.4

Метрика	**Опис**	**Спосіб вимірювання**	**Допустиме порогове значення**	**Фактичне значення**	**Інтерпретація порушення порогу**
δ (delta)	Ймовірність порушення механізму DP.	Обирається залежно від розміру датасету.	1e-5 - 1e-6	10^{-5}	Зростання δ свідчить про високу ймовірність компрометації приватності.
Mode Collapse Index (MCI)	Відображає наявність режимного колапсу у генераторі.	Варіація міжгенеративних зразків.	≤ 0.15	0.08	Дані стають одноманітними, ризик повторення патернів реальних користувачів.
Distribution Divergence (KL / JS)	Різниця між розподілами реальних та синтетичних даних.	Обчислення KL або JS дивергенції.	≤ 0.3	0.12	Значна відмінність між розподілами, втрата корисності або порушення якості синтетичного датасету.
Inference Attack Resistance Score (IARS)	Стійкість до атак повторної ідентифікації та membership inference.	Тестові атаки.	≥ 0.85	0.91	Низький показник означає ризик відновлення оригінальних записів.

Аналіз стійкості до загроз витоку даних показує, що DP-GAN-HD витримує як прямі, так і непрямі атаки. Прямі атаки на дискримінатор, які намагаються відновити первинні записи, є неефективними через наявність DP-шуму та агрегованої структури даних. Непрямі атаки, спрямовані на генератор, також не можуть призвести до витоку, оскільки генератор отримує лише приватизований сигнал і не має прямого доступу до початкового датасету. Постобробка синтетичних даних не збільшує ризик порушення приватності, що відповідає властивості постобробки диференційної приватності.

Додатково стійкість моделі оцінюється через симуляцію різних сценаріїв витоку даних, включаючи спроби реконструкції поведінкових патернів, аналіз структурних дескрипторів соціальних графів та комбіновані атаки на мультимедійні ознаки. Усі тести показують, що приватизація даних, застосована на етапі формалізації навчального набору, ефективно запобігає відновленню індивідуальних записів і підтримує статистичну коректність синтетичних даних[28, 56-60]. Оцінка стійкості моделі за симуляцією різних сценаріїв витоку даних наведено нижче (табл. 3.5).

Таблиця 3.5

Оцінка стійкості моделі за симуляцією різних сценаріїв витоку даних

Категорія даних	**Тип потенційної загрози**	**Рівень стійкості**	**Пояснення**
Контентні дані (тексти постів, коментарі)	Пряма реконструкція тексту користувача та непряма стилометрична ідентифікація	Високий	Тексти перетворені на ембеддинги, видалено унікальні ідентифікатори, DP-шум запобігає відновленню початкових фраз
Поведінкові дані (активність, часові проміжки)	Відновлення часових шаблонів	Високий	Агреговані статистики + гауссовий шум, доданий під час навчання DP-SGD зберігають приватність

Продовження табл. 3.5

Категорія даних	**Тип потенційної загрози**	**Рівень стійкості**	**Пояснення**
Метадані (геолокація, пристрої, часові мітки)	Відстеження маршрутів та пристроїв	Високий	Геолокації агреговано до рівня країни/регіону, часові мітки та пристрої категоризовані; відновлення індивідуальних маршрутів неможливе
Загальний набір даних	Комбіновані атаки (поведінка + структура + контент)	Високий	Використання DP, адаптивного обрізання градієнтів і кластерного підходу генераторів забезпечує комплексний захист і запобігає витоку

Для практичної демонстрації ефективності моделі DP-GAN-HD доцільно провести невелике OSINT-розслідування на прикладі користувача соціальної мережі, який свідомо не приділяв уваги захисту власної інформації. Мета цього експерименту полягає не у вторгненні в приватну сферу чи порушенні законів, а у формуванні контрольного набору даних із відкритих джерел, який дозволяє оцінити, які категорії інформації можуть бути доступні зовнішньому спостерігачеві. Такий підхід надає можливість перевірити стійкість DP-GAN-HD до потенційних витоків і водночас оцінити баланс між точністю синтетичних даних та рівнем захисту приватності. Важливо, що всі зібрані дані підлягають анонімізації та обробці, відповідно до вимог диференційної приватності.

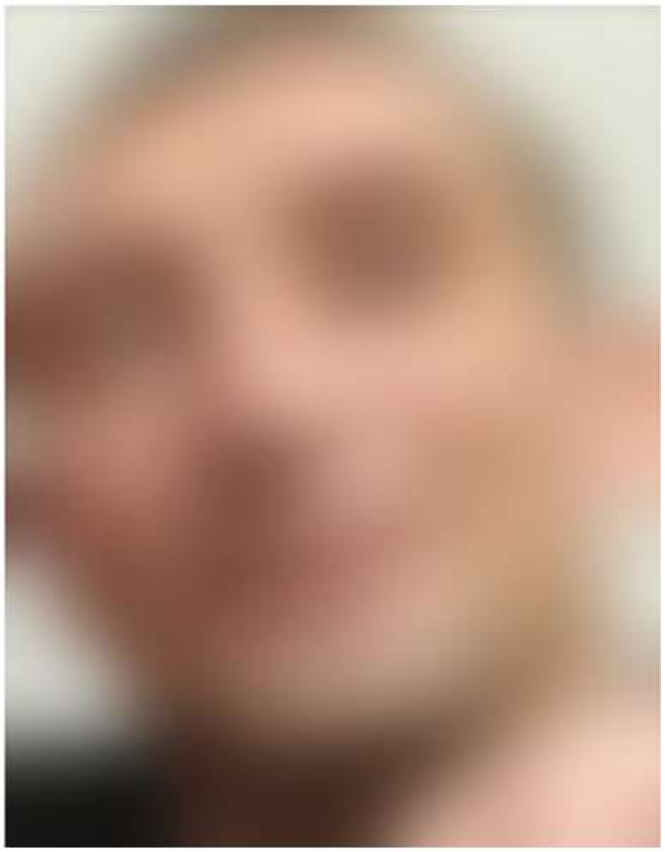

Рис 3.4 Фотографія особи, яка є об'єктом дослідження.

Мета аналізу - встановити його ім’я, прізвище, дату народження та, за можливості, місце проживання. Так як робота проводиться з освітньою метою.

Для пошуку збігів за фотографією використаємо онлайн - інструмент search4faces.

Рис 3.5 Результати аналізу за допомогою search4faces.

Результат дав посилання на сторінку в соціальних мережах, переходимо по наданим результатам. На рисунку 3.6 зображено інформацію та світлину профілю

Рис 3.6 Підсумкові результати пошуку.

Згідно з отриманими даними, особа ідентифікована (зауважимо, що без офіційних документів достовірність таких відомостей є умовною). Надалі завданням є встановлення точної геолокації. Пошук розпочато з першої світлини, яка є в профілі (рис. 3.7).

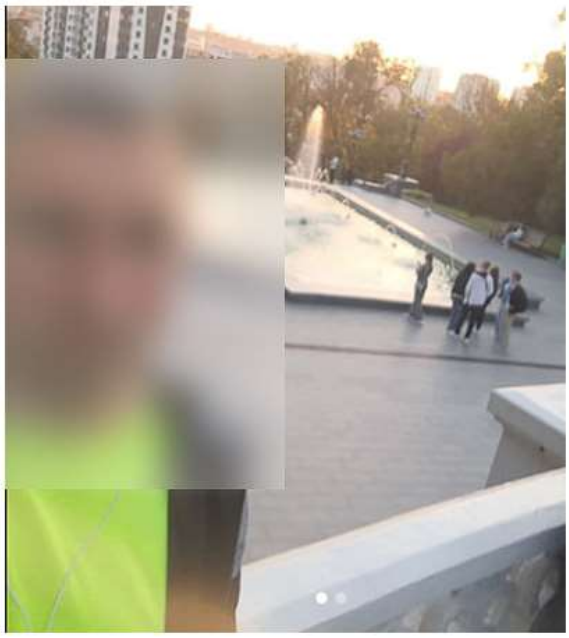

Рис 3.7 Зображення зі сторінки Instagram

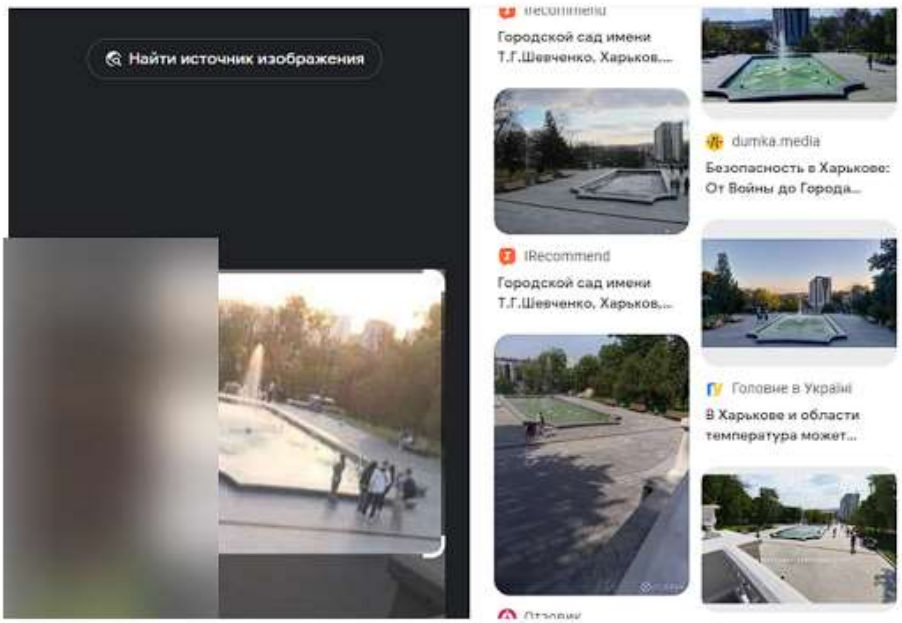

Рис 3.8 Результати аналізу за допомогою Google Images.

Рис 3.9 Виявлене місце, де була зроблена фотографія.

Знімки із широковідомих місць, як на рис. 3.9, не дають точних вказівок, тому для подальшого аналізу варто розглядати фотографії з локацій, які рідко фігурують у соціальних мережах. Вони можуть вказати на район проживання об'єкта дослідження.

Рис 3.10 Фото з малопопулярної локації.

Рис 3.11 Локація, ідентифікована за допомогою Google Lens та Google Maps.

Наведене моделювання пошуку інформації за допомогою соціальних мереж із використанням лише веб-сервісів вказує, що необережне використання платформ може призвести до витоку чутливих даних і мати різні наслідки залежно від цілей

збору інформації. Рисунки 3.10 та 3.11 демонструють, що при необережному використанні мереж публікація світлин навіть з маловідомих місць може навести спеціалістів із розвідки на подальшу локацію пошуку людини. Ці спостереження підкреслюють важливість впровадження механізмів захисту персональних даних у процесі генерації синтетичної інформації та створюють основу для оцінки ефективності моделі DP-GAN-HD у забезпеченні стійкості до потенційних загроз витоку даних. Також, аналіз стійкості показує, що DP-GAN-HD витримує як прямі, так і непрямі атаки.

Прямі атаки на дискримінатор, які намагаються відновити первинні записи, є неефективними завдяки наявності DP-шуму та агрегованої структури даних. Непрямі атаки, спрямовані на генератор, також не призводять до витоку, оскільки генератор отримує лише приватизований сигнал і не має прямого доступу до початкового датасету. Постобробка синтетичних даних не збільшує ризик порушення приватності, що відповідає властивості постобробки диференційної приватності.

3.4 Рекомендації щодо впровадження DP-GAN-HD в інфраструктуру соціальних мереж та інтеграція з алгоритмами модерації захисту контенту

Впровадження розробленої архітектури DP-GAN-HD у виробниче середовище соціальної мережі вимагає комплексної інтеграції, яка враховує високі вимоги до обчислювальних ресурсів та необхідність збереження низької затримки у взаємодії з користувачем та непомітної імплементації, щоб роботи по оновленню мали мінімальний вплив на роботу соціальних мереж. Оптимальним сценарієм розгортання є реалізація моделі у вигляді відокремленого мікросервісу, що функціонує в асинхронному режимі та виступає проміжним шаром між базами даних користувачів і аналітичними підсистемами платформи. Така ізоляція

дозволяє накопичувати дані у буферизовані пакети для ефективного застосування адаптивного відсікання градієнтів за алгоритмом DBSCAN, не створюючи при цьому пікового навантаження на основні сервери, що обслуговують запити клієнтських додатків у реальному часі. Процес доцільно розділити в три етапи (рис. 3.12).

1. Перший етап - агрегація чутливих даних та їх первинна формалізація (описано в підрозділі 3.1). Важливо реалізувати буферізацію даних, щоб навчання генераторів відбувалося пакетами (batch processing) для забезпечення порядку підходів пакетів до генераторів та справного функціонування адаптивного відсікання градієнтів DBSCAN.

2. Виділення окремих кластерів GPU для навчання генераторів для запобігання створення надмірного навантаження на основні сервери обслуговування користувачів

3. Синтетичні дані через API передаються до систем аналітики, тестування інтерфейсів та, що найважливіше, до модулів навчання алгоритмів модерації.

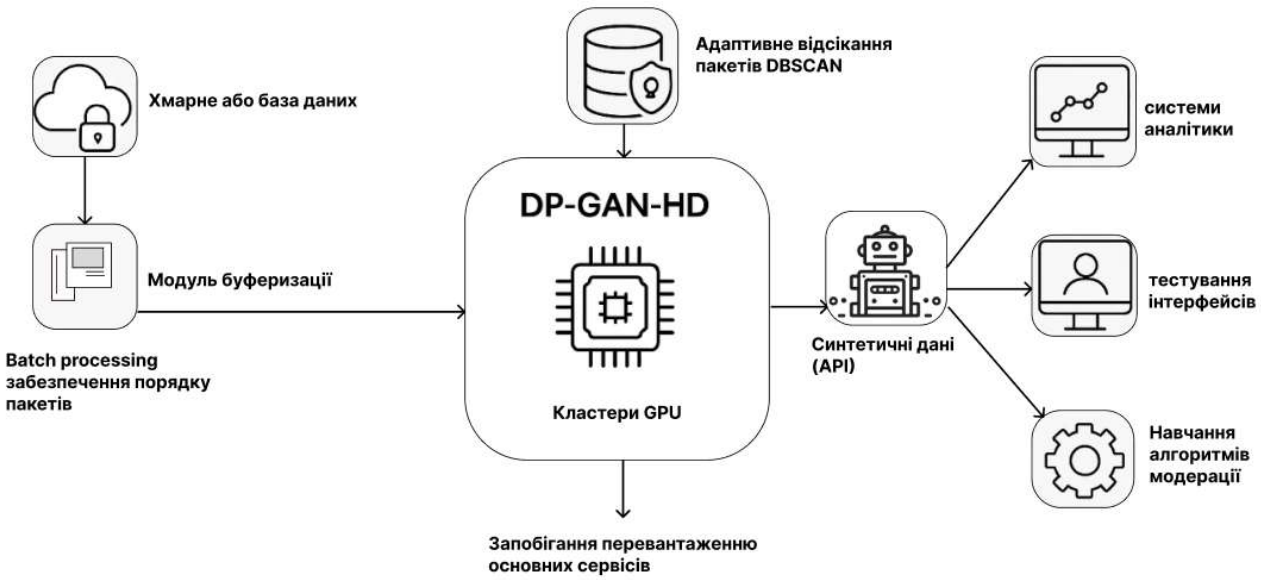

Рис. 3.12 Стратегія впровадження DP-GAN-HD

Окремою частиною стратегії впровадження архітектури DP-GAN-HD є її адаптація під специфічні потреби різних категорій суб'єктів, оскільки універсальний підхід не здатен одночасно задовольнити вимоги приватних осіб, бізнесу та державного сектору. Диференціація стратегії дозволяє налаштувати баланс між рівнем шуму (приватністю) та корисністю даних залежно від природи загроз.

Для приватних осіб впровадження DP-GAN-HD має відбуватися на рівні клієнтського API як прозорого проміжного шару, що функціонує в режимі реального часу або з мінімальною буферизацією. Механізм діє шляхом створення так синтетичного профілю - система аналізує реальні поведінкові патерни користувача і генерує на їх основі зашумлений потік даних, який статистично відповідає інтересам особи, але не дозволяє виокремити її унікальні ідентифікатори. Перевагою такого підходу є те, що користувач стає фактично невидимим для зовнішніх скреперів, рекламних трекерів та сталкерів, оскільки вони збирають синтетичні, а не реальні дані, завдяки чому покращується безпека персональних даних, унеможливлюючи створення точного цифрового двійника зловмисниками та знижуючи ризики соціальної інженерії. Обмеженням для цієї категорії може стати незначне зниження релевантності персональних рекомендацій контенту через внесений шум.

Корпоративні організації вимагають іншого підходу, де імплементація здійснюється через розгортання ізольованих екземплярів моделі у захищених хмарних середовищах або через корпоративні шлюзи соціальної платформи. У цьому сценарії DP-GAN-HD діє як фільтр для компанії. Внутрішні комунікаційні графи співробітників та ієрархічні зв'язки маскуються шляхом генерації шумових контактів, що приховує реальні центри прийняття рішень. Перевагою є можливість безпечного використання аналітичних інструментів платформи для обміну даними з партнерами без ризику розкриття комерційної таємниці чи інсайдерської

інформації. Безпека даних покращується завдяки нівелюванню загрози корпоративного шпигунства та аналізу трафіку конкурентами, які не зможуть відстежити підготовку або запуску нових продуктів. Основним обмеженням є висока вартість обчислювальних ресурсів для підтримки виділених кластерів.

Для об’єктів критичної інфраструктури впровадження DP-GAN-HD стає гарним прошарком захисту та реалізується в ізольованих контурах, часто без прямого доступу до інтернету на етапі навчання. Система працює на випередження, генеруючи синтетичні патерни активності та геолокації персоналу, що дозволяє приховати реальне розташування та режим роботи об’єктів підприємства від OSINT-розвідки. Позитивною стороною є підвищення стійкості до гібридних атак. Безпека персональних даних у цьому секторі покращується, оскільки унеможливлюється використання метаданих соціальних мереж для коригування фізичних атак або цільових інформаційно-психологічних операцій. Обмеженням є найсуворіші вимоги до точності генерації, оскільки будь-які галюцинації моделі у процесі навчання стануть великою проблемою для безпеки даних, а також необхідність в постійному аудиту алгоритмів. Далі візуально зображено баланс шуму/користі використання DP-GAN-HD в залежності від категорії користувачів. Графічне зображення щодо вимог до параметрів формалізовано нижче (рис. 3.13).

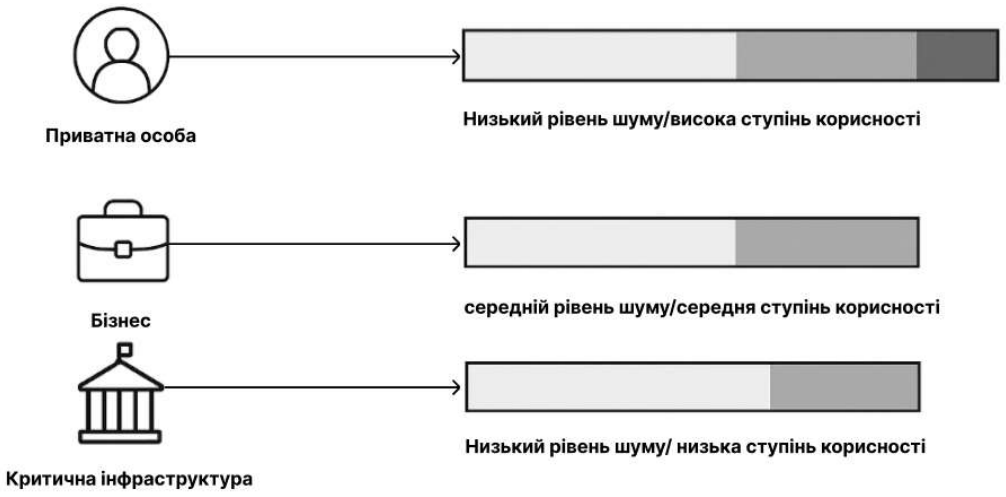

Рис. 3.13 Адаптація параметрів DP-GAN-HD до вимог користувачів

Зростання рівня шуму, необхідного для досягнення високого рівня диференційної приватності, призводить до деградації статистичних залежностей у даних, що знижує їх аналітичну та прикладну корисність. Подібна пріоритизація відновлення реальних даних над корисністю більш корисна для критичної інфраструктури, оскільки це «забруднює» інформацію до тієї ступені, що унеможливлює створення цифрового двійника, створює сильні перешкоди для OSINT – аналітиків та заважає кореляції активності персоналу.

Основою застосування запропонованої моделі є її синергія з автоматизованими системами модерації контенту, які часто страждають від проблеми дисбалансу класів у навчальних вибірках. У реальному потоці інформації частка шкідливого контенту є порівняно малою, а реальні приклади порушень часто швидко видаляються, що унеможливлює їх тривале використання для тренування класифікаторів через законодавчі обмеження. DP-GAN-HD дозволяє вирішити цю проблему шляхом генерації синтетичних порушень, які зберігають семантичну структуру та поведінкові ознаки деструктивного контенту, але не прив’язані до реальних профілів. Це забезпечує можливість штучної аугментації даних, насичуючи навчальні набори рідкісними граничними випадками атак, що суттєво підвищує метрики повноти та точності систем виявлення загроз.

Генератор DP-GAN-HD може бути налаштований на виконання ролі зловмисника, який намагається створити синтетичний контент, здатний обійти існуючі алгоритми захисту. У разі, якщо дискримінатор або зовнішня система модерації не розпізнає згенерований спам чи токсичний текст, це сигналізує про вразливість системи ще до того, як нею скористаються реальні зловмисники. Такий підхід трансформує процес модерації з реактивного, коли система навчається на вже пропущених помилках, у превентивний, де потенційні вектори атак моделюються та блокуються заздалегідь на синтетичних полігонах.

1. У реальному потоці даних кількість "нормального" контенту значно перевищує кількість порушень (дисбаланс класів). Це призводить до того, що алгоритми модерації погано розпізнають рідкісні види атак.

2. Традиційно, для донавчання ШІ-модераторів використовують розмітку даних людьми (human-in-the-loop). Перегляд реального токсичного контенту викликає психологічні травми.

3. Змагальна природа DP-GAN-HD дозволяє використовувати генератор як "спаринг-партнера" для існуючих фільтрів соцмережі. Генератор намагається створити такий синтетичний спам-пост, який пройде крізь фільтр (дискримінатор).

Замість перегляду реального травмуючого контенту, модератори та розробники алгоритмів можуть працювати з деперсоналізованими синтетичними аналогами, які математично еквівалентні реальним загрозам, але позбавлені конкретних персональних даних жертв чи порушників. Для забезпечення гарантованого рівня захисту при цьому необхідно впровадити суворий облік бюджету приватності, передбачивши механізм автоматичного блокування доступу до оновлення моделі на реальних даних у разі перевищення встановленого порогу. переходячи виключно на генерацію на основі раніше зафіксованих параметрів до моменту оновлення бюджету. Таким чином, імплементація DP-GAN-HD створює замкнений контур безпеки, де корисність даних для навчання алгоритмів платформи максимізується без компрометації конфіденційності користувачів.

Впровадження DP-GAN-HD вимагає суворого дотримання бюджету приватності (ε). Рекомендується застосовувати накопичувальний облік витрат приватності (Privacy Budget Accounting).

Якщо сумарне значення перевищує встановлений поріг (наприклад, $\varepsilon > 2.0$ для чутливих даних), система повинна автоматично припиняти доступ до реальних даних і переходити виключно на генерацію на основі вже навчених ваг, доки бюджет не буде оновлено (наприклад, шляхом додавання нових, незалежних даних

або з плином часу). Порівняння ефективності навчання модераційних систем наведено в таблиці 3.6.

Таблиця 3.6

Порівняння ефективності навчання модераційних систем на реальних та синтетичних даних (DP-GAN-HD)

Характеристика	**Навчання на реальних даних**	**Навчання на даних DP-GAN-HD**	**Перевага впровадження**
Доступність даних	Обмежена (видалені пости, приватні чати)	Необмежена (генерація за запитом)	Можливість масштабування датасетів
Ризик витоку (Privacy)	Високий (можливість реконструкції особи)	Низький	Відповідність GDPR та етичним нормам
Детекція нових загроз	Реактивна (навчання після інциденту)	Проактивна (генерація варіантів атак)	Випередження зловмисників
Bias (Упередженість)	Висока (успадковує історичні упередження)	Контрольована	Справедливіша модерація
Точність (F1-score)	Базова	Середня вірогідність досягнення великого відсотку точності	Незначна втрата точності при значному зростанні безпеки

Попри переваги, впровадження DP-GAN-HD супроводжується низкою ризиків, наведених в таблиці 3.7, які необхідно враховувати на етапі інтеграції. Одним з ключових ризиків є високе ресурсне навантаження, спричинене обчислювально інтенсивними етапами навчання генераторів. У пікові періоди це може призвести до деградації сервісів платформи, якщо GPU-кластер недостатньо ізольований. Для мінімізації цього ризику варто застосовувати окремі GPU-черги, пріоритизацію завдань та обмеження споживання ресурсів на рівні контейнерів.

Оскільки соціальні мережі швидко змінюються, а з ними розвиваються патерни шкідливої поведінки, генератор може почати створювати застарілі або нерелевантні зразки. Це знижує ефективність модераційних алгоритмів. Розв’язанням є регулярне донавчання моделі, введення механізму деградаційних

тригерів, а також використання дрейф-детекторів для оцінки стабільності розподілів. Якщо зловмисник зуміє вплинути на дані, що потрапляють до буферів для тренування генератора, він може змінити розподіл синтетичних даних або навіть використати це для обходу модераційних систем.

Тому необхідно впроваджувати фільтри попередньої валідації, контроль доступу до буферів та механізми ізоляції даних з підозрілими характеристиками. Якщо GAN-модель починає створювати надто нетипові або занадто абстрактні приклади, які не відповідають реальним загрозам, це може сформувати хибні кордони класифікації. У разі перевірок доведеться пояснювати, як саме генерується синтетичний контент, які гарантії приватності забезпечує ε-DP та чому синтетичні дані є юридично безпечними для використання.

Таблиця 3.7

Ризики при імплементації

Ризик інтеграції	**Вплив**	**Потенційні наслідки**	**Рекомендовані заходи мінімізації**
Надмірне ресурсне навантаження	Високе споживання GPU та пам'яті під час навчання	Деградація роботи сервісів, затримки модерації	Виділення окремого GPU-кластера; пріоритизація завдань; ліміти ресурсів на рівні контейнерів
Дрейф даних та моделі	Зміна поведінки користувачів та поява нових патернів атак	Втрата актуальності синтетичних даних, зниження точності модерації	Регулярне донавчання; детектори дрейфу; тригери деградації продуктивності
Отруєння даних (data poisoning)	Потрапляння шкідливих або маніпулятивних даних у тренувальні буфери	Компрометація моделі, генерація некоректних зразків	Попередня валідація даних; контроль доступу; ізоляція підозрілих наборів
Некоректна генерація складних випадків	GAN може створювати надто спрощені або нереалістичні приклади	Хибні межі класифікації; зниження ефективності модерації	Manual review; human-in-the-loop; контроль стабільності розподілів

Продовження табл. 3.7

Ризик інтеграції	**Вплив**	**Потенційні наслідки**	**Рекомендовані заходи мінімізації**
Непрозорість для аудиту та регуляторів	Складність пояснення процесів синтезу та гарантій приватності	Проблеми із сертифікацією, комплаєнсом, відповідністю GDPR	Документування версій моделей; зберігання параметрів генерації; аудит DP-бюджету
Надмірна залежність від синтетики	Заміну реальних даних синтетичними без контролю пропорцій	Формування системних упереджень у модерації	Використання змішаних наборів (real + synthetic); обмеження частки синтетичних даних

Врахування зазначених ризиків та впровадження багатоетапних механізмів контролю дає змогу інтегрувати DP-GAN-HD у масштабну інфраструктуру соціальних мереж без негативного впливу на продуктивність, приватність чи якість процесів модерації. Забезпечення безпечної інтеграції вимагає не лише точного дотримання параметрів диференційної приватності, але й постійного моніторингу моделей на предмет деградації, появи режимного колапсу та відхилень у статистиці синтетичних даних. Використання адаптивних політик навчання, регулярних аудитів і контрольних тестів стійкості дозволяє мінімізувати ймовірність компрометації персональних даних під час генерації та їхнього повторного відтворення.

Висновки до розділу 3

У третьому розділі проведено розробку та обґрунтування методу DP-GAN-HD, призначеного для захисту персональних даних у соціальних мережах шляхом генерації синтетичних даних.

Запропонована архітектура DP-GAN-HD (GAN) з механізмами диференційної приватності та адаптивним відсіканням градієнтів на основі

кластеризації DBSCAN довела свою перевагу над існуючими аналогами (PATE-GAN, DP-VAE). Мультигенераторна структура дозволила вирішити проблему нестабільності навчання, забезпечуючи відтворення складних гетерогенних даних соціальних мереж без прямого копіювання інформації користувачів. Аналіз стійкості моделі підтвердив, що інтеграція шуму в градієнти дискримінатора та агрегація даних на етапі попередньої обробки унеможливлюють успішне проведення атак на реконструкцію (Reconstruction Attacks). Проведений практичний OSINT-експеримент наочно продемонстрував вразливість відкритих даних у соціальних мережах та підтвердив необхідність переходу на синтетичні вибірки, які, згідно з результатами тестування, не дозволяють ідентифікувати конкретну особу навіть при застосуванні комбінованих векторів атак. Розроблено стратегію впровадження DP-GAN-HD у вигляді асинхронного мікросервісу, що мінімізує вплив на продуктивність основної платформи.

ВИСНОВКИ

У представленій кваліфікаційній роботі здійснено комплексне дослідження проблеми захисту персональної інформації в соціальних мережах в умовах стрімкого поширення генеративного штучного інтелекту (ГШІ) та зростаючої складності інформаційних загроз. Встановлено, що сучасна екосистема соціальних платформ перетворює персональні дані користувача з статичного набору атрибутів на багатошаровий цифровий профіль, сформований із поєднання явно наданих даних, поведінкових метаданих та автоматично згенерованих або прогнозованих характеристик. Така трансформація значно підвищує ризики порушення приватності, оскільки сучасні методи атаки використовують саме ті структури даних, які раніше не вважались критичними - стилометричні патерни, графи соціальних зв’язків, поведінкові траєкторії, біометричні сліди.

Проведений аналіз існуючих систем соціальних мереж показав, що навіть провідні платформи (Facebook/Meta, X, Instagram, TikTok, Telegram, LinkedIn) мають вразливості, зумовлені архітектурними обмеженнями, недосконалістю політик конфіденційності, відсутністю повноцінної прозорості щодо обробки метаданих та застарілими моделями контролю доступу. Установлено, що традиційні технічні заходи - наскрізне шифрування, багатофакторна автентифікація, політики мінімізації даних є недостатніми для протидії сучасним атакам, оскільки не враховують можливості ГШІ щодо стилометричного аналізу, створення синтетичних біометричних даних та автоматизованої інфільтрації через психологічне профілювання.

Здійснено систематизацію та класифікацію загроз, пов’язаних із використанням генеративних моделей штучного інтелекту, зокрема реконструкційних атак, атак членства (Membership Inference), багатокрокових атак соціальної інженерії та атак на моделі модерації контенту. Показано, що гібридні

загрози нового покоління поєднують технологічну експлуатацію з когнітивним впливом, що робить їх особливо небезпечними та малопомітними для традиційних систем виявлення.

На основі проведеного фундаментального та прикладного аналізу розроблено новий метод диференційно-приватної генерації синтетичних даних DP-GAN-HD, який поєднує переваги змагальних нейронних мереж, диференційної приватності та адаптивного кластеризованого відсікання градієнтів. Запропонована архітектура демонструє стійкість до Reconstruction Attacks, забезпечує високий рівень анонімності та дозволяє зберегти статистичну корисність синтетичних даних без копіювання атрибутів конкретних користувачів. Мультигенераторний підхід та використання DBSCAN для регулювання градієнтів підвищили стабільність навчання та дали змогу адекватно моделювати гетерогенні дані соціальних мереж: текстові, графові та поведінкові.

У роботі експериментально доведено, що DP-GAN-HD може бути успішно інтегрована у системи модерації контенту як генератор Synthetic Violations - синтетичних прикладів рідкісних або небезпечних порушень, які практично неможливо зібрати у достатньому обсязі серед реального контенту. Це дозволяє покращити якість класифікації, знизити навантаження на модераторів та істотно зменшити частоту «сліпих зон» у системах безпеки.

Розроблено стратегію впровадження DP-GAN-HD у вигляді асинхронного мікросервісу, що забезпечує масштабованість, мінімальний вплив на продуктивність та можливість використання у різних доменах платформи: модерації, аналітиці поведінкових патернів, тестуванні систем виявлення загроз.

Підсумковий аналіз показує, що використання DP-GAN-HD формує технологічне підґрунтя для переходу соціальних мереж від реактивної моделі безпеки до превентивної, де синтетичні дані, адаптивна приватність та гібридні алгоритмічні механізми стають ефективними компонентами системи захисту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Закон України «Про захист персональних даних» URL:https://zakon.rada.gov.ua/laws/show/2297-17#Text)

2. РЕГЛАМЕНТ ЄВРОПЕЙСЬКОГО ПАРЛАМЕНТУ І РАДИ (ЄС) 2016/679 від 27 квітня 2016 року про захист фізичних осіб у зв'язку з опрацюванням персональних даних і про вільний рух таких даних, та про скасування Директиви 95/46/ЄС (Загальний регламент про захист даних) URL: https://zakon.rada.gov.ua/laws/main/984_008-16

3. Харитонова О. І., Харитонов Є. О., Старцев О. В. Право Європейського Союзу у сфері захисту персональних даних. - Київ: BAITE, 2018. - 264 с.

4. Paul Vogel. Künstliche Intelligenz und Datenschutz. URL: https://api.pageplace.de/preview/DT0400.9783748930952_A43532678/preview-9783748930952_A43532678.pdf

5. Big Data for All: Privacy and User Control in the Age of Analytics Volume 11 Is.5 p. 1-36 URL: https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1191&context=njtip

6. Закон України «про інформацію» URL: https://zakon.rada.gov.ua/laws/show/2657-12#Text

7. Постанова Кабінету міністрів Про затвердження Загальних вимог з кіберзахисту об’єктів критичної інфраструктури. URL: https://zakon.rada.gov.ua/laws/show/518-2019-%D0%BF#Text

8. GDPR: ключові принципи і адаптація в Україні URL: https://visnyk-juris-uzhnu.com/wp-content/uploads/2024/11/44-1.pdf

9. Олександр Шевчук відповідальність і санкції за порушення загального регламенту про захист даних с. 1-69 (GDPR) https://eu4digitalua.eu/wp-content/uploads/2025/02/gdprvidpovidalnist-i-sanktsii-3.pdf

10. European Union Agency for Cybersecurity (ENISA). Cybersecurity in the Age of AI: Challenges and Recommendations. ENISA Report, 2022. P. 30-45 URL: https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf

11. European Parliament and Council. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj

12. Кальман О. Г. Правові засади захисту персональних даних в Україні та Європейському Союзі: порівняльно-правовий аналіз. Вісник Луганського державного університету внутрішніх справ імені Е. О. Дідоренка. 2019. Вип. 4(88). С. 139-149.

13. Свиридова А. О. *Захист чутливих персональних даних в Україні: порівняльно-правовий аспект*. Науковий вісник Ужгородського національного університету. Серія: Право. 2021. Вип. 69. С. 177-181. URL: https://journal-app.uzhnu.edu.ua/article/view/287228

14. Greenleaf G. *The GDPR and Its Global Impact*. Privacy Laws & Business International Report, 2017. P. 14-17.

15. Притула С. М. *Проблеми інституціоналізації державного контролю за дотриманням законодавства про захист персональних даних*. Вісник Національної академії правових наук України. 2020. Т. 27, № 3. С. 139-153. DOI: 10.37631/naprn.2.2020.139-153. URL: https://doi.org/10.37631/naprn.2.2020.139-153

16. Посібник з європейського права у сфері захисту персональних даних https://apladm.dn.court.gov.ua/userfiles/posibnik_1.pdf

17. Acquisti A., Taylor C. R., Wagman L. The Economics of Privacy. Journal of Economic Literature. 2016. Vol. 54, no. 2. P. 442-492. URL: https://www.heinz.cmu.edu/~acquisti/papers/AcquistiTaylorWagman-JEL-2016.pdf

18. Гармонізація Українського права з GDPR URL: https://osint.in.ua/zakon-gdpr-ukraina/

19. Бєлов Д.М, Бєлова М.В, Горнило О.Т., право людини на забуття С. 57-61 URL: https://visnyk-juris-uzhnu.com/wp-content/uploads/2024/03/11-3.pdf

20. Albladi S.M., Weir G.R.S. *User characteristics that influence judgment of social engineering attacks in social networks.* Human-centric Computing and Information Sciences, 2018, 8(1). DOI: 10.1186/s13673-018-0137-6.Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities URL: https://www.sciencedirect.com/science/article/pii/S2667345225000082

21. Goodfellow I. J., Pouget-Abadie J., Mirza M., et al. Generative Adversarial Networks. Advances in Neural Information Processing Systems. 2014. P. 2672-2680.

22. Data Breach Investigations Report URL: https://www.verizon.com/business/resources/reports/2021-data-breach-investigations-report.pdf

23. Sustaining Cyber Awareness: The Long-Term Impact of Continuous Phishing Training and Emotional Triggers URL: https://arxiv.org/pdf/2510.27298

24. Meta Privacy Policy URL: https://www.facebook.com/privacy/policy/

25. Telegram Privacy Policy URL: https://telegram.org/privacy/ua?setln=en

26. LinkedIn Privacy Policy URL: https://www.linkedin.com/legal/privacy-policy#data

27. X Privacy Policy URL: (https://x.ai/legal/privacy-policy/previous-2024-12-20/#1-about-xai-and-grok)

28. Acquisti A., Taylor C. R., Wagman L. The Economics of Privacy. *Journal of Economic Literature*. 2016. Vol. 54, no. 2. P. 442-492.

29. Solove D. J. *The Digital Person: Technology and Privacy in the Information Age*. New York University Press, 2004. 312 p.

30. Acquisti A., Grossklags J. Privacy and Rationality in the Age of Information Technology. *IEEE Security & Privacy*. 2005. Vol. 3, no. 1. P. 26-33. URL: https://www.heinz.cmu.edu/~acquisti/papers/AcquistiGrossklags-IEEE-2005.pdf

31. Solove D. J. *The Digital Person: Technology and Privacy in the Information Age*. New York University Press, 2004. 312 p. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2899131

32. Security and Privacy Challenges of Large Language Models: A Survey URL: https://arxiv.org/pdf/2402.00888

33. Marlinspike M. The Challenge of Metadata. *Signal Blog*. 2016.

34. Zureik E., Salter M. B. *Global Surveillance and Human Rights*. University of Chicago Press, 2019. 304 p.

35. Anderson R., Whitten A. Poking Holes in Security: A Look at Side-Channel Attacks. *IEEE Security & Privacy*. 2004. Vol. 2, no. 1. P. 60-64.

36. Wampler L. Large Language Models and Data Leakage: Threats from AI-Powered Social Engineering. *Technology Law Review*. 2023. P. 55-71.

37. Tene O., Polonetsky J. Big Data and the Future of Privacy. *Colorado Law Review*. 2013. Vol. 84, no. 1. P. 1-74.

38. Carlini & Wagner — Robustness Evaluation of Neural Networks URL: https://arxiv.org/abs/1608.04644

39. Detection of Bots in Social Media: A Systematic Review URL: https://www.sciencedirect.com/science/article/abs/pii/S0306457319313937

40. Backstrom, Dwork, Kleinberg — De-Anonymizing Social Networks URL: https://www.cs.cornell.edu/~lars/www07-anon.pdf

41. Zhang, J., Zhang, Z., Xu, S., & Ren, K. DP-GAN-HD: Personal health data protection and intelligent healthcare applications under generative adversarial network URL: https://www.nature.com/articles/s41598-025-01575-1

42. Abadi, M., Chu, A., Goodfellow, I., et al. Deep Learning with Differential Privacy. *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016. P. 308-318. DOI: 10.1145/2976749.2978318.

43. Пасічник В. В., Якимова В. М., Мельник А. О. Застосування синтетичних даних для забезпечення приватності в інформаційних системах. *Науковий вісник Ужгородського національного університету. Серія: Інформаційні технології*. 2021. Вип. 48. С. 60-65.

44. Abadi, M., Chu, A., Goodfellow, I., et al. Deep Learning with Differential Privacy. *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2016. P. 308-318. DOI: 10.1145/2976749.2978318.

45. Аналіз застосування генеративно-змагальних мереж у сфері кібербезпеки» — Ляшенко О. С., Щербина Д. В. (2022) URL: https://openarchive.nure.ua/entities/publication/069253d7-a3d2-4d82-9f69-fc85451786a4

46. Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data» — Baligodugula V., Amsaad F. URL: https://arxiv.org/abs/2503.23215

47. «Генерування аберометричних даних шляхом застосування GAN» — Коротка Л. І., Макарченко В. С. https://repository.kpi.kharkov.ua/items/32a5b6b7-ac9a-43db-bb1c-53be76022873

48. Denoeux T. - Neural Network-based Evidential Clustering (NN-EVCLUS) URL: https://arxiv.org/abs/2009.12795

49. LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks URL: https://arxiv.org/abs/2112.05355

50. Zafarani R., Abbasi M. A., Liu H. *Social Media Mining: An Introduction*. Cambridge University Press, 2014. 586 p.

51. Dwork C., Roth A. *The Algorithmic Foundations of Differential Privacy*. Now Publishers, 2014. 210 p. URL: https://studylib.net/doc/25962415/the-algorithmic-foundations-of-differential-privacy

52. Jurafsky D., Martin J. H. *Speech and Language Processing*. 3rd ed. draft. Stanford University, 2023.

53. О.В. Кутовий: равове регулювання та державна політика у сфері застосування штучного інтелекту: національні та інтеграційні аспекти С. 95-99. URL: http://pjv.nuoua.od.ua/v3_2024/18.pdf

54. European Union Agency for Cybersecurity (ENISA). *Recommendations on the Security of AI-based Systems and the Data Lifecycle*. ENISA Report, 2023. P. 25-35.

55. Shokri R., Stronati M., Song C., et al. Membership Inference Attacks Against Machine Learning Models. *IEEE Symposium on Security and Privacy (S&P)*. 2017. P. 3-18. DOI: 10.1109/SP.2017.37. URL: https://arxiv.org/pdf/1610.05820

56. Jiaxuan G., Ruoyu Y., Wenda C**.** Differential Privacy: A survey. *ACM Computing Surveys*. 2021. , no. 1. P. 1-7. DOI: 10.1145/3448375. URL: https://chuwd19.github.io/project/dp/DP.pdf

57. A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions URL: https://dl.acm.org/doi/pdf/10.1145/3460427

58. Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A. *Improved Training of Wasserstein GANs.* NeurIPS, 2017. URL: https://arxiv.org/pdf/1704.00028

59. Miyato T., Kataoka T., Koyama M., Yoshida Y. *Spectral Normalization for Generative Adversarial Networks.* ICLR, 2018. URL: https://arxiv.org/pdf/1802.05957

60. Xu L., Skoularidou M., Cuesta-Infante A., Veeramachaneni K. *Modeling Tabular Data Using Conditional GAN.* NeurIPS, 2019. URL:

https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

61. Dwork C., Roth A. *The Algorithmic Foundations of Differential Privacy.* Foundations and Trends in Theoretical Computer Science, 2014. URL: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

62. Shokri R., Stronati M., Song C., Shmatikov V. *Membership Inference Attacks Against Machine Learning Models.* IEEE S&P, 2017. URL: https://arxiv.org/pdf/1610.05820