УДК 378.046
DOI https://doi.org/10.32782/tnv-tech.2023.4.2

КЛАСИФІКАЦІЇ МОДЕЛЕЙ ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ У КІБЕРБЕЗПЕЦІ

Антоненко А. В. *– кандидат технічних наук, доцент,*
доцент кафедри стандартизації та сертифікації сільськогосподарської продукції
Національного університету біоресурсів і природокористування України
ORCID ID: 0000-0001-9397-1209

Бенедіко І. В. *– магістр*
Державного університету інформаційно-телекомунікаційних технологій
ORCID ID: 0009-0009-5544-8391

Вічкарук А. І. *– магістр*
Державного університету інформаційно-телекомунікаційних технологій
ORCID ID: 0009-0005-2531-3905

Лисенко К. В. *– магістр*
Державного університету інформаційно-телекомунікаційних технологій
ORCID ID: 0009-0005-1625-9679

Сижко О. Ю. *– магістр*
Державного університету інформаційно-телекомунікаційних технологій
ORCID ID: 0009-0004-8846-4041

В статті розглянуто взаємозв'язок між штучним інтелектом (ШІ) та кібербезпекою, аналізуючи важливі виклики та можливості, що виникають у зв'язку зі швидким розвитком цих двох сфер. У сучасному світі, де штучний інтелект стає все більш поширеним і використовується в різних галузях, кібербезпека стає одним з найважливіших аспектів забезпечення безпеки та захисту інформації. Стаття пояснює, що хоча ШІ може приносити значні переваги, він також створює нові загрози та ризики для кібербезпеки. В статті запропоновано комплексний огляд взаємозв'язку між цими двома сферами, зосереджуючись на викликах та можливостях, пов'язаних зі штучним інтелектом у контексті кібербезпеки. Вона ставить акцент на необхідності розробки ефективних заходів для захисту від загроз, що виникають у зв'язку зі штучним інтелектом, та наголошує на постійному вдосконаленні стратегій кібербезпеки для забезпечення безпеки та захисту інформації. У сучасному трактуванні системи штучного інтелекту – це системи машинного навчання, іноді це ще більше звужується до штучних нейронних мереж. Якщо ми говоримо про все ширше проникнення машинного навчання у різні сфери застосування інформаційних технологій, то, природно, що мають виникати перетини з кібербезпекою. Але проблема в тому, що такий перетин не може бути описаний якоюсь однією моделлю. Поєднання Штучний інтелект та кібербезпека мають безліч різних аспектів застосування. Загальним є, природно, використання методів машинного навчання, але завдання, і навіть досягнуті нині результати, є різними. Наприклад, якщо застосування машинного навчання виявлення атак і вторгнень показує реальні досягнення проти застосовувалися раніше підходами, то атаки самі системи машинного навчання поки повністю перемагають можливі захисти. Класифікації моделей застосування машинного навчання у кібербезпеці і присвячена дана стаття.

Ключові слова: *штучний інтелект, машинне навчання, кібербезпека, кібератака, автоматизація.*

Antonenko A. V., Benediko I. V., Vikarchuk A. I., Lysenko K. V., Syzhko O. Yu. Classifications of machine learning application models in cyber security

The article examines the relationship between artificial intelligence (AI) and cybersecurity, analyzing the important challenges and opportunities arising from the rapid development of these two fields. In today's world, where artificial intelligence is becoming more common and used in various industries, cyber security is becoming one of the most important aspects of information security and protection. The article explains that while AI can bring significant benefits, it also creates new threats and risks to cybersecurity. Overall, the article "Artificial Intelligence and Cybersecurity" offers a comprehensive overview of the relationship between these two fields, focusing on the challenges and opportunities associated with artificial intelligence in the context of cybersecurity. It emphasizes the need to develop effective measures to protect against threats arising from artificial intelligence and emphasizes the continuous improvement of cyber security strategies to ensure the security and protection of information. In the modern interpretation, artificial intelligence systems are machine learning systems, sometimes this is further narrowed down to artificial neural networks. If we are talking about the ever-widening penetration of machine learning into various areas of application of information technologies, then naturally there should be intersections with cyber security. But the problem is that such an intersection cannot be described by any one model. The combination of artificial intelligence and cyber security has many different application aspects. Common is, of course, the use of machine learning methods, but the tasks, and even the results achieved today, are different. For example, if the application of machine learning to detect attacks and intrusions shows real achievements against previously used approaches, then the attacks of the machine learning systems themselves completely defeat possible defenses. This article is devoted to the classification of machine learning application models in cyber security.

Key words: *artificial intelligence, machine learning, cyber security, cyber attack, automation.*

Вступ. Штучний інтелект сьогодні перевизначив те, як використовуються комп'ютери [1]. Штучний інтелект стає частиною повсякденного життя. Навіть такі абсолютно зрозумілі пристрої користувача, як мобільні телефони вже містять чіпи для штучного інтелекту (Pixel 6 від Google, iPhone). ШІ змінює те, як комп'ютери програмуються та як вони використовуються. Завдяки машинному навчанню програмісти більше не пишуть правил. Натомість вони створюють нейронну мережу, яка сама витягує ці правила у процесі навчання. Це інший спосіб мислення.

Постановка проблеми. Штучний інтелект (а на сьогоднішній день – це машинне навчання) всюди, комп'ютерна безпека повинна охоплювати всі процеси, відповідно, ці два поняття не могли не зустрітися. Саме взаємозв'язок штучного інтелекту та кібербезпеки і є темою цієї статті. Ці взаємозв'язки різні, рішення існують абсолютно різні, і рівень вирішення різних проблем також різний. Тема Штучний Інтелект і кібербезпека не може бути представлена як одне рішення (або навіть сукупність кількох рішень), оскільки вона описує різні завдання.

Метою дослідження є аналіз взаємозв'язку між штучним інтелектом та кібербезпекою з метою визначення важливості заходів забезпечення кібербезпеки у контексті розвитку штучного інтелекту.

Предметом дослідження є взаємозв'язок між штучним інтелектом та кібербезпекою, зокрема аналіз технологій штучного інтелекту, які впливають на кібербезпеку, і розгляд можливих загроз, які виникають у зв'язку з використанням штучного інтелекту.

Об'єктом дослідження є штучний інтелект та кібербезпека.

Аналіз останніх досліджень і публікацій. Важливості питань інформаційної безпеки нашої країни і формуванню механізму міжнародної кібербезпеки приділяли увагу численні науковці. Так, Безуглий Д.С. обґрунтував необхідність інформаційної безпеки як складової частини національної безпеки країни [1]. Аналіз останніх досліджень і публікацій свідчить про те, що певні аспекти вітчизняних проблем інформаційної безпеки у той чи інший спосіб досліджувались у наукових

працях Арістова І.В., Березовської І.Р., Дзьобаня О.П., Калюжного Р.А., Кормича Б.А., Ліпкана В.А., Марущак А.І., Цимбалюка В.С., Юдіна О.К. та інших.

Тему штучного інтелекту та кібербезпеки розгядється в таких джерелах, як "AI & Society" – науковий журнал, присвячений дослідженню взаємозв'язку між штучним інтелектом та суспільством; "Journal of Artificial Intelligence Research" – науковий журнал, що публікує оригінальні дослідження у галузі штучного інтелекту; "Conference on Neural Information Processing Systems" (NeurIPS) – одна з найбільш визначних конференцій у галузі штучного інтелекту; "ITEM – Інформаційно-технологічний економіко-математичний журнал" – український науковий журнал, присвячений інформаційним технологіям; "Data Science UA" – український портал про аналіз даних та штучний інтелект, який містить статті про кібербезпеку; "Communications of the ACM" – журнал, що охоплює широкий спектр тем, пов'язаних з інформаційними технологіями, включаючи штучний інтелект та кібербезпеку; "IEEE Security & Privacy" – журнал, присвячений кібербезпеці та приватності, "Communications of the ACM" – журнал, що охоплює широкий спектр тем, пов'язаних з інформаційними технологіями, включаючи штучний інтелект та кібербезпеку;); "The Global AI Index" – звіт, що оцінює глобальну ситуацію з розвитку штучного інтелекту та інновацій у цій галузі; "Cybersecurity Ventures" – аналітична компанія, що займається дослідженням та прогнозуванням трендів у сфері кібербезпеки.

Виклад основного матеріалу дослідження. Підвищення кібербезпеки за допомогою ШІ – це найбільш просунута на сьогоднішній день область. Цінність, яку привносить тут машинне навчання, полягає у визначенні атак, пошуку шаблонів та закономірностей, що відповідають вторгненням, швидкому аналізі та пріоритизації загроз, аналізі накопиченої інформації для адаптації методів виявлення вторгнення.

Перша відповідь на питання, навіщо тут ШІ, згідно [2], полягає у слові «автоматизація». Автор наводить американські дані Бюро статистики праці США про те, що можливості працевлаштування у сфері кібербезпеки зростуть на 33% з 2020 по 2030 рік, що у понад шість разів перевищує середній показник по країні. Навряд чи картина інших країнах відрізняється від наведеної. При цьому, згідно з дослідженням ринку праці в частині кібербезпеки ISC, опублікованому в жовтні 2021 року, у всьому світі не вистачає 2.72 мільйона фахівців з кібербезпеки. Відповідно, альтернативиавтоматизації вирішення завдань кібербезпеки просто немає.

Завдання кібербезпеки складаються з запобігання атакам, виявлення атак, проведення розслідувань, класифікації та аналізу загроз, а також навчання та моделювання систем кібербезпеки.

Запобігання атакам (профілактика) – це зусилля щодо зниження кількості вразливостей, що містяться в програмному забезпеченні. Типові приклади є, наприклад, в огляді, який описує системи машинного навчання, які виконують пошук шкідливих програм на Android.

У 2021 році Інститут AV-Test виявив понад 125 мільйонів нових шкідливих програм. Здатність методів машинного навчання узагальнювати минулі шаблони для виявлення нових варіантів шкідливих програм і є ключем до побудови системи захисту, що масштабується.

Можна відзначити, що пошук у Google Scholar робіт на запит "ML for malware detection" показує понад 20 000 статей.

Виявлення атак включає виявлення підозрілої поведінки та оповіщення про неї безпосередньо в міру її виникнення. Мета полягає в тому, щоб швидко реагувати

на атаки, включаючи визначення масштабу атаки, закриття входів для атакуючих та усунення вразливостей (бекдорів тощо), які міг експлуатувати зловмисник.

Очевидно, що пошук у загальному випадку невідомих шаблонів атак потенційно може призводити до великої кількості помилкових спрацьовувань (false positives) [3]. У літературі наголошується, що основна проблема при виявленні підозрілої активності якраз і полягає в тому, щоб знайти правильний баланс між забезпеченням достатнього охоплення за рахунок пошуку точних попереджень системи безпеки та кількості хибних спрацьовувань.

Можна виділити такі напрямки, що стосуються використання машинного навчання для попередження атак [2]:

– розстановка пріоритетів для попереджень про потенційні атаки;

– виявлення численних спроб злому з плином часу, які є частиною більших і тривалих кампаній зі злому;

– виявлення слідів дій шкідливих програм, як усередині комп'ютера, і у мережі;

– ідентифікація потоку шкідливого програмного забезпечення, що запроваджується через конкретну організацію. Це так звані Living off the Land (LotL) атаки – кібератаки, в яких атакуючий використовує легальне програмне забезпечення в організації для виконання атакуючих дій [4];

– визначення автоматизованих підходів до пом'якшення наслідків атак, коли потрібне швидке реагування, щоб запобігти поширенню атаки. Наприклад, автоматизована система може відключати мережне підключення та блокувати пристрій, якщо виявляється послідовність попереджень, яка, як відомо, пов'язана з діями програми-вимагача.

Розслідування та виправлення (відновлення після атак) – це методи, що використовуються після порушення безпеки, призначені для того, щоб надати клієнтам цілісне уявлення про порушення безпеки, включаючи ступінь порушення, список порушених пристроїв та даних, інформацію про поширення атаки та причини інциденту. У зв'язку з атаками використовується термін наступальний ШІ. Рис. 1 підсумовує напрями атак із використанням систем машинного навчання на матриці загроз MITRE.

Виділяються такі області атак із використанням ШІ: прогнозування, генерація, аналіз, пошук, ухвалення рішення.

Прогнозування – зробити деякий прогноз на основі даних, що раніше спостерігалися. Приклад атаки з використанням машинного навчання – ідентифікація натискання клавіш на смартфоні на основі руху (вібрації). Інші наведені приклади стосувалися передбачення чутливих даних для користувачів соціальних мереж (пошук слабкої ланки для атаки), пошуку вразливостей програмного забезпечення (рис. 1).

Генерація – створення контенту з використанням ШІ. Приклади такої генерації для наступальних цілей – фальсифікація медіа-даних, добір паролів, модифікацію трафіку. Останнє (в англомовній літературі – traffic-space attacks) є, фактично, змагальною атакою на систему машинного навчання, яка використовується для аналізу трафіку (визначення вторгнень). Мета атаки – приховати реальне вторгнення. Дипфейки – ще один приклад наступального ШІ у цій категорії. Діпфейк – це правдоподібний медіафайл. Створюються вони з глибокого навчання. Технологія може бути використана для того, щоб видавати себе за жертву, імітуючи її голос або особу під час фішингової атаки.

Аналіз – це завдання аналізу чи вилучення корисної інформації з даних чи моделі. Дослідження атакованої моделі ML з метою визначення реальних факторів,

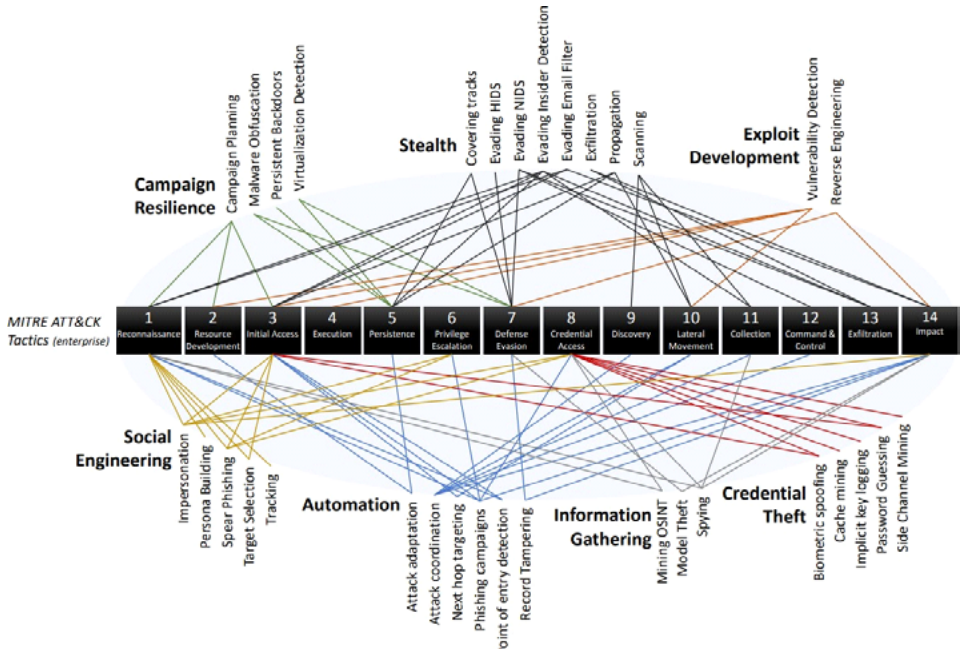

Рис. 1. Машинне навчання у кібератаках

що впливають, наприклад, на класифікацію. Мається на увазі використання пояснюючих підходів (LIME, SHAPLEY та ін.). Розуміння роботи атакованої моделі необхідне створення ефективних атак чи приховування вторгнень. Якщо модель, що атакується, недоступна, то такі експерименти можуть проводитися на її тіньовій копії;

Пошук – це завдання пошуку інформації або об’єктів для атаки за заданими критеріями. Наведені приклади – пошук (ідентифікація) людини за зображеннями на кількох зламаних камерах, пошук можливих інсайдерів з семантичного аналізу публікацій у соціальних мережах, анотування (реферування, сумаризація) документів при зборі даних із відкритих джерел (OSINT – відкрита розвідка) (останнє є приклад автоматизації).

Ухвалення рішення – це завдання розробки стратегічного плану чи координації операції (атаки). Приклади в ШІ – використання роєвого інтелекту для управління автономною мережею ботів та планування оптимальних атак на мережі [5]: Одна з найбільш успішно використовуваних систем автоматизації у наступальному ШІ – це боти у соціальних мережах. Інший приклад автоматизації наступальних процесів – автоматизований тест на проникнення (penetration test), використовує навчання з підкріпленням (рис. 2).

Машинне навчання використовується для атак на біометричні системи аутентифікації: підробка голосу тощо. Машинне навчання використовується і для генерації фішингових атак. Мета – обійти системи захисту, створити привабливіший контент і спонукати користувачів клікнути зловмисне посилання, встановити в системі програмне забезпечення і т.д. Приклади наступальних дій включають підбір паролів, заплутування вихідного коду програм, маскування трафіку, управління мережею роботів.

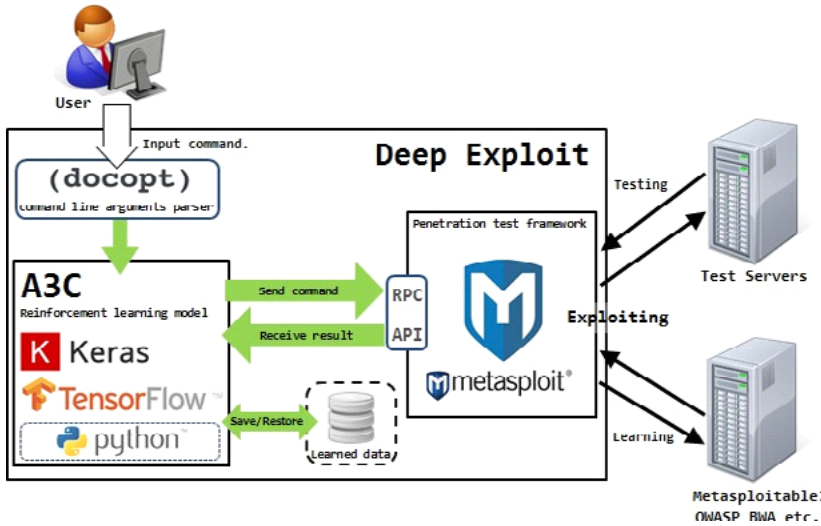

Рис. 2. Deep Exploit

Атаки на системи ШІ досить нова область для комп'ютерної безпеки. Атаки можуть бути спрямовані на самі ШІ системи (фактично – на системи машинного навчання). Будь-яка впроваджена система машинного навчання є, зрештою, програма. Але проблема полягає в тому, що для таких програм традиційні методи аналізу безпеки не застосовні. Проблеми з безпекою саме таких програм не можуть бути вирішені традиційними методами. Звичайно, скомпрометоване середовище виконання програми призводитиме до проблем. Але це не головне лихо [6].

Системи машинного навчання залежить від даних. На основі представлених тренувальних даних система виробляє деякі узагальнення, які потім використовуються для обробки реальних (тестових) даних. Так ось модифікації даних на різних етапах конвеєра машинного навчання і призводять до того, що такі системи можуть або не працювати, або навпаки, видавати потрібні атакуючому результати. При цьому спеціально модифіковані дані будуть, взагалі кажучи, такими ж, як і «чисті» дані. Загалом їх не можна буде розрізнити. Більше того, оскільки навчання завжди проводиться на деякому тренувальному наборі даних, генеральна сукупність залишається у загальному випадку невідомою. І «зміна» даних на етапі експлуатації може статися (і найчастіше трапляється) без будь-яких шкідливих дій. Просто тому, що так влаштовані дані. Атаками у разі називають саме спеціальну зміна даних чи спеціальну підстановку даних, у яких система працює неправильно (взагалі не працює). Загалом – це проблема стійкості систем машинного навчання. Цій проблемі зараз приділяється багато уваги, оскільки це основне, що перешкоджає використанню систем ШІ у критичних додатках (авіоніка, ядерна безпека тощо).

Інша назва атак на системи машинного навчання – змагальні приклади. Таким чином, ворожі дії на системи можуть здійснюватися у формі традиційних уразливостей, а також за допомогою нової категорії: змагальних прикладів. Як приклади традиційних уразливостей можна вказати, наприклад, звіт про уразливості в програмному пакеті Tensorflow [7], що, природно, означає наявність уразливостей у системах, що використовують його. Дослідники з Нью-Йоркського університету виявили, що більшість середовищ ШІ не перевіряють цілісність завантажених

моделей ШІ, на відміну від загальноприйнятої практики з традиційним програмним забезпеченням, де криптографічна перевірка файлів/бібліотек, що виконуються, є стандартною практикою вже більше десяти років. Громадські датсети можуть містити помилки в розмітці, що, природно, впливає на роботу навчених за їх допомогою систем [8].

Також атаки бувають цільові (наприклад, атакуючий хоче досягти певного результату від класифікатора) та нецільові (просто перешкодити правильній роботі класифікатора).

Модифікацію вхідних даних (за фактом найпоширеніший тип атаки) ще називають атаками ухилення. Крадіжка (IP stealing) включає отримання відомостей про модель (а це потрібно для організації атак) і так звані інверсні атаки, які спрямовані на відновлення лежать в основі приватних даних, використаних для навчання цільової системи.

Мікрософт [9] зазначає, що кількість таких атак зростає. Насамперед, це стосується, звісно, критичних застосувань. У роботі описуються зусилля США та Китаю щодо протидії системам ШІ один одного. Загалом, через відсутність повного захисту, такі атаки доводиться сприймати як певний універсальний ризик, пов'язаний із використанням систем машинного навчання. При цьому необхідно враховувати можливість здійснення атаки, так і практичну здійсненність таких атак.

Очевидно, що модифікувати вхідні дані можна практично завжди. Наприклад, так звані фізичні атаки (зміна форми подання), є одними з найбільш здійсненних і небезпечних для систем розпізнавання. Простий приклад фізичної атаки – камуфляж (захисне забарвлення) [10]. Для організації атак ухиленням використовують як прості модифікації даних (наприклад, атака Salt & Pepper – додавання чорних та білих точок до зображення, так і спеціальні рішення з використання машинного навчання, наприклад, моделей, що породжують. Отруєння даних можна, очевидно, уникнути, якщо використовувати власні перевірені набори даних, уникати використання даних з невідомих джерел або перевіряти такі дані.

Крадіжка даних і моделі технічно пов'язані з аналізом безлічі відгуків атакованої системи спеціальним чином підготовлені вхідні дані. Якщо це рішення ML as a service, то способу опитувати систему може просто бути. Але якщо не можна опитувати саму модель, можна спробувати створити її копію (shadow model) і відпрацьовувати атаки на ній. Звідси випливає висновок у тому, що на відміну класичного програмного забезпечення, де самі алгоритми найчастіше відкриті, для систем машинного навчання деталі реалізації моделей у критичних областях повинні ховатися, оскільки такі знання дозволять побудувати тіньову модель (копію моделі) для відпрацювання атак.

Загалом атаки ухиленням (тобто модифікація вхідних даних) є головною практичною проблемою. На сьогоднішній день атаки в цій галузі випереджають захист. І це є основною перешкодою для впровадження систем машинного навчання в критичні програми. В окремих випадках (залежно від даних та розміру моделі) можна говорити про формальні докази стійкості систем машинного навчання [11]. В інших випадках підходи до формального доказу стикатимуться з трендом на збільшення параметрів сучасних мереж. У більшості випадків «захист» складається з включення модифікованих даних до тренувальних наборів та обліку таким чином можливих модифікацій даних за рахунок точності системи. Питання, що це не всі можливі модифікації, як правило, ігнорується.

Як було зазначено вище, основний напрям робіт тут – це створення стійких систем (моделей) машинного навчання. З практичної точки зору, для розробки

систем машинного навчання для критичних застосувань необхідні так звані довірені середовища розробки, які гарантують відсутність компрометації інструментальних засобів та представляють інструменти для підвищення довіри до результатів роботи систем.

Слід зазначити, що проблеми із захистом систем ШІ повністю усвідомлюються, як у промисловому, і у індустріальному співтоваристві. Є широко відомий каталог MITRE, що підтримується Мікрософт та іншими організаціями, в якому збирається інформація щодо атак на системи ШІ. Зокрема, у ньому є так звана матриця загроз Adversarial ML для каталогізації загроз для ШІ систем. Для інженерів та політиків Microsoft у співпраці з Центром Беркмана Кляйна у Гарвардському університеті випустила таксономію режимів збоїв машинного навчання. DARPA пропонує безкоштовні ресурси з метою оцінки безпеки систем машинного навчання. Мікрософт пропонує свій продукт з відкритим кодом Counterfit, як інструмент оцінки безпеки систем ШІ. Міністерство оборони США включило безпеку систем ШІ до свого списку основних принципів побудови ШІ. Американський інститут стандартів NIST працює над схемою оцінки ризиків ШІ, спрямованої на вирішення безлічі аспектів систем ШІ, включаючи надійність та безпеку [12].

Досягнення в галузі машинного навчання та комп'ютерної графіки розширили можливості державних та недержавних суб'єктів з виробництва та розповсюдження високоякісного аудіовізуального контенту, званого синтетичними медіа та дипфейками. Технології штучного інтелекту для створення дипфейків тепер можуть створювати контент, який не відрізняється від реальних людей, сцен та подій. Такий контент може реально загрожувати національній безпеці. Розширення можливостей генеративних методів штучного інтелекту для синтезу різних сигналів, включаючи високоякісні аудіовізуальні зображення, має значення для кібербезпеки. При персоналізації використання ШІ для створення дипфейків може підвищити ефективність операцій соціальної інженерії (програма видає себе за деяку реальну особу) та переконати, наприклад, кінцевих користувачів надати зловмисникам доступ до систем та інформації [13].

У більш широкому масштабі генеруюча міць методів штучного інтелекту та синтетичних середовищ має важливі наслідки для оборони та національної безпеки. Ці методи можуть використовуватись противниками для створення правдоподібних заяв світових лідерів та командувачів, для фабрикації переконливих операцій під хибним прапором та створення фальшивих новин [14]. Дослідження університету Georgia Tech показує, що поширення синтетичних медіа мало ще один тривожний ефект: зловмисні суб'єкти назвали реальні події «фальшивими», скориставшись новими формами заперечення, які приходять із втратою довіри в епоху дипфейків. Відео- та фото-докази, наприклад, зображення звірств, називають фейком. Поширення синтетичних ЗМІ, відоме як «дивіденд брехуна», спонукає людей називати справжні ЗМІ «фальшивими» і створює правдоподібне заперечення їхніх дій. У презентації Мікрософт [2] зазначається, що можна очікувати, що синтетичні медіа та області їх застосування будуть згодом ставати все більш витонченими, включаючи переконливе чергування дипфейків з подіями у світі, що реально відбуваються, і синтез дипфейків у реальному часі. Генерації в реальному часі можна використовувати для створення переконливих інтерактивних самозванців (наприклад, що з'являються на телеконференціях і керуються людиною-контролером), які, здається, мають природну позу голови, вираз обличчя та висловлювання. Зазначимо, що нам, можливо, доведеться зіткнутися з проблемою штучно створених людей, які можуть автономно брати участь у переконливих

розмовах в реальному часі з аудіо та візуальних каналів. Природно, що за таких умов визначення дипфейків стає дуже актуальним завданням.

Приклад – програма DARPA Semantic Forensics (SemaFor) [15]. Програма SemaFor спрямована на розробку інноваційних семантичних технологій для аналізу медіа. Ці технології включають алгоритми семантичного виявлення, які визначають, були створені мультимодальні медіаактиви або ними маніпулювали. Алгоритми атрибуції зроблять висновок, чи мультимодальне медіа виходить від конкретної організації або окремої особи. Алгоритми характеризації будуть міркувати про те, чи були створені мультимодальні медіа чи ними маніпулювали в зловмисних цілях. Ці технології SemaFor допоможуть виявляти, стримувати та розуміти кампанії супротивника з дезінформації.

Інша програма – DARPA MediaForensics (MediaFor) [16]. Презентація визначає Media Forensic як наукове дослідження в галузі збору, аналізу, інтерпретації, та подання аудіо-, відео– та графічних доказів, отриманих у ході розслідування та судового розгляду. Поставлена мета – розробити технології автоматизованої оцінки цілісності зображення чи відео (рис. 3).

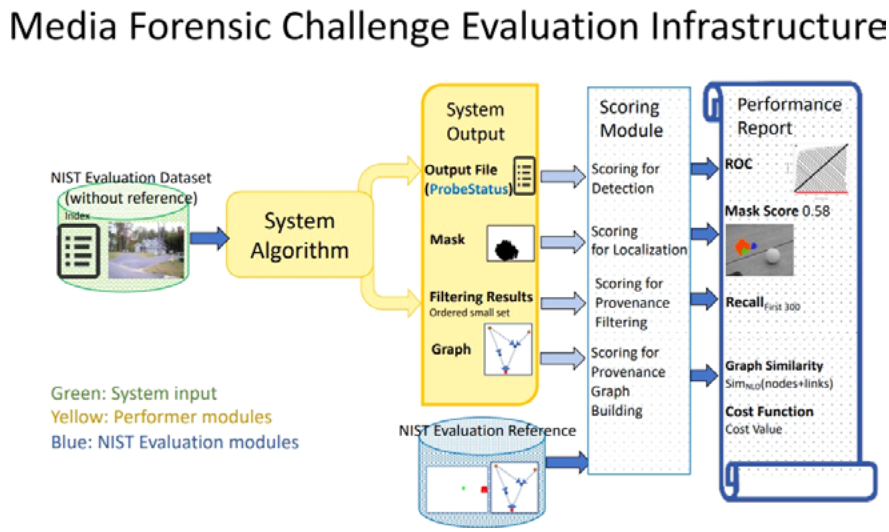

Рис. 3. Оцінка цілісності контенту

Мікрософт у презентації [2] вважає перспективним підхід до протидії загрозі синтетичних носіїв на основі технології походження цифрового контенту. Походження цифрового контенту використовує криптографію та технології баз даних для підтвердження джерела та історії змін (походження) будь-яких цифрових носіїв. Це пов'язано з тим, що у довгостроковій перспективі ні люди, ні методи ШІ не зможуть надійно відрізнити факти від вигадок, створених ШІ, і, відповідно, ми повинні терміново підготуватися до очікуваної траєкторії все більш реалістичних та переконливих дипфейків. У частині створення технологій сертифікації аудіо-візуального контенту з'явилися міжгалузеві партнерства Project Origin, Content Authenticity Initiative (CAI) та Coalition to Content Provenance and Authenticity (C2PA).

У січні 2022 року C2PA випустила специфікацію стандарту, що забезпечує сумісність систем походження цифрового контенту [17]. Це дозволяє випускати

комерційні інструменти виробництва контенту відповідно до стандарту C2PA, які дозволятимуть авторам та мовникам повідомляти глядачів про вихідне джерело та історію редагування фото- та аудіовізуальних матеріалів. У заключному звіті NSCAI рекомендується використовувати технології походження цифрового контенту, щоб пом'якшити проблему синтетичних медіа. У Конгресі США двопартійний Закон про цільову групу з дипфейків пропонує створити Національну цільову групу з дипфейків та цифрового походження. Технології блокчейн пропонується використовувати для підтвердження авторства медіа даних [18].

Висновки. У праці розглянуто галузі перетину кібербезпеки та штучного інтелекту (машинного навчання). Ці області включають атаки з використання штучного інтелекту, захист від атак з використанням штучного інтелекту, захист самих систем машинного навчання і виробництво контенту за допомогою систем машинного навчання. Необхідно відзначити, що здатності систем машинного навчання поки дозволяють домагатися кращих результатів ніж використання дискримінантних моделей, де змагальні атаки залишаються невирішеною проблемою. Системи штучного інтелекту демонструють вражаючі здібності по створенню контенту, що на рівні кібербезпеки відбивається у здатності створювати невиразні дипфейки, тому єдиним реальним способом боротьби тут є сертифікація (підтвердження походження) контенту. У сфері кібербезпеки самих систем штучного інтелекту атаки поки що також домінують над захистом. Деяким (слабким і тимчасовим) «захистом» тут поки що є те, що кількість реально здійснених атак менша за кількість потенційно можливих. У цій галузі також зосереджено найбільшу кількість досліджень. У сфері організації та управління атаками роль штучного інтелекту полягає в розумній автоматизації процесу. У частині використання машинного навчання для детектування атак є набагато більше успіхів, порівняно з іншими областями. Тут працюють механізми нейронних мереж з пошуку та виявлення шаблонів у даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Kouliaridis, Vasileios, and Georgios Kambourakis. A comprehensive survey on machine learning techniques for android malware detection. Information 12.5.2021.

2. Yuan, Zhenlong, et al. Droid-sec: deep learning in android malware detection. Proceedings of the 2014 ACM conference on SIGCOMM. 2014.

3. Vinayakumar, R., et al. Robust intelligent malware detection using deep learning. IEEE Access 7. 2019.

4. Tajaddodianfar, Farid, Jack W. Stokes, and Arun Gururajan. Texception: a character/word-level deep learning model for phishing URL detection. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2020.

5. Basnet, Ram, Srinivas Mukkamala, and Andrew H. Sung. Detection of phishing attacks: A machine learning approach. Soft computing applications in industry. Springer, Berlin, Heidelberg. 2008.

6. Divakaran, Dinil Mon, and Adam Oest. Phishing Detection Leveraging Machine Learning and Deep Learning: A Review. arXiv preprint arXiv:2205.07411. 2022.

7. Твердохліб А.О., Коротін Д.С. Ефективність функціонування комп'ютерних систем при використанні технології блокчейн і баз данних. *Таврійський науковий вісник. Серія: Технічні науки*. 2022.

8. Цвик О.С. Аналіз і особливості програмного забезпечення для контролю трафіку. *Вісник Хмельницького національного університету. Серія: Технічні науки*, (1). 2023,

9. Новіченко Є.О. Актуальні засади створення алгоритмів обробки інформації для логістичних центрів. *Таврійський науковий вісник. Серія: Технічні науки*, (1). 2023.

10. Зайцев Є.О. Smart засоби визначення аварійних станів у розподільних електричних мережах міст. *Таврійський науковий вісник. Серія: Технічні науки*, (5). 2022.

11. Shenfield, Alex, David Day, and Aladdin Ayesh. Intelligent intrusion detection systems using artificial neural networks. Ict Express 4.2. 2018.

12. Mishra, Preeti, et al. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials 21.1. 2018.

13. Alsaheel, Abdulellah, et al. {ATLAS}: A sequence-based learning approach for attack investigation. 30th USENIX Security Symposium (USENIX Security 21). 2021.

14. Ongun, Talha, et al. Living-Off-The-Land Command Detection Using Active Learning. 24th International Symposium on Research in Attacks, Intrusions and Defenses. 2021.

15. Kok, S., et al. Ransomware, threat and detection techniques: A review. Int. J. Comput. Sci. Netw. Secur 19.2. 2019.

16. Wu, Yirui, Dabao Wei, and Jun Feng. Network attacks detection methods based on deep learning techniques: a survey. Security and Communication Networks. 2020.

17. Xin, Yang, et al. Machine learning and deep learning methods for cybersecurity. IEEE Access 6.2018.

18. Noor, Umara, et al. A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories. Future Generation Computer Systems 95. 2019.

REFERENCES:

1. Kouliaridis, Vasileios, and Georgios Kambourakis. A comprehensive survey on machine learning techniques for android malware detection. Information 12.5. 2021. [in English].

2. Yuan, Zhenlong, et al. (2014) Droid-sec: deep learning in android malware detection. Proceedings of the 2014 ACM conference on SIGCOMM. [in English].

3. Vinayakumar, R., et al. (2019) Robust intelligent malware detection using deep learning. IEEE Access 7. [in English].

4. Tajaddodianfar, Farid, Jack W. Stokes, and Arun Gururajan (2020) Texception: a character/word-level deep learning model for phishing URL detection. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

5. Basnet, Ram, Srinivas Mukkamala, and Andrew H. Sung. (2008) Detection of phishing attacks: A machine learning approach. Soft computing applications in industry. Springer, Berlin, Heidelberg. [in English].

6. Divakaran, Dinil Mon, and Adam Oest. Phishing Detection Leveraging Machine Learning and Deep Learning: A Review. arXiv preprint arXiv:2205.07411. 2022. [in English].

7. Tverdokhlib A.O., Korotin D.S. (2022) Efektyvnist funktsionuvannia kompiuternykh system pry vykorystanni tekhnolohii blokchein i baz dannykh. Tavriiskyi naukovyi visnyk. Seriia: Tekhnichni nauky, (6) [in Ukrainian].

8. Tsvyk O.S. (2023) Analiz i osoblyvosti prohramnoho zabezpechennia dlia kontroliu trafiku. Visnyk Khmelnytskoho natsionalnoho universytetu. Ceriia: Tekhnichni nauky, (1) [in Ukrainian].

9. Novichenko Ye.O. (2023) Aktualni zasady stvorennia alhorytmiv obrobky informatsii dlia lohistychnykh tsentriv. Tavriiskyi naukovyi visnyk. Seriia: Tekhnichni nauky, (1) [in Ukrainian].

10. Zaitsev Ye.O. (2022) Smart zasoby vyznachennia avariinykh staniv u rozpodilnykh elektrychnykh merezhakh mist. Tavriiskyi naukovyi visnyk. Seriia: Tekhnichni nauky, (5) [in Ukrainian].

11. Shenfield, Alex, David Day, and Aladdin Ayesh. (2018) Intelligent intrusion detection systems using artificial neural networks. Ict Express 4.2. [in English].

12. Mishra, Preeti, et al. (2018) A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials 21.1. [in English].

13. Alsaheel, Abdulellah, et al. (2021) {ATLAS}: A sequence-based learning approach for attack investigation. 30th USENIX Security Symposium (USENIX Security 21). [in English].

14. Ongun, Talha, et al. (2021) Living-Off-The-Land Command Detection Using Active Learning. 24th International Symposium on Research in Attacks, Intrusions and Defenses. [in English].

15. Kok, S., et al. (2019) Ransomware, threat and detection techniques: A review. Int. J. Comput. Sci. Netw. Secur 19.2. [in English].

16. Wu, Yirui, Dabao Wei, and Jun Feng. (2020) Network attacks detection methods based on deep learning techniques: a survey. Security and Communication Networks. [in English].

17. Xin, Yang, et al. (2018) Machine learning and deep learning methods for cybersecurity. IEEE Access 6. [in English].

18. Noor, Umara, et al. (2019) A machine learning framework for investigating data breaches based on semantic analysis of adversary's attack patterns in threat intelligence repositories. Future Generation Computer Systems 95. [in English].