УДК 004.89:004.056.53:681.5

ВИКОРИСТАННЯ АГЕНТНОГО ШІ В ЗАДАЧАХ ЗАХИСТУ ОБ'ЄКТІВ КРИТИЧНОЇ ІНФРАСТРУКТУРИ

М. Б. Осінній[1,a], І. В. Стьопочкіна[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

В роботі запропоновано метод виявлення кіберзагроз в критичній інфраструктурі , який спирається на виявлення аномалій на основі агентів штучного інтелекту. Запропоновано загальну архітектуру системи та прототип програмного рішення, яке збирає дані з попередніх інцидентів безпеки, для встановлення нормальних та аномальних патернів поведінки системи. Навчені агенти працюють як класифікатори на основі обгрунтовано обраного алгоритму. Система здатна інтелектуально обробляти одержані результати класифікації та надавати можливість взаємодій із користувачем.

Ключові слова: Агенти ШІ, ОКІ

Вступ

Сучасний цифровий світ стикається з безпрецедентним **зростанням кількості та складності кіберзагроз та кіберінцидентів**. Ця тенденція зумовлена кількома факторами [1]:

1. **Посилення цифровізації:** Все більше аспектів нашого життя та бізнесу переходить в онлайн, створюючи ширшу поверхню для атак. Критична інфраструктура, фінансові системи, державні установи та приватні компанії стають мішенями.
2. **Розвиток технологій:** Зловмисники постійно вдосконалюють свої інструменти та тактики, використовуючи новітні технології, такі як штучний інтелект, для автоматизації атак та обходу традиційних засобів захисту. Пропонуються рішення які вже здатні проводити роботу з тестування на проникнення [2, 3] та навіть експлуатувати вразливості нульового дня [4].
3. **Геополітична напруженість:** Кіберпростір все частіше стає ареною для міждержавних конфліктів, шпигунства та диверсій. Спонсоровані державами хакерські групи здійснюють складні та цілеспрямовані атаки (APT).
4. **Комерціалізація кіберзлочинності:** Існує ціла "індустрія"кіберзлочинності, де інструменти для атак, викрадені дані та послуги зловмисників продаються та купуються на чорному ринку (наприклад, програми-вимагачі як послуга - Ransomware-as-a-Service).

Наслідки кіберінцидентів можуть бути руйнівними: фінансові збитки, втрата конфіденційних даних, пошкодження репутації, порушення роботи критично важливих сервісів, загроза національній безпеці.

Виклик: Величезний обсяг та швидкість сучасних кібератак перевантажують традиційні підходи до безпеки та можливості аналітиків-людей. Ручний аналіз кожного підозрілого сигналу стає неможливим і неефективним. **Нагальна потреба в швидкому інтелектуальному аналізі:** Саме тут на перший план виходить **необхідність у швидкому та інтелектуальному аналізі кіберінцидентів**. Це означає використання передових технологій та методологій для:

1. **Миттєвого виявлення:** Автоматизовані системи на базі машинного навчання (ML) та штучного інтелекту (AI) здатні аналізувати величезні потоки даних в реальному часі, виявляючи аномалії та підозрілу активність значно швидше за людину.
2. **Глибокого аналізу:** Інтелектуальні інструменти допомагають швидко зрозуміти природу атаки, її джерело, використані вектори, потенційний масштаб ураження та цілі зловмисників. Вони можуть корелювати події з різних джерел (логи, мережевий трафік, дані про загрози) для побудови повної картини інциденту.
3. **Пріоритезації реагування:** Автоматизований аналіз дозволяє швидко оцінити критичність інциденту та визначити пріоритетність дій, спрямовуючи ресурси на найсерйозніші загрози.
4. **Ефективного реагування:** Системи оркестрації, автоматизації та реагування на інциденти безпеки (SOAR) можуть автоматизувати рутинні завдання реагування (наприклад, блокування шкідливих IP-адрес, ізоляцію уражених систем), звільняючи час аналітиків для складніших завдань.
5. **Прогнозування та профілактики:** Аналіз минулих інцидентів та глобальних даних про

[a]osinnii.maksym@lll.kpi.ua

загрози (Threat Intelligence) за допомогою інтелектуальних систем дозволяє виявляти нові тенденції, передбачати майбутні атаки та проактивно зміцнювати захист.

1. Мета дослідження

Метою даного дослідження є розробка та реалізація інтелектуального агента (ШІ-агента), призначеного для автоматизації процесів обробки та аналізу інцидентів кібербезпеки. Основні завдання в рамках досягнення цієї мети включають:

- Побудову моделі ШІ-агента, здатної аналізувати вхідні дані про інциденти кібербезпеки для виявлення ключових характеристик та закономірностей.
- Створення системи для ефективного та структурованого зберігання інформації про проаналізовані інциденти у базі даних.
- Розробку комунікаційного модуля, що дозволить ШІ-агенту взаємодіяти з дослідником інцидентів (аналітиком безпеки) у форматі діалогу або через запити.
- Забезпечення функціональності агента для надання релевантної та своєчасної інформації щодо конкретних інцидентів (тікетів) за запитом дослідника, підтримуючи процес прийняття рішень.

2. Матеріали та методи

В ході дослідження для розробки та оцінки функціональності ШІ-агента застосовувалися методи **експерименту** та **комп'ютерного моделювання**. Було спроектовано та програмно реалізовано модульну систему, що включає наступні ключові компоненти:

- **Система виявлення аномалій:** Для ідентифікації відхилень від нормальної поведінки в аналізованих даних (наприклад, мережевому трафіку) була використана модель на базі *ансамблю автокодувальників* (Autoencoder Ensemble). Цей підхід дозволяє виявляти складні нелінійні залежності та нетипові патерни.
- **Система генерації правил:** Для автоматичної генерації класифікаційних правил формату "ЯКЩО-ТОДІ"використовується підсистема на основі алгоритму *RIPPER* (Repeated Incremental Pruning to Produce Error Reduction) [5]. Алгоритм навчається на даних, де числова оцінка аномальності бінаризована (напр., 'висока'/'низька' аномальність), і формує набір легко інтерпретованих правил для класифікації нових подій, придатних для оперативного застосування у системах фільтрації або реагування.
- **Підсистема аналізу та суммаризації:** Реалізовано модуль для агрегації, узагальнення та кількісної оцінки активності елементів інфраструктури (напр., хостів) та характеру їх взаємодії впродовж певного часового вікна.
- **Підсистема графічного відображення:** Для полегшення аналізу та інтерпретації результатів дослідником інцидентів було створено модуль візуалізації, який представляє інформацію про взаємодії елементів, виявлені аномалії та їх зв'язки у вигляді *графів* або інших інтуїтивно зрозумілих діаграм.
- *Інтелектуальний агент (ШІ-агент):* Центральний компонент системи, що інтегрує дані з інших підсистем та зовнішніх джерел. Агент має доступ до актуальних баз даних *Threat Intelligence* для збагачення інформації про інциденти. Він здатен автоматично додавати релевантні відомості **тікетів** – звітів про кіберінцидентb. Крім того, агент реалізує функціонал для взаємодії з користувачем (дослідником інцидентів) у форматі діалогу, відповідаючи на запитання щодо конкретних тікетів та надаючи зібрану аналітичну інформацію.

Експериментальна перевірка розроблених модулів та всієї системи в цілому проводилася на спеціально підготовлених наборах даних, що імітують різні сценарії кіберінцидентів.

2.1. Система виявлення аномалій

Ключовим компонентом первинного аналізу є підсистема виявлення аномалій, реалізована на основі системи **KitNet** [6]. Її завданням є ідентифікація нетипової або підозрілої активності шляхом аналізу *кореляційно-зв'язаних між собою ознак (фіч)*, екстрагованих з пакетів мережевого трафіку в режимі реального часу або пакетної обробки. В основі системи KitNet лежить модель, побудована на базі *ансамблю автокодувальників* (Autoencoder Ensemble). Автокодувальники, що є типом нейронних мереж, навчаються відтворювати вектор вхідних ознак на виході. У контексті виявлення аномалій, модель Kitnet навчається на даних, що характеризують нормальну поведінку системи – тобто типову поведінку мережевого трафіку. При обробці нових векторів ознак, висока помилка реконструкції (значна відмінність між вхідними та реконструйованими на виході даними) свідчить про те, що поточний стан мережевого трафіку не відповідає вивченій моделі нормальної поведінки і, отже, є аномальним.

2.2. Автоматична генерація класифікаційних правил за допомогою алгоритму RIPPER

Для автоматизації процесу виявлення та класифікації мережевих аномалій було реалізовано підхід, що базується на алгоритмі генерації правил RIPPER (Repeated Incremental Pruning to Produce Error Reduction) з використанням бібліотеки `wittgenstein`. Метою є створення набору легко інтерпретованих правил формату "ЯКЩО-ТОДІ які класифікують мережеві події як такі, що мають високий або низький рівень аномальності. Створені правила додаються до тікету з метою надати оператору ефективні засоби фільтрації мережевого трафіку для зниження аномальності в системі.

Таблиця 1. Table of Conditions (Wider Rule ID Column)

Rule id	Conditions	
	Type	**Value**
f4f36e1e-5782-439c-8445-8e4d1d1bdf0c	dstMac	ff:ff:ff:ff:ff:ff
	srcMac	3c:33:00:98:ee:fd
32fb7729-c7c3-4d72-9818-17ab17505e22	srcproto	123
	dstIP	211.115.194.21

2.3. Загальна архітектура системи

Розроблена система являє собою інтегрований програмний комплекс для обробки інцидентів кібербезпеки, архітектура якого відображає послідовний конвеєр аналізу даних. Система включає наступні функціональні блоки та потоки інформації:

1. **Вхідні дані:** Початковою точкою є отримання мережевих пакетів (`Packets`) з аналізованої мережі (`Network`).
2. **Модуль виявлення аномалій (`Anomaly Detection Engine`):** Цей блок відповідає за первинний аналіз мережевого трафіку.
   - Центральний компонент `Anomaly` ідентифікує аномальні патерни або відхилення від базової лінії нормальної поведінки.
   - На основі виявлених аномалій ініціюються процеси генерації правил (`Rules`), створення зведеного опису (`Summary`) та візуалізації даних (`Plot`).
   - Результати цих процесів формують первинний опис виявленої аномалії.
3. **Створення інциденту (`Ticket`):** Зібрана інформація (правила, опис, візуалізація) агрегується та формалізується у вигляді запису про інцидент – `Тікету`.
4. **Інтелектуальна обробка (`Reporting AI`):** Цей модуль, що функціонує як ШІ-агент, виконує збагачення та попередній аналіз інциденту.
   - Агент отримує на вхід створені `Тікети` (`Read Tickets`).
   - Здійснюється інтеграція з зовнішньою платформою кіберрозвідки (`Threat Intelligence Platform`) для отримання додаткового контексту (`Provide Intel`).
   - Агент аналізує зібрану інформацію та генерує `Перевірений Тікет` (`Reviewed Ticket`), що містить консолідовані та збагачені дані.
5. **Платформа аналізу та зберігання (`Review Platform`):** Центральне сховище та інтерфейс для аналітиків.
   - Приймає та зберігає `Reviewed Tickets`.
   - Окрім традиційного зберігання тікетів, використовує `Векторну Базу Даних` (`Vector Database`) для зберігання векторних представлень інцидентів, що уможливлює ефективний семантичний пошук, виявлення схожих інцидентів та інші аналітичні опе-

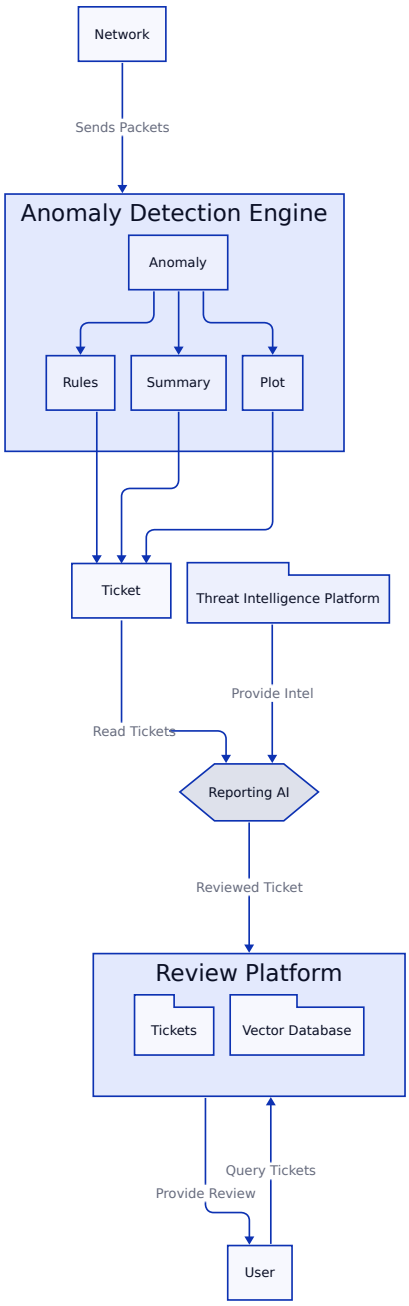

Рис. 1. Схема роботи моделі

рації.

6. **Інтерфейс користувача (`User`):** Забезпечує взаємодію аналітика безпеки з системою.
   - Користувач може надсилати запити до `Review Platform` (`Query Tickets`) для пошуку, фільтрації та отримання даних про інциденти.
   - Система надає відповідну інформацію користувачеві (`Provide Review`), використовуючи дані з обох сховищ (тікети та векторна БД).

Таким чином, архітектура системи забезпечує наскрізний процес від моніторингу мережі до надання аналітику комплексної, збагаченої інформації про інциденти з використанням як класичних, так і векторних методів обробки та зберігання даних.

Висновок

У даній роботі вирішено актуальну науково-практичну задачу розробки інтелектуальної системи автоматизації аналізу кіберінцидентів, спрямованої на захист Об'єктів Критичної Інфраструктури в умовах зростаючої складності загроз. Метою було створення комплексного рішення для ефективного виявлення, аналізу та збагачення інформації про інциденти. Запропонована модульна архітектура системи успішно інтегрує низку сучасних підходів. Реалізовано виявлення аномалій на основі ансамблю автокодувальників (Kitnet), автоматичну генерацію інтерпретованих класифікаційних правил за допомогою алгоритму RIPPER, а також збагачення даних про інциденти за рахунок інтеграції з платформами Threat Intelligence. Центральним елементом виступає ШІ-агент, який координує ці процеси, агрегує інформацію у вигляді тікетів та забезпечує діалогову взаємодію з аналітиком. Впровадження векторної бази даних дозволяє здійснювати поглиблений семантичний аналіз та виявляти неочевидні зв'язки між інцидентами. Розроблений підхід має значний потенціал для підвищення ефективності центрів моніторингу та реагування на кіберінциденти (SOC), дозволяючи автоматизувати рутинний аналіз, прискорити час реакції та надати аналітикам більш повну та контекстуалізовану інформацію. Подальший розвиток включатиме тестування системи на ширшому спектрі даних, вдосконалення алгоритмів ШІ-агента для кращого розуміння природної мови та генерації звітів, а також дослідження методів пояснюваного ШІ. Представлена робота демонструє життєздатність та перспективність інтеграції передових методів машинного навчання та зберігання даних для створення ефективних інструментів у сфері кібербезпеки ОКІ.

2.4. Унікальність та наукова новизна підходу

Унікальність та наукова новизна запропонованого підходу до аналізу кіберінцидентів для Об'єктів Критичної Інфраструктури (ОКІ) полягає у створенні синергетичної системи, центральним елементом якої є інтелектуальний агент. На відміну від багатьох існуючих рішень, що фокусуються на окремих аспектах безпеки, наш підхід інтегрує різні етапи аналізу під управлінням ШІ-агента. Ключовими елементами новизни є інтеграція методів глибокого виявлення аномалій (на базі ансамблю автокодувальників KitNet) з автоматичною генерацією інтерпретованих класифікаційних правил за алгоритмом RIPPER. Це забезпечує не лише виявлення складних патернів, але й надання аналітику зрозумілих пояснень та готових для використання правил. Додаткову новизну вносить використання векторної бази даних для поглибленого семантичного аналізу інцидентів та виявлення прихованих зв'язків, а також реалізація інтерактивної діалогової взаємодії ШІ-агента з користувачем. Таким чином, запропонована архітектура та реалізовані компоненти представляють інтегроване, більш інтелектуальне та ефективне рішення для управління кіберінцидентами порівняно зі стандартними системами, значно підвищуючи можливості аналітиків SOC.

Перелік використаних джерел

1. *Google.* Adversarial Misuse of Generative AI. — 01.2025. — URL: https://services.google.com/fh/files/misc/adversarial-misuse-generative-ai.pdf ; Accessed on May 5, 2025. Report.
2. PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing / G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, S. Rass // 33rd USENIX Security Symposium (USENIX Security 24). — USENIX Association, 2024. — С. 847—864.
3. VulnBot: Autonomous Penetration Testing for A Multi-Agent Collaborative Framework / H. Kong, D. Hu, J. Ge, L. Li, T. Li, B. Wu. — 2025. — arXiv: 2501.13411 [cs.SE]. — URL: https://arxiv.org/abs/2501.13411 ; Accessed on May 5, 2025. arXiv preprint arXiv:2501.13411.
4. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities / Y. Zhu, A. Kellermann, A. Gupta, P. Li, R. Fang, R. Bindu, D. Kang. — 2024. — arXiv: 2406.01637 [cs.MA]. — URL: https://arxiv.org/abs/2406.01637 ; Accessed on May 5, 2025.
5. *Cohen W. W.* Fast Effective Rule Induction // International Conference on Machine Learning. — 1995. — С. 115—123.
6. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection / Y. Mirsky, T. Doitshman, Y. Elovici, A. Shabtai. — 2018. — arXiv: 1802.09089 [cs.CR]. — URL: https://arxiv.org/abs/1802.09089.