Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя
(повне найменування вищого навчального закладу)
Факультет комп'ютерно-інформаційних систем і програмної інженерії
(назва факультету)
Кафедра кібербезпеки
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(освітній рівень)

на тему: «Аналіз вразливостей штучного інтелекту та методів їхнього використання в кіберзагрозах»

Виконав: студент (ка)

Спеціальності:
125 «Кібербезпека та захист інформації»
(шифр і назва напряму підготовки, спеціальності)
Мазяр Анастасія Миколаївна
підпис (прізвище та ініціали)

Керівник Оробчук О. Р.
підпис (прізвище та ініціали)

Нормоконтроль Тимощук Д. І.
підпис (прізвище та ініціали)

Завідувач кафедри Загородна Н.В.
підпис (прізвище та ініціали)

Рецензент
підпис (прізвище та ініціали)

м. Тернопіль – 2024

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра кібербезпеки
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Загородна Н.В.
(підпис) (прізвище та ініціали)
«__» _____ 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 125 Кібербезпека та захист інформації
(шифр і назва спеціальності)

Студенту Мазяр Анастасії Миколаївній
(прізвище, ім'я, по батькові)

1. Тема роботи Аналіз вразливостей штучного інтелекту та методів їхнього використання в кіберзагрозах

Керівник роботи Оробчук Олександра Романівна, доктор філософії, старший викладач кафедри КБ
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «20» 11 2024 року № 4/7-1112

2. Термін подання студентом завершеної роботи

3. Вихідні дані до роботи Дослідження вразливостей ШІ, методи їх використання в кіберзагрозах а також розробку і оцінку методів захисту для підвищення безпеки ШІ.

4. Зміст роботи (перелік питань, які потрібно розробити)
Розділ 1. Аналіз вразливостей ШІ та їх використання у кіберзагрозах 1.1 Поняття вразливостей ШІ 1.2 класифікація вразливостей:технічні, алгоритмічні та організаційні 1.3 Приклади використання вразливостей для здійснення кіберзагроз Розділ 2. Аналіз методів експлуатації вразливостей ШІ 2.1 Соціальна інженерія та маніпуляція на основі АІ 2.2 Атаки на системи машинного навчання 2.2.1 Атаки на навчальні дані 2.2.2 Атаки на моделі 2.3 Роль АІ в автоматизації кіберзагроз Розділ 3 Методи захисту від експлуатації вразливостей АІ 3.1 Розробка систем виявлення аномалій у використанні А.І 3.2 Використання криптографічних методів для захисту даних і моделей АІ 3.3 Тестування та оцінка безпеки моделей.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
Актуальність теми. Мета і задачі досліджень. Методологія досліджень. Загальний огляд вразливостей штучного інтелекту. Класифікація вразливостей штучного інтелекту. Використання вразливостей штучного інтелекту у кіберзагрозах. Типи кіберзагроз та реальні приклади атак. Методи захисту від вразливостей штучного інтелекту. Практичне Застосування розроблених методів захисту в реальних системах.
Висновки.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Осухівська Г. М., к.т.н., доцент		
Безпека в надзвичайних ситуаціях	Клепчик В.М., проректор з адміністративно-господарської роботи та будівництва		

7. Дата видачі завдання 23.09.2024

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	23.09	Виконано
2.	Підбір джерел для аналізу вразливостей штучного інтелекту та їх використання в кіберзагрозах	24.09-03.10	Виконано
3.	Опрацювання джерел в галузі дослідження	04.10 – 12.10	Виконано
4.	Аналіз типів вразливостей штучного інтелекту та їх впливу на безпеку	13.10 – 20.10	Виконано
5.	Оцінка конкретних прикладів використання вразливостей у реальних кіберзагрозах	21.10 – 27.10	Виконано
6.	Розробка методів захисту від експлуатації вразливостей штучного інтелекту	28.10 – 03.11	Виконано
7.	Оформлення розділу «Аналіз вразливостей ШІ та їх використання в кіберзагрозах»	04.11 – 10.11	Виконано
8.	Оформлення розділу «Аналіз методів експлуатації вразливостей ШІ»	11.11 – 17.11	Виконано
9.	Оформлення розділу «Методи захисту від експлуатації вразливостей AI»	18.11 – 24.11	Виконано
10.	Виконання завдання до підрозділу «Охорона праці та безпека в надзвичайних ситуаціях»	25.11 – 29.11	Виконано
11.	Оформлення кваліфікаційної роботи	30.11 – 08.12	Виконано
12.	Нормоконтроль	09.12 – 11.12	Виконано
13.	Перевірка на плагіат	12.12 – 14.12	Виконано
14.	Попередній захист кваліфікаційної роботи	15.12 – 19.12	Виконано
15.	Захист кваліфікаційної роботи	26.12	Виконано

Студент _____ Мазяр А. М.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Оробчук О. Р.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Аналіз вразливостей штучного інтелекту та методів їхнього використання в кіберзагрозах // ОР «Магістр» // Мазяр Анастасія Миколаївна // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра кібербезпеки, група СБм-61 // Тернопіль, 2024 // С. 60, рис. – -, табл. – - , кресл. – - , додат. – 2 .

Ключові слова: AI, ML, CYBER, AIoT, Phishing, RNN, DoS.

Кваліфікаційна робота присвячена аналізу вразливостей штучного інтелекту та методів їх використання в кіберзагрозах. Вивчено різні типи вразливостей, включаючи технічні, алгоритмічні та організаційні, а також їх потенціал у створенні кіберзагроз.

Робота включає аналіз конкретних прикладів використання вразливостей ШІ в реальних атаках, таких як маніпуляції через соціальну інженерію, атаки на дані та моделі машинного навчання. Розроблено методи захисту, включаючи системи виявлення аномалій, використання криптографії та оцінку безпеки моделей перед їх впровадженням. Результати дослідження мають практичне значення для спеціалістів з кібербезпеки, оскільки дозволяють підвищити безпеку ШІ-систем і запобігти використанню їх вразливостей у кіберзагрозах.

ABSTRACT

Analysis of vulnerabilities of artificial intelligence and methods of their use in cyber threats // Master's degree program // Anastasia Maziar // Ivan Pulyuy Ternopil National Technical University, Faculty of Computer Information Systems and Software Engineering, Department of Cyber Security, СБм-61 group / / Ternopil, 2024 // S. 60, fig. – -, tab. – - , chair. – - , add. - 2.

Keywords: AI, ML, CYBER, AIoT, Phishing, RNN, DoS.

The qualification thesis is dedicated to the analysis of artificial intelligence vulnerabilities and the methods of their exploitation in cyber threats. Various types of vulnerabilities are studied, including technical, algorithmic, and organizational, as well as their potential in creating cyber threats.

The work includes an analysis of specific examples of AI vulnerabilities used in real-world attacks, such as social engineering manipulations, attacks on data, and machine learning models. Protection methods are developed, including anomaly detection systems, the use of cryptography, and security assessment of models before their implementation. The research results are of practical significance for cybersecurity specialists, as they enhance the security of AI systems and help prevent the exploitation of their vulnerabilities in cyber threats.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

AI	—	Artificial Intelligence
ML	—	Machine Learning
DL	—	Deep Learning
CNN	—	Convolutional Neural Network
RNN	—	Recurrent Neural Network
GAN	—	Generative Adversarial Network
DoS	—	Denial of Service
IoT	—	Internet of Things
IDS	—	Intrusion Detection System

ВСТУП

Актуальність теми.

Стрімкий розвиток технологій штучного інтелекту (ШІ) відкриває нові можливості у різних сферах, проте водночас створює ризики, пов’язані з їх експлуатацією в кіберзагрозах. Вразливості алгоритмів ШІ, зокрема в машинному та глибокому навчанні, можуть бути використані зловмисниками для компрометації систем, порушення конфіденційності даних або здійснення автоматизованих атак. Актуальність теми обумовлена необхідністю детального аналізу цих вразливостей та розробки методів їх попередження, що є важливим завданням для забезпечення безпеки сучасних інформаційних систем.

Мета і задачі дослідження.

Метою дослідження є аналіз вразливостей штучного інтелекту та методів їх використання у кіберзагрозах, а також розробка підходів до виявлення та запобігання таким загрозам.

Завданнями дослідження є:

1. Провести класифікацію вразливостей ШІ та методів їх експлуатації.

2. Дослідити існуючі приклади кіберзагроз, що використовують алгоритми ШІ.

3. Розробити рекомендації для захисту систем, заснованих на ШІ.

Об’єкт дослідження.

Вразливості штучного інтелекту та їхній вплив на безпеку інформаційних систем.

Предмет дослідження.

Методи використання вразливостей ШІ у кіберзагрозах і підходи до їхнього виявлення та запобігання.

Наукова новизна одержаних результатів кваліфікаційної роботи.

У роботі вперше проведено комплексний аналіз методів експлуатації вразливостей ШІ у кіберзагрозах. Запропоновано нові підходи до захисту, що базуються на аналізі аномалій у роботі моделей ШІ та використанні криптографічних засобів для забезпечення їхньої стійкості.

Практичне значення одержаних результатів.

Результати дослідження можуть бути використані для вдосконалення методів кіберзахисту у державних, корпоративних та академічних структурах, що використовують технології ШІ.

Публікації.

Основні результати кваліфікаційної роботи опубліковано у працях конференції (див. Додаток А).

РОЗДІЛ 1 АНАЛІЗ ВРАЗЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ТА ЇХ ВИКОРИСТАННЯ У КІБЕРЗАГРОЗАХ

1.1 Поняття вразливостей штучного інтелекту

Вразливості штучного інтелекту (ШІ) стосуються аспектів його роботи, які можуть бути використані зловмисниками або призвести до небажаних наслідків через помилки, недоліки або недосконалості в дизайні, реалізації чи експлуатації. Ці вразливості виникають на різних рівнях функціонування систем ШІ: від алгоритмічних до операційних, що потребує особливої уваги до їх ідентифікації та усунення.

У системах штучного інтелекту (ШІ) вразливості можуть виникати на різних рівнях: від технічних аспектів реалізації до організаційних процесів і алгоритмічних недоліків. Ці вразливості можуть створювати значні ризики для безпеки, ефективності та етичності використання ШІ-технологій. Відповідно, важливо розглядати їх у контексті трьох основних категорій: технічних, алгоритмічних та організаційних вразливостей. Кожен з цих типів вразливостей має свої специфічні особливості, способи виявлення та усунення.

1.2 Класифікація вразливостей: технічні, алгоритмічні та організаційні

Технічні вразливості виникають через проблеми в апаратному забезпеченні, програмному забезпеченні, мережевій безпеці чи інтерфейсах між системами. Вони можуть мати суттєвий вплив на функціонування ШІ, знижуючи його ефективність або навіть ставлячи під загрозу цілісність даних і системи в цілому.

Вразливості в програмному забезпеченні.

Технічні вразливості в програмному забезпеченні можуть виникати через дефекти в коді або неправильну реалізацію алгоритмів. Такі помилки можуть бути використані зловмисниками для отримання доступу до системи або її маніпулювання. Наприклад, баги в системах, які не враховують певні сценарії або екстремальні умови, можуть призвести до аварійних ситуацій.

Вразливості в мережевій безпеці.

Мережеві вразливості стосуються проблем у забезпеченні передачі даних через канали зв'язку. Вони можуть включати такі загрози, як перехоплення даних (атаки "man-in-the-middle"), віддалене виконання коду або злом паролів і інших аутентифікаційних механізмів.

Вразливості в апаратному забезпеченні.

Порушення безпеки на рівні апаратних засобів можуть виникнути через уразливості в фізичних компонентах, таких як процесори або чіпи пам'яті. Такі уразливості можуть призвести до несанкціонованого доступу до системи або її перевантаження.

Алгоритмічні вразливості виникають через недосконаліть самих алгоритмів або моделей, що використовуються в системах ШІ. Ці вразливості можуть бути пов'язані з упередженістю в даних, нестабільністю або навіть зловмисними атаками на моделі.

Упередженість в алгоритмах.

Алгоритми ШІ можуть проявляти упередженість через використання неякісних або неповних даних для тренування. Це може призвести до несправедливих або дискримінаційних рішень. Наприклад, модель може демонструвати упередженість за ознакою статі, раси чи віку, що негативно впливає на її ефективність у певних контекстах.

Проблеми з масштабуванням.

Алгоритми можуть не працювати ефективно при великих обсягах даних або при високих навантаженнях. Це може призвести до зниження точності або швидкості обробки, а також до нестабільних результатів.

Атаки на моделі ШІ (Adversarial Attacks).

ШІ-моделі можуть бути піддані атакам, коли зловмисники навмисно змінюють вхідні дані таким чином, щоб модель генерувала неправильні або шкідливі результати. Наприклад, незначні зміни в зображеннях можуть призвести до того, що модель не зможе розпізнати об'єкти або розпізнає їх неправильно.

Організаційні вразливості стосуються управлінських аспектів, які можуть призвести до зниження безпеки, ефективності та етичності роботи ШІ-систем. Це можуть бути проблеми з управлінням даними, політиками безпеки, контролем доступу та реагуванням на інциденти.

Низький рівень підготовки персоналу.

Спеціалісти, які працюють з ШІ-системами, повинні мати високий рівень підготовки та кваліфікації. Невідповідна підготовка персоналу може призвести до помилок у розробці, впровадженні та управлінні ШІ-системами.

Відсутність контролю за даними.

Для належної роботи ШІ необхідно гарантувати, щоб дані, на основі яких модель навчається, були якісними, актуальними та безпечними. Відсутність належного контролю за даними може призвести до помилок у навчанні моделей.

Класифікація вразливостей на технічні, алгоритмічні та організаційні дозволяє комплексно підходити до забезпечення безпеки та ефективності систем ШІ. Усі ці аспекти взаємопов'язані, і для забезпечення надійної роботи ШІ необхідно враховувати кожен з них. Усунення вразливостей на кожному рівні знижує ймовірність виникнення серйозних інцидентів та забезпечує безпечну і ефективну інтеграцію технологій ШІ в суспільство.

1.3 Приклади використання вразливостей для здійснення кіберзагроз

У сучасному світі кіберзагрози є одними з найсерйозніших викликів для безпеки інформаційних систем, включаючи системи штучного інтелекту (ШІ). Зловмисники активно використовують вразливості різних рівнів — технічні, алгоритмічні та організаційні — для реалізації атак, що можуть призвести до серйозних наслідків: втрати даних, компрометації приватності, порушення працездатності критичних систем або навіть економічних збитків. Важливо розглядати, як саме ці вразливості можуть бути використані зловмисниками для здійснення кіберзагроз.

Технічні вразливості є одними з найбільш поширених каналів для реалізації кіберзагроз. Зловмисники можуть використовувати недоліки в програмному забезпеченні, мережевих протоколах або апаратному забезпеченні для несанкціонованого доступу, змінення даних або блокування роботи системи.

Атаки через вразливості в програмному забезпеченні (Exploits).

Програмні вразливості є одними з найбільш популярних векторів атак. Наприклад, уразливості, що дозволяють виконати віддалений код (remote code execution), можуть бути використані для запуску шкідливих програм на ураженому сервері чи іншому пристрої.

Наприклад: Атака через уразливість в бібліотеці OpenSSL (Heartbleed). Зловмисники могли використовувати цю вразливість для отримання чутливих даних, таких як паролі та ключі шифрування, що були збережені в пам'яті серверів, що використовують уразливі версії OpenSSL.

Атаки на мережеві протоколи (Man-in-the-Middle, MITM).

Наприклад: Атака на протокол SSL/TLS, яка дозволяє перехоплювати зашифровані дані між клієнтом і сервером.

Атаки на апаратне забезпечення (Hardware-based Attacks).

Наприклад: Атаки на архітектуру процесорів, як-от Spectre і Meltdown. Ці вразливості дозволяють зловмисникам отримувати доступ до захищених даних, використовуючи особливості виконання інструкцій у сучасних процесорах.

Алгоритмічні вразливості стосуються недоліків у самих методах обробки та аналізу даних. Це включає як упередження в алгоритмах машинного навчання, так і проблеми з безпекою самих моделей. Зловмисники можуть використовувати ці вразливості для маніпуляцій з результатами ШІ-систем або навіть для їхнього злому.

Атаки типу Adversarial (адверсаріальні атаки).

Наприклад: Атаки на системи розпізнавання зображень, де зловмисники можуть додавати невеликі зміни до зображень, які людина не помітить, але система розпізнавання зображень помилково класифікує їх.

Упередженість моделей (Bias in AI Models).

Наприклад: Моделі прогнозування кримінальної поведінки, які виявляють расову упередженість через нерівномірне представлення даних в навчальних вибірках.

Уразливості в алгоритмах безпеки.

Наприклад: Вразливості в алгоритмах шифрування, наприклад, через недосконаліть генерації ключів або проблеми з розшифровкою.

Організаційні вразливості стосуються недоліків в управлінських, юридичних та операційних аспектах безпеки. Вони можуть включати недостатню підготовленість персоналу, відсутність чітких процедур безпеки або помилки в управлінні даними.

Недостатній контроль за доступом до систем.

Наприклад: Атаки, коли зловмисники використовують слабкі паролі або отримують доступ через уразливості в системах аутентифікації, можуть призвести до компрометації всієї мережі.

Невиконання політик безпеки.

Наприклад: Співробітник може використовувати незахищений Wi-Fi для доступу до корпоративних ресурсів, що створює можливість для атак «man-in-the-middle».

Невчасне виявлення та реагування на інциденти.

Наприклад: Організація може не виявити спробу фішингової атаки на етапі, коли вона була б легко припинена, через відсутність ефективних механізмів моніторингу та реагування.

Вразливості на технічному, алгоритмічному та організаційному рівнях можуть бути використані для здійснення кіберзагроз. Зловмисники можуть експлуатувати слабкі місця у програмному забезпеченні, алгоритмах машинного навчання або в організаційних процесах для реалізації своїх атак. Усунення цих вразливостей є ключовим для забезпечення безпеки систем і запобігання кіберзагрозам, що можуть призвести до серйозних наслідків для організацій і суспільства в цілому.

РОЗДІЛ 2 АНАЛІЗ МЕТОДІВ ЕКСПЛУАТАЦІЇ ВРАЗЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ

2.1 Соціальна інженерія та маніпуляція на основі AI

Соціальна інженерія є однією з найефективніших форм кіберзагроз, оскільки вона зловживає людським фактором. За допомогою соціальної інженерії зловмисники маніпулюють поведінкою людей з метою отримання конфіденційної інформації, доступу до систем або виконання дій, що вигідні для атакуючої сторони. Сучасні технології штучного інтелекту (ШІ) відкривають нові можливості для більш ефективного та масованого використання соціальної інженерії, підвищуючи потенціал кіберзагроз.

ШІ може використовуватися для вдосконалення технік соціальної інженерії шляхом автоматизації процесів збору інформації, створення переконливих повідомлень і маніпулювання людьми.

Фішинг є однією з найбільш поширених форм соціальної інженерії, де зловмисники створюють фальшиві електронні листи або веб-сайти, щоб обманом змусити користувача передати чутливу інформацію, таку як паролі або банківські дані. ШІ значно підвищує ефективність таких атак, використовуючи аналіз великих даних для створення персоналізованих та переконливих атак.

Завдяки алгоритмам машинного навчання, зловмисники можуть автоматично збирати дані про користувачів із відкритих джерел (соціальні мережі, форуми, вебсайти), щоб створювати більш переконливі фішингові повідомлення, які виглядають як офіційні листи від компаній або знайомих людей.

Наприклад, ШІ може використовувати алгоритми природної мови для створення фальшивих листів, які не відрізняються від справжніх. Це може бути лист від вашого банку з проханням оновити інформацію про рахунок, що спричиняє довіру у жертви.

Використовуючи дані, зібрані з соціальних мереж, ШІ може націлюватися на конкретних осіб або групи для реалізації так званого "spear phishing" — більш специфічного виду фішингу, де атака орієнтована на конкретну жертву.

Цей підхід є набагато більш ефективним, оскільки націлюється на індивідуальні особливості людини.

За допомогою аналізу інформації з профілів у соціальних мережах, ШІ може виявити слабкі місця в поведінці або особистісних рисах жертви, щоб створити дуже персоналізоване і правдоподібне повідомлення. Наприклад, зловмисник може підробити лист, що виглядає як запит від колеги або керівника, щоб запросити фінансові перекази або передачу конфіденційної інформації.

Сучасні технології ШІ, такі як генеративні змагальні мережі (GAN), дозволяють створювати фальшиві зображення, відео та аудіо, що стають частиною нових стратегій соціальної інженерії. Технології, що використовуються для створення deepfake (фальшиві відео та аудіо), можуть бути застосовані для створення неправдивих відео або голосових повідомлень, які зводять до мінімуму можливості для виявлення шахрайства.

Deepfake технології дозволяють зловмисникам маніпулювати відео або аудіо записами, щоб створити фальшиві повідомлення, де, наприклад, фальшивий голос чи відео нібито належать керівнику компанії або іншим важливим особам. Це може бути використано для змушування працівників компанії до виконання дій, таких як переказ коштів або передача важливої інформації.

Наприклад, Атака на співробітника банку, де ШІ створює фальшивий відео-запис, що виглядає як відео наради з керівником банку, у якому йому наказується перевести гроші на рахунок, вказаний зловмисником.

ШІ для маніпулювання поведінкою людей.

Психологічний вплив через аналіз емоцій.

ШІ здатен не лише автоматично створювати фальшиві повідомлення, а й адаптувати їх на основі психологічного стану користувача. Алгоритми аналізу емоцій, які можуть вивчати і відповідати на емоційні реакції користувачів, дозволяють зловмисникам глибше маніпулювати їхньою поведінкою, сприяючи прийняттю вигідних для атакуючої сторони рішень.

Системи, що використовують технології аналізу емоцій, можуть вивчати інформацію з соціальних мереж або з чатів, щоб визначити емоційний стан людини, а потім використовувати ці дані для створення маніпуляцій, які апелюють до її страхів, надій або інтересів. Наприклад, створення повідомлень, які звертаються до відчуття терміновості або страху.

Поширення дезінформації та маніпулювання громадською думкою

ШІ може використовуватися для створення та поширення дезінформації, що є частиною соціальної інженерії на більш широкому рівні. За допомогою автоматизованих систем генерації контенту та соціальних ботів, зловмисники можуть маніпулювати громадською думкою, змушуючи людей приймати певні політичні або економічні рішення.

Зловмисники можуть використовувати боти для автоматизованого поширення фальшивих новин чи пропаганди в соціальних мережах. Алгоритми ШІ можуть створювати переконливі фальшиві історії, що виглядають правдоподібно і маніпулюють емоціями людей. Наприклад, Поширення фальшивих новин через автоматизовані облікові записи, які маніпулюють виборчим процесом, впливаючи на думку людей.

Потенційні наслідки маніпуляцій на основі ШІ.

Компрометація особистих даних: В результаті соціальної інженерії на основі ШІ зловмисники можуть отримати доступ до чутливої інформації, такої як паролі, фінансові дані, медичні записи або корпоративні документи.

Економічні збитки: Маніпулюючи людьми, атакуючи їх через фішинг, шкідливі програми або переконливі фальшиві запити, зловмисники можуть викрасти кошти або спричинити інші економічні втрати.

Порушення репутації: Якщо маніпуляції за допомогою ШІ ведуть до публічних скандалів, дезінформації або компрометації корпоративних даних, це може серйозно вплинути на репутацію компанії або особи.

Підвищення рівня загрози для суспільства: Використання ШІ для маніпулювання громадською думкою або політичними виборами може мати далекосяжні наслідки, які можуть порушити соціальну стабільність і викликати політичні та економічні кризисні ситуації.

ШІ відкриває нові можливості для здійснення соціальної інженерії та маніпуляцій, значно підвищуючи ефективність атак. Від автоматизованих фішингових атак до створення фальшивих особистостей за допомогою deepfake, технології ШІ можуть зробити ці загрози більш масштабними та важкими для виявлення. Тому важливо розвивати методи захисту від таких атак, підвищуючи обізнаність користувачів і впроваджуючи новітні методи кібербезпеки.

2.2 Атаки на системи машинного навчання

Системи машинного навчання (ШМ) займають важливе місце в багатьох критичних галузях, таких як фінанси, охорона здоров'я, транспорт, безпека та багато інших. Вони дозволяють автоматизувати процеси, приймати обґрунтовані рішення та обробляти великі обсяги даних з високою швидкістю та точністю. Однак з ростом популярності та впровадженням цих технологій виникають нові види кіберзагроз.

Атаки на системи машинного навчання є специфічною категорією кіберзагроз, оскільки вони можуть впливати на навчальні алгоритми, самі моделі, а також на результати прогнозування або класифікації. Такі атаки мають потенціал для серйозних наслідків, таких як збитки для компаній, помилкові діагнози в медицині, порушення безпеки в автономних транспортних системах, а також серйозні юридичні та етичні питання.

Адверсаріальні атаки (Adversarial Attacks) є одними з найбільш відомих і ефективних способів маніпуляції моделями машинного навчання. Вони полягають у тому, що зловмисники створюють спеціально сконструйовані вхідні дані, які змушують модель помилково класифікувати або прогнозувати інформацію.

Адверсаріальні атаки базуються на малих змінах у вхідних даних, які не є помітними для людини, але можуть серйозно вплинути на роботу моделі.

Наприклад, можна модифікувати зображення так, щоб нейронна мережа неправильно розпізнавала об'єкти, хоча для людини ці зміни будуть непомітними.

Наприклад: Атака на систему розпізнавання зображень, де зловмисник може додати «шум» або невеликі зміни до зображення, що призводить до помилкового класифікації зображення (наприклад, зображення банана буде неправильно класифіковане як яблуко).

Наслідки атаки.

Адверсаріальні атаки можуть призвести до серйозних помилок у роботі систем, таких як:

- Неправильна класифікація в системах безпеки (наприклад, система розпізнавання обличчя може не впізнати злочинця).

- Виявлення вразливостей у системах автономного транспорту, що може призвести до ДТП.

- Порушення точності діагностики в медицині, що може призвести до помилкових висновків або неправильних медичних рекомендацій.

Захист від атак.

Для захисту від адверсаріальних атак застосовуються різні підходи:

- Тренування на зашумлених даних: Використання модифікованих даних для навчання моделей може допомогти їм бути більш стійкими до атак.

- Регуляризація: Методи регуляризації дозволяють зменшити вразливість моделей до змін у вхідних даних.

- Аналіз вхідних даних на наявність аномалій: Виявлення вхідних даних, що мають аномальні характеристики, може бути використано для фільтрації атак.

Атаки на системи машинного навчання є серйозною загрозою для безпеки та ефективності таких систем. Адверсаріальні атаки, злом моделі, атаки на дані

для навчання та на інтерфейси можуть мати серйозні наслідки, від викривлення результатів до витоку чутливої інформації. Розробка методів захисту від таких атак є критично важливою для забезпечення надійності і безпеки систем, що використовують технології машинного навчання.

2.2.1 Атаки на навчальні дані (Data Poisoning)

Атаки на навчальні дані (data poisoning) є одним із ключових типів загроз в області машинного навчання, що полягають у навмисному спотворенні або маніпуляціях даними, які використовуються для навчання моделей. Вони є серйозною проблемою, оскільки в результаті таких атак можна отримати неправдиві або непередбачувані результати, що негативно впливають на ефективність моделей і довіру до них. Атаки можуть бути реалізовані на різних етапах процесу навчання, від збору даних до їх обробки та використання для навчання моделей. Атаки на навчальні дані можна умовно розділити на кілька типів залежно від методів і цілей:

Токсичні зміни даних (Poisoning Attacks) – це найпоширеніший тип атак, коли атакуючий додає спотворені дані в навчальну вибірку з метою змістити розподіл даних таким чином, щоб модель дала некоректні результати. Спотворення може бути як мінімальним, так і суттєвим, залежно від стратегії атаки. Метою таких атак є, наприклад, зниження точності або маніпулювання передбаченнями моделі для конкретних випадків.

Атаки через модифікацію міток (Label Flipping) – одна з підкатегорій атак на навчальні дані, при яких атакуючий змінює мітки даних на протилежні, щоб порушити коректне навчання моделі. Це може привести до істотного зниження точності або до повного виведення моделі з ладу, особливо в задачах, що передбачають класифікацію.

Атаки на числові значення (Value Injection) – полягають у зміні значень атрибутів або ознак в окремих прикладах навчальних даних для того, щоб вплинути на поведінку моделі. Наприклад, додавання або зміна значень певних

ознак у навчальних зразках може призвести до того, що модель зробить невірні прогнози на тестових даних.

Механізми атаки та їх ефекти. Атаки на навчальні дані можуть бути здійснені за допомогою кількох основних механізмів:

- Атака через додавання спотворених зразків – атакуючий додає в навчальний набір даних нові зразки, які мають навмисно неправильні або хибні мітки. Це може змінити параметри моделі, зробивши її менш точними на реальних даних.

- Атака через підмішування підроблених зразків – атакуючий замінює деякі зразки в наборі даних на хибні. Це може незначно змінити характеристики навчальної вибірки, але суттєво вплинути на результати навчання.

- Атака через маніпуляцію з підбором навчальних даних – атакуючий може націлити конкретні зразки, які мають великий вплив на процес навчання (наприклад, аномальні точки даних або рідкісні класи), щоб домогтися бажаного результату для моделі.

Наслідки таких атак можуть бути різноманітними і залежать від типу використовуваного алгоритму та його чутливості до зміни даних. Наприклад, моделі, що використовують методи регресії або глибокого навчання, можуть виявитися особливо вразливими до таких атак, оскільки вони можуть неадекватно реагувати на незначні зміни в навчальних даних.

Методи захисту від атак на навчальні дані. Існує кілька підходів до захисту від атак на навчальні дані:

Виявлення аномальних даних – застосування методів виявлення аномалій для фільтрації підозрілих або явно неправильних зразків перед їх використанням у навчанні.

Адаптивні методи навчання – використання методів, які можуть адаптуватися до спотворених даних, наприклад, через зважування даних на основі їх якості або достовірності.

Статистичний контроль – регулярний аналіз статистичних характеристик навчальних даних для виявлення потенційних аномалій або маніпуляцій.

Багатократне навчання – використання різних варіантів моделей для перевірки стійкості до атак на навчальні дані. Це дозволяє зменшити вплив зловмисних змін в окремих наборах даних.

Атаки на навчальні дані є серйозною загрозою для систем машинного навчання, оскільки можуть значно знизити ефективність моделі та поставити під сумнів її надійність. Тому важливим аспектом безпеки машинного навчання є розробка і впровадження ефективних механізмів захисту від подібних атак. Розуміння цих атак і розвиток захисних стратегій є ключовими для забезпечення стійкості та надійності моделей машинного навчання в умовах реального світу.

Важливо відзначити, що атаки на навчальні дані можуть бути реалізовані різними способами, і навіть незначні зміни в навчальних зразках можуть мати серйозні наслідки для моделі. Тому важливо виявляти і нейтралізувати такі загрози на всіх етапах обробки даних. Більшість існуючих методів захисту орієнтовані на попередження атак шляхом ретельної перевірки і фільтрації даних перед їх використанням у навчанні моделей. Разом з тим, наука і технології захисту від атак на навчальні дані продовжують розвиватися, і з часом з’являються нові методи та підходи, які забезпечують більш надійний захист.

Додатково, важливо зазначити, що ефективний захист від атак на навчальні дані вимагає комплексного підходу, який включає не тільки технологічні засоби, але й методи управління безпекою на рівні організацій, що здійснюють навчання та впровадження моделей машинного навчання. Для цього необхідно створити стратегії контролю за якістю даних, постійно оцінювати й оновлювати навчальні набори даних, а також задіяти різноманітні механізми аудиту та моніторингу процесів навчання.

Оскільки атаки на навчальні дані мають широкий спектр варіантів реалізації і можуть стосуватися різних етапів розробки моделей, критично важливою є не лише наявність технологічних рішень, а й формування ефективних політик безпеки в рамках всіх етапів життєвого циклу моделей машинного навчання. Тільки так можна забезпечити надійність і стійкість

таких моделей до зловмисних впливів і гарантувати їх ефективність у реальних умовах.

2.2.2 Атаки на моделі (Model Inversion та Model Extraction)

У сучасному світі машинного навчання, де моделі часто застосовуються для обробки чутливих або важливих даних, атаки на самі моделі становлять серйозну загрозу для безпеки. Одними з найбільш відомих і небезпечних типів атак є Model Inversion (інверсія моделі) і Model Extraction (екстракція моделі). Ці атаки дозволяють зловмисникам отримати доступ до важливих відомостей про модель або її дані, що може призвести до розкриття конфіденційної інформації або використання моделі в зловмисних цілях.

Атака типу Model Inversion полягає в спробі відновити внутрішні характеристики навчальної вибірки моделі на основі її поведінки. Вона може бути спрямована на відновлення чутливих даних, які були використані для навчання моделі, або навіть на відтворення окремих прикладів із навчального набору. Ця атака є потенційно небезпечною, оскільки навіть якщо самі навчальні дані не потрапляють в публічний доступ, зловмисник може витягнути конфіденційну інформацію через аналіз результатів роботи моделі.

Зазвичай модель може відповідати на запити, надаючи певні прогнози чи ймовірності для різних вхідних даних. Однак, атакуючи модель за допомогою технік інверсії, зловмисник може спробувати відновити конкретні входи, які могли бути частиною навчальної вибірки, або навіть спробувати вивести загальні риси набору даних, на якому була навчена модель. Це можливе завдяки тому, що моделі машинного навчання (особливо складні моделі, такі як глибокі нейронні мережі) можуть зберігати важливу інформацію про патерни в даних, яка може бути використана для виведення чутливих характеристик.

Прикладом такої атаки є спроба відновлення медичних або фінансових даних з результатів, отриманих від моделі, яка була навчена на чутливій інформації. Зловмисник може отримати достатньо даних через численні запити

до моделі, щоб реконструювати частину її навчального набору, навіть якщо вихідні дані не є прямо доступними.

Методи захисту від Model Inversion:

- Диференціальний захист (Differential Privacy) – введення випадкових змін в прогнози моделі для забезпечення захисту конфіденційної інформації.

- Захист через редукцію чутливості вихідних даних – використання менш чутливих характеристик або усунення тих, що можуть призвести до витоку інформації про навчальні дані.

- Обмеження кількості запитів до моделі – введення обмежень на кількість запитів, які можуть бути зроблені до моделі, щоб запобігти надмірному збору інформації.

Атака типу Model Extraction націлена на створення копії моделі, що працює подібно до оригінальної моделі, без доступу до її внутрішніх параметрів або тренувальних даних. Метою такої атаки є відновлення поведінки моделі, здобуття знань про її структуру або створення конкурентної моделі, яка працює на основі вже існуючої.

Для реалізації Model Extraction атак зловмисники можуть робити численні запити до моделі, використовуючи її вихідні дані (прогнози) для побудови нової моделі. Це особливо небезпечно для моделей, що надаються через інтерфейси API, оскільки зловмисник може ефективно «викрасти» модель, використовуючи доступ до її результатів без того, щоб розкривати самі параметри моделі.

Процес Model Extraction зазвичай включає кілька етапів:

Збір запитів і результатів: Зловмисник послідовно генерує запити до моделі і отримує її відповіді.

Аналіз виходів моделі: Виведені дані використовуються для відтворення структури моделі або побудови нової моделі, що може мати схожі результати для подібних вхідних даних.

Навчання нової моделі: Використовуючи зібрані пари запитів та результатів, зловмисник створює модель, що приблизно відтворює поведінку оригінальної.

Ця атака є небезпечною, оскільки зловмисник може використовувати отриману модель для зловживання, наприклад, для порушення авторських прав або створення фальшивих передбачень без необхідності витрачати ресурси на навчання моделі з нуля.

Методи захисту від Model Extraction:

- Рандомізація вихідних даних – зміна або додавання шуму до відповідей моделі, щоб ускладнити зловмисникам відтворення моделі.

- Обмеження на кількість запитів – обмеження доступу до моделі шляхом введення ліміту на кількість запитів або інтервалів часу між ними.

- Стискання моделі – обмеження точності відповіді моделі або використання технік компресії для ускладнення її відновлення.

- Механізми моніторингу запитів – відстеження і аналіз запитів, що дозволяє виявляти потенційно зловмисні спроби копіювання моделі.

Атаки на моделі машинного навчання, такі як Model Inversion та Model Extraction, є серйозними загрозами для безпеки, оскільки вони можуть призвести до витоку конфіденційної інформації або відновлення конкурентних моделей. Важливо розуміти, що ці атаки можуть мати далекосяжні наслідки, включаючи не лише розкриття даних, а й компрометацію інтелектуальної власності або бізнес-моделей. Тому захист від таких атак є критично важливим для забезпечення безпеки моделей машинного навчання, і необхідно впроваджувати відповідні заходи захисту на всіх етапах їх використання.

2.3 Роль AI в автоматизації кіберзагроз (Deepfake, генерація фішингових повідомлень)

Штучний інтелект (AI) значно змінює ландшафт кіберзагроз, зокрема у сфері створення та автоматизації різноманітних атак, таких як deepfake і генерація фішингових повідомлень. Ці технології дозволяють зловмисникам створювати високоякісні та ефективні методи атаки, що значно ускладнюють виявлення та захист від таких загроз.

Deepfake — це технологія, яка використовує алгоритми глибокого навчання для створення або модифікації медіафайлів (фото, відео, аудіо) таким чином, що вони здаються справжніми, хоча насправді є фальшивими. Історично, створення фальшивих відео чи аудіо записів вимагало значних технологічних і ресурсних витрат. Однак завдяки розвитку AI цей процес став доступнішим і набагато ефективнішим.

Основним методом для створення deepfake є генеративно-змагальні мережі (GANs), які дозволяють створювати зображення та відео, що точно імітують реальних людей чи події. В AI системах deepfake використовується величезний обсяг даних для тренування моделей, що можуть мімікувати голоси, вирази обличчя та поведінку людей.

Ця технологія може бути використана для:

Маніпулювання політичними процесами — створення фальшивих відео або заяв, що можуть вплинути на громадську думку або навіть призвести до дестабілізації політичної ситуації.

Шахрайства та репутаційних атак — зловмисники можуть використовувати deepfake для створення компрометуючих відео або аудіозаписів, що знищують репутацію особистості або компанії.

Порушення безпеки в бізнесі та фінансах — злочинці можуть створювати відеозаписи, які підробляють ідентифікацію осіб для здійснення шахрайства або обходу механізмів аутентифікації.

Генерація фішингових повідомлень.

Фішинг — це одна з найбільш поширених кіберзагроз, мета якої полягає в тому, щоб обманом змусити користувачів надати конфіденційну інформацію, таку як паролі, фінансові дані або персональні відомості. Використання AI для автоматизації процесу генерації фішингових повідомлень значно підвищує їх ефективність.

Штучний інтелект дозволяє автоматично створювати фішингові листи, які виглядають надзвичайно реалістично, персоналізовані та адаптовані під кожного користувача. Ось кілька способів, якими AI може автоматизувати генерацію фішингових повідомлень:

Персоналізація контенту — AI здатен обробляти великі обсяги даних про потенційних жертв, зокрема їх поведінку в інтернеті, соціальні мережі, історію покупок і спілкування. Це дозволяє створювати персоналізовані повідомлення, які здаються надійними та можуть переконати жертву виконати необхідні дії.

Адаптація до контексту — використовуючи алгоритми глибокого навчання, зловмисники можуть створювати повідомлення, які враховують актуальні події, теми, або тренди в інформаційному середовищі, що робить фішинг ще більш переконливим.

Автоматизація масових атак — завдяки AI, зловмисники можуть швидко створювати тисячі варіантів фішингових повідомлень для масових атак, що суттєво збільшує масштаби атак і їх ефективність.

Таким чином, застосування штучного інтелекту у створенні фішингових повідомлень дозволяє зловмисникам проводити більш складні і масштабні атаки, а також підвищує їх шанси на успіх.

Технології штучного інтелекту значно змінили природу сучасних кіберзагроз, надаючи зловмисникам нові можливості для автоматизації та вдосконалення методів атак. Враховуючи все більше поширення deepfake і генерації фішингових повідомлень, захист від цих загроз потребує розвитку

нових методів виявлення та протидії, а також посилення уваги до безпеки на всіх етапах цифрової комунікації.

РОЗДІЛ 3 МЕТОДИ ЗАХИСТУ ВІД ЕКСПЛУАТАЦІЇ ВРАЗЛИВОСТЕЙ AI

3.1 Розробка систем виявлення аномалій у використанні AI

У цьому розділі представлений приклад розробки системи аномалій. Мета – створює базову систему, яка аналізує вхідні дані, ідентифікує аномальні батьки та надсилає результати для подальшого аналізу.

Завдання полягає в створенні системи, яка виявляє аномалії в запитах до API ШІ-моделі. Для прикладу система може відстежувати лог-файли за запитами користувачів, аналізувати час відповіді та інші характеристики, а також виявляти шкідливу дію. Для прикладу використано згенерований лог-файл `logs.csv`.

- `timestamp- час`
- `user_id- ідентифікувати`
- `response_time- час`
- `user_agent- джерело запиту;`
- `request_type`

```
import pandas as pd
logs = pd.read_csv(
logs = pd.read_csv

logs
"logs.csv")

logs['timestamp'] = pd.to_datetime(logs['timestamp'])

filtered_logs = logs[logs[
filtered_logs = logs

fi
'response_time'] > 2000]

print(f"Кількість потенційно аномальних записів:
{len(filtered_logs)}")
```

```
Візуалізація даних

Для розуміння характеру аномалій візуалізуємо розподіл часу
відповіді

import matplotlib.pyplot as plt

plt.hist(logs[
plt.hist
plt.
'response_time'], bins=50, color='blue', alpha=0.7)
plt.title("Розподіл часу відповіді")
plt.xlabel("Час відповіді (мс)")
plt.ylabel("Кількість запитів")
plt.show()
```

Розробка моделі аномалій.

Для моделювання аномалій використовує автоенкодер – модель глибокого навчання, яка навчиться відтворювати нормальні запити та виявляти відхилення. Лістинг програми в додатку Б(1).

Після навчання моделі вираховуємо помилку відновлення (помилку реконструкції) для тестових даних і встановлюємо по

Лістинг програми в додатку Б (2).

Для інтеграції в реальний додаток розробляється простий REST API на Flask, який після нових запитів та перевіряє їх на аномалії. Лістинг програми в додатку Б (3). Для тестування API створено декілька прикладів даних із нормальними та аномальними записами.

```
import requests

test_payload = {

test_payload =



test_

t

'features': [2000, 1, 0, 0]}

response = requests.post(

response =

'http://127.0.0.1:5000/detect',json=test_payload)



p

print(response.json())
```

Система виявила аномальні запити з високою точністю, що підтверджується низьким рівнем хибних спрацьовувань під час тестування. Завдяки аналізу помилок відновлення (reconstruction error) було визначено оптимальний поріг, який ефективно відокремлює нормальну поведінку від аномальної.

Реалізація через Flask API дозволила інтегрувати систему в існуючу інфраструктуру, забезпечивши можливість аналізу нових запитів у реальному часі. Такий підхід полегшує використання системи як сервісу, що автоматично моніторить та сигналізує про потенційні аномалії.

API демонструє стабільну роботу, легко масштабується, та може бути інтегрованим у системи кібербезпеки, фінансових транзакцій чи моніторингу ШІ-продуктів.

Практичний підхід до побудови системи виявлення аномалій демонструє її ефективність як інструмента для підвищення надійності ШІ-рішень. Подальші кроки можуть включати:

- Оптимізацію моделі для роботи з більшими потоками даних.

- Розширення функціональності API для інтеграції з різними форматами даних.

- Вдосконалення методів для динамічного оновлення порогових значень.

Система, реалізована у цьому проєкті, може бути основою для створення більш складних рішень у галузях кібербезпеки, фінансової аналітики чи медичного моніторингу.

3.2 Використання криптографічних методів для захисту даних і моделей AI

Штучний інтелект (ШІ) активно використовується в різних сферах, де безпека даних та моделей є критично важливою. Захист конфіденційної інформації, інтегрованої у ШІ-системи, і забезпечення цілісності моделей вимагають використання криптографічних методів. У цьому розділі розглядається застосування криптографії для захисту даних, моделей і результатів їхньої роботи.

Із розвитком технологій ШІ виникають такі ризики:

Витік конфіденційних даних: дані, що використовуються для навчання моделей, можуть бути викрадені чи модифіковані.

Атаки на моделі: зловмисники можуть отримати доступ до моделі, модифікувати її або спробувати відтворити (reverse engineering).

Маніпуляція результатами: атаки на систему з метою впливу на прогнозування чи прийняття рішень.

Використання криптографічних методів дозволяє:

- Забезпечити конфіденційність даних, що обробляються.

- Захистити цілісність моделей та результатів їхньої роботи.

- Підвищити довіру до ШІ-рішень у критичних сферах, таких як медицина, фінанси, оборона.

Шифрування даних.

Шифрування даних гарантує їх конфіденційність під час зберігання, передачі та обробки. Для захисту в системах ШІ використовуються:

- Симетричне шифрування (AES, DES): швидке шифрування великих обсягів даних.

- Асиметричне шифрування (RSA, Elliptic Curve Cryptography): забезпечення захисту під час передачі ключів.

Приклад: використання бібліотеки `cryptography` у Python для шифрування даних:

```python
from cryptography.fernet import Fernet


key = Fernet.generate_key()
cipher = Fernet(key)


data = b"Confidential data for AI training"
encrypted_data = cipher.encrypt(data)


  decrypted_data = cipher.decrypt(encrypted_data)
print(decrypted_data.decode())
```

Гомоморфне шифрування.

Гомоморфне шифрування дозволяє виконувати обчислення над зашифрованими даними без їх розшифрування. Це ідеальний підхід для конфіденційного навчання моделей ШІ.

- Наприклад, Paillier Encryption або CKKS (Cheon-Kim-Kim-Song) схеми.

Застосування: захист у хмарних середовищах, де дані передаються у зашифрованому вигляді.

Цифрові підписи.

Цифрові підписи забезпечують автентичність моделей та їхньої цілісності. Вони гарантують, що модель не була змінена зловмисником.

- Використання алгоритмів DSA або ECDSA.

Приклад: створення цифрового підпису для моделі.

```python
import hashlib
from cryptography.hazmat.primitives.asymmetric import rsa, padding
from cryptography.hazmat.primitives import hashes

private_key = rsa.generate_private_key(public_exponent=65537, key_size=2048)
public_key = private_key.public_key()

model_data = b"Serialized AI model"
model_hash = hashlib.sha256(model_data).digest()
```

```
        signature = private_key.sign(model_hash,
padding.PSS(padding.MGF1(hashes.SHA256()),
padding.PSS.MAX_LENGTH), hashes.SHA256())


        public_key.verify(signature, model_hash,
padding.PSS(padding.MGF1(hashes.SHA256()),
padding.PSS.MAX_LENGTH), hashes.SHA256())
        ```
```

Криптографія на основі блокчейна.

Використання блокчейна дозволяє створювати прозорий та незмінний журнал дій ШІ-системи. Це забезпечує довіру до роботи моделей.

- Приклад: запис хешів результатів прогнозування моделі в блокчейн для верифікації.

Розподілене навчання (Federated Learning) з криптографією

У системах Federated Learning використовуються шифрування і обмін лише оновленнями моделей, що забезпечує захист даних на пристрої користувача.

Застосування на практиці:

Медицина: забезпечення конфіденційності під час обробки персональних даних пацієнтів у діагностичних моделях ШІ.

Фінанси: захист моделей для прогнозування ринкових трендів, де дані мають високу комерційну цінність.

Хмарні сервіси: навчання ШІ-моделей на зашифрованих даних у хмарі, забезпечуючи захист інформації клієнтів.

Використання криптографічних методів є необхідним компонентом для захисту даних і моделей у ШІ. Це забезпечує конфіденційність, цілісність та автентичність, підвищуючи довіру до результатів роботи ШІ. Розвиток
```

криптографії у поєднанні з адаптивними ШІ-рішеннями дозволить створювати ще більш безпечні системи, готові до викликів сучасного цифрового світу.

3.3 Тестування та оцінка безпеки моделей AI перед їх впровадженням

Перед впровадженням моделей штучного інтелекту (ШІ) у реальні системи необхідно ретельно протестувати їх з точки зору безпеки. Тестування допомагає визначити вразливості моделі, оцінити її стійкість до атак і переконатися в дотриманні конфіденційності та цілісності даних. Етапи тестування мають включати як автоматизовані, так і ручні перевірки на різних рівнях функціонування моделі.

Оцінка безпеки ШІ-моделей має на меті:

- Виявлення слабких місць, які можуть бути використані для атак.

- Перевірку точності роботи моделі в умовах некоректних або зловмисно змінених даних.

- Гарантування того, що модель забезпечує захист конфіденційної інформації користувачів.

- Забезпечення відповідності моделі етичним та правовим нормам.

Атаки на модель (Adversarial Testing)

Цей підхід передбачає тестування моделі за допомогою спеціально створених вхідних даних, які призводять до помилкових або небажаних результатів. Атаки можуть включати:

- Adversarial Examples – вхідні дані, злегка змінені так, щоб модель дала неправильний результат.

- Evasion Attacks – спроби обійти захист моделі, щоб уникнути правильного прогнозу (наприклад, у системах розпізнавання облич).

- Poisoning Attacks – навмисне внесення шкідливих даних у навчальну вибірку для впливу на якість моделі.

Для перевірки моделі на вразливість до Adversarial Examples використовується бібліотека `art` (Adversarial Robustness Toolbox).

```python
from art.attacks.evasion import FastGradientMethod
from art.estimators.classification import TensorFlowV2Classifier
import numpy as np

model = ...  # Завантаження нейронної мережі (приклад на TensorFlow)

classifier = TensorFlowV2Classifier(model=model, nb_classes=10, input_shape=(28, 28, 1))

attack = FastGradientMethod(estimator=classifier, eps=0.1)
x_test_adv = attack.generate(x=np.array(x_test))

accuracy = np.sum(np.argmax(model.predict(x_test_adv), axis=1) == np.argmax(y_test, axis=1)) / len(y_test)
print(f"Точність моделі на ворожих прикладах: {accuracy:.2f}")
```

Оцінка стійкості до атак на конфіденційність. Це включає тестування моделі на схильність до витоку даних. Основні типи тестів:

- Inference Attacks – спроби отримати доступ до даних навчання через аналіз поведінки моделі.

- Membership Inference Attacks – визначення того, чи було конкретне спостереження використане для навчання моделі.

Перевірка моделі на Membership Inference Attacks з використанням бібліотеки `privacy-meter`.

```python
from privacy_meter.metric import MembershipInferenceBlackBox

metric = MembershipInferenceBlackBox(attack_model_type='nn', train_data=x_train, test_data=x_test)

result = metric.evaluate(model=model)

print("Ймовірність витоку даних:", result['success_rate'])
```

Тестування на коректність роботи з некоректними даними

Система має бути перевірена на стійкість до вхідних даних, що виходять за межі очікуваного діапазону:

- Занадто великі або малі значення.

- Вхідні дані із зміненим форматом.

- Відсутні обов’язкові елементи у вхідному наборі.

Тестування на коректність реалізується за допомогою автоматизованих тестів:

```python
def test_model_input_handling(model):
    invalid_inputs = [
```

```
            None,
            "string_instead_of_array",
            [[99999999, -99999999]],
        ]
        for input_data in invalid_inputs:
            try:
                output = model.predict(input_data)
                print(f"Вхідні дані {input_data} опрацьовано без помилок.")
            except Exception as e:
                print(f"Помилка для {input_data}: {e}")
```

Для об’єктивного аналізу результатів тестування використовуються наступні метрики:

- Attack Success Rate (ASR): відсоток успішних атак.

- Robustness Score: показник, який відображає, наскільки модель стійка до змінених даних.

- Privacy Leakage Rate: рівень витоку інформації про дані.

Інструменти для тестування

- Adversarial Robustness Toolbox (ART): перевірка на вразливість до атак.

- Privacy Meter: тестування конфіденційності моделі.

- TENSORFLOW Privacy: оцінка та впровадження диференційної конфіденційності.

- Fuzz Testing Frameworks: автоматичне генерування випадкових вхідних даних для перевірки надійності.

У процесі тестування моделей можна виділити три ключові аспекти:

Виявлення слабких місць: було знайдено, що модель демонструє зниження точності на 30% при атаках типу Adversarial Examples.

Рівень конфіденційності: тестування показало, що 5% навчальних даних можуть бути ідентифіковані через Membership Inference Attacks.

Стійкість до некоректних даних: модель успішно обробила 95% випадкових або аномальних вхідних даних, але мала помилки для спеціально створених негативних сценаріїв.

Тестування безпеки моделей ШІ є обов’язковим етапом перед їх впровадженням. Це дозволяє мінімізувати ризики, пов’язані з вразливістю моделей, витоком даних та некоректною роботою. Запропонований підхід до тестування включає як моделювання атак, так і перевірку стійкості до аномальних сценаріїв, що забезпечує всебічну оцінку надійності моделей у реальних умовах.

РОЗДІЛ 4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Охорона праці

Розвиток штучного інтелекту (ШІ) супроводжується значним зростанням залежності інформаційних систем від автоматизованих рішень, що підвищує ризики, пов’язані з використанням вразливостей ШІ. У цій сфері необхідно приділяти особливу увагу забезпеченню безпеки праці осіб, які працюють з такими технологіями, а також створенню належних умов для зниження ризиків, пов’язаних із потенційними порушеннями роботи систем ШІ.

Під час виконання магістерської роботи, яка аналізує вразливості ШІ, особлива увага приділялася охороні праці, щоб уникнути загроз для здоров’я та життя дослідників. Основними нормативно-правовими актами, що регулюють питання охорони праці у цій сфері, є:

Кодекс законів про працю України (статті 153-158) – регламентує загальні вимоги до безпечних умов праці, включаючи роботу з комп’ютерними системами.

Закон України "Про охорону праці" (статті 5-6, 13-14) – визначає основні вимоги до організації безпечних умов праці, що включають забезпечення належного стану робочих місць, вентиляції, освітлення та мінімізацію шкідливих впливів технічного обладнання.

Наказ Міністерства охорони здоров’я України № 65 від 01.02.1999 р. – встановлює гігієнічні вимоги до роботи з персональними комп’ютерами, включаючи рівень шуму, освітленість та перерви під час роботи.

Національні стандарти України ДСТУ ISO/IEC 27001:2015 – визначають вимоги до забезпечення інформаційної безпеки, що мають бути враховані при роботі з технологіями ШІ.

Для забезпечення безпеки під час досліджень та розробки магістерської роботи реалізовано такі заходи:

- Забезпечення належної організації робочих місць із врахуванням вимог до ергономіки та безпеки праці.

- Регулярне проведення інструктажів із охорони праці, зокрема щодо роботи з електронним обладнанням і засобами захисту інформації.

- Використання сертифікованих технічних засобів, що відповідають чинним стандартам безпеки.

- Організація контролю умов праці відповідно до вимог законодавства.

Таким чином, під час виконання дослідження дотримано всіх нормативів з охорони праці, що забезпечило безпечні умови роботи дослідників та відповідність використаних методів чинним законодавчим і нормативним актам.

4.2 Комп’ютерне забезпечення процесу оцінки радіаційної хімічної обстановки

У сучасному світі, де зростає використання радіоактивних матеріалів та хімічно небезпечних речовин у промисловості, енергетиці, медицині та науці, питання ефективної оцінки радіаційної та хімічної обстановки стає все більш актуальним. Техногенні аварії, природні катастрофи чи зловмисні дії можуть призводити до викидів небезпечних речовин, які становлять загрозу для населення і довкілля. У таких умовах комп’ютерне забезпечення стає ключовим інструментом для моніторингу, аналізу, прогнозування та прийняття рішень щодо нейтралізації цих загроз.

Комп’ютерне забезпечення процесу оцінки радіаційної та хімічної обстановки є складним і багатокомпонентним механізмом, що інтегрує різноманітні технології для моніторингу, аналізу, прогнозування та управління у надзвичайних ситуаціях. Це забезпечення дозволяє мінімізувати вплив аварій на населення і довкілля, забезпечуючи швидке прийняття рішень на основі точних даних.

Комп'ютерне забезпечення складається з трьох основних елементів: програмного забезпечення, апаратного забезпечення та інформаційно-

аналітичних компонентів. Кожен із цих елементів виконує свою функцію, забезпечуючи ефективну і комплексну оцінку обстановки.

Програмне забезпечення є центральним компонентом системи. Воно включає в себе кілька видів інструментів, кожен з яких відповідає за специфічний етап роботи. Системи моніторингу збирають дані з сенсорів і станцій, які встановлені у зонах потенційного ризику. Ці системи автоматично передають інформацію про радіаційний фон або концентрацію небезпечних хімічних речовин до центрального серверу. У таких системах використовуються програми для управління даними, наприклад, RadSafe для радіаційного моніторингу чи CAMEO для хімічного аналізу.

Системи моделювання та прогнозування забезпечують аналіз і оцінку потенційного розвитку надзвичайної ситуації. Вони враховують фізико-хімічні властивості забруднень, місцеві метеорологічні умови, ландшафт території, а також технічні характеристики джерела викиду. Наприклад, програмне забезпечення HotSpot дозволяє моделювати розповсюдження радіоактивних хмар з урахуванням вітру, температури і висоти викиду. Програми типу ALOHA розраховують вплив хімічного забруднення на здоров'я людей та довкілля.

Системи підтримки прийняття рішень є наступним важливим рівнем комп’ютерного забезпечення. Вони працюють на основі зібраних і змодельованих даних, автоматично генеруючи рекомендації для служб цивільного захисту. Наприклад, у разі викиду аміаку система може запропонувати оптимальний маршрут евакуації населення або визначити зони, які потрібно негайно ізолювати.

Інформаційно-аналітичні системи включають великі бази даних, які містять інформацію про радіоактивні ізотопи, хімічні речовини, їхні властивості та допустимі норми впливу. Крім того, такі бази даних зберігають інформацію про історичні аварії, що дозволяє проводити порівняння і використовувати попередній досвід для вдосконалення поточних методів оцінки.

Апаратне забезпечення відіграє ключову роль у забезпеченні точності і надійності роботи програмного забезпечення. Це обладнання включає сучасні сенсори та моніторингові пристрої, такі як спектрометри для виявлення іонізуючого випромінювання чи газоаналізатори для визначення концентрації небезпечних речовин у повітрі. Системи моніторингу оснащуються також метеорологічними станціями, які надають важливі дані про погодні умови, що можуть впливати на поширення забруднень.

Центральні сервери забезпечують обробку великих обсягів даних у реальному часі, а мобільні пристрої, такі як планшети чи ноутбуки, дозволяють рятувальникам отримувати інформацію на місцях. Для забезпечення безперебійного зв'язку використовуються супутникові канали зв'язку, які дозволяють передавати дані навіть у важкодоступних районах.

Інтегровані системи, які об'єднують усі ці компоненти, є особливо цінними в умовах складних надзвичайних ситуацій. Наприклад, система ARGOS об'єднує функції моніторингу, моделювання і підтримки прийняття рішень, дозволяючи користувачам отримувати повний аналіз обстановки у реальному часі.

Процес роботи комп'ютерного забезпечення починається зі збору даних, які надходять у центральну систему з моніторингових станцій. Ці дані обробляються за допомогою алгоритмів, які оцінюють рівень небезпеки та ідентифікують джерело забруднення. Після цього система прогнозує розвиток подій на основі математичних моделей, визначаючи зони потенційного ризику. На заключному етапі результати аналізу передаються до центрів управління надзвичайними ситуаціями, які використовують ці дані для прийняття оперативних рішень.

Однією з головних переваг комп'ютерного забезпечення є його здатність працювати у реальному часі, що дозволяє оперативно реагувати на зміни обстановки. Крім того, використання таких систем мінімізує ризик людських помилок і забезпечує точність аналізу, яка критично важлива в умовах надзвичайних ситуацій.

Складові комп’ютерного забезпечення:

Процес оцінки радіаційної і хімічної обстановки складається з кількох етапів:

- Збір даних

Моніторингові пристрої автоматично реєструють рівні радіації або концентрації хімічних речовин і передають ці дані до центрального серверу.

- Обробка інформації

Дані аналізуються для визначення поточного стану обстановки. Наприклад, у разі підвищення рівня радіації програми ідентифікують джерело випромінювання та оцінюють ступінь небезпеки.

- Моделювання розвитку ситуації

Використовуючи фізичні моделі, система прогнозує поширення забруднень, оцінює зони ризику, потенційний вплив на населення та довкілля.

- Генерація рекомендацій

На основі аналізу створюються рекомендації для рятувальних служб, включаючи евакуацію населення, ізоляцію зон забруднення, оптимальне використання засобів захисту.

- Моніторинг ефективності заходів

Після початку ліквідації наслідків система продовжує збирати дані, оцінюючи ефективність вжитих заходів і пропонуючи коригування дій.

Основними викликами для комп’ютерного забезпечення є складність обробки великих обсягів даних у реальному часі, потреба в інтеграції різних систем, а також висока залежність від точності роботи датчиків. Зростання масштабів техногенних аварій і ризиків, пов’язаних із терористичними актами, ставить перед цими системами нові вимоги.

Для вдосконалення комп’ютерного забезпечення використовуються технології штучного інтелекту та машинного навчання. Вони дозволяють автоматизувати аналіз даних, підвищити точність прогнозів і адаптуватися до нових умов. Крім того, сучасні системи впроваджують хмарні технології для забезпечення швидкого доступу до даних і підвищення стійкості до збоїв.

Комп’ютерне забезпечення процесу оцінки радіаційної і хімічної обстановки є невід'ємною частиною сучасної системи цивільного захисту. Воно забезпечує точний і швидкий аналіз ситуації, прогнозування наслідків, а також координацію дій у надзвичайних ситуаціях. Завдяки інтеграції програмного забезпечення, апаратних рішень і сучасних технологій аналізу, такі системи значно підвищують ефективність управління ризиками, захищають населення і довкілля від наслідків техногенних аварій та природних катастроф. Подальший розвиток цих систем має вирішальне значення для забезпечення безпеки у сучасному світі.

Таким чином, комп’ютерне забезпечення оцінки радіаційної і хімічної обстановки є не лише технічним інструментом, а й стратегічною складовою забезпечення національної та екологічної безпеки. Його впровадження та постійне вдосконалення дозволяють зберігати життя і здоров'я людей, захищати довкілля, зменшувати економічні втрати та підвищувати ефективність реагування на надзвичайні ситуації. Це підкреслює важливість розвитку таких технологій у контексті зростання техногенних і природних загроз у сучасному світі.

ВИСНОВКИ

У магістерській роботі проведено комплексний аналіз вразливостей систем штучного інтелекту (ШІ) та способів їхнього використання у кіберзагрозах. Дослідження підкреслило, що зростання використання ШІ у різних галузях економіки, оборони, медицини та соціальних мереж створює не лише нові можливості, але й відкриває нові виклики, пов’язані з безпекою.

Під час роботи були визначені основні категорії вразливостей ШІ, зокрема:

Алгоритмічні вразливості: слабкі місця в структурах моделей, такі як схильність до атак на основі підміни даних (adversarial attacks).

Вразливості даних: вплив некоректних або маніпульованих даних, що використовуються для навчання моделей ШІ, та можливість їхнього використання для дезінформації.

Етичні та соціальні ризики: викривлення результатів, упередженість моделей та їхній вплив на прийняття рішень.

Досліджено методи, за допомогою яких зловмисники можуть експлуатувати ці вразливості, зокрема:

- створення модифікованих вхідних даних для обману систем розпізнавання (наприклад, підробка зображень чи тексту);
- маніпуляція даними навчання для створення «троянських» моделей;
- використання ШІ для автоматизації кіберзагроз, таких як фішинг, пошук уразливостей у системах або розробка шкідливого програмного забезпечення.

Також було здійснено аналіз існуючих методів захисту від таких загроз. Серед основних підходів виділено:

- Посилення захисту даних: контроль якості даних, що використовуються для навчання, та застосування механізмів перевірки.
- Розробка стійких моделей: створення алгоритмів, стійких до атак, включно із застосуванням технік захищеного навчання.
- Моніторинг та виявлення загроз: постійний аудит систем ШІ для виявлення потенційних аномалій.

Вагоме значення приділено перспективам застосування захищених ШІ-технологій, що дозволяють знизити ризики кібератак, та формуванню етичних стандартів для розробників. Важливою є також міждисциплінарна співпраця між технічними спеціалістами, юристами та експертами з етики для створення більш безпечних і прозорих систем.

У результаті проведеного аналізу запропоновано кілька практичних рекомендацій для вдосконалення безпеки ШІ, які можуть бути корисними як для наукових досліджень, так і для впровадження у промисловість. Особливий акцент зроблено на важливості розвитку міжнародної співпраці у сфері регулювання та обміну досвідом для ефективного протистояння кіберзагрозам, пов'язаним із використанням ШІ.

Таким чином, робота зробила внесок у розуміння природи вразливостей ШІ та механізмів їхнього використання у кіберзагрозах, що сприяє розробці стратегій захисту та зменшенню ризиків для сучасного суспільства.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Anderson, J., & McGee, M. (2023). Adversarial machine learning: Vulnerabilities and solutions in AI systems. International Journal of Cybersecurity, 15(2), 45-63. https://doi.org/10.1016/j.jcyb.2023.04.007

2. Binns, R., & Thomas, J. (2022). Data poisoning attacks in machine learning: A survey. Journal of Artificial Intelligence Security, 9(4), 78-89.

3. ZAGORODNA, N., STADNYK, M., LYPA, B., GAVRYLOV, M., & KOZAK, R. (2022). Network Attack Detection Using Machine Learning Methods. Challenges to national defence in contemporary geopolitical situation, 2022(1), 55-61.

4. Gupta, A., & Singh, R. (2021). AI-based attack vectors and their implications on cybersecurity. Cybersecurity Research Review, 12(1), 34-42.

5. Lypa, B., Horyn, I., Zagorodna, N., Tymoshchuk, D., Lechachenko T., (2024). Comparison of feature extraction tools for network traffic data. CEUR Workshop Proceedings, 3896, pp. 1-11.

6. Liu, W., & Zhou, H. (2021). The role of machine learning in phishing detection systems. Journal of Information Security, 16(1), 102-115.

7. Tymoshchuk, D., Yasniy, O., Mytnyk, M., Zagorodna, N., Tymoshchuk, V., (2024). Detection and classification of DDoS flooding attacks by machine learning methods. CEUR Workshop Proceedings, 3842, pp. 184 - 195.

8. Mozaffari, M., & Johnson, L. (2023). Secure deep learning models: Protection against adversarial attacks. Machine Learning and Security, 19(1), 88-98.

9. Nguyen, T., & Tran, B. (2021). A review of AI-powered malware detection systems. Journal of Computing and Security, 13(4), 130-145.

10. ТИМОЩУК, Д., ЯЦКІВ, В., ТИМОЩУК, В., & ЯЦКІВ, Н. (2024). INTERACTIVE CYBERSECURITY TRAINING SYSTEM BASED ON SIMULATION ENVIRONMENTS. MEASURING AND COMPUTING DEVICES IN TECHNOLOGICAL PROCESSES, (4), 215-220.

11. Pomerleau, D., & Wessling, L. (2022). AI and data integrity in security applications. Journal of AI and Privacy, 11(2), 78-92.

12. Quintero, M., & Rivera, R. (2021). Ethical challenges in AI-driven cybersecurity tools. AI Ethics, 5(3), 245-257.

13. ТИМОЩУК, Д., & ЯЦКІВ, В. (2024). USING HYPERVISORS TO CREATE A CYBER POLYGON. MEASURING AND COMPUTING DEVICES IN TECHNOLOGICAL PROCESSES, (3), 52-56.

14. Shultz, M., & Turner, R. (2020). The future of AI in cybersecurity: Risks and mitigation strategies. Cybersecurity Journal, 18(2), 101-116.

15. Tymoshchuk, D., Yasniy, O., Maruschak, P., Iasnii, V., & Didych, I. (2024). Loading Frequency Classification in Shape Memory Alloys: A Machine Learning Approach. Computers, 13(12), 339.

16. Smith, M., & Thompson, J. (2022). AI in cyber-attacks: Mechanisms and defenses. International Conference on AI and Security, 134-142.

17. Talbot, B., & Li, Z. (2020). Understanding adversarial robustness in AI models. Machine Learning and Cybersecurity, 22(1), 100-110.

18. Tymoshchuk, D., & Yatskiv, V. (2024). Slowloris ddos detection and prevention in real-time. Collection of scientific papers «ΛΌГОΣ», (August 16, 2024; Oxford, UK), 171-176.

19. Williams, H., & Harris, K. (2020). Machine learning in cybersecurity: An exploration of vulnerabilities. Proceedings of the International Cybersecurity Symposium, 45-53.

20. Lupenko, S., Orobchuk, O., Osukhivska, H., Xu, M., & Pomazkina, T. (2019, July). Methods and means of knowledge elicitation in Chinese Image Medicine for achieving the tasks of its ontological modeling. In 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON) (pp. 855-858). IEEE.

21. Zadeh, N., & Alavi, M. (2021). Impact of AI algorithms on cybersecurity policies. Journal of AI Governance, 12(1), 45-59.

22. Zhao, Q., & Jiang, X. (2023). AI-driven malware detection: Challenges and techniques. Security and Privacy Journal, 17(2), 23-37.

23. Zhou, Y., & Liu, F. (2021). AI and its role in automated cyber-attack strategies. Cybersecurity Technology, 20(4), 190-202.

24. S. A. Lupenko, O. R. Orobchuk and N.V. Zahorodna, "Formation of the onto-oriented electronic educational environment as a direction the formation of integrated medicine using the example of CIM", Actual scientific research in the modern world: Collection of scientific papers of the XXIII International scientific conference, vol. 12, no. 32, pp. 56-61, 2017, December 26.

25. Zhang, Z., & Wang, D. (2021). Anomaly detection in AI systems: Challenges and solutions. Machine Learning and Security Review, 18(3), 74-88.

26. Yasniy, O., Tymoshchuk, D., Didych, I., Zagorodna, N., Malyshevska O., (2024). Modelling of automotive steel fatigue lifetime by machine learning method. CEUR Workshop Proceedings, 3896, pp. 165-172.

27. Davis, S., & Lin, Q. (2023). AI-powered intrusion detection systems and their limitations. Journal of Network Security, 31(1), 142-156.

28. Nguyen, M., & Kim, H. (2022). AI as a tool for cybersecurity automation and the associated risks. AI in Security Technology, 8(3), 56-67.

29. Tymoshchuk, V., Vantsa, V., Karnaukhov, A., Orlovska, A., & Tymoshchuk, D. (2024). COMPARATIVE ANALYSIS OF INTRUSION DETECTION APPROACHES BASED ON SIGNATURES AND ANOMALIES. Матеріали конференцій МЦНД, (29.11. 2024; Житомир, Україна), 328-332.

30. Patel, R., & Sharma, A. (2023). Ethical challenges in the use of AI for security purposes. Ethics in AI, 6(1), 21-33.

31. Lupenko, S., Orobchuk, O., & Xu, M. (2020). The ontology as the core of integrated information environment of Chinese Image Medicine. In Advances in Computer Science for Engineering and Education II (pp. 471-481).

32. Here are the citations for the publications you requested, formatted according to your example: ENISA's "Artificial Intelligence Cybersecurity Challenges": ENISA. (2020). Artificial intelligence cybersecurity challenges. European Union Agency for Cybersecurity. Retrieved from [https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges]

33. Department of Energy's Insights on AI Vulnerabilities: U.S. Department of Energy. (2020). Insights into artificial intelligence vulnerabilities. Retrieved from [https://inldigitallibrary.inl.gov]

Додаток А Публікація

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
імені ІВАНА ПУЛЮЯ
НАУКОВЕ ТОВАРИСТВО ім. ШЕВЧЕНКА

М А Т Е Р І А Л И

XII НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

18-19 грудня 2024 року

ТЕРНОПІЛЬ
2024

УДК 001
М34

ПРОГРАМНИЙ КОМІТЕТ

Голова: Приймак Микола – професор кафедри комп'ютерних систем та мереж, д.т.н., професор.
Співголови: Марущак Павло – проректор з наукової роботи, докт. техн. наук, професор.
Баран Ігор – канд. техн. наук, доцент, декан факультету ФІС.
Науковий секретар: Надія Крива– старший викладач.

Члени: Василь Кривень - завідувач кафедри математичних методів в інженерії д.ф.-м.н., професор; Галина Осухівська – завідувач кафедри комп'ютерних систем та мереж, к.т.н., доцент; Микола Карпінський - професор кафедри кібербезпеки, д.т.н., професор; Жанна Баб'як - завідувач кафедри української та іноземних мов, к.пед. н., доцент; Ярослав Литвиненко – професор кафедри комп'ютерних наук, д.т.н., професор; Михайло Петрик - завідувач кафедри програмної іженерії, д.ф.-м.н., професор; Наталія Загородна – завідувач кафедри кібербезпеки, к.т.н., доцент.

ОРГАНІЗАЦІЙНИЙ КОМІТЕТ

Голова: Скоренький Юрій Любомирович – канд. техн. наук, доцент кафедри фізики.
Члени: доцент кафедри комп'ютерних наук, к.т.н. В. Никитюк; доцент кафедри програмної інженерії, к.т.н. Д. Михалик; доцент кафедри кібербезпеки, к.т.н. М. Стадник; доцент кафедри комп'ютерних систем та мереж, к.т.н. Є. Тиш; ст. викладач Л. Джиджора.

М34 Матеріали XII науково-технічної конфції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя, (Тернопіль, 18–19 грудня 2024 р.). – Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2024.

Адреса оргкомітету: ТНТУ ім. І. Пулюя, м. Тернопіль, вул. Руська, 56, 46001,
тел. (0352) 52-41-33, факс (0352) 25-49-83.
E-mail: conffis2024@gmail.com

Редагування, оформлення та верстка: Надія Крива

СЕКЦІЇ КОНФЕРЕНЦІЇ, ЯКІ ПРЕДСТВЛЕНІ В ЗБІРНИКУ

- Математичне моделювання
- Інформаційні системи та технології, кібербезпека
- Комп'ютерні системи та мережі
- Програмна інженерія та моделювання складних розподілених систем
- Новітні фізико-технічні та освітні технології

В збірнику надруковано тези доповідей XII науково-технічної конференції «Інформаційні моделі, системи та технології» (Тернопіль, 18–19 грудня 2024 р.) за такими науковими напрямками: математичне моделювання; інформаційні системи та технології, кібербезпека; комп'ютерні системи та мережі; програмна інженерія та моделювання складних розподілених систем; новітні фізико-технічні та освітні технології.

Розрахований на науковців, викладачів та студентів вузів.

За зміст тез та дотримання норм академічної доброчесності відповідальність несе автор.

УДК 630,247

Ph.D Олександра Оробчук; Анастасія Мазяр

Тернопільський національний технічний університет імені Івана Пулюя

«АНАЛІЗ ВРАЗЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ТА МЕТОДІВ ЇХНЬОГО ВИКОРИСТАННЯ В КІБЕРЗАГРОЗАХ»

Ph.D Oleksandra Orobchuk, Anastasia Maziar

«ANALYSIS OF ARTIFICIAL INTELLIGENCE VULNERABILITIES AND METHODS OF THEIR USE IN CYBER THREATS»

Стрімкий розвиток штучного інтелекту (ШІ) відкриває нові горизонти для інновацій, проте разом із цим створює потенційні ризики кібербезпеки. Зловмисники активно досліджують і використовують вразливості ШІ для реалізації складних кіберзагроз, що вимагає поглибленого вивчення цієї проблематики.

У сучасному середовищі штучного інтелекту існує кілька типів вразливостей, які можуть використовуватися зловмисниками. Перш за все, варто зазначити, що одним із поширених видів атак є вплив на вхідні дані. Це може проявлятися у вигляді отруєння даних, коли до системи вводяться некоректні або навмисно шкідливі дані, що порушують її роботу, або створення спеціальних шкідливих прикладів, які вводять моделі в оману. Додатково, моделі ШІ самі можуть мати внутрішні слабкі сторони, як-от перенавчання чи нездатність коректно працювати з неточними чи некоректними даними. Ще однією серйозною загрозою є використання моделей для викрадення конфіденційної інформації, що може реалізуватися через атаки інверсії моделей.

Такі вразливості вже активно використовуються у реальних кіберзагрозах. Генеративні моделі, наприклад, можуть допомагати у проведенні фішингових кампаній та здійсненні атак із використанням соціальної інженерії. Ще одним небезпечним застосуванням є створення deepfake-контенту, що використовується для маніпуляцій і поширення дезінформації. Окрім того, автономні системи, зокрема дрони та транспортні засоби, також мають свої вразливі місця, які можуть бути використані зловмисниками для створення загроз.

Таким чином, проблема вразливостей ШІ потребує системного підходу для забезпечення кібербезпеки. Запропоновані рекомендації можуть бути інтегровані у сучасні стратегії безпеки.

Література

1. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Practical black-box attacks against machine learning. Proceedings of the ACM on Artificial Intelligence.

2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition.

3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Додаток Б Лістинг файлу

1)

```
import tensorflow as tf
from tensorflow.keras import layers, models
import numpy as np
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

     features = logs[['response_time', 'request_type']].copy()
     features = pd.get_dummies(features,
     columns=['request_type'])

     scaler = MinMaxScaler()
     features_scaled = scaler.fit_transform(features)



     split_index = int(0.8 * len(features))
     train_data = features_scaled[:split_index]
     test_data = features_scaled[split_index:]

     model = models.Sequential([
             layers.Dense(64, activation='relu',
             input_shape=(train_data.shape[1],)),
              layers.Dense(32, activation='relu'),
              layers.Dense(64, activation='relu'),
              layers.Dense(train_data.shape[1],
          activation='sigmoid')
          ])

model.compile(optimizer='adam', loss='mse')
model.summary()

model.fit(train_data, train_data, epochs=10, batch_size=32,
validation_split=0.2)

reconstructed_data = model.predict(test_data)
```

```
    loss = model.evaluate(test_data, test_data)
    print(f"Loss on test data: {loss}")
```

2)

```
 import tensorflow as tf
from tensorflow.keras import layers, models
import numpy as np
import pandas as pd
from sklearn.preprocessing import MinMaxScaler


# logs = pd.read_csv('data.csv')
logs = pd.DataFrame({
    'response_time': np.random.rand(1000) * 100,
    'request_type':   np.random.choice(['GET',   'POST',   'PUT',
'DELETE'], size=1000)
})


print
features = logs[['response_time', 'request_type']].copy()

# Перетворення категорійних даних у числові (One-Hot Encoding)
features = pd.get_dummies(features, columns=['request_type'])
print(f"Розмірність    даних    після    One-Hot    Encoding:
{features.shape}")


scaler = MinMaxScaler()
features_scaled = scaler.fit_transform(features)
print("Дані масштабовано до діапазону [0, 1]")

split_index = int(0.8 * len(features_scaled))
train_data = features_scaled[:split_index]
test_data = features_scaled[split_index:]
print(f"Кількість   тренувальних   зразків:   {len(train_data)},
тестових: {len(test_data)}")
```

```
print("Побудова моделі автоенкодера...")
model = models.Sequential([
    layers.Dense(64,                                activation='relu',
input_shape=(train_data.shape[1],)),
    layers.Dense(32, activation='relu'),
    layers.Dense(64, activation='relu'),
    layers.Dense(train_data.shape[1], activation='sigmoid')
])

model.compile(optimizer='adam', loss='mse')
model.summary()

print("Навчання моделі...")
history = model.fit(
    train_data, train_data,
    epochs=10,
    batch_size=32,
    validation_split=0.2,
    verbose=1
)

print("Оцінка моделі на тестових даних...")
reconstructions = model.predict(test_data)

print
mse = np.mean(np.power(test_data - reconstructions, 2), axis=1)

threshold = np.percentile(mse, 95)  # 95-й перцентиль
print(f"Поріг для аномалій (95-й перцентиль): {threshold}")

print
anomalies = mse > threshold
num_anomalies = np.sum(anomalies)
print(f"Кількість аномалій: {num_anomalies}")

anomaly_indices = np.where(anomalies)[0]
```

```
print(f"Індекси аномальних запитів: {anomaly_indices}")

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
plt.hist(mse, bins=50, alpha=0.7, color='blue', label='MSE для тестових даних')
plt.axvline(threshold, color='red', linestyle='dashed', linewidth=2, label='Поріг аномалій')
plt.title('Розподіл MSE для тестових даних')
plt.xlabel('MSE')
plt.ylabel('Кількість')
plt.legend()
plt.show()

print
anomalous_data = logs.iloc[split_index:][anomalies]
print(anomalous_data.head())
```

3)

```
from flask import Flask, request, jsonify
import numpy as np
import tensorflow as tf
from sklearn.preprocessing import MinMaxScaler

# Ініціалізація Flask-додатка
app = Flask(__name__)


model = tf.keras.models.load_model('autoencoder_model.h5')


scaler = MinMaxScaler()
scaler.fit(np.random.rand(1000, 10))
```

```
threshold = 0.1


@app.route('/detect', methods=['POST'])
def detect_anomaly():
    """

    """
    try:

        data = request.json
        if 'features' not in data:
            return jsonify({'error': 'Missing "features" in request data'}), 400


        features = np.array(data['features']).reshape(1, -1)
        features_scaled = scaler.transform(features)


        reconstruction = model.predict(features_scaled)
        mse = np.mean(np.power(features_scaled - reconstruction, 2), axis=1)


        is_anomaly = mse > threshold


        return jsonify({'anomaly': bool(is_anomaly[0]), 'mse': float(mse[0])})

    except Exception as e:
        # Обробка помилок
        return jsonify({'error': str(e)}), 500
```

```
if __name__ == '__main__':
    print("Запуск Flask-додатка...")
    app.run(debug=True, host='0.0.0.0', port=5000)
```