ISSN 2519-2310 (Online)

КОМП'ЮТЕРНІ НАУКИ ТА КІБЕРБЕЗПЕКА, 2024, № 2 (26)

COMPUTER SCIENCE AND CYBERSECURITY, 2024, Iss. 2 (26)

DOI: https://doi.org/10.26565/2519-2310-2024-2-01
УДК 004.056.5

ДОСЛІДЖЕННЯ ТА КЛАСИФІКАЦІЯ ОСНОВНИХ ТИПІВ АТАК НА СИСТЕМИ ШТУЧНОГО ІНТЕЛЕКТУ В КІБЕРБЕЗПЕЦІ

Владислав Вілігура[1], PhD, старший викладач кафедри КІСМТ, e-mail: v.v.vilihura@karazin.ua, ORCID: https://orcid.org/0000-0002-1137-2382

Єлизавета Острянська[1], молодший науковий співробітник, e-mail: antelizza@gmail.com, ORCID: https://orcid.org/0000-0003-1412-8470

[1]*Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків, 61022, Україна*

Рукопис надійшов 1 листопада 2024 р. Отримано після рецензування 2 грудня 2024 р. Прийнято 23 грудня 2024 р.

Анотація: Сучасний розвиток штучного інтелекту (ШІ) та машинного навчання (ML) відкриває нові можливості у сфері кібербезпеки, проте водночас створює серйозні виклики у вигляді інтелектуальних кібератак. Дослідження присвячене аналізу та класифікації способів використання ШІ у зловмисних цілях та вивченню ефективних методів протидії таким загрозам. Зокрема, стаття охоплює основні види атак, що використовують технології ML, які демонструють, як зловмисники можуть маніпулювати алгоритмами машинного навчання, підривати довіру до даних та обходити системи захисту. Окрему увагу приділено механізмам атак отруєння даних, так як вони вважаються найбільш небезпечними при машинному навчанні, які передбачають внесення шкідливих даних у процес навчання моделей, що призводить до викривлення результатів та підриву ефективності алгоритмів безпеки. Також розглядаються атаки проникнення, у яких атакуючі створюють унікальні зразки даних, що можуть залишатися невидимими для традиційних систем виявлення загроз. Атаки на конфіденційність аналізуються як спосіб отримання конфіденційної інформації з ML-моделей, що може використовуватися для викрадення даних користувачів. Атаки зловживання демонструють, як зловмисники можуть використовувати ШІ-інструменти для автоматизації атак, масштабування фішингових кампаній та аналізу слабких місць захисних систем. Актуальність дослідження зумовлена тим, що традиційні підходи до кіберзахисту вже не здатні ефективно протистояти загрозам, які адаптуються та еволюціонують завдяки машинному навчанню. Стаття наголошує на критичній важливості дослідження методів захисту, зокрема побудови надійних систем машинного навчання, що мають вбудовані механізми виявлення аномалій та адаптацію до нових загроз. Одним із ключових підходів є федеративне навчання, яке дозволяє тренувати моделі без централізованого зберігання даних, зменшуючи ризик витоку інформації. Також розглянуто розвиток глибокого навчання у сфері кіберзахисту, що дозволяє аналізувати поведінкові патерни загроз у режимі реального часу. Важливим аспектом залишається поєднання технологічних заходів із людським контролем, оскільки, незважаючи на потужність ШІ-інструментів, людський фактор залишається ключовим у процесі забезпечення кібербезпеки. Отже, стаття демонструє баланс між можливостями та загрозами ШІ у сфері кібербезпеки, підкреслюючи необхідність подальших досліджень у напрямку стійких ML-моделей, які можуть ефективно протистояти атакам. Без належного регулювання та контролю ШІ може стати не лише

захисником, а й інструментом зловмисників, що вимагає розробки нових стратегій безпеки та міжнародного врегулювання у сфері кіберзахисту.

Ключові слова: *штучний інтелект, кібератаки, машинне навчання, кібербезпека, федеративне навчання*

Як цитувати: Вілігура В., Острянська Є. Дослідження та класифікація основних типів атак на системи штучного інтелекту в кібербезпеці. *Комп'ютерні науки та кібербезпека.* 2024; № 2(26): С. 6–18. https://doi.org/10.26565/2519-2310-2024-2-01

In cites: Vilihura V., Ostrianska Ye. (2024). Research and classification of the main types of attacks on artificial intelligence systems in cybersecurity. *Computer Science and Cybersecurity.* 2(26): 6–18. https://doi.org/10.26565/2519- 2310-2024-2-01 (in Ukrainian)

1. Вступ

У сучасному світі штучний інтелект (ШІ) дедалі активніше інтегрується в різні сфери людської діяльності – від фінансів і медицини до автономного транспорту та кібербезпеки. Разом із розвитком технологій ШІ виникають нові виклики, зокрема загрози, пов'язані з атаками на штучний інтелект. Такі атаки можуть мати серйозні наслідки, включаючи компрометацію безпеки даних, маніпулювання алгоритмами та створення вразливостей у критично важливих системах.

Машинне навчання, як основа більшості сучасних ШІ-систем, дозволяє їм самостійно знаходити закономірності в даних, адаптуватися до змін і вдосконалюватися з часом. Однак, оскільки моделі ML значною мірою покладаються на якість вхідних даних та гіпотези, на яких вони ґрунтуються, вони стають вразливими до певних типів атак. Зловмисники можуть маніпулювати навчальними даними, підмінювати вхідні параметри або експлуатувати слабкі місця алгоритмів для досягнення своїх цілей.

По суті, методологія машинного навчання, яка використовується в сучасних системах штучного інтелекту, сприйнятлива до атак через загальнодоступні API, які розкривають модель, і проти платформ, на яких вони розгорнуті. Для атак на моделі безпеки зловмисники можуть порушити захист конфіденційності та захист даних і моделі, просто використовуючи загальнодоступні інтерфейси та надаючи вхідні дані, які знаходяться в межах прийнятного діапазону. У цьому сенсі виклики, з якими стикається AML, подібні до тих, що постають перед криптографією. Сучасна криптографія спирається на безпечні алгоритми в теоретичному сенсі інформації. Таким чином, люди повинні зосередитися лише на їх надійному та безпечному впровадженні, що і є першочерговою задачею для спільноти науковців-криптологів. На відміну від криптографії, немає інформаційно-теоретичних доказів безпеки для широко використовуваних алгоритмів машинного навчання. Як наслідок, багато досягнень у розробці засобів пом'якшення різних класів атак мають емпіричний та обмежений характер.

Але багато компаній та структур в останні роки активно працюють над врегулюванням використання ШІ в свої системах. Так, наприклад, серед широкого спектру діяльності NIST робить внесок у дослідження, стандарти, оцінки та дані, необхідні для розвитку, використання та забезпечення надійного штучного інтелекту (ШІ). У 2024 NIST опублікував звіт [1] щодо загроз на основі машинного навчання, а 2025 доповнив його [25]. У цьому звіті розроблено таксономію понять і визначено термінологію у сфері змагального машинного навчання (AML).

Атаки на системи AI/ML можна поділити на кілька категорій залежно від цілей зловмисника, методів реалізації та рівня впливу на модель. Серед найбільш поширених типів

атак – атаки на навчання (poisoning attacks), атаки на проникнення (evasion attacks), атаки на конфіденційність (privacy attacks) та атаки на доступність (denial-of-service attacks). Кожен із цих типів атак використовує різні механізми впливу, від підміни навчальних даних до експлуатації вразливостей у вже навчених моделях.

Актуальність дослідження обумовлена зростаючою кількістю випадків злому та маніпуляцій ШІ-системами, що може призвести до фінансових втрат, загроз безпеці та втрати довіри до технологій. Тому, метою цієї статті є дослідження загроз та ризиків пов’язаних з використанням ШІ. В статті визначаються типи атак, які можуть бути спрямовані на AI/ML-моделі, етапи життєвого циклу атаки, цілі та завдання зловмисника, а також можливості зловмисника та знання процесу навчання. Дослідження дозволяє краще зрозуміти сучасні ризики у сфері штучного інтелекту та визначити заходи для підвищення безпеки таких систем.

2. Класифікація атак на основі ШІ

Атаки на основі ШІ можна класифікувати за багатьма різними параметрами. Так, наприклад, кілька систем класифікації атак були представлені в роботах [7, 8]. Також NIST в своєму звіті щодо атак на машинне навчання ШІ представив різні типи класифікації [1, 25]. В цьому розділі будуть розглянуті деякі типи класифікацій.

2.1. Основні типи атак

Атаки на основі машинного навчання та ШІ прийнято класифікувати за такими загальними параметрами [1]:

1) метод навчання та етап процесу навчання, коли атаку встановлюють;
2) цілі та завдання зловмисника;
3) можливості зловмисника;
4) знання зловмисника про процес навчання.

На рис. 1 представлено загальну класифікацію атак, що спрямовані на AI/ML-моделі.

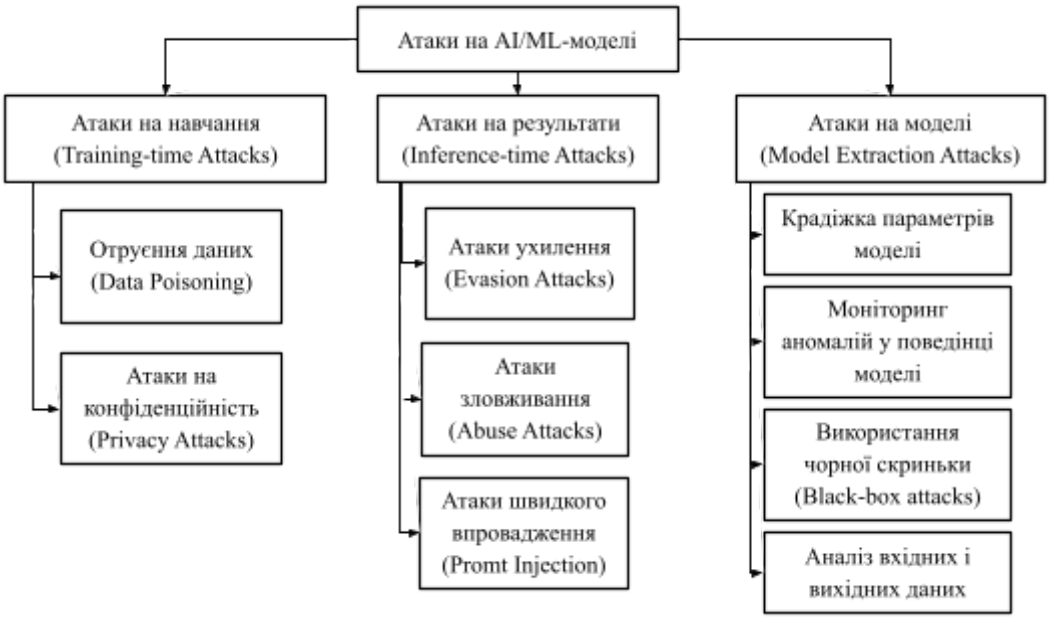

Рис. 1 – Класифікація атак на системи ШІ

Fig. 1 – Classification of attacks on AI systems

На рис. 1 зображено схему класифікації атак на системи ШІ. В свою чергу кожну з зазначених атак можна поділити на наступні основні підгрупи:

- Отруєння даних (Data Poisoning):
  - Внесення шкідливих даних у навчальний набір.
  - Підміна міток класів (Label-flipping Attack).
  - Використання прихованих тригерів (Backdoor Attack).
  - Маніпуляція вагами нейромережі.
- Атаки на конфіденційність (Privacy Attacks):
  - Витяг даних з навчального набору.
  - Інверсія моделі (Model Inversion Attack).
  - Атака на членство (Membership Inference Attack).
  - Злам параметрів моделі.
- Атаки ухилення (Evasion Attacks):
  - Маніпуляція вхідними даними.
  - Атака на зображення/текст/аудіо.
  - Обхід антивірусних-систем.
  - Зміна параметрів розпізнавання.
- Атаки зловживання (Abuse Attacks):
  - Використання генеративного ШІ для створення фальшивого контенту.
  - Deepfake (відео, аудіо, зображення).
  - Атаки соціальної інженерії.
  - Створення шкідливого коду.
- Атаки швидкого впровадження (Prompt Injection):
  - Вплив на текстові моделі (ChatGPT, Bard тощо).
  - Нав’язування небажаних відповідей.
  - Обхід обмежень моделі.
  - Генерація фейкової інформації.

2.2. Класифікація атак за метою зловмисника

Цілі зловмисника класифікуються за трьома критеріями відповідно до трьох основних типів порушень безпеки [1], які розглядаються під час аналізу безпеки системи: порушення доступності, порушення цілісності та компрометація конфіденційності даних. Відповідно, успіх зловмисника вказує на досягнення однієї або кількох із цих цілей.

Атака на доступність – це невибіркова атака на машинне навчання (ML), під час якої зловмисник намагається порушити продуктивність моделі під час розгортання. Атаки на доступність можуть бути влаштовані через отруєння даних, коли зловмисник контролює частину навчального набору або шляхом отруєння моделі, коли зловмисник контролює параметри моделі.

Атака на цілісність спрямована на цілісність вихідних даних моделі ML, що призводить до неправильних прогнозів, які виконує модель ML. Зловмисник може спричинити порушення цілісності, здійснивши атаку ухилення під час розгортання або атаку отруєння під час навчання. Атаки ухилення вимагають модифікації тестових зразків для створення змагальних прикладів, які неправильно класифікуються моделлю до іншого класу, залишаючись прихованими та непомітними для людей. Приклади таких атак можна знайти в роботах [12, 13].

При атаках спрямованих на порушення конфіденційності, зловмисники можуть бути зацікавлені в отриманні інформації про навчальні дані або про модель ML (що призводить до атак щодо конфіденційності даних та моделі відповідно). Зловмисник може мати різні цілі для

компрометації конфіденційності навчальних даних, наприклад, зміна даних (виведення вмісту або особливостей навчальних даних), викрадення даних [14, 15] (можливість витягувати навчальні дані з генеративних моделей) і викрадення властивостей щодо розподілу навчальних даних [16].

2.3. Класифікація атак за знанням зловмисника

Ще одним критерієм для класифікації атак є те, наскільки зловмисник має знання про систему машинного навчання. Існує три основних типи атак за цим критерієм [1]: біла скринька, чорна скринька та сіра скринька, див. рис. 2.

● Атаки білої скриньки. Вони припускають, що зловмисник працює з повним знанням системи машинного навчання, включаючи дані навчання, архітектуру моделі та додаткові параметри моделі. Хоча ці атаки діють на основі дуже сильних припущень, головною причиною їх аналізу є перевірка вразливості системи від найгірших зловмисників і оцінка потенційних засобів пом’якшення.

● Атаки чорної скриньки. Ці атаки передбачають мінімальні знання про систему ML. Зловмисник може отримати доступ для запитів до моделі, але він не має іншої інформації про те, як модель навчена. Ці атаки є найбільш практичними, оскільки вони припускають, що зловмисник не знає системи ШІ та використовує системні інтерфейси, що легко доступні для звичайного використання.

● Атаки сірої скриньки. Існує цілий ряд атак сірої скриньки, які фіксують суперечливі знання між атаками чорної скриньки та білої скриньки. В роботі [17] представлено структуру для класифікації атак сірого ящика. Зловмисник може знати архітектуру моделі, але не знати її параметри, або зловмисник може знати модель та її параметри, але не знати навчальні дані. Інші поширені припущення для атак сірого ящика полягають у тому, що зловмисник має доступ до даних, розподілених ідентично навчальним даним, і знає представлення функції. Останнє припущення важливе в додатках, де вилучення функцій використовується перед навчанням моделі ML, таких як кібербезпека, фінанси та охорона здоров’я.

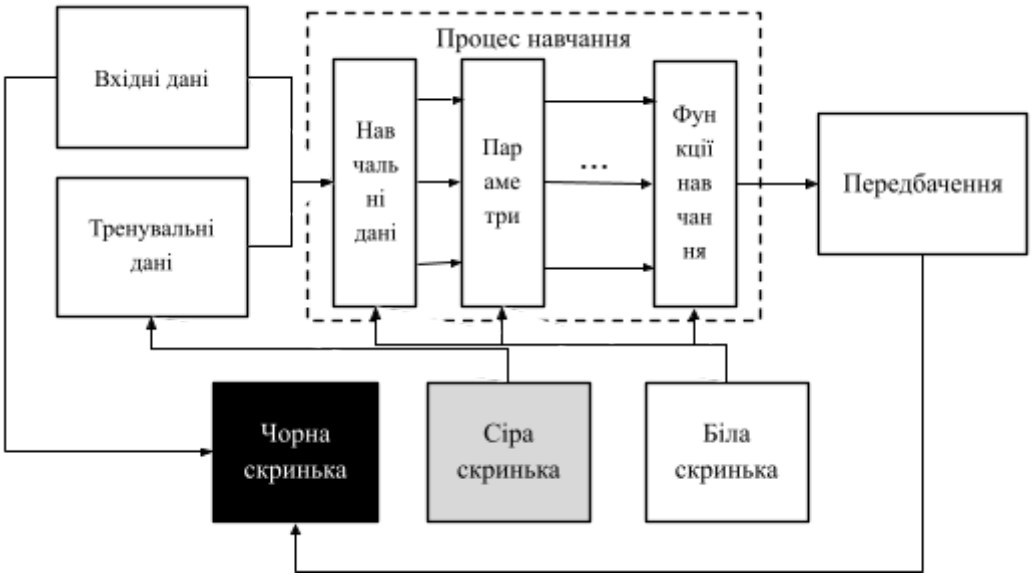

Рис. 2 – Схема ступеню обізнаності зловмисника
Fig. 2 – Scheme of the degree of awareness of the attacker

Тобто, загально кажучи, з точки зору інформації, якщо зловмисник має повні знання про модель, такі як параметри, функції та навчальні дані, ми говоримо про атаку білого ящика. І навпаки, якщо зловмисник не має жодних знань про внутрішню роботу моделі та має лише доступ до її прогнозів, ми називаємо це атакою чорної скриньки. Все, що знаходиться між цими двома, потрапляє в категорію сірої скриньки [22]. Схематично це зображено на рис. 2.

На практиці зловмисник часто починає з точки зору чорної скриньки та намагається підвищити свої знання, наприклад, виконуючи логічні висновки або оракул-атаки, коли зловмисник запитує модель, щоб отримати підказки про внутрішні елементи моделі або дані навчання. Часто конфіденційну інформацію про цільову модель можна отримати більш традиційними засобами, такими як розвідка з відкритим кодом (OSINT), соціальна інженерія, кібершпигунство тощо.

3. Атака отруєння даних

Атаки на етапі навчання ML називаються атаками отруєння [1, 9]. Під час атаки отруєння даних [5, 9] зловмисник контролює підмножину навчальних даних, вставляючи або змінюючи навчальні зразки. У атаці отруєння моделі [10] зловмисник контролює модель та її параметри. Атаки з отруєнням даних можуть застосовуватися до всіх парадигм навчання, тоді як атаки з отруєнням моделі є найбільш поширеними у федеративному навчанні [11], де клієнти надсилають локальні оновлення моделі на сервер, що обробляє вхідні дані, і в атаках на ланцюг поставок, де шкідливий код може бути доданий до моделі постачальниками технології моделі. Під федеративним навчанням тут мається на увазі метод машинного навчання, орієнтований на умови, в яких кілька суб'єктів (часто званих клієнтами) спільно навчають модель, при цьому дані, які використовуються для навчання, розподіляються децентралізовано. Це відрізняє його від машинного навчання, в якому дані зберігаються централізовано.

Перші атаки отруєння, виявлені в додатках кібербезпеки, були атаками на доступність проти генерації профілів хробака та класифікаторів спаму, які невибірково впливають на всю модель машинного навчання та, по суті, спричиняють атаку типу «відмова в обслуговуванні» для користувачів системи ШІ.

Атаки отруєння вважаються одними з найнебезпечніших серед атак на ШІ та можуть спричинити або порушення доступності, або порушення цілісності. Зокрема, атаки з порушенням доступності спричиняють деградацію моделі машинного навчання на всіх етапах, тоді як цільові та бекдорні атаки з отруєнням є більш прихованими та викликають порушення цілісності на невеликому наборі даних. Атаки отруєння використовують широкий спектр конкурентних можливостей, таких як отруєння даних, отруєння моделі, контроль міток, контроль вихідного коду та контроль тестових даних, що призводить до кількох підкатегорій атак отруєння. За моделлю загрози вони можуть використовуватись як у сценаріях білої скриньки, так і чорної скриньки [18], що були розглянуті в розділі 1.3 цієї статті.

Серед методів запобігання атакам отруєння даних виділяють два найефективніші [1]:

● Очищення навчальних даних. Ці методи використовують той факт, що отруєні набори зазвичай відрізняються від звичайних навчальних наборів, які не контролюються зловмисниками. Таким чином, методи очищення даних призначені для очищення навчального набору та видалення отруєних наборів перед виконанням навчання машинного навчання.

● Надійне навчання. Альтернативним підходом до пом'якшення атак з порушенням доступності є модифікація алгоритму навчання ML і проведення надійного навчання замість звичайного навчання. У кількох статтях визначено методи надійної оптимізації, такі як використання функції скорочених втрат [20] або випадкове згладжування для додавання шуму під час навчання [21].

4. Атака ухилення

Evasion Attacks (атаки ухилення) – це тип кібератак, при яких зловмисники модифікують вхідні дані, щоб обійти систему машинного навчання та спричинити неправильну класифікацію або прийняття хибних рішень. Такі атаки зазвичай відбуваються на етапі використання моделі, коли вона вже навчена та розгорнута. Зловмисники вносять незначні, часто непомітні для людини зміни у вхідні дані, які, однак, суттєво впливають на результат роботи моделі.

Серед типів атак ухилення за класифікацією NIST AI 100-2e2025 [25] можна виділити наступні:

1. Атаки з використанням градієнта (Gradient-based attacks): Зловмисники використовують інформацію про градієнти функції втрат моделі для визначення найефективніших змін вхідних даних, які призведуть до помилкової класифікації [22].
2. Атаки на основі балів (Score-based attacks): Атакуючі отримують доступ до оцінок впевненості моделі (confidence scores) і використовують методи оптимізації для створення підроблених прикладів, які модель класифікує неправильно.
3. Атаки на основі рішень (Decision-based attacks): Зловмисники мають доступ лише до кінцевих рішень моделі (наприклад, міток класів) і застосовують методи оптимізації для створення підроблених прикладів, які змушують модель робити помилки.
4. Атаки переміщення (Transfer attacks) [23]: Атакуючі тренують замінну модель, генерують на ній підроблені приклади та переносять ці атаки на цільову модель, використовуючи схожість між моделями.

Атаки ухилення становлять серйозну загрозу для систем кібербезпеки. Зловмисники можуть змінювати характеристики шкідливого програмного забезпечення, щоб обійти антивірусні системи, або модифікувати мережевий трафік для уникнення виявлення системами вторгнень.

На даний момент основними методами захисту [1] від атак ухилення є:

1. Змагальне навчання (Adversarial training): Включення підроблених прикладів у процес навчання моделі для підвищення її стійкості до атак.
2. Використання ансамблів моделей (Ensemble methods) [24]: Комбінування кількох моделей для зменшення ймовірності успішної атаки на всі моделі одночасно.
3. Моніторинг та оновлення (Continuous monitoring and updating): Регулярне оновлення моделей та систем виявлення для адаптації до нових стратегій атак та покращення стійкості.

Тим не менш, ці методи мають різні обмеження, такі як знижена точність для змагального навчання та випадкового згладжування, а також обчислювальна складність для формальних методів. Тому потрібно завжди шукати компроміс між надійністю та точністю. Розуміння та впровадження цих методів захисту є критично важливим для забезпечення безпеки та надійності систем машинного навчання в умовах зростаючих загроз атак ухилення.

5. Атака на конфіденційність

Атаки на конфіденційність (Privacy Attacks) – це тип кібератак, спрямованих на отримання конфіденційної інформації з моделей штучного інтелекту (ШІ), їхніх навчальних даних або вихідних даних. Ці атаки можуть бути використані для крадіжки персональних даних, компрометації комерційної інформації або зламу моделей машинного навчання.

Нижче розглянемо основні типи атак на конфіденційність:

1. Атаки з відновленням даних (Data Reconstruction Attacks): Зловмисники намагаються відтворити вихідні дані, використовуючи доступ до моделі ШІ або її вихідних

результатів. Це може статися, коли модель видає занадто детальні відповіді або має вразливості, що дозволяють відновити частини навчальних даних.

2. Атаки з інверсією моделі (Model Inversion Attacks): У цьому випадку зловмисники використовують вихідні дані моделі для відтворення вхідних даних або характеристик, що використовувалися під час навчання. Це може розкрити конфіденційну інформацію про сторін, представлених у навчальних даних.
3. Атаки з визначення принадності (Membership Inference Attacks): Зловмисники намагаються визначити, чи були конкретні записи включені в навчальний набір даних моделі. Це може бути використано для розкриття участі особи в певних заходах або її належності до певних груп.
4. Атаки на основі метаданих (Metadata-based Attacks): Навіть, якщо самі дані залишаються захищеними, зловмисники можуть використовувати метадані (наприклад, час доступу, розмір файлів) для отримання конфіденційної інформації або встановлення патернів, які можуть бути використані в подальших атаках.
5. Атаки через бічні канали (Side-channel Attacks): Зловмисники можуть аналізувати поведінку системи ШІ, наприклад, час відгуку або споживання енергії, щоб отримати інформацію про внутрішні процеси або дані моделі.

Для захисту від таких атак рекомендується:

- Підвищення анонімності та агрегування даних. Використання методів, які зменшують ризик ідентифікації індивідуальних записів у навчальних даних.
- Федеративне навчання (Federated Learning) – навчання моделей на локальних пристроях без передачі даних на сервер.
- Диференційна приватність. Додавання контрольованого шуму до даних або результатів моделі, щоб запобігти відновленню вихідних даних без значного впливу на точність моделі.
- Обмеження доступу та моніторинг: Контроль доступу до моделей та даних, а також постійний моніторинг використання для виявлення підозрілої активності.
- Оцінка вразливостей: Регулярне тестування моделей на стійкість до атак на конфіденційність та впровадження відповідних заходів захисту.

Атаки на конфіденційність є серйозною загрозою для систем ШІ, в тому числі і в сфері кібербезпеки. Тому захист конфіденційності в системах ШІ є критично важливим для збереження довіри користувачів та дотримання нормативних вимог. Зловмисники використовують різні методи, щоб отримати доступ до конфіденційних даних або параметрів моделей. Захист таких систем вимагає комплексного підходу, включаючи сучасні криптографічні методи, моніторинг активності та підвищення обізнаності користувачів про ризики.

6. Атака на зловживання

Ще одним типом атак на системи ШІ є атаки зловживання (abuse attacks) [25]. Ці атаки спрямовані на зловживання або маніпуляцію системами штучного інтелекту (ШІ) з метою отримання небажаних або шкідливих результатів. Ці атаки використовують вразливості в структурі або реалізації моделей ШІ, щоб змусити систему поводитися неналежним чином.

Приклади атак зловживання:

1. Використання упередженості моделі (Bias Exploitation): Атака, при якій зловмисник використовує існуючі упередження або слабкі місця в моделі ШІ, щоб отримати певні результати або посилити дискримінаційні тенденції.
2. Зловживання функціональністю (Functionality Misuse): Маніпуляція системою ШІ для

виконання дій, які не були передбачені розробниками, наприклад, використання чат-бота для генерації небажаного контенту або спаму.

3. Атаки на основі підказок (Prompt Injection Attacks): Введення спеціально створених запитів або команд, які змушують модель ШІ генерувати небажаний або шкідливий контент.

Для упередження таким атакам NIST рекомендує наступні заходи безпеки:

- Удосконалення алгоритмів виявлення аномалій – розвиток механізмів, які можуть розпізнавати підозрілі взаємодії зі ШІ.
- Жорсткіші політики перевірки даних – аналіз вхідних даних для виявлення можливих маніпуляцій.
- Підвищення прозорості ШІ-систем – покращення документації процесів прийняття рішень у моделях.
- Механізми протидії зловмисного проникнення – наприклад, введення додаткових рівнів перевірки в моделях безпеки.
- Розробка стандартів етичного використання ШІ – активне регулювання та контроль за впровадженням таких технологій.

Таким чином, атаки зловживання є серйозною загрозою для систем ШІ, оскільки вони дозволяють зловмисникам використовувати їх у несподіваний спосіб. Використання ШІ для автоматизації шахрайства, маніпуляції суспільною думкою або обходу обмежень створює нові виклики для кібербезпеки. Запобігання таким атакам вимагає комплексного підходу, включаючи вдосконалення алгоритмів безпеки, розробку політик відповідального використання ШІ та постійний моніторинг загроз.

7. Загальний підсумок щодо атак на системи ШІ

В розділах 2-5 цієї статті було розглянуто 4 найвпливовіші типи атак на системи ML/AI. Підсумки аналізу розглянутих атак наведено в таблиці 1 нижче.

Таблиця 1 – Порівняльна характеристика атак на AI/ML-системи
Table 1 – Comparative characteristics of attacks on AI/ML systems

Тип атаки	Мета атаки	Фаза атаки	Методи	Наслідки
Poisoning Attack (атака отруєння даних)	Вплив на якість навчання	Навчання моделі	Додавання шкідливих даних до навчального набору	Погіршення якості прийнятих рішень, хибне спрацьовування, зниження безпеки для подальших атак
Evasion Attack (атака ухилення)	Обхід механізмів безпеки моделі	Виконання моделі	Маніпуляція вихідними даними, генерація спеціальних зловмисних даних	Обхід захисних механізмів, зниження точності моделі
Privacy Attack (атака на конфіденційність)	Викрадення даних, на яких навчалася модель	Виконання моделі	Аналіз відповідей моделі, відновлення даних	Витік конфіденційних даних
Abuse Attacks (атаки зловживання)	Використання ШІ для створення шкідливих даних	Після розгортання моделі	Генеративні моделі для атак, маніпуляції	Маніпуляція інформацією, автоматизація шахрайства

Таким чином, було розглянуто:

Атаки отруєння даних – проводяться на етапі навчання та можуть довгостроково впливати на точність моделі, змушуючи її робити неправильні висновки.

Атаки ухилення – це атаки на фазі виконання, коли зловмисники намагаються обдурити модель, вводячи в неї спеціально створені дані.

Атаки на конфіденційність – спрямовані на витяг конфіденційних даних із моделі, що ставить під загрозу безпеку особистої інформації користувачів.

Атаки зловживання – пов'язані з неправомірним використанням можливостей ШІ, наприклад, для створення фейкових відео, зловмисного коду чи маніпуляцій у соціальних мережах.

8. Висновки

1. ШІ є гарним інструментом для автоматизації процесів виявлення та реагування на атаки та загрози, і значно підвищує ефективність захисту систем та компаній. Використання алгоритмів машинного навчання сприяє швидкому аналізу великих обсягів даних та ідентифікації аномалій у поведінці користувачів і систем.

2. Проте ШІ не тільки забезпечує ефективний захист, а й створює нові загрози через його можливе використання зловмисниками. Саме тому питання щодо виявлення атак та протидії їм є дуже важливим питанням в сучасному кіберпросторі. Тому у цій статті було розглянуто та проведено всебічний аналіз атак на сучасні моделі ML/AI.

3. Надійність системи ШІ залежить від усіх атрибутів, які її характеризують. Наприклад, система штучного інтелекту, яка є точною, але легко сприйнятливою до агресивних дій, навряд чи заслуговує на довіру. Так само навряд чи можна довіряти системі ШІ, яка дає шкідливо упереджені або несправедливі результати, навіть, якщо вона надійна. Існують також компроміси між прозорістю та конкурентоспроможністю. На жаль, неможливо одночасно максимізувати продуктивність системи ШІ щодо цих атрибутів. Наприклад, системи ШІ, оптимізовані тільки для точності, як правило, мають недостатню ефективність з точки зору конкурентної надійності та справедливості. І навпаки, система ШІ, оптимізована для змагальності, може продемонструвати нижчу точність і погіршити результати надійності.

4. Атаки отруєння даних є найнебезпечнішим видом атак на основі ШІ у довгостроковій перспективі, оскільки вони впливають на саму модель і можуть лишатися непомітними протягом тривалого часу.

5. Атаки на конфіденційність та атаки зловживання особливо небезпечні через можливість витоку конфіденційних даних та створення шкідливого контенту.

6. В результаті дослідження було виявлено загальні наслідки атак на основі ШІ:

- Підрив довіри до ШІ – постійні атаки та маніпуляції можуть зробити ШІ менш надійним інструментом для ухвалення рішень.
- Витік конфіденційної інформації – атаки на конфіденційність призводять до масштабних втрат персональних та корпоративних даних.
- Обхід засобів захисту – ухилення та отруєння моделей ставлять під загрозу сучасні системи кібербезпеки, знижуючи їхню ефективність.
- Масштабованість атак – зловмисники можуть використовувати ШІ для автоматизації та прискорення атак, що збільшує їхній вплив.
- Ризики на державному рівні – атаки на основі ШІ можуть загрожувати національній безпеці, економіці та критичній інфраструктурі.

7. Тому наразі основними шляхами захисту від атак на основі ШІ є:

- Розробка стійких ШІ-моделей, які менш вразливі до отруєння або ухилення.

- Впровадження механізмів виявлення атак, таких як моніторинг змін у навчальних даних та алгоритмах.
- Підвищення прозорості ШІ – створення пояснюваних моделей, які можна перевірити на наявність маніпуляцій.
- Посилення регулювання та стандартів – держави та організації повинні встановлювати правила щодо використання ШІ.

8. Кожен розглянутий тип атак на системи ШІ має різні механізми впливу, але всі вони можуть суттєво знизити ефективність та безпеку моделей машинного навчання. Захист від таких атак вимагає комплексного підходу. Окрім визначених вище методів протидії атакам, важливим залишається людський фактор – навчання спеціалістів та користувачів, адже багато атак базуються саме на соціальній інженерії.

9. У більшості випадків організаціям доведеться прийняти компроміс між бажаними властивостями та вирішити, яким із них віддати пріоритет залежно від системи ШІ, варіанту використання та потенційно багатьох інших міркувань щодо економічних, екологічних, соціальних, культурних, політичних і глобальних наслідків технології ШІ.

10. Важливо зазначити, що з розвитком технологій ШІ з’являються нові типи атак, тому постійний моніторинг та оновлення знань у цій сфері є необхідними для забезпечення кібербезпеки.

11. Підсумовуючи важливо зазначити, що безпечне використання ШІ у кібербезпеці є балансом між технологічними інноваціями та загрозами, що виникають внаслідок їхнього розвитку. Розробка адаптивних методів захисту та вдосконалення моделей машинного навчання допоможуть зменшити ризики, пов’язані з атаками, та забезпечити надійний рівень кіберзахисту в майбутньому.

Конфлікт інтересів

Автори повідомляють про відсутність конфлікту інтересів.

References

1. Vassilev A., Oprea A., Fordyce A., Anderson H. (2024) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) *NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023*. – Access mode: https://doi.org/10.6028/NIST.AI.100-2e2023
2. Booth H., Souppaya M., Vassilev A., Ogata M., Stanley M., Scarfone K. (2024) Secure Development Practices for Generative AI and Dual-Use Foundation AI Models: An SSDF Community Profile. (National Institute of Standards and Technology, Gaithersburg, MD), *NIST Special Publication (SP) NIST SP 800-218A*. – Access mode: https://doi.org/10.6028/NIST.SP.800-218A
3. Oprea A., Singhal A. and Vassilev A.. (2022) Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy?, *in Computer*, vol. 55, no. 11, pp. 94-99, Nov. URL: https://doi.org/10.1109/MC.2022.3190787
4. Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin’ichi Satoh, Luc Van Gool, Zheng Wang. (2024) Physical Adversarial Attack Meets Computer Vision: A Decade Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.46, no.12, pp.9797-9817, URL: https://doi.org/10.48550/arXiv.2209.15179
5. Koundinya A., Patil S., Chandu B. (2024) Data Poisoning Attacks in Cognitive Computing, *IEEE 9th International Conference for Convergence in Technology (I2CT)*, pp.1-4, https://doi.org/10.1109/I2CT61223.2024.10544345
6. National Institute of Standards and Technology. (2023) *Artificial Intelligence Risk Management Framework. (AI RMF 1.0)*. – Access mode: https://doi.org/10.6028/NIST.AI.100-1

7. Biggio B., Roli F. (2018) Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition,* 84:317–331 https://doi.org/10.48550/arXiv.1712.03141
8. Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. (2018) When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In 27th USENIX Security Symposium (USENIX Security 18), pp. 1299–1316, https://www.usenix.org/conference/usenixsecurity18/presentation/suciu
9. Biggio B., Nelson B., Laskov P. (2012) Poisoning attacks against support vector machines. *In Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML*, URL: https://doi.org/10.48550/arXiv.1206.6389
10. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. (2018) Trojaning attack on neural networks. *In NDSS*. The Internet Society, URL: https://dx.doi.org/10.14722/ndss.2018.23291
11. Kairouz, Peter; McMahan, H. Brendan; Avent, Brendan; Bellet, Aurélien; Bennis, Mehdi; Bhagoji, Arjun Nitin; Bonawitz, Kallista; Charles, Zachary; Cormode, Graham (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning 14 (1–2)*: https://doi.org/10.1561/2200000083. ISSN 1935-8237
12. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. (2014) Intriguing properties of neural networks. *In International Conference on Learning Representations*, URL: https://doi.org/10.48550/arXiv.1312.6199
13. Ian Goodfellow, Jonathon Shlens, Christian Szegedy. (2015) Explaining and harnessing adversarial examples. *In International Conference on Learning Representations*, URL: https://doi.org/10.48550/arXiv.1412.6572
14. Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, Dawn Song. (2019) The Secret Sharer: Evaluating and testing unintended memorization in neural networks. *In USENIX Security Symposium, USENIX 19*, pages 267–284. – URL: https://arxiv. org/abs/1802.08232
15. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert - Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel. (2021) Extracting training data from large language models. *In 30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, URL: https://doi.org/10.48550/arXiv.2012.07805
16. Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. (2018) Property inference attacks on fully connected neural networks using permutation invariant representations. *In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18*, page 619–633, New York, NY, USA. Association for Computing Machinery. URL: https://doi.org/10.1145/3243734.3243834
17. Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, Tudor Dumitras.(2018) When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. *In 27th USENIX Security Symposium (USENIX Security 18)*, pages 1299–1316. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/suciu
18. Battista Biggio, Blaine Nelson, Pavel Laskov. (2012) Poisoning attacks against support vector machines. *In Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML*, URL: https://doi.org/10.48550/arXiv.1206.6389
19. Nihad Hassan. (2024) What is data poisoning (AI poisoning) and how does it work? *Search Enterprise AI, TechTarget*. – URL: https://www.techtarget.com/searchenterpriseai/definition/data-poisoning-AI-poisoning
20. Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. (2019) Sever: A robust meta-algorithm for stochastic optimization. *In International Conference on Machine Learning*, pages 1596–1606. PMLR, URL: https://doi.org/10.48550/arXiv.1803.02815
21. Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. (2020) Certifed robustness to label-flipping attacks via randomized smoothing. *In International Conference on Machine Learning*, pages 8230–8241. PMLR, URL: https://doi.org/10.48550/arXiv.2002.03018
22. The Tactics & Techniques of Adversarial Machine Learning. HiddenLayer/ (2022). – URL: https://hiddenlayer.com/innovation-hub/the-tactics-and-techniques-of-adversarial-ml
23. Chi Zhang, Zifan Wang, Ravi Mangal, Matt Fredrikson, Limin Jia, Corina Pasareanu. (2023) Transfer Attacks and Defenses for Large Language Models on Coding Tasks. – URL: https://doi.org/10.48550/arXiv.2311.13445

24. D. Li and Q. Li,(2023) Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection, in *IEEE Transactions on Information Forensics and Security*. – URL: https://doi.org/10.48550/arXiv.2006.16545
25. Vassilev A, Oprea A, Fordyce A, Anderson H (2025) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. *(National Institute of Standards and Technology, Gaithersburg, MD) NIST Artificial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2025*. – URL: https://doi.org/10.6028/NIST.AI.100-2e2025

RESEARCH AND CLASSIFICATION OF THE MAIN TYPES OF ATTACKS ON ARTIFICIAL INTELLIGENCE SYSTEMS IN CYBERSECURITY

Vladyslav Vilihura[1], PhD, Senior Lecture; e-mail: v.v.vilihura@karazin.ua;
ORCID: https://orcid.org/0000-0002-1137-2382
Yelyzaveta Ostrianska[1], junior researcher; e-mail: antelizza@gmail.com;
ORCID: https://orcid.org/0000-0003-1412-8470
[1] *V. N. Karazin Kharkiv National University, Ukraine*
Manuscript was received November 1, 2024; Received after review December 2, 2024;
Accepted December 23, 2024

Abstract. The modern development of artificial intelligence (AI) and machine learning (ML) opens up new opportunities in the field of cybersecurity, but at the same time creates serious challenges in the form of intelligent cyberattacks. The study is devoted to the analysis and classification of ways to use AI for malicious purposes and the study of effective methods to counter such threats. In particular, the article covers the main types of attacks using ML technologies, which demonstrate how attackers can manipulate machine learning algorithms, undermine trust in data, and bypass protection systems. Special attention is paid to the mechanisms of data poisoning attacks, as they are considered the most influential in machine learning, which involve introducing malicious data into the process of training models, which leads to distortion of results and undermines the effectiveness of security algorithms. Privacy attacks are analyzed as a way to obtain confidential information from ML models, which can be used to steal user data. Abuse attacks demonstrate how attackers can use AI tools to automate attacks, scale phishing campaigns, and analyze vulnerabilities in defense systems. The relevance of the study is due to the fact that traditional approaches to cyber defense are no longer able to effectively counter threats that adapt and evolve due to machine learning. The article emphasizes the critical importance of researching defense methods, in particular, building reliable machine learning systems that have built-in mechanisms for detecting anomalies and adapting to new threats. One of the key approaches is federated learning, which allows training models without centralized data storage, reducing the risk of information leakage. The development of deep learning in the field of cyber defense is also considered, which allows analyzing behavioral patterns of threats in real time. The combination of technological measures with human control remains an important aspect, since, despite the power of AI tools, the human factor remains key in the process of ensuring cybersecurity. Thus, the article demonstrates the balance between the opportunities and threats of AI in the field of cybersecurity, emphasizing the need for further research in the direction of resilient ML models that can effectively resist attacks. Without proper regulation and control, AI can become not only a defender, but also a tool for attackers, which requires the development of new security strategies and international regulation in the field of cybersecurity.

Keywords: *artificial intelligence, cyberattacks, machine learning, cybersecurity, federated learning*

Conflicts of Interest: the authors declare no conflict of interest.