Міністерство освіти і науки України
Криворізький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних систем та мереж

Пояснювальна записка
до кваліфікаційної роботи бакалавра
за спеціальністю 123 «Комп'ютерна інженерія»

на тему: Підходи інтеграції штучного інтелекту в системи кібербезпеки

Проектував _____ Я. В. Сидорук

Керівник роботи _____ А. О. Сенько

Консультант _____ А. О. Сенько

Нормоконтроль _____ Д. І. Кузнєцов

Завідувач кафедри _____ А. І. Купін

Кривий Ріг
2026

Криворізький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних систем та мереж

Ступінь вищої освіти бакалавр
Спеціальність 123 «Комп'ютерна інженерія»

ЗАТВЕРДЖУЮ
Завідувач кафедри, голова циклової комісії

_____ А. І. Купін

"____" _____ 20__ року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

_____
(прізвище, ім'я, по батькові)

1. Тема роботи _____
_____
_____

керівник роботи _____,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом вищого навчального закладу від "___"_____ 20__року №__

2. Строк подання студентом роботи _____

3. Вихідні дані до роботи _____
_____
_____
_____
_____
_____

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)_____
_____
_____
_____
_____
_____

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)_____
_____
_____
_____
_____

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання_____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк виконання етапів роботи	Примітка

Студент _____ _____
(підпис) (прізвище та ініціали)

Керівник роботи _____ _____
(підпис) (прізвище та ініціали)

Зміст

Вступ

Актуальність теми дослідження. Сучасний цифровий простір перебуває в умовах безперервної трансформації, яка супроводжується стрімким зростанням кількості кіберзагроз, їх ускладненням та диверсифікацією. Традиційні підходи до забезпечення інформаційної безпеки, засновані на сигнатурному аналізі та статичних правилах виявлення загроз, дедалі частіше виявляються неспроможними протистояти сучасним атакам, зокрема атакам нульового дня, поліморфному шкідливому програмному забезпеченню та цільовим атакам підвищеної складності (APT). У цьому контексті інтеграція технологій штучного інтелекту в системи кібербезпеки набуває не лише теоретичного, а й гостро практичного значення.

За даними аналітичної компанії Cybersecurity Ventures, глобальні збитки від кіберзлочинності у 2024 році перевищили 9,5 трильйона доларів США, і прогнозується подальше зростання цього показника. Водночас кількість непокритих вакансій у сфері кібербезпеки у світі вимірюється мільйонами, що робить автоматизацію та інтелектуалізацію процесів захисту об'єктивною необхідністю. Штучний інтелект дозволяє не лише скоротити час виявлення загроз із годин до секунд, але й забезпечити адаптивне реагування на принципово нові вектори атак без необхідності постійного оновлення правил вручну.

Аналіз сучасного стану проблеми. Питання застосування методів машинного навчання та штучного інтелекту в задачах інформаційної безпеки активно досліджуються в науковій спільноті. Значний внесок у розвиток цього напряму зробили такі дослідники, як А. Гудфеллоу, Й. Бенджіо та А. Курвілл, чиї праці з глибокого навчання заклали теоретичне підґрунтя для сучасних систем виявлення аномалій. Роботи Дж. Андерсона присвячені раннім концепціям поведінкового аналізу в мережах, тоді як дослідження Дж. Сомайї та В. Ліса стосуються класифікації мережевих вторгнень методами машинного навчання. Серед сучасних авторів варто відзначити праці Р. Сіддіквї щодо застосування глибоких нейронних мереж для виявлення шкідливого програмного забезпечення, а також дослідження групи авторів під керівництвом Т. Ліма, присвячені ансамблевим методам класифікації кіберзагроз. Теоретичні аспекти UEBA-систем розроблялися М. Дагг та Р. Кентом, тоді як практичні підходи до автоматизованого реагування на інциденти (SOAR) знайшли відображення у дослідженнях фахівців компаній Gartner, IBM та Palo Alto Networks. Незважаючи на значний обсяг наявних публікацій, питання комплексної інтеграції різних підходів ШІ в єдину архітектуру кібербезпеки залишаються недостатньо дослідженими, особливо з урахуванням специфіки сучасних гібридних інфраструктур.

Проблема дослідження полягає у відсутності систематизованого підходу до інтеграції методів штучного інтелекту в системи кібербезпеки, що враховував би як технічні аспекти реалізації, так і оперативні вимоги реальних середовищ експлуатації.

Мета дослідження — проаналізувати існуючі підходи інтеграції штучного інтелекту в системи кібербезпеки, розробити та практично реалізувати модель виявлення кіберзагроз на основі машинного навчання з оцінкою її ефективності.

Об'єкт дослідження — процес виявлення та нейтралізації кіберзагроз в інформаційних системах.

Предмет дослідження — методи, підходи та інструменти інтеграції технологій штучного інтелекту в системи кібербезпеки.

Завдання дослідження:

1. Здійснити огляд технологій штучного інтелекту, що застосовуються у сфері кібербезпеки, та проаналізувати сучасний ландшафт кіберзагроз.
2. Дослідити методи машинного навчання для виявлення аномалій, вторгнень та аналізу поведінки користувачів.
3. Розглянути підходи до автоматизованого реагування на інциденти безпеки.
4. Розробити та налаштувати модель виявлення загроз на основі машинного навчання.
5. Провести тестування розробленої системи та сформулювати рекомендації щодо її впровадження.

Практична значущість роботи визначається тим, що розроблена модель виявлення загроз може бути адаптована та впроваджена в реальних корпоративних середовищах для підвищення рівня автоматизації процесів захисту інформації. Сформульовані рекомендації щодо впровадження матимуть цінність для спеціалістів із кібербезпеки, системних адміністраторів та осіб, відповідальних за прийняття рішень у сфері інформаційної безпеки організацій.

Методи дослідження. У роботі використано комплекс загальнонаукових та спеціальних методів: системний аналіз — для дослідження архітектурних підходів до інтеграції ШІ; порівняльний аналіз — для зіставлення існуючих рішень і підходів; методи машинного навчання (класифікація, виявлення аномалій, кластеризація) — для побудови моделі виявлення загроз; експериментальний метод — для тестування розробленої системи; методи статистичного аналізу — для оцінювання ефективності моделі.

Структура роботи. Дипломна робота складається зі вступу, трьох розділів, висновків та списку використаних джерел.

1. Аналіз сучасного стану інтеграції штучного інтелекту в системи кібербезпеки

1.1 Огляд технологій штучного інтелекту, що застосовуються в кібербезпеці

Стрімкий розвиток цифрових технологій у XXI столітті породив принципово нові виклики в галузі захисту інформації, що закономірно зумовив інтенсивне впровадження інструментів штучного інтелекту (ШІ) у системи кібербезпеки. Сьогодні ШІ розглядається не як допоміжний засіб, а як центральна складова сучасної архітектури інформаційного захисту, здатна забезпечити адаптивну, проактивну та інтелектуальну протидію загрозам, що постійно еволюціонують [6]. Розуміння технологічного базису ШІ, що застосовується у кібербезпеці, є необхідною передумовою для аналізу можливостей та обмежень наявних рішень.

Серед ключових технологій ШІ, що знайшли практичне застосування у кібербезпеці, провідне місце посідає машинне навчання (Machine Learning, ML). Ця технологія дозволяє системам автоматично навчатися на основі даних без явного програмування, виявляти статистичні закономірності в масивах інформації та формувати прогностичні моделі поведінки. В контексті кібербезпеки методи ML використовуються передусім для класифікації мережевого трафіку, ідентифікації шкідливого програмного забезпечення, виявлення аномальних патернів доступу та оцінки ризиків [12]. Дослідження підтверджують, що моделі на основі алгоритму Random Forest здатні класифікувати мережеві атаки з точністю понад 97%, що суттєво перевищує можливості традиційних сигнатурних систем [33].

Глибоке навчання (Deep Learning, DL) як підгалузь машинного навчання забезпечує ще вищий рівень абстракції при обробці даних завдяки багатошаровим нейронним мережам. Згорткові нейронні мережі (CNN) ефективно застосовуються для аналізу бінарних образів шкідливих програм, виявляючи структурні подібності між невідомими зразками та відомими сімействами малварі навіть за відсутності ідентичного сигнатурного збігу. Рекурентні нейронні мережі (RNN) та їх модифікація LSTM (Long Short-Term Memory) демонструють високу ефективність у виявленні аномалій у часових рядах даних, зокрема при аналізі послідовностей мережевих подій та журналів системних викликів [34]. Автоенкодери, що є різновидом нейронних мереж без учителя, широко використовуються для моделювання нормальної поведінки систем та виявлення відхилень від неї, що особливо цінно при виявленні раніше невідомих типів атак.

Обробка природної мови (Natural Language Processing, NLP) відкриває нові можливості для кібербезпеки у сфері аналізу загроз та інтелекту. Системи на основі NLP здатні автоматично обробляти звіти про кіберзагрози, технічні бюлетені безпеки, повідомлення у даркнеті та соціальних мережах для раннього виявлення ознак підготовки атак. Великі мовні моделі (LLM), зокрема сімейства GPT та їх аналоги, починають активно застосовуватись для автоматизованого аналізу коду на наявність вразливостей, генерації сценаріїв реагування на інциденти та підтримки аналітиків безпеки [23]. Концепція розвитку штучного інтелекту в Україні, затверджена розпорядженням Кабінету Міністрів у 2020 році, прямо вказує на пріоритетність розвитку NLP-технологій як частини національної ШІ-стратегії [13].

Комп'ютерний зір як технологія ШІ знаходить застосування у специфічних задачах кібербезпеки, пов'язаних з аналізом зображень, виявленням фішингових сайтів за їх візуальними характеристиками, ідентифікацією підроблених документів та верифікацією особистості в системах контролю доступу. Зокрема, системи виявлення фішингових сторінок на основі комп'ютерного зору досягають точності виявлення до 96%, аналізуючи структуру та оформлення сторінок замість перевірки URL за чорними списками, що дозволяє ефективно протидіяти новим фішинговим кампаніям [11].

Окремого розгляду заслуговують технології генеративного ШІ та їх двоїста роль у кібербезпеці. З одного боку, генеративні моделі дозволяють автоматизувати створення синтетичних навчальних наборів даних для тренування систем виявлення загроз, подолавши проблему дефіциту розмічених прикладів атак. З іншого боку, ті самі технології використовуються зловмисниками для автоматизованого створення переконливих фішингових повідомлень, генерації поліморфного шкідливого коду та проведення атак соціальної інженерії небаченого раніше масштабу [4]. Ця двоїста природа генеративного ШІ ставить нові виклики перед дослідниками та практиками кібербезпеки і потребує розробки специфічних контрзаходів.

Важливою технологічною складовою є системи автоматизації та оркестрування безпеки (SOAR — Security Orchestration, Automation and Response), що інтегрують методи ШІ для координації реагування на інциденти. Такі системи здатні автоматично виконувати заздалегідь визначені сценарії реагування (playbooks) при виявленні загроз, скорочуючи час між виявленням інциденту та початком протидії з годин до секунд [24]. Дослідження показують, що впровадження SOAR-платформ з елементами ШІ дозволяє скоротити витрати на реагування на інциденти в середньому на 37% завдяки автоматизації рутинних завдань.

Технологія федеративного навчання (Federated Learning) набуває дедалі більшого значення у кібербезпеці, оскільки дозволяє навчати моделі ШІ на розподілених наборах даних без їх централізації, що критично важливо з огляду на вимоги конфіденційності та захисту чутливої інформації.

Організації можуть спільно навчати моделі виявлення загроз, ділячись знаннями, але не чутливими даними, що дозволяє формувати більш потужні моделі при дотриманні нормативних вимог [18].

Усі перелічені технології не функціонують ізольовано, а формують взаємопов'язану екосистему інтелектуальних засобів захисту. Їх спільне застосування в рамках єдиної архітектури безпеки забезпечує синергетичний ефект, коли вихідні дані одних систем слугують вхідними для інших, формуючи багаторівневий інтелектуальний захист [6]. Розуміння технологічних можливостей та обмежень кожного із зазначених підходів є необхідною умовою для побудови ефективної стратегії інтеграції ШІ в корпоративні системи безпеки.

1.2 Аналіз сучасних кіберзагроз та вимог до систем захисту

Сучасний ландшафт кіберзагроз характеризується безпрецедентним рівнем складності, динамічності та цілеспрямованості. Якщо ще 10–15 років тому переважна більшість кібератак мала опортуністичний характер і була спрямована проти випадкових жертв, то нині домінує модель цільових атак підвищеної складності, що плануються та реалізуються протягом місяців із задіянням значних людських та технічних ресурсів [22]. Ця трансформація кардинально змінює вимоги до систем захисту та визначає необхідність переходу від реактивних до проактивних підходів.

Цільові атаки підвищеної стійкості (Advanced Persistent Threats, APT) є однією з найбільш серйозних загроз для державних установ, критичної інфраструктури та великих корпорацій. Характерними рисами APT-атак є тривале непомітне перебування в мережі жертви (median dwell time у 2023 році становив 16 днів за даними Mandiant), поетапне просування вглиб інфраструктури (lateral movement), ретельне планування та використання 0-day вразливостей. Традиційні засоби захисту виявляються принципово неспроможними виявляти APT-атаки, оскільки зловмисники навмисно імітують легітимну активність і уникають тригерів сигнатурного виявлення [34]. Саме для протидії цьому типу загроз особливо цінними є поведінкові моделі на основі ШІ, здатні виявляти тонкі аномалії в патернах мережевої активності.

Програми-вимагачі (ransomware) залишаються однією з найбільш фінансово руйнівних категорій кіберзагроз. За даними аналітичних компаній, у 2023 році загальна сума виплачених викупів перевищила 1,1 млрд доларів США, а середній розмір виплати для великих організацій сягнув 1,54 млн доларів. Сучасні ransomware-угруповання практикують так звану «подвійну extortion» — шифрування даних поєднується з погрозою їх публікації, що суттєво посилює тиск на жертву. Технічна еволюція ransomware включає використання легітимних системних інструментів (living-off-the-land), шифрування у пам'яті без запису на диск та атаки на системи резервного

копіювання [27]. Виявлення ransomware на ранніх стадіях до початку масового шифрування є критичним завданням, з яким системи на основі ШІ справляються значно ефективніше традиційних засобів.

Атаки на ланцюг постачання (supply chain attacks) набули особливої небезпечності після резонансних інцидентів останніх років. Принцип таких атак полягає у компрометації надійного постачальника програмного забезпечення або сервісу з метою подальшого проникнення до мережі цільової організації через довірені канали. Зловмисники використовують той факт, що організації, як правило, не перевіряють безпеку оновлень від перевірених постачальників. Масштаб потенційного ураження атак на ланцюг постачання надзвичайно великий: 1 скомпрометований постачальник може відкрити доступ до тисяч кінцевих жертв одночасно [23]. Протидія цьому типу загроз вимагає впровадження систем моніторингу поведінки програмного забезпечення на основі ШІ, здатних виявляти нетипову активність навіть від авторизованих джерел.

Соціальна інженерія та фішинг продовжують залишатися домінуючим вектором початкового проникнення, відповідальним за понад 80% успішних кібератак. Сучасні фішингові кампанії досягли якісно нового рівня завдяки використанню генеративного ШІ для створення персоналізованих повідомлень, deepfake-технологій для імітації голосу та відео керівників організацій, а також спіар-фішингу на основі аналізу відкритих джерел інформації про конкретних співробітників [4]. Традиційні фільтри спаму та фішингу, засновані на чорних списках та сигнатурному аналізі, виявляються неефективними проти персоналізованих кампаній — тут необхідні системи контентного аналізу та виявлення аномалій на основі NLP та машинного навчання.

Загрози з боку інсайдерів (insider threats) є особливо складними для виявлення, оскільки їх автори мають легітимний доступ до ресурсів організації та знайомі зі структурою її захисту. Зловмисні інсайдери можуть поступово і непомітно виводити конфіденційні дані протягом тривалого часу, уникаючи порогових значень систем DLP. Необережні інсайдери, своєю чергою, можуть ненавмисно створювати вразливості через неуважність або недотримання протоколів безпеки [25]. Для виявлення обох типів інсайдерських загроз системи UEBA (User and Entity Behavior Analytics) на основі ШІ є найбільш ефективним інструментом, здатним виявляти тонкі відхилення від встановленого для кожного користувача базового рівня нормальної поведінки.

DDoS-атаки (Distributed Denial of Service) еволюціонували від простих флуд-атак до складних багатовекторних кампаній, що поєднують об'ємні атаки (volumetric), атаки на рівні протоколу та атаки на рівні застосунку. Розміри атак продовжують зростати: у 2023 році були зафіксовані атаки потужністю понад 71 Тбіт/с. Крім того, з'явились так звані «розумні» DDoS-атаки, що адаптуються в реальному часі до засобів захисту жертви, змінюючи вектори

та інтенсивність трафіку. Ефективне відбиття таких атак можливе лише за допомогою систем автоматичного виявлення та класифікації аномальних трафікових патернів на основі ML [33].

Вразливості у хмарних середовищах стали критичною точкою ризику у зв'язку з масовою міграцією організацій до хмарних платформ. Неправильні конфігурації хмарних ресурсів залишаються провідною причиною витоків даних у хмарних середовищах — за різними оцінками, вони відповідальні за понад 65% інцидентів у хмарі. Захист хмарних інфраструктур вимагає нових підходів до моніторингу та аналізу, оскільки традиційні периметральні методи тут непридатні — необхідний постійний аналіз конфігурацій, мережевого трафіку та API-активності, що є природним середовищем застосування ШІ [18].

З урахуванням описаного ландшафту загроз формуються конкретні вимоги до сучасних систем захисту. По-перше, вони повинні забезпечувати виявлення загроз у реальному часі з мінімальним рівнем хибно позитивних спрацьовувань. По-друге, системи захисту мають бути здатні до виявлення раніше невідомих загроз — 0-day атак та нових різновидів шкідливого програмного забезпечення. По-третє, вкрай важливою є здатність до кореляції подій з різних джерел телеметрії для формування цілісної картини атаки. По-четверте, системи повинні забезпечувати автоматизоване реагування для скорочення часу між виявленням та нейтралізацією загрози. По-п'яте, необхідна масштабованість для обробки великих обсягів даних телеметрії в хмарних та гібридних середовищах [27]. Усі ці вимоги фактично визначають ШІ як необхідну технологічну основу сучасних систем кібербезпеки.

Стратегія кібербезпеки України, затверджена Указом Президента від 26 серпня 2021 р. № 447, прямо визначає необхідність розвитку національного потенціалу у сфері кіберзахисту із застосуванням передових технологій, у тому числі штучного інтелекту [27]. В умовах триваючого збройного конфлікту, що супроводжується безпрецедентним тиском на українську цифрову інфраструктуру з боку державних кібергруп противника, значущість інтелектуальних систем кіберзахисту набуває особливого стратегічного виміру.

1.3 Огляд існуючих підходів і рішень на ринку кібербезпеки

Ринок рішень кібербезпеки з елементами штучного інтелекту переживає стрімке зростання та трансформацію. За оцінками провідних аналітичних агентств, обсяг глобального ринку ШІ у кібербезпеці у 2023 році становив близько 22,4 млрд доларів США і, за прогнозами, досягне 60,6 млрд доларів до 2028 року при середньорічному темпі зростання 22,3%. Ця динаміка відображає як зростання попиту на інтелектуальні засоби захисту, так і стрімкий технологічний прогрес у сфері ШІ [23]. Розуміння ринкового

ландшафту є необхідною умовою для обґрунтованого вибору підходів до інтеграції ШІ в конкретних організаційних контекстах.

Клас рішень SIEM (Security Information and Event Management) із інтегрованими можливостями ШІ представляє найбільш зрілий сегмент ринку. Сучасні SIEM-платформи, такі як IBM QRadar, Splunk Enterprise Security та Microsoft Sentinel, інтегрують можливості машинного навчання для кореляції подій безпеки, виявлення аномалій та пріоритизації інцидентів. Ключовою інновацією порівняно з традиційними SIEM є здатність автоматично виявляти раніше невідомі патерни атак замість простого застосування заздалегідь визначених правил кореляції [24]. Microsoft Sentinel, наприклад, використовує моделі машинного навчання для злиття даних з різних джерел та виявлення прихованих зв'язків між, здавалося б, непов'язаними подіями безпеки.

Платформи захисту кінцевих точок нового покоління (Next-Generation Endpoint Detection and Response, EDR/XDR) є іншим значущим сегментом ринку. Рішення таких компаній, як CrowdStrike Falcon, SentinelOne та Cortex XDR від Palo Alto Networks, базуються на постійному збиранні телеметрії з кінцевих точок та її аналізі за допомогою ШІ в хмарі. Ключовою перевагою цих платформ є можливість ретроспективного полювання на загрози (threat hunting) — виявлення ознак минулих компрометацій у вже зібраній телеметрії після отримання нових індикаторів компрометації [11]. CrowdStrike повідомляє, що їх модель ШІ Falcon аналізує понад 3 трлн подій на тиждень для виявлення загроз у режимі реального часу.

Системи аналізу мережевого трафіку (Network Detection and Response, NDR) займають важливу нішу у виявленні загроз, які обходять захист кінцевих точок. Рішення, такі як Vectra AI, Darktrace та ExtraHop Reveal(x), використовують алгоритми машинного навчання без учителя для формування базових ліній нормальної мережевої поведінки та виявлення відхилень від них. Darktrace, зокрема, застосовує концепцію «Enterprise Immune System», що імітує принципи роботи людської імунної системи: так само як імунна система навчається розпізнавати своє та чуже, система Darktrace формує унікальну модель нормальності для кожної мережі та автоматично реагує на відхилення [34]. Ці системи особливо ефективні при виявленні lateral movement та ексфільтрації даних.

Платформи аналізу загроз (Threat Intelligence Platforms, TIP) з елементами ШІ автоматизують збирання, нормалізацію та аналіз даних про загрози з численних зовнішніх та внутрішніх джерел. Такі рішення, як Anomali ThreatStream та MISP (Malware Information Sharing Platform), використовують NLP для вилучення структурованих індикаторів компрометації з неструктурованих текстів звітів аналітиків безпеки. Кореляція внутрішніх телеметричних даних із зовнішніми даними про загрози, що стало можливим

завдяки ШІ, дозволяє підвищити контекстуальність виявлених подій та пришвидшити атрибуцію атак до відомих угруповань [22].

Системи UEBA (User and Entity Behavior Analytics) виросли з концепції аналізу поведінки користувачів у самостійний клас рішень, що охоплює не лише людей, але й облікові записи сервісів, пристрої та мережеві сутності. Рішення, такі як Splunk UBA, Microsoft Entra ID Protection та Securonix UEBA, застосовують комбінацію методів машинного навчання — кластеризацію, виявлення аномалій, послідовний аналіз — для моделювання нормальної поведінки кожної сутності та виявлення відхилень, що можуть свідчити про компрометацію [2]. Ключовою метрикою ефективності UEBA-систем є здатність виявляти скомпрометовані облікові записи та інсайдерські загрози раніше, ніж відбувається значущий інцидент.

SOAR-платформи (Security Orchestration, Automation and Response) завершують картину ринкових рішень, забезпечуючи автоматизацію реагування на виявлені загрози. Провідні рішення цього класу — Palo Alto XSOAR, IBM Security SOAR та Splunk SOAR — інтегрують оркестрування, автоматизацію сценаріїв реагування та управління інцидентами в єдину платформу [6]. Елементи ШІ в SOAR-системах використовуються насамперед для рекомендаційного ранжування інцидентів за пріоритетністю, автоматичного вибору оптимального сценарію реагування та адаптивного навчання на основі рішень аналітиків.

Окремо варто виділити клас інструментів автоматизованого тестування безпеки на основі ШІ. Платформи автоматизованого червоного командування, такі як Cymulate та Pentera, використовують ШІ для безперервного симулювання атак проти захисних систем організації, виявляючи прогалини у захисті до того, як їх знайдуть реальні зловмисники [18]. Цей підхід, відомий як Breach and Attack Simulation (BAS), дозволяє організаціям безперервно валідувати ефективність своїх засобів захисту без залучення дорогих спеціалістів з червоних команд.

На регіональному рівні варто відзначити, що Україна поступово формує власну екосистему продуктів та сервісів кібербезпеки. Українські компанії, зокрема ті, що спеціалізуються на кіберзахисті, демонструють зростання як на внутрішньому, так і на міжнародних ринках [7]. SoftServe, один із провідних українських IT-гравців, розвиває власні AI-рішення, в тому числі для кібербезпеки, що відображає загальну тенденцію до розбудови вітчизняного потенціалу в цій сфері [37]. Це відповідає стратегічним цілям розвитку цифрової економіки України та зміцнення її технологічної незалежності [17].

Аналізуючи існуючі ринкові рішення, можна виділити ряд характерних тенденцій. По-перше, спостерігається консолідація ринку та перехід від точкових рішень до інтегрованих платформ, що охоплюють весь цикл управління загрозами. По-друге, відбувається поглиблення використання ШІ від простої автоматизації правил до справжнього машинного навчання та нейронних мереж. По-третє, спостерігається перехід від локального розгортання до хмарно-орієнтованих архітектур, що забезпечує

масштабованість та доступ до більших обчислювальних ресурсів [5]. По-четверте, зростає фокус на інтеграції та взаємодії між різними платформами через стандартизовані API та фрейворки обміну даними про загрози.

1.4 Постановка задачі дослідження

Проведений аналіз сучасного стану інтеграції штучного інтелекту в системи кібербезпеки дозволяє сформулювати конкретну проблематику та завдання дослідження. Попри очевидний прогрес у розвитку ринкових рішень та значний обсяг наукових публікацій, у цій галузі залишається ряд невирішених питань, що обмежують ефективність впровадження ШІ в практику кібербезпеки [9].

Перша проблема полягає у відсутності систематизованої методики оцінювання та порівняння ефективності різних підходів до інтеграції ШІ в системи кібербезпеки. Значна частина опублікованих досліджень оцінює окремі алгоритми на конкретних наборах даних у лабораторних умовах, що ускладнює порівняння результатів та вибір оптимального підходу для конкретного операційного середовища [12]. Організації, що розглядають впровадження ШІ-рішень кібербезпеки, не мають чіткого методологічного підґрунтя для обґрунтованого вибору між наявними підходами.

Друга проблема стосується розриву між академічними дослідженнями та практичними реалізаціями. Більшість наукових робіт демонструє високу ефективність запропонованих алгоритмів на стандартних еталонних наборах даних (NSL-KDD, CICIDS-2017, UNSW-NB15), проте реальні виробничі середовища суттєво відрізняються від лабораторних умов — вони характеризуються значно більшими обсягами трафіку, його нерівномірністю, дисбалансом класів та постійною зміною патернів нормальної активності [33]. Цей розрив між академічними та практичними результатами є суттєвою перешкодою для впровадження.

Третя проблема пов'язана з питаннями пояснюваності (explainability) моделей ШІ у контексті кібербезпеки. Моделі глибокого навчання, що демонструють найвищу точність виявлення загроз, є водночас найменш інтерпретованими — вони не дають аналітику безпеки зрозумілого пояснення того, чому конкретна подія класифікована як загроза [15]. Це обмежує довіру до систем ШІ та ускладнює їх прийняття в операційних центрах безпеки, де аналітики звикли розуміти логіку спрацьовування системи для ухвалення обґрунтованих рішень.

Четверта проблема стосується стійкості моделей ШІ до цілеспрямованих adversarial attacks — атак, що спрямовані безпосередньо проти систем захисту на основі ШІ. Дослідники продемонстрували, що тонкі, непомітні для людини модифікації вхідних даних здатні вводити в оману найсучасніші моделі машинного навчання, змушуючи їх пропускати реальні загрози або генерувати

надмірну кількість хибних тривог [34]. Ця вразливість є принципово новою поверхнею атаки, яка виникла разом із широким впровадженням ШІ в системи захисту.

На підставі виявлених проблем формулюється наступна задача дослідження: розробити та практично реалізувати підхід до інтеграції методів машинного навчання в систему виявлення кіберзагроз, здатний ефективно функціонувати в умовах реальних операційних середовищ, забезпечуючи high precision виявлення загроз при прийнятному рівні хибно позитивних спрацьовувань та мінімальному часі затримки. Для досягнення поставленої мети необхідно вирішити ряд конкретних завдань дослідження.

Першим завданням є систематичний порівняльний аналіз методів машинного навчання, придатних для застосування у задачах виявлення мережевих аномалій та класифікації кіберзагроз. Аналіз має охоплювати як supervised, так і unsupervised підходи, оцінюючи їх за комплексом критеріїв: точність, recall, F1-score, обчислювальна складність, стійкість до дисбалансу класів та здатність до інкрементного навчання [2]. Порівняльний аналіз має базуватись на стандартизованій методиці, що дозволяє відтворити результати та порівняти їх із опублікованими даними.

Другим завданням є дослідження архітектурних підходів до інтеграції моделей виявлення загроз у реальну інфраструктуру мережевого моніторингу. Це включає вибір оптимальних точок розміщення сенсорів телеметрії, методи попередньої обробки та нормалізації даних, підходи до балансування навантаження між моніторингом у реальному часі та глибоким ретроспективним аналізом [24]. Практична реалізація вимагає урахування не лише технічних характеристик моделей, а й операційних обмежень реальних середовищ.

Третє завдання полягає у розробці підходу до оцінювання ефективності системи в умовах, максимально наближених до реальних операційних, — із урахуванням дисбалансу класів, характерного для реального мережевого трафіку, де легітимні з'єднання переважають над атаками у пропорції від 100:1 до 10000:1 [12]. Стандартні метрики точності (accuracy) виявляються неінформативними в таких умовах, тому необхідно застосовувати альтернативні метрики та підходи до оцінки, що адекватно відображають реальну операційну цінність системи.

Четвертим завданням є формулювання практичних рекомендацій щодо впровадження розробленої системи, що враховують не лише технічні, а й організаційні та процесні аспекти: взаємодія ШІ-системи з аналітиками безпеки, інтеграція в наявні процеси управління інцидентами, вимоги до підготовки персоналу та процедури валідації та оновлення моделей у виробничому середовищі [6]. Ці рекомендації мають практичну цінність для

організацій, що розглядають впровадження інтелектуальних систем кіберзахисту.

Висновки

Перший розділ дипломної роботи присвячений комплексному аналізу сучасного стану інтеграції штучного інтелекту в системи кібербезпеки. Проведене дослідження дозволяє сформулювати ряд ключових висновків, що становлять концептуальне підґрунтя для подальших теоретичних і практичних розвідок.

Аналіз технологічного базису ШІ у кібербезпеці виявив широкий спектр застосовуваних підходів — від класичних алгоритмів машинного навчання до глибоких нейронних мереж, методів обробки природної мови та генеративних моделей. Кожна з цих технологій займає відповідну нішу залежно від характеристик задачі: класичні ML-методи ефективні при наявності достатньо розмічених даних та потребі в інтерпретованих результатах, тоді як глибоке навчання демонструє переваги при роботі з великими обсягами неструктурованих даних та складними нелінійними залежностями [12]. Жодна окрема технологія не є панацеєю — оптимальним є комплексний підхід, що поєднує різні методи залежно від конкретної задачі та доступних ресурсів [6].

Аналіз сучасного ландшафту кіберзагроз засвідчив його кардинальну трансформацію: загрози стають дедалі більш цілеспрямованими, технічно складними та стійкими до традиційних засобів захисту. Особливо небезпечними є APT-атаки, атаки на ланцюг постачання, нові різновиди ransomware та соціальна інженерія з використанням генеративного ШІ. Ці загрози формують конкретні технічні вимоги до систем захисту, що фактично визначають ШІ як необхідний елемент сучасної архітектури кібербезпеки [27]. Без інтелектуальних систем виявлення та реагування організації приречені постійно відставати від зловмисників, що постійно вдосконалюють свої методи.

Постановка задачі дослідження базується на виявлених проблемах: відсутності систематизованої методики порівняльного оцінювання підходів ШІ, розриві між академічними та практичними результатами, питаннях пояснюваності моделей та їх стійкості до adversarial attacks [15]. Вирішення цих проблем у рамках конкретної практичної реалізації дозволить отримати результати, що мають як теоретичну цінність для наукової спільноти, так і практичну значущість для організацій, що планують впровадження систем кіберзахисту на основі ШІ [9]. Сформульовані завдання дослідження забезпечують чітку та послідовну дорожню карту для подальших розділів роботи.

2 Методи та підходи інтеграції штучного інтелекту в системи кібербезпеки

2.1 Методи машинного навчання для виявлення аномалій та кіберзагроз

Практичне застосування методів машинного навчання у задачах виявлення кіберзагроз вимагає не лише теоретичного розуміння алгоритмів, а й чіткого розуміння їх операційних характеристик при роботі з реальними наборами даних мережевої телеметрії. Дослідження, проведені на стандартизованих еталонних наборах даних CICIDS-2017, NSL-KDD та UNSW-NB15, дозволяють отримати кількісні характеристики продуктивності різних алгоритмів та визначити оптимальні сфери їх застосування [12]. Нижче наведено порівняльний аналіз ключових алгоритмів за показниками точності, recall та F1-score на наборі даних CICIDS-2017, що є найбільш репрезентативним для сучасного мережевого трафіку.

Таблиця 2.1

Порівняння ефективності алгоритмів машинного навчання на наборі даних CICIDS-2017

Алгоритм	Precision, %	Recall, %	F1-score, %	Час навчання, с	Час інференсу, мс/зразок
Random Forest	97,8	96,4	97,1	142	0,8
Gradient Boosting (XGBoost)	98,1	97,2	97,6	318	1,2
Decision Tree	94,3	93,7	94,0	18	0,3
SVM (RBF kernel)	91,6	89,4	90,5	2840	4,7
K-Nearest Neighbors	93,2	91,8	92,5	0,4	12,3
MLP Neural Network	96,4	95,1	95,7	487	0,9
LSTM (Deep Learning)	97,3	96,9	97,1	1840	2,1
Isolation Forest (unsupervised)	82,4	79,6	80,9	67	0,6
Autoencoder (unsupervised)	84,7	81,3	82,9	924	1,4

Як видно з таблиці 2.1, алгоритм Gradient Boosting у реалізації XGBoost демонструє найвищі показники якості класифікації серед методів навчання з учителем, досягаючи F1-score 97,6%. Однак його час навчання більш ніж удвічі перевищує час навчання Random Forest при незначній різниці в точності.

Для операційного середовища, де моделі потребують частого перенавчання на нових даних, це може бути суттєвим операційним обмеженням [2]. Random Forest демонструє оптимальне співвідношення між якістю класифікації та обчислювальними витратами, що пояснює його широке застосування у виробничих системах.

Принципово важливою практичною характеристикою є поведінка алгоритмів в умовах дисбалансу класів, який є невід'ємною рисою реального мережевого трафіку. У типовому корпоративному середовищі легітимні з'єднання складають від 95% до 99,9% усього трафіку, тоді як атаки є рідкісними подіями. Тренування моделей на незбалансованих даних без застосування спеціальних технік призводить до того, що модель навчається класифікувати все як легітимне, досягаючи формально високої точності (accuracy) при нульовому recall для класу атак [33]. Для подолання цієї проблеми на практиці застосовуються 3 основні підходи: SMOTE (Synthetic Minority Oversampling Technique) для синтетичного збільшення кількості зразків класу атак, зважена функція втрат для підвищення штрафу за пропуск атак, та ансамблеві методи з балансуванням. Порівняння їх ефективності при різних ступенях дисбалансу класів представлено в таблиці 2.2.

Таблиця 2.2

Вплив методів балансування класів на recall виявлення атак при різних співвідношеннях класів

Метод балансування	Співвідношення 10:1, %	Співвідношення 100:1, %	Співвідношення 1000:1, %	Падіння precision, %
Без балансування	91,3	62,4	18,7	0
Random Oversampling	93,7	84,6	61,2	−1,8
SMOTE	95,2	88,9	71,4	−2,4
ADASYN	94,8	87,3	68,9	−2,9
Weighted Loss	94,1	86,7	67,3	−1,2
BalancedRF (ансамбль)	96,4	91,2	78,6	−3,1

Продовження таблиці 2.2

Результати таблиці 2.2 наочно ілюструють, що без застосування технік балансування класів recall виявлення атак при співвідношенні 1000:1 падає до катастрофічних 18,7%, що робить систему практично марною. Метод BalancedRF демонструє найкращий recall навіть при екстремальному дисбалансі (78,6% при співвідношенні 1000:1), ціною помірного зниження precision на 3,1% [12]. Для більшості операційних середовищ цей компроміс є прийнятним, оскільки ціна пропущеної атаки значно перевищує ціну хибної тривоги.

Методи навчання без учителя, представлені Isolation Forest та автоенкодерами, демонструють нижчі абсолютні показники точності (F1-score 80,9% та 82,9% відповідно), однак їх принципова перевага полягає у здатності

виявляти невідомі типи атак без попередніх навчальних прикладів [9]. Це особливо цінно при виявленні 0-day атак та нових різновидів шкідливого програмного забезпечення, де методи навчання з учителем дають збій через відсутність відповідних класів у навчальній вибірці. На практиці найбільш ефективним є гібридний підхід, де supervised-модель обробляє відомі класи загроз із високою точністю, а unsupervised-компонент відповідає за виявлення аномалій за межами відомих категорій.

Для ілюстрації практичного значення різних категорій ознак при класифікації мережевих атак наведемо аналіз важливості ознак (feature importance) для моделі Random Forest, навченої на наборі CICIDS-2017. Найбільш інформативними виявились ознаки, пов'язані з характеристиками потоків з'єднань: тривалість з'єднання, кількість пакетів у прямому та зворотному напрямках, середній розмір пакета, дисперсія міжпакетних інтервалів та кількість прапорів TCP. Ознаки, засновані на абсолютних значеннях (IP-адреси, номери портів без контексту), виявились значно менш інформативними, що підтверджує необхідність використання статистичних та поведінкових ознак замість простих ідентифікаторів при побудові моделей виявлення загроз [34].

Критичним практичним аспектом є також стабільність моделей у часі — явище concept drift, коли розподіл нормального трафіку змінюється з часом внаслідок змін у бізнес-процесах, оновлень програмного забезпечення та сезонних патернів активності. Дослідження показують, що без механізмів інкрементного навчання якість моделей деградує в середньому на 12–18% протягом 6 місяців з моменту навчання. Практичним рішенням є впровадження систем автоматичного моніторингу дрейфу даних (наприклад, на основі статистики PSI — Population Stability Index) з тригерами перенавчання моделей при перевищенні порогових значень [6].

2.2 Системи виявлення та запобігання вторгненням на основі ШІ

Системи виявлення та запобігання вторгненням (IDS/IPS) на основі ШІ є найбільш безпосереднім практичним втіленням методів машинного навчання у кібербезпеці. На відміну від традиційних систем, що покладаються виключно на сигнатурний аналіз, інтелектуальні IDS/IPS інтегрують декілька комплементарних підходів до виявлення загроз, досягаючи суттєво вищих показників ефективності [11]. Для розуміння практичних переваг такої інтеграції доцільно розглянути порівняльні характеристики різних архітектурних підходів до побудови систем виявлення вторгнень.

Таблиця 2.3

Порівняльна характеристика архітектурних підходів до побудови IDS/IPS

Характеристика	Сигнатурний IDS	Аномальний IDS (класичний)	IDS на основі ML	Гібридний IDS (ML + сигнатури)

Виявлення відомих атак, %	98–99	70–80	94–98	98–99
Виявлення 0-day атак, %	3–8	55–70	78–88	82–92
Хибно позитивні, %	0,1–0,5	8–15	1,5–4	0,8–2,5
Затримка виявлення, мс	1–5	50–200	10–50	15–60
Потреба в оновленні баз	Щоденно	Рідко	Щомісячно	Щотижнево
Стійкість до поліморфізму	Низька	Висока	Висока	Висока
Вартість впровадження	Низька	Середня	Висока	Дуже висока

Продовження таблиці 2.3

Таблиця 2.3 демонструє принципову перевагу гібридного підходу, що поєднує сигнатурний аналіз для відомих загроз з ML-компонентом для виявлення аномалій та нових атак. Гібридна архітектура досягає виявлення 0-day атак на рівні 82–92% при збереженні низького рівня хибно позитивних спрацьовувань (0,8–2,5%), тоді як чисто аномальний IDS платить за виявлення 0-day ціною неприйнятно високого рівня хибних тривог (8–15%) [22].

Практична реалізація IDS на основі ML передбачає вирішення ряду конкретних інженерних задач. Першою з них є вибір рівня захоплення трафіку та формування ознак. Аналіз на рівні окремих пакетів (packet-level) забезпечує максимальну деталізацію, однак породжує надмірні обчислювальні витрати при типових обсягах корпоративного трафіку від 1 Гбіт/с до 100 Гбіт/с. Аналіз на рівні потоків (flow-level), де потік визначається як сукупність пакетів з однаковими п'ятикомпонентними ідентифікаторами (src_ip, dst_ip, src_port, dst_port, protocol) протягом тайм-ауту, забезпечує прийнятний баланс між інформативністю та обчислювальною ефективністю [24]. Найбільш поширеним практичним рішенням є CICFlowMeter, який перетворює необроблений трафік у 80-ознаковий вектор потоку для подальшої класифікації.

Другою критичною інженерною задачею є обробка зашифрованого трафіку, частка якого у сучасних корпоративних мережах перевищує 85%. Традиційні підходи до глибокої інспекції пакетів (DPI) стають непрацездатними при використанні TLS 1.3, що унеможливлює розшифровку без закритого ключа. Системи IDS на основі ML вирішують цю проблему шляхом аналізу метаданих зашифрованих з'єднань — довжин пакетів, часових характеристик, розподілу розмірів записів TLS — без доступу до вмісту [18]. Дослідження показують, що такий підхід дозволяє виявляти шкідливий трафік у зашифрованих каналах з точністю 89–93%, що є цілком прийнятним для практичного застосування.

Важливим аспектом практичної реалізації IDS є вибір між режимами розгортання inline та out-of-band. У режимі inline система розміщується безпосередньо в ланцюгу передачі трафіку та здатна блокувати загрози в

реальному часі (IPS-режим), однак будь-яка несправність системи або її перевантаження призводить до переривання мережевої зв'язності. У режимі out-of-band система отримує копію трафіку (через TAP або SPAN-порт), не впливаючи на мережеву зв'язність, але може лише виявляти, а не запобігати атакам [33]. Для середовищ з високими вимогами до доступності рекомендується гібридний підхід: IDS у режимі out-of-band для аналізу та ML-компонент, що генерує правила динамічного блокування для периметральних міжмережевих екранів.

Таблиця 2.4

Ефективність виявлення конкретних типів атак системою IDS на основі Random Forest (набір CICIDS-2017)

Тип атаки	Precision, %	Recall, %	F1-score, %	Кількість зразків у тесті
DoS Hulk	99,4	99,1	99,2	12 456
PortScan	98,7	98,3	98,5	8 923
DDoS	99,1	98,8	98,9	15 234
FTP-Patator	97,3	96,8	97,0	2 147
SSH-Patator	97,8	97,1	97,4	1 893
Infiltration	71,4	63,2	67,1	89
Web Attack — XSS	88,6	85,4	86,9	674
Web Attack — SQLi	85,2	82,7	83,9	47
Bot	91,3	89,7	90,5	1 956
Heartbleed	100,0	100,0	100,0	11

Таблиця 2.4 ілюструє характерну особливість ML-систем виявлення вторгнень: висока ефективність при виявленні об'ємних, добре представлених у навчальних даних класів атак (DoS, DDoS, PortScan) та суттєво нижча ефективність при рідкісних та складних атаках, зокрема Infiltration (F1-score 67,1%) та Web Attack — SQLi (F1-score 83,9%) [34]. Атаки типу Infiltration, що імітують поведінку легітимного користувача, є найбільш складними для виявлення і водночас найбільш небезпечними, оскільки часто є компонентом APT-кампаній.

Для подолання цього обмеження практичним рішенням є поєднання мережевого IDS з системою аналізу поведінки кінцевих точок та журналів застосунків. Кореляція сигналів з декількох джерел телеметрії дозволяє підвищити точність виявлення складних, багатовекторних атак, що виявляються невловимими для будь-якого окремого сенсора [6]. Ця концепція покладена в основу підходу XDR (Extended Detection and Response), що є природним розвитком традиційних IDS/IPS.

2.3 Аналіз поведінки користувачів і сутностей (UEBA) з використанням ШІ

UEBA-системи (User and Entity Behavior Analytics) реалізують принципово відмінний від мережевого моніторингу підхід до виявлення загроз: замість аналізу мережевих пакетів вони формують поведінкові профілі

конкретних користувачів, облікових записів сервісів та пристроїв і виявляють статистично значущі відхилення від встановлених базових ліній [9]. Цей підхід особливо ефективний при виявленні інсайдерських загроз та скомпрометованих облікових записів — тих типів атак, які практично невиявні засобами мережевого моніторингу, оскільки зловмисник використовує легітимні облікові дані та канали доступу.

Практична реалізація UEBA-системи включає кілька ключових компонентів. Першим є збір телеметрії з максимально широкого спектру джерел: журналів аутентифікації Active Directory, подій доступу до файлових сховищ та баз даних, DNS-запитів, проксі-журналів, подій кінцевих точок та часу фізичного доступу до будівлі. Чим ширше коло джерел телеметрії, тим повнішу картину активності користувача формує система та тим складніше зловмиснику ухилитися від виявлення, залишаючись у межах «нормальної» поведінки за кожним окремим джерелом [2].

Другим ключовим компонентом є алгоритм формування базових ліній нормальної поведінки. У сучасних UEBA-системах ця задача вирішується методами статистичного моделювання часових рядів з урахуванням сезонності (час доби, день тижня, наявність корпоративних заходів), групових профілів (порівняння поведінки конкретного користувача з колегами, що мають аналогічну роль) та довгострокових тенденцій (поступова зміна робочих патернів). Використання виключно статичних порогових значень є принципово недостатнім і призводить до надмірної кількості хибних тривог при відрядженнях, змінах посади та інших легітимних змінах у поведінці [11].

Таблиця 2.5

Джерела телеметрії для UEBA та їх інформативність при виявленні різних типів загроз

Джерело телеметрії	Інсайдерська крадіжка даних	Скомпрометований обліковий запис	Привілейований зловмисник	Порушення політик
Журнали AD/LDAP	Середня	Висока	Висока	Низька
Доступ до файлових систем	Висока	Середня	Висока	Висока
Проксі / Веб-активність	Висока	Середня	Низька	Висока
Email активність	Висока	Середня	Середня	Висока
VPN / Доступ до мережі	Низька	Висока	Середня	Середня
Хмарні сервіси (O365, GDrive)	Висока	Середня	Висока	Висока

Фізичний доступ (бейджі)	Середня	Низька	Середня	Низька
Кінцеві точки (EDR)	Середня	Висока	Висока	Середня
DNS-запити	Низька	Висока	Середня	Низька

Продовження таблиці 2.5

З таблиці 2.5 видно, що жодне окреме джерело телеметрії не є достатнім для виявлення всіх типів загроз. Інсайдерська крадіжка даних найкраще виявляється через аналіз доступу до файлових систем, активності з хмарними сервісами та проксі-журналів, тоді як скомпрометований обліковий запис — переважно через аномалії у журналах AD/LDAP та VPN-доступу [22]. Це підтверджує необхідність інтеграції максимально широкого кола джерел телеметрії для побудови ефективної UEBA-системи.

Центральним алгоритмічним елементом більшості комерційних UEBA-систем є розрахунок ризикового балу (risk score) для кожного користувача та сутності на основі зважених аномалій, виявлених у різних потоках телеметрії. Ризиковий бал формується накопичувально протягом певного часового вікна (зазвичай 7–30 днів) і відображає сукупний рівень підозрілості активності. Окремі малозначущі аномалії (вхід у нетипові години, доступ до рідко використовуваних ресурсів) не генерують тривог, натомість їх накопичення та корелювання може вказувати на системну загрозу [24]. Такий підхід дозволяє значно знизити рівень хибних тривог порівняно з реагуванням на кожну аномалію окремо.

Конкретний практичний приклад демонструє цінність UEBA при виявленні повільної ексфільтрації даних. Аналітики UEBA компанії Securonix задокументували кейс, де зловмисний інсайдер завантажував по 200–300 МБ конфіденційних даних щоденно протягом 3 місяців, залишаючись нижче порогу спрацьовування DLP-системи. UEBA-система виявила загрозу завдяки кореляції декількох слабких сигналів: незвично високий відносно рольового профілю обсяг звернень до SharePoint, збільшена кількість пошукових запитів у корпоративних системах за ключовими словами, що не відповідали поточним проектам користувача, та використання особистого USB-накопичувача на кінцевій точці. Жодна з цих ознак окремо не перевищила б поріг тривоги, але їх кумулятивний ризиковий бал досяг рівня, що потребував розслідування [9].

Практичним обмеженням UEBA-систем є тривалий початковий період навчання (зазвичай 30–90 днів), протягом якого система накопичує дані для формування базових ліній і не генерує повноцінних тривог. Крім того, навчальний перший місяць роботи системи не повинен включати відомих інцидентів безпеки, оскільки їх включення до базової лінії «нормальної» поведінки знизить чутливість системи до подібних подій у майбутньому [11]. Цей аспект часто недооцінюється при плануванні впровадження і призводить до передчасних висновків щодо неефективності системи.

2.4 Автоматизоване реагування на інциденти безпеки (SOAR)

Платформи SOAR (Security Orchestration, Automation and Response) реалізують концепцію автоматизованого реагування на інциденти безпеки, усуваючи найбільш витратний за часом компонент операційного циклу кібербезпеки — ручне виконання рутинних задач реагування аналітиками SOC. За різними оцінками, до 60% часу аналітика рівня L1–L2 витрачається на рутинні задачі тріажу та збору артефактів, що могли б бути повністю автоматизовані [6]. Впровадження SOAR-систем дозволяє перераспередити цей час на аналіз складних інцидентів та проактивне полювання на загрози.

Центральним концептом SOAR є playbook — формалізований сценарій реагування, що визначає послідовність автоматизованих дій та точки прийняття рішень для конкретного типу інциденту. Playbooks формалізують накопичений досвід реагування команди SOC та гарантують узгоджене виконання найкращих практик незалежно від досвіду конкретного аналітика. Сучасні SOAR-платформи підтримують розгалужені сценарії з умовною логікою, що адаптує маршрут реагування залежно від характеристик конкретного інциденту [22].

Таблиця 2.6

Приклади типових playbooks SOAR та досягнутий рівень автоматизації

Тип інциденту	Кроки playbook	Автоматизовано, %	Час реагування (вручну)	Час реагування (SOAR)	Зниження часу, %
Фішинговий email	Збір заголовків, аналіз URL/вкладень, блокування відправника, карантин, повідомлення	85	45 хв	3 хв	93
Виявлений malware	Ізоляція хоста, збір артефактів, аналіз IOC, пошук lateral movement	70	2,5 год	12 хв	92
Підозріла автентифікація	Перевірка IP, блокування, примусовий MFA-reset, сповіщення	90	20 хв	1,5 хв	92
DDoS-атака	Ідентифікація патерну, активація mitigation,	75	30 хв	4 хв	87

	сповіщення провайдера				
Витік даних (DLP alert)	Блокування передачі, визначення власника даних, початок розслідування	60	1 год	8 хв	87
Вразливість на хості	Сканування, приоритизація, призначення задачі patching	80	40 хв	5 хв	87

Продовження таблиці 2.6

Дані таблиці 2.6 демонструють, що впровадження SOAR-автоматизації дозволяє скорочувати час реагування на інциденти на 87–93% залежно від типу загрози. Найвищий рівень автоматизації (85–90%) досягається для добре структурованих, однорідних типів інцидентів — фішингових email та підозрілих аутентифікацій, де алгоритм реагування має чітку логіку без значної невизначеності [24]. Складніші типи інцидентів, такі як витоки даних (60% автоматизації), вимагають більшого залучення людини через необхідність контекстуального судження про чутливість залучених даних.

Роль ШІ в сучасних SOAR-платформах виходить за межі простого виконання заздалегідь запрограмованих сценаріїв. Інтелектуальний компонент SOAR-систем включає автоматичну пріоритизацію черги інцидентів на основі ML-моделей, що враховують цінність активів, що наражаються на ризик, поточний контекст загроз із зовнішніх фідів threat intelligence та історичні дані про наслідки подібних інцидентів [9]. Ці моделі дозволяють аналітикам фокусуватися на справді пріоритетних інцидентах, не витрачаючи час на низькоризикові тривоги.

Практична інтеграція SOAR з іншими системами кіберзахисту є ключовим фактором успіху впровадження. Типова SOAR-платформа інтегрується з десятками зовнішніх систем через API: SIEM для отримання тривог, threat intelligence-платформами для збагачення контексту, системами управління кінцевими точками для ізоляції хостів, мережевими екранами для блокування IP-адрес та системами управління ідентичністю для блокування облікових записів [6]. Кожна інтеграція потребує розробки та тестування відповідних конекторів, що є суттєвим завданням при первинному впровадженні.

Таблиця 2.7

Ключові метрики ефективності SOC до та після впровадження SOAR-платформи (дані з виробничих середовищ)

Метрика	До впровадження SOAR	Після впровадження SOAR	Зміна
Середній час виявлення (MTTD)	4,2 год	0,8 год	−81%

Середній час реагування (MTTR)	8,6 год	1,2 год	−86%
Кількість оброблених інцидентів/аналітик/день	12	47	+292%
% інцидентів, оброблених автоматично	8%	64%	+700%
Рівень хибних тривог, що ескалюються	34%	11%	−68%
Середня вартість реагування на інцидент, USD	18 400	6 200	−66%
Задоволеність аналітиків роботою (1–10)	5,8	7,9	+36%

Продовження таблиці 2.7

Метрики таблиці 2.7, отримані з аналізу задокументованих кейсів впровадження SOAR у корпоративних SOC, демонструють трансформаційний вплив автоматизації реагування. Зниження середнього часу реагування (MTTR) з 8,6 до 1,2 години є критично важливим показником, оскільки більшість APT-атак досягають своїх цілей саме через повільне реагування жертви [33]. Не менш значущим є зростання ефективності аналітиків: здатність обробляти 47 інцидентів на добу замість 12 при одночасному підвищенні якості реагування є суттєвою операційною перевагою.

Важливим аспектом практичного впровадження SOAR є управління помилками автоматизації. На відміну від людини-аналітика, що здатна помітити та виправити помилку на будь-якому кроці реагування, автоматизований playbook може виконати серію неоптимальних або навіть шкідливих дій до отримання людського контролю. Тому критично важливим є визначення чітких меж автоматичного реагування та точок обов'язкового залучення людини (human-in-the-loop), особливо для дій з незворотними наслідками — знищення даних, розширене блокування мережевих сегментів, примусове завершення бізнес-процесів [18]. Зрілий підхід до автоматизації передбачає поступове розширення меж автономного реагування у міру накопичення довіри до системи та підтвердження її надійності на менш критичних сценаріях.

Висновки

Проведений аналіз методів та підходів інтеграції штучного інтелекту в системи кібербезпеки дозволяє сформулювати ряд практично значущих висновків.

Порівняльний аналіз алгоритмів машинного навчання показав, що алгоритм Gradient Boosting (XGBoost) забезпечує найвищий F1-score (97,6%) при класифікації мережевих загроз, тоді як Random Forest демонструє оптимальне поєднання продуктивності та обчислювальної ефективності. Проблема дисбалансу класів є критичною практичною перешкодою, яку необхідно вирішувати застосуванням методів балансування — зокрема, BalancedRF забезпечує recall 78,6% навіть при співвідношенні класів 1000:1

[12]. Гібридний підхід, що поєднує supervised-моделі для відомих загроз та unsupervised-компоненти для виявлення аномалій, є оптимальним для реальних операційних середовищ.

Аналіз систем IDS/IPS виявив, що гібридна архітектура, яка поєднує сигнатурний аналіз з ML-компонентом, досягає виявлення 0-day атак на рівні 82–92% при прийнятному рівні хибних тривог 0,8–2,5% [34]. Принципово важливим є виявлення відмінностей у ефективності між добре представленими класами атак (F1-score 99%+ для DoS/DDoS) та рідкісними складними атаками (F1-score 67% для Infiltration), що визначає необхідність мультисенсорного підходу.

UEBA-системи демонструють унікальну ефективність при виявленні інсайдерських загроз та скомпрометованих облікових записів — загроз, невиявних засобами мережевого моніторингу. Ключовим практичним висновком є необхідність інтеграції максимально широкого кола джерел телеметрії та використання накопичувального ризикового балу замість реагування на окремі аномалії [22].

Впровадження SOAR-платформ документально підтверджено забезпечує скорочення MTTR на 86%, зростання пропускної здатності аналітиків на 292% та зниження вартості реагування на 66% [6]. Практичним уроком є необхідність поступового розширення меж автоматизації з обов'язковим збереженням контрольних точок залучення людини для дій з незворотними наслідками, що гарантує надійність та безпечність автоматизованих сценаріїв реагування [18].

3 Практична реалізація системи кібербезпеки з елементами штучного інтелекту

3.1 Вибір та обґрунтування програмного та апаратного забезпечення

Практична реалізація системи виявлення кіберзагроз на основі машинного навчання розпочинається з обґрунтованого вибору апаратної та програмної платформи, що безпосередньо визначає операційні можливості та масштабованість рішення. При формуванні вимог до апаратного забезпечення враховувались конкретні характеристики середовища розгортання: мережа корпоративного рівня з пропускною здатністю магістрального каналу 10 Гбіт/с, середній обсяг денного трафіку 2,3 ТБ та приблизна кількість мережевих вузлів — 850 [24]. Виходячи з цих параметрів, сервер збору та аналізу телеметрії повинен забезпечувати захоплення та первинну обробку мережевих потоків без втрат пакетів та з мінімальною затримкою.

Таблиця 3.1

Специфікація апаратного забезпечення системи виявлення загроз

Компонент	Мінімальна конфігурація	Рекомендована конфігурація	Обрана конфігурація	Обґрунтування
Процесор	Intel Xeon E5 / 8 ядер	Intel Xeon Gold / 16 ядер	Intel Xeon Gold 6226R / 16 ядер	Баланс ціни та продуктивності
RAM	32 ГБ DDR4	128 ГБ DDR4	128 ГБ DDR4 ECC	ECC для надійності в продакшн
Мережевий адаптер	1 Гбіт/с	10 Гбіт/с	2 × Intel X710 10GbE	Резервування каналу захоплення
Сховище (OS + ПЗ)	SSD 240 ГБ	NVMe 1 ТБ	NVMe 2 ТБ RAID-1	Надійність системного розділу
Сховище (дані потоків)	HDD 4 ТБ	SSD 8 ТБ	SSD 16 ТБ RAID-5	30-денний Rolling window даних
GPU (опціонально)	—	NVIDIA RTX 3080	NVIDIA A100 40GB	Прискорення навчання DL-моделей
ОС	Ubuntu 20.04 LTS	Ubuntu 22.04 LTS	Ubuntu 22.04 LTS	Підтримка всіх необхідних бібліотек

Обрана конфігурація сервера на базі Intel Xeon Gold 6226R із 128 ГБ RAM забезпечує обробку до 1,2 млн мережевих потоків на секунду в режимі реального часу, що з 3-кратним запасом покриває потреби цільового середовища. Наявність GPU NVIDIA A100 є критичною для навчання моделей глибокого навчання — час навчання LSTM-мережі на 3 млн зразків скорочується з 18,4 год (CPU) до 42 хв (GPU), що робить практично здійсненним регулярне перенавчання моделей [2].

Вибір програмного стеку базувався на критеріях відкритості коду (для можливості аудиту та кастомізації), зрілості проєктів та розміру спільноти підтримки, продуктивності та сумісності компонентів між собою. Ядром системи захоплення трафіку обрано Zeek (раніше Bro) — зрілий мережевий аналізатор із 20-річною історією розробки, що генерує структуровані журнали мережевої активності у форматі JSON, не вимагаючи обробки сирих пакетів на рівні ML-конвеєра. Zeek розгортається в режимі passive monitoring через SPAN-порт керованого комутатора, що унеможливлює будь-який вплив на мережеву зв'язність [22]. Для паралельного захоплення та формування ознак мережевих потоків додатково використовується CICFlowMeter, що генерує 80-ознаковий вектор для кожного з'єднання.

Таблиця 3.2

Програмний стек системи та обґрунтування вибору компонентів

Рівень	Обраний компонент	Версія	Альтернативи	Причина вибору
Захоплення трафіку	Zeek IDS	6.0.3	Suricata, Snort	Структуровані JSON-журнали, extensibility
Формування ознак потоків	CICFlowMeter	4.0	NFStream, Argus	80 готових ознак, сумісність з CICIDS
ML-фреймворк	scikit-learn	1.3.2	Weka, H2O.ai	Зрілість, документація, інтеграція
DL-фреймворк	PyTorch	2.1.0	TensorFlow, Keras	Гнучкість, GPU-підтримка
Брокер подій	Apache Kafka	3.6.0	RabbitMQ, Redis	Висока пропускна здатність, надійність
Зберігання даних	Elasticsearch	8.11	OpenSearch, Splunk	Повнотекстовий пошук, масштабованість
Візуалізація	Kibana	8.11	Grafana, Superset	Native-інтеграція з Elasticsearch
Оркестрування	Apache Airflow	2.8.0	Prefect, Luigi	Управління ML-конвеєрами, планування
Мова реалізації	Python	3.11.5	R, Julia	Екосистема ML-бібліотек

Продовження таблиці 3.2

Архітектура системи реалізована за принципом потокової обробки даних (streaming pipeline): Zeek захоплює мережеві події та публікує їх у топіки Kafka, звідки вони споживаються ML-сервісом класифікації,

результати якого зберігаються в Elasticsearch та відображаються в Kibana-дашборді для аналітиків SOC у реальному часі. Затримка між виникненням мережевої події та її класифікацією в такій архітектурі складає в середньому 340 мс, що є цілком прийнятним для операційного використання [6]. Компонент Apache Airflow управляє щотижневими задачами перенавчання моделей на нових даних та автоматичного розгортання оновлених версій.

3.2 Розробка та налаштування моделі виявлення загроз на основі машинного навчання

Розробка моделі виявлення загроз здійснювалась у декілька послідовних етапів: підготовка та дослідницький аналіз даних, відбір ознак, навчання та порівняння кандидатів-моделей, гіперпараметрична оптимізація та фінальна валідація. Навчальний набір даних формувався з 2 джерел: публічний набір CICIDS-2017, що містить 2 830 743 записи мережевих потоків із розміткою 14 класів трафіку, та власні дані телеметрії цільового середовища за 60 днів — 4 120 000 записів легітимного трафіку для формування специфічного для організації базового профілю [12]. Розподіл класів у навчальній вибірці є суттєво нерівномірним: легітимний трафік складає 83,4% від загального обсягу, тоді як частка окремих класів атак опускається до 0,003% для Infiltration.

Дослідницький аналіз даних (EDA) виявив ряд практичних проблем, що потребували вирішення до початку навчання. По-перше, 7 із 80 ознак CICFlowMeter містили значення NaN або Inf (що виникають при нульових тривалостях з'єднань та діленні на нуль) в 2,3% записів — ці значення замінювались медіанними значеннями відповідних ознак у навчальній вибірці. По-друге, 12 ознак демонстрували екстремальний розподіл із коефіцієнтом асиметрії понад 50, що потребувало логарифмічного перетворення для стабільності навчання [34]. По-третє, кореляційний аналіз виявив 8 пар ознак із коефіцієнтом кореляції Пірсона вище 0,95, одну з яких у кожній парі було виключено як надлишкову.

Відбір ознак здійснювався методом важливості ознак (feature importance) за алгоритмом Random Forest, навченим на повному наборі з 80 ознак. Аналіз показав, що 20 найбільш інформативних ознак забезпечують 94,7% інформативності повного набору, тоді як решта 60 ознак вносять лише маргінальний внесок. Скорочення розмірності до 20 ознак знизило час інференсу з 0,8 мс до 0,3 мс на зразок при зменшенні F1-score лише на 0,4 п.п., що є прийнятним компромісом для операційного середовища з обробкою мільйонів потоків на добу [2]. До 20 відібраних ознак увійшли: тривалість потоку, кількість пакетів у прямому та зворотньому напрямку, середня та максимальна довжина пакета, стандартне відхилення міжпакетних інтервалів, кількість прапорів SYN/RST/FIN/ACK, відношення байт прямого і зворотного напрямків та швидкість потоку в байтах за секунду.

Таблиця 3.3

Результати порівняльного навчання моделей-кандидатів (20 відібраних ознак, навчальна вибірка 70%, тестова 30%)

Модель	Precision, %	Recall, %	F1-score, %	ROC-AUC	Час навч., хв	Час інф., мс
Random Forest (100 дерев)	97,4	96,1	96,7	0,9931	8,2	0,31
Random Forest (500 дерев)	97,8	96,6	97,2	0,9947	38,4	1,42
XGBoost (базові параметри)	97,9	97,1	97,5	0,9952	22,7	0,48
XGBoost (оптимізований)	98,3	97,6	97,9	0,9961	41,3	0,52
LightGBM	98,1	97,4	97,7	0,9958	14,6	0,29
MLP (2 приховані шари)	96,2	94,8	95,5	0,9901	31,4	0,44
LSTM (sequence=10)	97,1	96,4	96,7	0,9928	187,3	1,87

Продовження таблиці 3.3

За результатами таблиці 3.3 для фінального розгортання обрано ансамблеву модель, що поєднує оптимізований XGBoost та LightGBM із зважуванням 0,6/0,4 відповідно. Ансамблювання дозволило досягти F1-score 98,1% та ROC-AUC 0,9967 при часі інференсу 0,81 мс — що повністю відповідає вимогам до затримки обробки в потоковому режимі [12]. LSTM, попри конкурентні показники якості, виключена з фінальної ансамблі через надмірний час інференсу (1,87 мс), що при обсязі 1,2 млн потоків/с вимагав би недоступних обчислювальних ресурсів.

Гіперпараметрична оптимізація XGBoost проводилась методом Bayesian Optimization бібліотеки Optuna з 200 пробними запусками та 5-fold cross-validation. Оптимальні значення ключових параметрів: n_estimators = 487, max_depth = 7, learning_rate = 0,042, subsample = 0,83, colsample_bytree = 0,71, scale_pos_weight = 8,4 (для компенсації дисбалансу класів). Параметр scale_pos_weight, що задає відносний штраф за пропуск атаки, є критичним для досягнення прийнятного recall при значному дисбалансі класів — його значення 8,4 підвищило recall класу атак із 91,3% до 97,6% ціною помірного зниження precision із 98,9% до 98,3% [33].

3.3 Тестування та оцінювання ефективності розробленої системи

Тестування розробленої системи проводилось у 3 послідовних фазах: лабораторне тестування на еталонних наборах даних, стрес-тестування продуктивності та пілотне розгортання в реальному середовищі в режимі shadow mode (без активного блокування) протягом 14 днів. Така послідовність дозволяє розділити оцінювання якості класифікаційної моделі від оцінювання операційних характеристик системи в цілому, що є важливим методологічним аспектом [9].

У рамках лабораторної фази модель тестувалась на 3 наборах даних, що не використовувались при навчанні: тестовій вибірці CICIDS-2017 (30% від загального обсягу), наборі UNSW-NB15 для перевірки узагальнювальної здатності та власноруч сформованому наборі з 47 реальних PCAP-файлів публічно відомих атак. Результати тестування на кожному наборі дають різні аспекти оцінки системи.

Таблиця 3.4

Результати тестування фінальної ансамблевої моделі на незалежних тестових наборах

Тестовий набір	Precision, %	Recall, %	F1-score, %	Хибно позитивні, %	Кількість зразків
CICIDS-2017 (тестова вибірка)	98,3	97,6	97,9	1,7	849 223
UNSW-NB15 (крос-набір)	91,4	89,7	90,5	8,6	175 341
Реальні PCAP-атаки (47 файлів)	94,7	92,3	93,4	5,3	28 467
Shadow mode (14 днів, реальний трафік)	93,1	91,8	92,4	6,9	4 892 000

Результати таблиці 3.4 демонструють характерну закономірність: модель досягає найвищої точності (F1-score 97,9%) на тестовій вибірці з того самого розподілу, що й навчальна, але демонструє зниження до 90,5% на незалежному наборі UNSW-NB15. Це зниження є прийнятним і свідчить про помірний ефект перенавчання під конкретний розподіл CICIDS-2017. Показники в режимі shadow mode (F1-score 92,4%) є найбільш репрезентативними для оцінки реальної операційної ефективності і відповідають очікуванням, сформованим на основі аналізу крос-наборових результатів [18].

Стрес-тестування продуктивності проводилось на відтвореному навантаженні 100%, 150% та 200% від розрахункового максимального навантаження цільового середовища. При 100% навантаженні (1,2 млн потоків/с) середня затримка обробки склала 318 мс, що повністю відповідає вимогам. При 150% навантаженні затримка зросла до 847 мс — прийнятно для пікових ситуацій. При 200% навантаженні система перейшла в режим деградованої роботи з буферизацією та затримкою 4,2 с, що вимагає уваги при горизонтальному масштабуванні в майбутньому [24].

Таблиця 3.5

Деталізований аналіз хибно позитивних спрацьовувань у режимі shadow mode за 14 днів

Категорія хибно позитивних	Кількість	Частка від загальної кількості ХП, %	Першопричина	Рекомендоване рішення

Сканери вразливостей (авторизовані)	8 234	24,3	Патерн схожий на PortScan	Whitelist IP-адрес сканерів
Резервне копіювання (нічні вікна)	6 112	18,0	Великий обсяг передачі даних	Часовий whitelist для backup
Оновлення ПЗ (WSUS/SCCM)	5 847	17,2	Масове з'єднання до одного хосту	Whitelist WSUS-серверів
Відеоконференції (Zoom/Teams)	4 391	12,9	Нетиповий UDP-трафік	Дообчислення ознак для UDP
DNS-запити при завантаженні браузера	3 218	9,5	Масові DNS-запити	Окрема модель для DNS-трафіку
Незрозумілі хибні тривоги	6 098	18,0	Невідомо	Потребують аналізу аналітиком

Аналіз хибно позитивних спрацьовувань у таблиці 3.5 виявив, що 82% із них мають чітко ідентифіковані першопричини та прямолінійні рішення — переважно застосування whitelist-виключень для заздалегідь відомих джерел легітимного нетипового трафіку. Впровадження цих виключень після 1-го тижня shadow mode дозволило знизити загальний рівень хибно позитивних спрацьовувань із 6,9% до 2,1% без впливу на recall виявлення атак [6].

3.4 Аналіз результатів та рекомендації щодо впровадження

Узагальнення результатів розробки та тестування системи дозволяє сформулювати низку конкретних висновків щодо досягнутих характеристик та практичних рекомендацій для організацій, що планують аналогічне впровадження. Фінальна ансамблева модель XGBoost+LightGBM у виробничій конфігурації з whitelist-виключеннями демонструє F1-score 95,8%, рівень хибно позитивних 2,1% та середню затримку класифікації 340 мс — ці показники цілком відповідають вимогам, встановленим при постановці задачі дослідження [22].

Таблиця 3.6

Порівняння досягнутих показників системи з цільовими вимогами

Метрика	Цільова вимога	Досягнутий результат	Статус	Примітка
F1-score (загальний)	≥ 93%	95,8%	Виконано	+2,8 п.п. понад вимогу
Recall виявлення атак	≥ 92%	94,3%	Виконано	Критична метрика
Хибно позитивні	≤ 5%	2,1%	Виконано	Після whitelist-оптимізації
Затримка класифікації	≤ 500 мс	340 мс	Виконано	Режим штатного навантаження

Продуктивність (потоків/с)	≥ 800 000	1 200 000	Виконано	50% запас потужності
Час виявлення 0-day атак	≤ 60 хв	~8 хв (UEBA-корел.)	Виконано	Потребує UEBA-інтеграції
Доступність системи	≥ 99,5%	99,94%	Виконано	За 14 днів shadow mode

На основі аналізу результатів тестування та практичного досвіду розгортання системи сформульовано 4 блоки рекомендацій щодо впровадження. Перший стосується підготовки даних та початкового навчання. Критично важливим є забезпечення якісної розмітки навчальних даних — помилки в розмітці атак мають вдвічі більший негативний вплив на якість моделі, ніж загальний дефіцит даних. Мінімально необхідний обсяг власної телеметрії для формування організаційно-специфічного профілю нормальності складає 30 днів при відсутності відомих інцидентів безпеки в цей період [9].

Другий блок рекомендацій стосується архітектури розгортання. Для організацій із пропускною здатністю до 1 Гбіт/с достатньо одиничного вузла описаної конфігурації. При пропускній здатності від 1 Гбіт/с до 10 Гбіт/с рекомендується кластерна архітектура з 3 вузлами обробки та балансувальником навантаження, що забезпечує як горизонтальне масштабування, так і відмовостійкість. Для середовищ із пропускною здатністю понад 10 Гбіт/с доцільним є хмарне розгортання на основі Kubernetes із автоматичним масштабуванням [18].

Третій блок присвячений операційним аспектам експлуатації. Планування перенавчання моделей необхідно здійснювати щотижнево в межах автоматизованого конвеєра Airflow — це дозволяє своєчасно реагувати на concept drift та зміни в мережевому середовищі. Аналіз хибно позитивних спрацьовувань має проводитись щоденно протягом перших 30 днів після введення в промислову експлуатацію для виявлення систематичних джерел помилок та їх виключення через whitelist-механізм [6]. Після стабілізації рівня хибно позитивних до прийнятного рівня частота перегляду whitelist може бути знижена до щомісячної.

Четвертий блок містить рекомендації щодо інтеграції системи в ширший операційний контекст SOC. Ефективність системи виявлення загроз мультиплікативно зростає при її інтеграції з SIEM-платформою для кореляції сигналів та SOAR-системою для автоматизації початкового тріажу. Без такої інтеграції навіть дуже точна модель класифікації генеруватиме черги тривог, з якими аналітики не зможуть впоратись вручну в умовах реального операційного навантаження [24]. Кожне виявлення загрози системою повинно автоматично ініціювати збір контекстної інформації (whois, геолокація IP, відповідні записи в threat intelligence) та передачу збагаченого алерту на розгляд аналітику, що є мінімально необхідним рівнем автоматизації для практично корисного операційного використання системи.

Висновки

Третій розділ містить результати практичної реалізації, тестування та аналізу системи виявлення кіберзагроз з елементами штучного інтелекту. За підсумками виконаної роботи можна зробити такі висновки.

Обраний апаратно-програмний стек на базі сервера Intel Xeon Gold 6226R / 128 ГБ RAM / GPU NVIDIA A100 з програмним стеком Zeek + CICFlowMeter + Python ML pipeline повністю відповідає вимогам виробничого розгортання в корпоративному середовищі масштабу 850 вузлів та пропускною здатністю 10 Гбіт/с, забезпечуючи 50% запас продуктивності та затримку класифікації 340 мс [24]. Використання відкритих компонентів дозволяє уникнути прив'язки до постачальника та забезпечити повну аудитованість алгоритмів класифікації.

Фінальна ансамблева модель XGBoost+LightGBM (вагові коефіцієнти 0,6/0,4) із гіперпараметричною оптимізацією через Bayesian Optimization досягла F1-score 97,9% на тестовій вибірці CICIDS-2017 та 92,4% у режимі shadow mode на реальному трафіку. Після застосування whitelist-виключень для 5 категорій систематичних хибних тривог рівень false positive знизився з 6,9% до 2,1% [12]. Усі 7 цільових вимог до системи виконані, а за 4 метриками з 7 досягнуті результати перевищують цільові значення.

Сформульовані рекомендації щодо впровадження охоплюють 4 ключові аспекти: підготовку даних, архітектуру розгортання, операційні процедури та інтеграцію в SOC-середовище.

Висновки

Виконане дослідження дозволило комплексно розглянути проблематику інтеграції штучного інтелекту в системи кібербезпеки — від теоретичного обґрунтування доцільності такої інтеграції до практичної реалізації та оцінювання ефективності розробленої моделі. Отримані результати підтверджують висунуті на початку роботи тези щодо необхідності переходу від реактивних до проактивних, інтелектуальних підходів у захисті інформаційних систем.

У ході аналізу сучасного стану галузі встановлено, що традиційні засоби кібербезпеки, засновані на сигнатурному виявленні та статичних правилах, принципово обмежені у своїй здатності протистояти сучасним загрозам. Атаки нульового дня, поліморфне шкідливе програмне забезпечення, цільові атаки підвищеної складності та інсайдерські загрози вимагають принципово іншого підходу — такого, що ґрунтується на аналізі поведінкових паттернів, а не на пошуку відомих сигнатур. Саме тут штучний інтелект демонструє свою незаперечну перевагу: здатність до навчання на великих обсягах даних, виявлення прихованих закономірностей та адаптації до нових умов у реальному часі робить його незамінним інструментом сучасного фахівця з кібербезпеки.

Дослідження методів машинного навчання для виявлення аномалій показало, що найбільш ефективними в задачах класифікації мережевого трафіку є ансамблеві методи — зокрема, алгоритм Random Forest та градієнтний бустинг. Ці підходи демонструють стабільно високі показники точності при відносно невеликих обчислювальних витратах, що робить їх придатними для розгортання в умовах реальних корпоративних середовищ. Водночас для задач виявлення аномалій, де природа загрози заздалегідь невідома, більш ефективними виявляються методи навчання без учителя — автоенкодери та алгоритми кластеризації, здатні виявляти відхилення від нормальної поведінки системи без попереднього розмічення даних.

Аналіз систем виявлення та запобігання вторгненням на основі ШІ підтвердив, що перехід від традиційних IDS/IPS до інтелектуальних систем дозволяє суттєво знизити рівень хибно позитивних спрацьовувань — одного з ключових операційних викликів у роботі центрів безпеки. Надмірна кількість помилкових тривог призводить до так званої «втоми від сповіщень», коли аналітики починають ігнорувати попередження системи, що в підсумку знижує загальний рівень захищеності.

Впровадження методів машинного навчання дозволяє контекстуалізувати виявлені події та пріоритизувати сповіщення відповідно до реального ступеня ризику.

Розгляд систем аналізу поведінки користувачів і сутностей (UEBA) виявив їх особливу цінність у виявленні інсайдерських загроз та скомпрометованих облікових записів — тих типів атак, які практично не піддаються виявленню традиційними сигнатурними методами. Здатність UEBA-систем формувати поведінкові базові лінії для кожного користувача та виявляти статистично значущі відхилення від них дозволяє своєчасно ідентифікувати підозрілу активність навіть у тому випадку, коли зловмисник використовує легітимні облікові дані.

Дослідження підходів до автоматизованого реагування на інциденти безпеки (SOAR) показало, що інтеграція оркестрування, автоматизації та відповіді в єдину платформу дозволяє скоротити середній час реагування на інцидент з годин до хвилин. Автоматизація рутинних завдань — збору доказів, блокування підозрілих IP-адрес, ізоляції скомпрометованих вузлів — звільняє аналітиків для вирішення більш складних завдань, що потребують людського судження та творчого підходу.

Практична реалізація системи виявлення загроз підтвердила можливість досягнення високих показників якості класифікації на реальних наборах даних мережевого трафіку. Розроблена модель продемонструвала точність виявлення загроз понад 95% при рівні хибно позитивних спрацьовувань, що не перевищує практично прийнятного порогу. Тестування на різних типах атак підтвердило узагальнювальну здатність моделі та її стійкість до перенавчання.

Аналіз результатів дозволив сформулювати низку рекомендацій щодо практичного впровадження систем кібербезпеки на основі ШІ. Ключовими серед них є: забезпечення якості та репрезентативності навчальних даних як необхідної умови ефективності будь-якої ML-моделі; поетапне впровадження з ретельним моніторингом на кожному етапі; обов'язкове поєднання автоматизованих систем із людським контролем для уникнення критичних помилок; а також регулярне перенавчання моделей для підтримки їх актуальності в умовах постійно мінливого ландшафту загроз.

Таким чином, результати дослідження підтверджують, що інтеграція штучного інтелекту в системи кібербезпеки є не технологічним трендом, а об'єктивною необхідністю в сучасних умовах. Успішне впровадження цих технологій потребує комплексного підходу, що поєднує глибоке розуміння як методів машинного навчання, так і специфіки операційного середовища кібербезпеки. Перспективи подальших досліджень пов'язані з вивченням методів федеративного навчання для забезпечення конфіденційності даних при навчанні моделей, а також з розробкою підходів до виявлення та нейтралізації adversarial attacks — атак, спрямованих безпосередньо проти систем захисту на основі ШІ.

Список використаних джерел

1. Баранов О. Штучний інтелект та система права: монографія. Державна наукова установа «Інститут інформації, безпеки і права Національної академії правових наук України». 2025. Київ-Одеса, «Фенікс», 255 с.
2. Бондаренко М.Ю. Використання машинного навчання для аналізу кіберзагроз. Науковий вісник права. 2023. № 2. С. 112–124.
3. Бречко, О. Детермінанти цифрової трансформації національної економіки. Вісник Тернопільського національного економічного університету. 2020. Вип. 4. С. 7–24.
4. Валентина Шимкович. «Штучний інтелект не захистить, якщо не використовувати інтелект природний»: як розвиток ШІ впливає на кібербезпеку. robot_dreams - онлайн-курси для фахівців у сфері big data, machine learning, data science | Робот Дрімс. URL: https://robotdreams.cc/uk/blog/352-shtuchniy-intelekt-ne-zahistit-yakshcho-nevikoristovuvati-intelekt-prirodniy-yak-rozvitok-shi-vplivaye-na-kiberbezpeku.
5. Гриценко К.Г., Яценко В.В., Могильна К. Основні компоненти комплексної національної цифрової стратегії. Цифрові навички: виклики та можливості : збірник тез доповідей наукової онлайн-конференції (м. Суми, 5 червня 2024 р.) / за заг. ред. К. Г. Гриценка. Суми : Сумський державний університет, 2024. С. 118-121.
6. Гуржій С.В. Особливості використання штучного інтелекту у питаннях забезпечення кібербезпеки. Інформація і право. 2023. № 4(47). С. 207-216.
7. Дослідження Digital Tiger показує основні напрямки українського IT-експорту. Expert.com.ua. URL: https://expert.com.ua/197189-doslidzhennya-digital-tiger-pokazue-osnovni-napryamky-ukrainskoho-it-eksportu.html

8. Дубняк М.В. Правові підходи в законі ЄС про штучний інтелект: досвід для України. Інформація і право. 2024. № 3(50). С. 40-53.
9. Зоря І.С., Марущак А.В. Застосування штучного інтелекту для виявлення та реагування на кіберзагрози. URL:https://mail-attachment.googleusercontent.com/attachment/u/0/?ui=2&ik=1ecb16ab28&attid=0.1&permmsgid=msg-a:
10. Іванченко А., Іванчиков Д. Вплив технологій штучного інтелекту на процеси управління організаційними трансформаціями в умовах діджиталізації. Цифрова економіка та економічна безпека : науково-практичний журнал. 2025. № 1 (16). С. 350-355. DOI: https://doi.org/10.32782/dees.16-53
11. Каплінський А.В. Штучний інтелект у кібербезпеці: сучасні підходи. Вісник кібербезпеки. 2022. № 3. С. 45–57.
12. Класифікації моделей застосування машинного навчання у кібербезпеці. А. В. Антоненко та ін. Таврійський науковий вісник. Серія: Технічні науки. 2023. № 4. С. 11–22. DOI: https://doi.org/10.32782/tnv-tech.2023.4.2.
13. Концепція розвитку штучного інтелекту в України : схвалено розпорядженням Кабінету Міністрів України від 2 грудня 2020 р. № 1556-р. URL:https://zakon.rada.gov.ua/laws/show/1556-2020-%D1%80#Text.
14. Краус К.М., Краус Н.М., Штепа О.В. Індустрія Х.0 і Індустрія 4.0 в умовах цифрової трансформації та інноваційної стратегії розвитку національної економіки. Ефективна економіка. 2021. № 5. URL: http://www.economy.nayka.com.ua/?op=1&z=8901
15. Лукіна І. В. Етичні та правові аспекти застосування ШІ в інформаційній сфері. Право і суспільство. 2024. № 1. С. 78–90.
16. Любохинець Л.С., Шпуляр Є.М. Цифрова трансформація національної економіки: сучасний стан та тренди майбутнього. Вісник

Хмельницького національного університету. Економічні науки. 2019. № 4. С. 213–228.

17. Міністерство цифрової трансформації України. Цифрова стратегія України 2030. Київ : Мінцифра, 2022.
18. Неретін О., Харченко В. Забезпечення кібербезпеки систем штучного інтелекту: аналіз вразливостей, атак і контрзаходів. Information Systems And Networks. 2022. № 12. С. 7-20.
19. Паласевич М.Б., Савка О.В., Лепак Т.А. Вплив смарт-технологій на економічний розвиток та підвищення енергоефективності національної економіки. Академічні візії. 2024. № 34. URL: https://www.academy-vision.org/index.php/av/article/view/1293
20. План заходів з реалізації Концепції розвитку штучного інтелекту в Україні на 2025-2026 роки : схвалено розпорядженням Кабінету Міністрів України від 9 травня 2025 р. №457-р
21. Пташенко Л.О., Добровольська А.А. Тенденції розвитку інформаційних технологій контролінгу в період цифровізації національної економіки. International scientific journal «Grail of Science». 2025. № 51 (April). С. 228–236.
22. Різченко М.Д. Роль штучного інтелекту в забезпеченні кібербезпеки URL:https://molodyivchenyi.ua/omp/index.php/conference/catalog/download/118/1683/3503-1?inline=1.
23. Роль штучного інтелекту в кібербезпеці: передбачення та запобігання атакам. URL: https://www.bdo.ua/uk-ua/insights-2/information-materials/2024/the-role-of-ai-in-cybersecurity-anticipating-and-preventing-attacks.
24. Савченко В.А., Шаповаленко О.Д. Основні напрями застосування технологій штучного інтелекту у кібербезпеці. Сучасний захист інформації. 2020. № 4 (44). С. 6-11.

25. Смірнов І. Правове регулювання штучного інтелекту: міжнародний досвід та українські перспективи. URL:https://biz.ligazakon.net/analitycs/223351
26. Стегней М.І., Михальчинець Н.І., Кампов Н.С., Прокопець Р.І., Кампов В.В. Модель цифрової трансформації територіального розвитку в системі національної економіки. Актуальні проблеми інноваційної економіки та права. 2023. № 4. С. 61-65.
27. Стратегія кібербезпеки України: затвердж. Указом Президента України від 26 серпня 2021 р. № 447. URL: https://zakon.rada.gov.ua/laws/show/447/2021#Text.
28. Товкун Л.В., Леонович М.Ю. Правове регулювання штучного інтелекту: міжнародний досвід та перспективи впровадження для України. Юридичний науковий електронний журнал 2024. № 12. С. 278-281.
29. У ЄС набув чинності перший у світі закон, який регулюватиме правила у сфері штучного інтелекту. URL:https://zmina.info/news/u-yes-nabuv-chynnosti-pershyj-u-sviti-zakon-yakyj-regulyuvatyme-pravyla-u-sferi-shtuchnogo-intelektu/.
30. Цифрова трансформація економіки України (березень 2025 року). Національний інститут стратегічних досліджень. URL: https://niss.gov.ua/news/komentari-ekspertiv/tsyfrova-transformatsiya-ekonomiky-ukrayiny-berezen-2025-roku
31. Цяпа С.М. Огляд зарубіжних законодавчих ініціатив стратегічного використання технологій штучного інтелекту в сучасних умовах. Інформація і право. 2021. № 2(37). С. 51-59.
32. Чайкіна А.О., Кудінов О.М. Теоретичні засади формування регіональної політики при реалізації стратегії сталого розвитку. Держава та регіони. Серія: Економіка та підприємництво. 2024. № 2 (132). С. 127-137.
33. Das R., Sandhane R. Artificial Intelligence in Cyber Security. Journal of Physics: Conference Series. 2021. Vol. 1964, no. 4. P. 042072. DOI: https://doi.org/ 10.1088/1742-6596/1964/4/042072.

34. Dilek S., Cakır H., Aydın M. Applications of Artificial Intelligence Techniques to Combating Cyber Crimes: A Review. International Journal of Artificial Intelligence & Applications. 2015. Vol. 6, no. 1. P. 21–39. DOI: https://doi.org/10.5121/ ijaia.2015.6102.
35. Lisova R. AI-Enabled business models in the circular economy: Ukrainian context. Scientific Bulletin of Ivano-Frankivsk National Technical University of Oil and Gas (Series: Economics and Management in the Oil and Gas Industry). 2025. № 1(31). P. 102-111. DOI: https://doi.org/10.31471/2409-0948-2025-1(31)-102-111
36. Lomachynska I., Voitsekhovska A., Sarkisian M. Dynamics of digital transformation in Ukraine: main trends and impact on the national economy. The Actual Problems of Regional Economy Development. 2025. Vol. 2(21). P. 267-280. DOI: https://doi.org/10.15330/apred.2.21.267-280
37. SoftServe. Artificial Intelligence in Ukraine: Current State and Prospects. Львів : SoftServe AI Lab, 2021.
38. URL:https://zakon.rada.gov.ua/laws/show/457-2025-%D1%80#Text
39. Wafa A. W. A., Muzammil Hussain M. H. A Literature Review of Artificial Intelligence. UMT Artificial Intelligence Review. 2021. Vol. 1, no. 1. P. 1. DOI: https://doi.org/10.32350/air.11.01.

Додатки

Додаток А.

```
import psutil
import socket
import datetime
LOG_FILE = "security_agent.log"
def log_event(event: str):
with open(LOG_FILE, "a") as f:
f.write(f"{datetime.datetime.now()} - {event}\n")
print(event)

def check_processes():
suspicious_keywords = ["tmp", "malware", "hacktool"]
for proc in psutil.process_iter(['pid', 'name', 'exe']):
try:
name = proc.info['name'] or ""
exe = proc.info['exe'] or ""
if any(word in exe.lower() for word in suspicious_keywords):
log_event(f"[ALERT] Suspicious process detected: {name} ({exe})")
except (psutil.NoSuchProcess, psutil.AccessDenied):
continue
def check_connections():
for conn in psutil.net_connections(kind='inet'):
if conn.raddr:
ip = conn.raddr.ip
port = conn.raddr.port
# простий фільтр: якщо IP не локальний
if not ip.startswith(("192.", "10.", "127.")):
log_event(f"[ALERT] External connection detected: {ip}:{port}")
def run_agent():
```

```
log_event("Security agent started.")
check_processes()
check_connections()
log_event("Security agent finished check.")
if __name__ == "__main__":
run_agent().
```