КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.33.1286
УДК 004.056:004.89

Жданова Юлія Дмитрівна
кандидат фізико-математичних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID: 0000-0002-9277-4972
y.zhdanova@kubg.edu.ua

Шевченко Світлана Миколаївна
кандидат педагогічних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID: 0000-0002-9736-8623
s.shevchenko@kubg.edu.ua

Золотухіна Оксана Анатоліївна
кандидат технічних наук, доцент,
доцент кафедри інтелектуальних технологій
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0000-0002-3314-417X
oksana.zolotukhina@knu.ua

Негоденко Олена Василівна
кандидат технічних наук, доцент,
доцент кафедри комп'ютерних наук
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID: 0000-0001-6645-1566
o.nehodenko@kubg.edu.ua

ЗАСТОСУВАННЯ ПОЯСНЮВАНОГО ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОЦІНЮВАННЯ РИЗИКІВ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Анотація. У статті розглядаються сучасні підходи до аналізу та оцінювання ризиків інформаційної безпеки з акцентом на застосуванні методів штучного інтелекту (ШІ).
Проведено системний огляд класичних якісних, кількісних та гібридних методів ризик-менеджменту в умовах зростання складності кіберзагроз, динамічних змін у векторах атак та швидкої адаптації зловмисників. На основі аналізу наукової літератури обґрунтовано необхідність переходу від статичних процедур до адаптивних моделей, що спираються на об'єктивні дані та аналітику та забезпечують неперервний моніторинг, самонавчання та оперативне оновлення оцінок ризику. Розглянуто класи інтелектуальних методів – експертні інтелектуальні системи, ймовірнісно-статистичні моделі, нейромережеві підходи, гібридні ШІ-системи та системи поведінкового аналізу – їхні переваги, обмеження та сфери застосування у задачах виявлення аномалій, прогнозування інцидентів та автоматизації реагування. Окрему увагу приділено ролі пояснюваного ШІ (XAI) для підвищення прозорості прийняття рішень, можливості аудиту моделей та довіри з боку користувачів і регуляторів. Проаналізовано специфічні ризики, пов'язані з використанням ШІ у кібербезпеці, зокрема уразливості самих інтелектуальних систем до спеціалізованих атак, і запропоновано напрями їхнього пом'якшення через комбіновані технічні та організаційні заходи. Наведено рекомендації щодо інтеграції ШІ-компонентів у освітні програми підготовки фахівців з кібербезпеки, що включають формування компетенцій у сфері машинного навчання, інтерпретованості моделей та практик безпечного розгортання. На основі порівняльного аналізу запропоновано концептуальні положення для побудови адаптивного методу оцінювання ризиків, який поєднує автоматизоване виявлення загроз, ймовірнісну оцінку наслідків та механізми пояснюваності результатів для прийняття обґрунтованих управлінських рішень. Для верифікації методу розроблено п'ять прикладних сценаріїв, які дозволяють протестувати функціональну

спроможність методу при ідентифікації прихованих загроз, ранжуванні факторів впливу та використанні у навчальних кейсах. Практична значущість роботи полягає у формуванні методологічної бази для впровадження інтелектуальних систем управління ризиками в критично важливих інформаційних інфраструктурах та організаціях різного рівня, а також у визначенні пріоритетів подальших досліджень у сфері безпечного та прозорого застосування ШІ в кіберпросторі.

Ключові слова: інформаційна безпека; кібербезпека; штучний інтелект; пояснюваний ШІ; оцінювання ризиків; інтелектуальний аналіз ризиків; адаптивні моделі.

ВСТУП

Постановка проблеми. Сучасні інформаційно-комунікаційні системи є невід'ємною складовою діяльності організацій, підприємств і державних установ. Їх широке впровадження супроводжується зростанням кількості кіберзагроз, які стають дедалі складнішими та використовують нові механізми впливу на інформаційні ресурси. За таких умов підвищується значущість ефективного аналізу та оцінювання ризиків інформаційної безпеки як основи для прийняття обґрунтованих рішень щодо захисту інформаційних систем [1, 2]. Традиційні методи, що використовують переважно експертне оцінювання, статичні моделі та формалізовані процедури, не завжди забезпечують необхідний рівень точності та гнучкості у динамічному та невизначеному середовищі [3, 4]. Це спричиняє необхідність переходу від статичних до динамічних, адаптивних моделей управління інформаційною безпекою, здатних до неперервного аналізу даних, самонавчання та оперативного оновлення оцінок ризику. Використання таких моделей створює передумови для трансформації управління ризиками від реактивного до проактивного. Одним із перспективних напрямів розв'язання зазначеної проблеми є застосування методів штучного інтелекту (ШІ) для аналізу, оцінювання та управління ризиками інформаційної безпеки. Використання ШІ дозволяє автоматизувати процеси виявлення загроз, аналізу уразливостей та оцінювання ймовірності реалізації атак, що сприяє підвищенню ефективності систем захисту [5-7].

Водночас, впровадження ШІ породжує нові типи ризиків, пов'язаних із уразливістю самих інтелектуальних систем до спеціалізованих атак, таких як отруєння даних (data poisoning), інверсія моделі (model inversion) та змагальні атаки (adversarial attacks) [8, 9]. Це зумовлює необхідність розробки нових підходів до оцінювання ризиків, які враховують специфіку функціонування інтелектуальних систем.

Разом із тим, важливим завданням стає інтеграція сучасних інтелектуальних технологій у процес підготовки фахівців з кібербезпеки [10]. Освітні програми мають формувати компетентності, пов'язані із використанням штучного інтелекту для аналізу ризиків, обробки даних та прийняття обґрунтованих рішень. Використання інтелектуальних моделей у навчальному процесі сприяє розвитку аналітичного мислення, когнітивних навичок роботи з даними та глибшому розумінню природи сучасних кіберзагроз.

Таким чином, актуальною науковою задачею є розробка методу аналізу та оцінювання ризиків інформаційної безпеки на основі моделей штучного інтелекту, який забезпечить поєднання адаптивності, точності та інтерпретованості результатів.

Аналіз останніх досліджень і публікацій. Сучасні наукові дослідження у сфері інформаційної безпеки демонструють зростаючий інтерес до використання методів штучного інтелекту для аналізу, оцінювання та управління ризиками. Це зумовлено необхідністю підвищення ефективності систем захисту в умовах постійного ускладнення кіберзагроз [3, 5, 11]. В науковій літературі останніх років прослідковуються кілька взаємопов'язаних напрямів досліджень, які формують сучасний контекст інтеграції ШІ у процеси управління кіберризиками.

За першим напрямом значна увага приділяється розробці універсальних підходів до оцінювання ризиків на основі ШІ. Дослідження демонструють можливість використання ШІ для побудови інтегрованих систем управління ризиками, які охоплюють усі етапи – від виявлення загроз до реагування на інциденти [6]. Такі підходи націлені на те, щоби поєднати в реальному часі аналіз загроз з механізмом пріоритезації кіберінцидентів та системою прийняття рішень.

Другим напрямом сучасних досліджень є впровадження принципів пояснюваного ШІ (Explainable AI, XAI) у системи оцінювання ризиків. Пояснюваність моделей дозволяє забезпечити прозорість процесу прийняття рішень, підвищити довіру до результатів аналізу та забезпечити можливість аудиту ШІ-систем, що є особливо значущим для інформаційних систем критичної інфраструктури, зокрема державних установ та освітнього середовища. Низка наукових робіт пропонують гібридні моделі, що поєднують машинне навчання та пояснюваний штучний інтелект, що дозволяє підвищити одночасно точність і прозорість результатів оцінювання ризиків [12].

Третій напрям досліджень присвячений впливу генеративного штучного інтелекту на кібербезпеку. Дослідження в цьому напрямі зазначають, що генеративний ШІ значно підвищує як ефективність захисту (наприклад, автоматизація розслідувань, генерація правил виявлення), так і потенціал атак (створення фішингових повідомлень, автоматизоване генерування експлойтів), створюючи нові виклики для систем оцінювання ризиків [1, 2, 13]. Аналогічні висновки містяться у [7], де зазначається, що ШІ трансформує традиційні підходи до управління ризиками, переводячи їх у площину data-driven підходів до прийняття рішень. Наукові доробки [14, 15] підтримують ідею переходу до адаптивних моделей та розглядають ШІ як основу сучасних систем кіберзахисту.

Наукові роботи четвертого напряму присвячені розробці формалізованих моделей оцінювання ризиків на основі ШІ. Окремі дослідження пропонують використання та адаптацію ймовірнісних методів, баєсівських підходів до специфіки ШІ-систем [16] та спеціалізованих фреймворків, що враховують уразливість моделей до нових типів атак та специфіку поведінкових характеристик загроз [8, 9].

Дослідження п’ятого напряму приділяють значну увагу практичному застосуванню штучного інтелекту у кібербезпеці. В [17] окреслюється роль ШІ як перспективного, але ще не сформованого інструмента кібербезпеки. В [18, 19] досліджуються питання використання нейромережевих моделей для оцінювання ризиків, виявлення аномалій та прогнозування кіберзагроз. Застосування ШІ у захисті критичної інфраструктури та бізнес-середовища досліджуються у [20, 21, 22], де розглядаються технічні та організаційні підходи до впровадження інтелектуальних технологій у системи кіберзахисту.

Окрему групу становлять праці, що вивчають ризики, пов’язані із самим використанням ШІ-технологій. Ці роботи вказують на те, що створюючи нові можливості, ШІ створює водночас нові загрози і ризики, що потребує нових методів оцінювання ризиків інформаційної безпеки [23, 24, 25]. Важливо відзначити, що разом із динамічним поступом штучного інтелекту мають синхронно змінюватися й освітньо-професійні програми підготовки фахівців з кібербезпеки та захисту інформації [26]. Від таких фахівців роботодавці вимагають володіти компетенціями як у сфері машинного навчання, так і у безпечному застосуванні технологій ШІ. Це потрібно не лише для захисту інформаційних систем, що працюють на основі ШІ, але й для використання ШІ-інструментів для виявлення ризиків та нейтралізації загроз. Виявилося, що такі вимоги виявляють істотну розбіжність між результатами навчання в університеті та запитами сучасного ринку праці. Для подолання цієї розбіжності потрібна системна інтеграція ШІ в освітні програми [27], причому, як стверджують автори [28], починати треба з середньої школи.

У цілому, аналіз наукових публікацій дозволяє виділити наступні ключові напрямки сучасних досліджень такі, як розвиток динамічних та адаптивних моделей оцінювання ризиків; інтеграція штучного інтелекту у процеси управління кіберризиками; врахування нових типів загроз, пов’язаних із використанням ШІ; підвищення точності та автоматизації аналізу та оцінювання ризиків. Огляд наукової літератури засвідчує високий ступінь розробленості загальних аспектів застосування ШІ в кібербезпеці, що обґрунтовує доцільність даного дослідження як такого, що доповнює попередні розвідки внаслідок інтеграції принципів адаптивності, інтерпретованості та валідації на реальних наборах даних.

Мета статті. Метою статті є розробка та обґрунтування концептуальних засад пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки, який поєднує адаптивні моделі штучного інтелекту для автоматизації аналізу загроз із механізмами інтерпретованості результатів для підтримки прийняття управлінських рішень.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Інформаційно-комунікаційні системи забезпечують процеси збору, обробки, зберігання та передавання даних у більшості сфер діяльності людини, бізнесу та державного управління. Розширення їх функціональних можливостей, поширення хмарних сервісів, технологій Інтернету речей і засобів дистанційної взаємодії супроводжується зростанням залежності суспільства від цифрової інфраструктури. За таких умов зростає і вплив кіберінцидентів на безперервність функціонування систем, збереження інформаційних ресурсів та захист конфіденційних даних, що актуалізує завдання аналізу й оцінювання ризиків інформаційної безпеки. У міжнародних стандартах ISO/IEC 27005, NIST RMF та інших підходах ризик інформаційної безпеки розглядається як поєднання ймовірності реалізації загрози та величини потенційних збитків. Відповідно, процес оцінювання ризиків ґрунтується на аналізі взаємозв’язків між інформаційними активами, уразливостями, загрозами та можливими наслідками інцидентів безпеки. Класичні підходи до оцінювання ризиків інформаційної безпеки базуються на визначенні взаємозв’язку між загрозами, уразливостями, інформаційними активами та можливими наслідками реалізації кіберзагроз (рис.1).

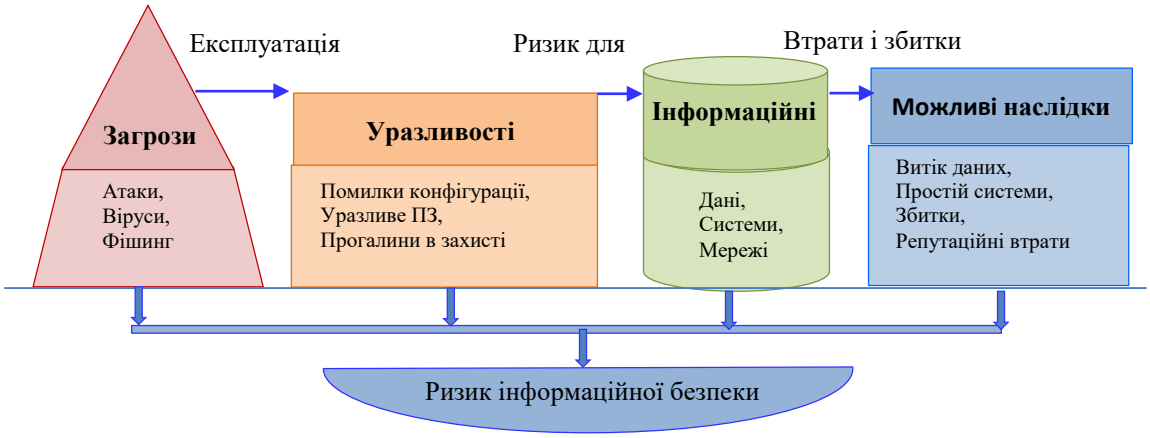

Рис.1. Концептуальна модель взаємозв'язку компонентів оцінювання ризиків

Класичні методи оцінювання ризиків інформаційної безпеки умовно поділяються на якісні, кількісні та гібридні, стислий порівняльний аналіз яких представлений у таблиці 1.

Таблиця 1

Порівняльний аналіз класичних методи оцінювання ризиків інформаційної безпеки

Методи	Зміст	Застосування	Переваги	Недоліки
Якісні	Ґрунтуються на суб'єктивних експертних оцінках та досвіді фахівців. Рівень ризику визначається за допомогою лінгвістичних або бальних шкал без точних математичних розрахунків.	Попереднє оцінювання безпеки, первинне ранжування ризиків.	-Прості в організації та швидкі у виконанні. -Не потребують точної статистики інцидентів.	-Високий рівень суб'єктивності оцінок. -Неможливо точно розрахувати фінансову доцільність витрат на безпеку.
Кількісні	Орієнтовані на отримання точних числових та грошових показників. Рівень ризику розраховується за допомогою теоретико-ймовірнісних та статистичних методів.	Планування бюджетних витрат на управління кіберризиками.	-Результати виражені в конкретних грошових одиницях. -Дозволяють чітко й об'єктивно обґрунтувати бюджетні витрати перед керівництвом.	-Висока трудомісткість та вартість проведення. -Потребують великої кількості історичних даних та статистики інцидентів.
Гібридні	Поєднують обидва підходи: якісні оцінки експертів трансформуються у числові коефіцієнти, бали або індекси за чітко визначеними математичними правилами.	Аудит кібербезпеки, оцінка захищеності.	-Гнучкі (працюють за дефіциту фінансових даних). -Менш суб'єктивні за рахунок математичної формалізації шкал.	-Результати є умовними індексами, а не реальними фінансовими втратами. -Складність правильного налаштування бальних шкал.

Серед якісних методів, що набули найбільшого поширення, можна виділити експертне опитування, анкетування, складання матриці ризиків, SWOT-аналіз [29]; серед кількісних – розрахунок показників ALE та SLE, моделювання методом Монте-Карло, застосування баєсівських мереж [30], графів атак, дерев атак; серед гібридних – оцінювання уразливостей за допомогою загальної системи

CVSS (Common Vulnerability Scoring System), оцінювання на основі нечіткої логіки, зокрема за допомогою нечітких когнітивних карт [31].

Збільшення числа кіберінцидентів, поява складних багатовекторних атак та активне застосування штучного інтелекту в кіберпросторі виявили неефективність класичних статичних моделей захисту. Сучасні загрози відрізняються такими рисами як: швидка модифікація, адаптивна поведінка, автоматизація атак, використання ШІ для обходу систем захисту.

Застосування технологій ШІ у сфері інформаційної та кібербезпеки суттєво трансформувало підходи до захисту цифрових систем. Якщо традиційні методи аналізу даних ґрунтувалися на статичних правилах та ручній обробці, то сучасні алгоритми ШІ забезпечують можливість обробки гігабайтів логів і мережевого трафіку за мілісекунди. Це дозволяє не лише оперативно виявляти відомі загрози, але й прогнозувати появу нових ще до того, як вони завдадуть шкоди інформаційним ресурсам.

До найбільш поширених моделей штучного інтелекту, що застосовуються у задачах інформаційної та кібербезпеки, можна віднести [32]:

1. Моделі для виявлення аномалій та загроз – класичне машинне навчання (Machine Learning). Ці моделі становлять основу сучасних систем керування інформацією та подіями безпеки (SIEM) і систем виявлення мережевих загроз (NDR). Вони призначені для класифікації трафіку, виявлення відомих і невідомих загроз та швидкої фільтрації великого обсягу подій. Прикладами є: дерева рішень (Decision Trees) та випадковий ліс (Random Forest), що використовуються для класифікація трафіку та виявлення шкідливого програмного забезпечення; працюють як ансамблеві класифікатори за набором ознак; метод опорних векторів (SVM) та ізоляційний ліс (Isolation Forest), що використовуються для виявлення невідомих загроз і аномалій.

2. Моделі глибокого навчання (Deep Learning) використовуються для аналізу складних, високовимірних або часових даних, де традиційні методи показують обмежену ефективність. Прикладами є: рекурентні нейронні мережі (RNN), що застосовуються для аналізу часових рядів і системних логів для виявлення підозрілих послідовностей подій; згорткові нейронні мережі (CNN), що застосовуються для аналізу бінарних артефактів і структури трафіку через представлення даних у матричному вигляді.

3. Генеративні моделі та великі мовні моделі (LLM) використовуються для автоматизація рутинних завдань аналітиків безпеки, прискорення розслідувань інцидентів та підтримки при аналізі коду.

4. Графові нейронні мережі (Graph Neural Networks, GNN) використовуються для моделювання та аналізу складних, розподілених атак і інфраструктури зловмисників через представлення мережі як графа для виявлення багатокрокових атак і латерального руху.

Саме розвиток технологій штучного інтелекту започаткував новий напрям у сфері кібербезпеки – інтелектуальний аналіз ризиків (Intelligent Risk Analysis, IRA). Основу такого підходу становлять методи машинного та глибинного навчання та аналіз великих даних – усі ці методи сприяють появі нових можливостей для підвищення ефективності захисту інформаційних систем. Їх застосування дозволяє не тільки автоматизувати процес виявлення загроз, аналізувати аномальні сценарії поведінки систем та оцінювати рівень ризику у режимі реального часу, але й здійснювати прогнозування кіберінцидентів та адаптувати механізми захисту до змін умов функціонування [33].

Таким чином, інтеграція інтелектуальних технологій у процеси управління ризиками формує підґрунтя для створення більш гнучких, точних та адаптивних систем кіберзахисту.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Порівняльний аналіз інтелектуальних методів оцінювання ризиків інформаційної безпеки. З огляду на результати проведеного аналізу сучасних підходів до оцінювання ризиків інформаційної безпеки можна зробити висновок, що методи, які спираються на експертні оцінки і статичні правила, виявляються обмеженими при роботі з великими обсягами гетерогенних даних, векторами атак, що швидко змінюються, і складними взаємозв'язками між компонентами інфраструктури. Подолати зазначені обмеження дозволяє перехід від статичного, періодичного оцінювання ризиків до інтелектуального аналізу, тобто до динамічного моделювання загроз в реальному часі за допомогою моделей ШІ.

Серед сучасних інтелектуальних методів оцінювання ризиків можна умовно виокремити наступні основні класи:

- експертні інтелектуальні системи;
- ймовірнісно-статистичні моделі;
- нейромережеві методи;

– гібридні ШІ-системи;
– системи поведінкового аналізу.

Результати порівняльного аналізу цих методів оцінювання ризиків наведено у таблиці 2.

Таблиця 2

Порівняльний аналіз інтелектуальних методів оцінювання ризиків інформаційної безпеки

Методи	Зміст	Застосування	Переваги	Недоліки
Експертні інтелектуальні системи	Використання евристичних правил та накопиченого досвіду фахівців у сфері інформаційної безпеки.	-Класичний ризик-менеджмент у стабільному середовищі; -оцінювання відповідності стандартам.	-Висока точність та об'єктивність; -результативність навіть в умовах дефіциту даних; -моделювання сценаріїв; -накопичення та збереження знань.	-Висока залежність від людського фактора; -складність оновлення бази знань; -низька адаптивність до нових типів атак
Ймовірнісно-статистичні моделі	Формалізація процесу оцінювання ризиків і визначення ймовірність реалізації загроз.	-Аналіз ризиків відмов систем, оцінка ймовірності кіберзагроз, -оцінка ймовірності та збитків від витоків даних.	-Математичне моделювання процесу оцінювання ризиків; -можливість кількісного оцінювання ризику; -обґрунтованість результату.	-Потребують значних обсягів апріорних даних; -складно адаптуються до нестабільних середовищ; -неефективні для аналізу великих потоків неструктурованих даних.
Нейромережеві методи	Виявлення, класифікація, кількісне вимірювання ймовірностей ризиків та прогнозування ймовірності їх настання.	-Класифікації кіберзагроз; -виявлення аномалій; -прогнозування атак.	-Ефективність при роботі з великими даними; -висока точність; -здатність до самоадаптації.	-Висока обчислювальна складність; -відсутність пояснюваності; -складність інтеграції в системи підтримки прийняття рішень.
Гібридні ШІ-методи	Поєднання класичних експертних систем та математичних моделей з алгоритмами ШІ.	-Виявлення складних цілеспрямованих кібератак; -динамічне оцінювання загроз.	-Здатність одночасної обробки числових даних та експертних оцінок; -висока точність та швидкодія; -здатність до самоадаптації.	-Складність впровадження; -спрямованість переважно на виявлення загроз; -не забезпечують комплексного оцінювання ризиків.
Системи поведінкового аналізу	Виявлення аномалій у роботі пристроїв та у діях користувачів.	-Викриття інсайдерських загроз; -розпізнавання прихованих кібератак; -виявлення скомпрометованих облікових записів.	-Ефективність при роботі з внутрішніми порушниками; -здатність до самоадаптації.	-Складність впровадження; -велика кількість хибних реагувань; -можливість виникнення правових питань.

Кожен з розглянутих класів інтелектуальних методів має свій математичний та логічний фундамент, який обумовлює ефективність застосування в конкретних умовах. Відмінності між цими методами визначаються в основному вимогами до обсягу та якості вхідних даних, стохастичністю середовища, а також інтерпретованістю результатів.

Зазначимо, що в практиці інтелектуального аналізу та оцінювання ризиків розглянуті вище методи застосовуються не відокремлено, а потоково, неперервно змінюючи один одного. Наприклад, якщо системою поведінкового аналізу зафіксовано аномалію у діях користувача, то далі ця аномалія аналізується нейромережевим методом, а ймовірнісно-статистична модель (зокрема, баєсівська мережа) оцінює ймовірність ризику для всієї мережі, наприкінці експертна система приймає рішення про блокування користувача.

Проблема прозорості та інтерпретованості рішень, отриманих за допомогою інтелектуальних методів оцінювання ризиків інформаційної безпеки.

У контексті інтелектуальних методів оцінювання ризиків прозорість означає розуміння користувачем логічних та алгоритмічних механізмів прийняття рішень системою в цілому, а інтерпретованість характеризує здатність системи пояснити своє рішення у зрозумілих для людини термінах.

Саме проблема прозорості та інтерпретованості рішень штучного інтелекту є одною з найважливіших у задачах інтелектуального оцінювання ризиків інформаційної безпеки, адже неправильне рішення може привести до фінансових збитків, витоку конфіденційних даних або порушень у функціонуванні критичної інфраструктури.

Рівень прозорості основних класів інтелектуальних методів оцінювання ризиків суттєво різний (рис. 2). Так, експерті системи є абсолютно прозорими (білий ящик), ймовірнісно-статистичні моделі прозорі з математичної точки зору, системи поведінкового аналізу мають середню прозорість (сірий ящик), а гібридні та нейромережеві методи мають дуже низьку, майже нульову прозорість (чорний ящик).

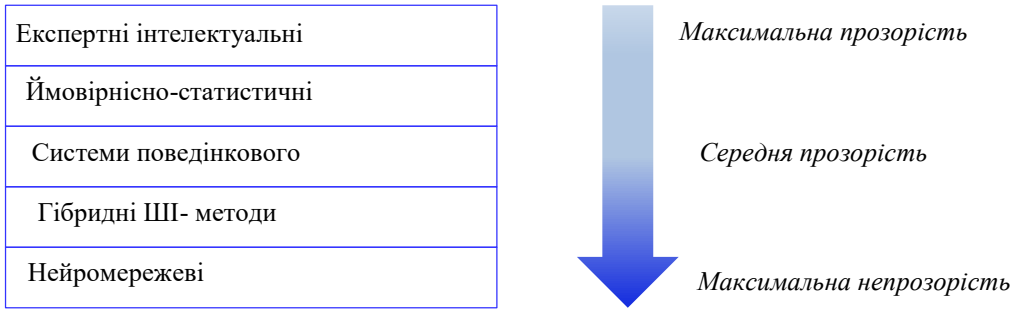

Рис 2. Рівні прозорості інтелектуальних методів оцінювання ризиків

Низький рівень прозорості викликає низку проблем, серед яких:

– хибні спрацьовування**,** коли ШІ постійно підвищує рівень ризику без пояснення причин, що приводить до ігнорування сповіщень про дійсно важливі інциденти;

– юридична та аудиторська невизначеність, коли відсутність пояснень ускладнює обґрунтування рішень відповідно до міжнародних стандартів (ISO/IEC 27001) та регуляторних актів (зокрема, Європейський AI Act і GDPR);

– уразливість до змагальних атак, коли мінімальні модифікації вхідних даних зловмисниками можуть привести до помилок класифікації подій.

Для вирішення цих проблем у сучасні гібридні та нейромережеві методи оцінювання для інтерпретації результатів впроваджується пояснюваний ШІ, за допомогою якого внутрішні стани та механізми прийняття рішень ШІ перекладаються на мову, зрозумілу для фахівця з кібербезпеки і для особи, що приймає рішення.

На практиці для цього використовують методи post-hoc пояснення, серед яких основними є SHAP (SHapley Additive exPlanations) на основі значень Шеплі з теорії ігор для кількісного розподілу внеску кожної ознаки у фінальне рішення та LIME (Local Interpretable Model-agnostic Explanations), що будує просту сурогатну модель (наприклад, лінійну або дерево рішень) для пояснення рішення в окремій точці простору ознак.

Загальна концепція пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

Основна ідея пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки полягає в тому, щоб використовувати алгоритми ШІ для автоматизації процесів обробки безпекових подій, підвищення чутливості до аномалій та прогнозування розвитку ризику, а також застосовувати пояснюваний ШІ для інтерпретованості та обґрунтованості рішень.Робочими процесами моделі методу є:

- автоматизований аналіз подій інформаційної безпеки;
- виявлення аномальної поведінки;
- прогнозування потенційних кіберзагроз;
- формування інтегральної оцінки ризику.

Пояснюваний інтелектуальний підхід до оцінювання ризиків інформаційної безпеки передбачає інтеграцію моделей ШІ різної природи (класифікаційних, часових, генеративних, графових та пояснюваних) в єдину аналітичну платформу з метою підвищення точності виявлення загроз, адаптивності до нових сценаріїв атак та обґрунтованого пріоритезування ризиків.

Метод орієнтований на впровадження в таких класах інформаційних систем як корпоративні інформаційні системи, системи моніторингу безпеки, SIEM-платформи, системи критичної інфраструктури, а також навчальні та дослідницькі середовища. Така спрямованість забезпечує як практичну застосовність у виробничих умовах, так і можливість подальшого наукового дослідження та валідації підходу.

Структурна модель пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

Структурно пояснюваний інтелектуальний метод оцінювання ризиків інформаційної безпеки може бути представлений як багаторівнева система, що складається з трьох взаємопов’язаних шарів (рис. 2).

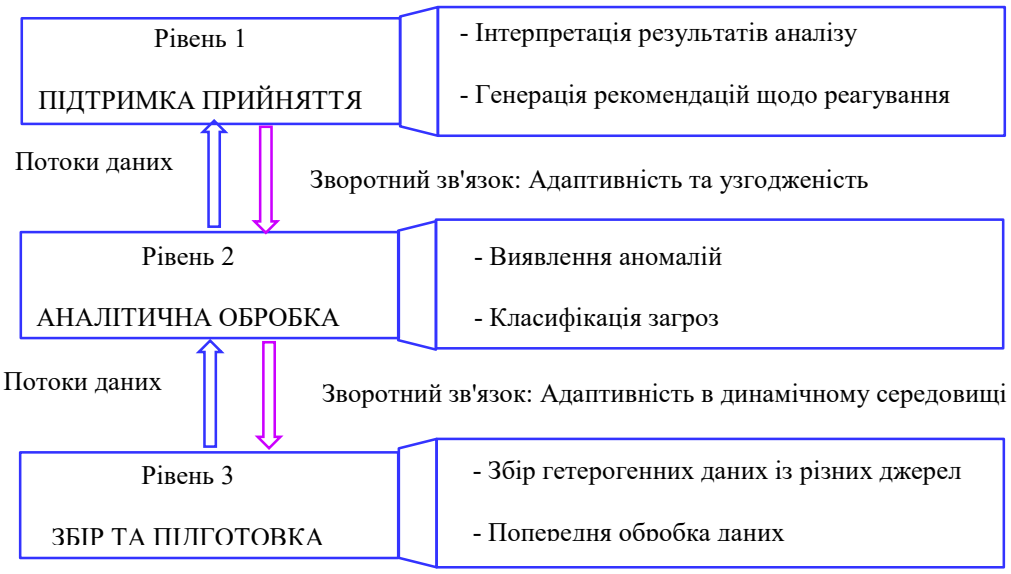

Рис.3. Структура моделі пояснюваного інтелектуального методу оцінювання ризиків

Взаємодія рівнів у структурі є системною, ієрархічною та динамічною. Принципи функціонування системи полягають у послідовній трансформації інформації, детермінованому міжрівневому обміні та неперервній адаптації до змін зовнішнього середовища.

Процес інтелектуального аналізу та управління ризиками реалізується через два ключові засоби взаємодії – механізми прямого і зворотного зв’язку.

Прямий зв’язок створюють потоки даних та метаданих під час руху знизу вгору. Очищені та перероблені первинні дані та контекстні метадані з рівня 3 надходять на рівень 2 у вигляді структурованих потоків подій безпеки, що є вхідними векторами параметрів для математичних моделей та алгоритмів ШІ. Результати інтелектуальної обробки (аналітичні висновки ШІ щодо оцінки ризиків) з рівня 2 передаються на рівень 1, де формують простір для прийняття рішень. Для підтримки прийняття рішень передбачено додатковий шар пояснення з використанням елементів пояснюваного ШІ, який надає рішенням прозорості та обґрунтованості.

Зворотний зв'язок забезпечує адаптивність й узгодженість системи в умовах динамічного загрозливого середовища. На основі прийнятих на рівні 1 рішень здійснюються керуючі впливи на рівень 2. Під дією цих впливів відбувається донавчання моделей ШІ, зміна правил класифікації загроз та коригування вагових коефіцієнтів критеріїв оцінювання ризиків. Результати змін на рівні 2 дозволяють коригувати правила фільтрації на рівні 3.

Таким чином, межі між рівнями структурної моделі методу інтелектуального оцінювання ризиків не є жорсткими, а діють як динамічні інтерфейси обміну.

Формальний опис пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

Нехай $X = \{x_1, x_2, \dots, x_n\}$ – множина вхідних даних, де через x_i , $i = 1,2, \dots, n$, позначені параметри системи (уразливості, активи, події безпеки тощо).

Модель M оцінювання ризику складається з трьох основних модулів:

– модуль виявлення аномалій – класифікатор $f_1(X)$, який визначає ймовірність появи загрози P_t ;

– модуль оцінювання наслідків – функція $f_2(X)$, що обчислює потенційні збитки L_t на основі історичних даних і вагових коефіцієнтів активів;

– модуль пояснюваності – механізм $f_3(M, X)$, який генерує інтерпретацію результатів через методи пояснюваного ШІ (наприклад, SHAP або LIME).

В конкретний момент часу t динамічна зважена модель інтегрального ризику фокусується на балансі між ймовірністю загрози та потенційними збитками і визначається за формулою:

$$R_t = w_1 \cdot P_t + w_2 \cdot L_t$$

де w_1, w_2 – вагові коефіцієнти, що відображають пріоритети захисту безпеки.

Модель M оновлює вагові коефіцієнти за допомогою механізму адаптації:

$$w_i^{(t+1)} = w_i^{(t)} + \eta \cdot \Delta w_i,$$

де η – швидкість навчання, Δw_i – зміна ваги на основі нових даних.

Модель M генерує:

– числову оцінку динамічного показнику ризику R_t;

– пояснення E_t у вигляді набору ключових факторів, що вплинули на оцінку;

– рекомендації щодо реагування (блокування, аудит, посилення контролю доступу, тощо).

Неперервне навчання моделі M реалізується наступним механізмом на основі нових кіберінцидентів:

$$M^{(t+1)} = M^{(t)} + \lambda \cdot \nabla L\left(M^{(t)}, X_t\right),$$

де λ – коефіцієнт оновлення, ∇L – градієнт функції втрат.

Як засіб оцінювання ефективності моделі можна використати метрики Precision, Recall для оцінки якості класифікації та індекс інтерпретованості $I_X(AI)$ для виявлення частки рішень ШІ, зрозумілих для людини.

Етапи пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

Пояснюваний інтелектуальний метод реалізується послідовністю взаємопов'язаних етапів, що забезпечують збір, аналіз та оцінювання ризиків інформаційної безпеки (таблиця 3).

Таблиця 3

Етапи пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки

Етап	Назва	Зміст
1	Збір та попередня обробка даних.	Акумуляція гетерогенних джерел, очищення, нормалізація та агрегація даних.
2	Формування множини параметрів ризику.	Визначення та інженерія ознак, що відображають уразливості й контекст системи.
3	Аналіз поведінкових характеристик системи.	Побудова моделей нормальної поведінки на основі часових рядів і взаємодій компонентів.
4	Виявлення аномалій із застосуванням ШІ-моделі.	Виявлення відхилень від норми за допомогою алгоритмів машинного навчання.

5	Оцінювання ймовірності реалізації загрози.	Кількісна інтерпретація виявлень для визначення ймовірності експлуатації індикаторів.
6	Обчислення інтегрального показника ризику.	Агрегація компонентних оцінок у єдину метрику з урахуванням ваг і критичності ресурсів.
7	Формування рекомендацій щодо реагування із застосуванням моделі пояснюваного ШІ.	Генерація пріоритетних заходів реагування та пом'якшення ризиків на основі оцінок.

На рис. 4 представлено діаграму послідовності одного циклу роботи пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

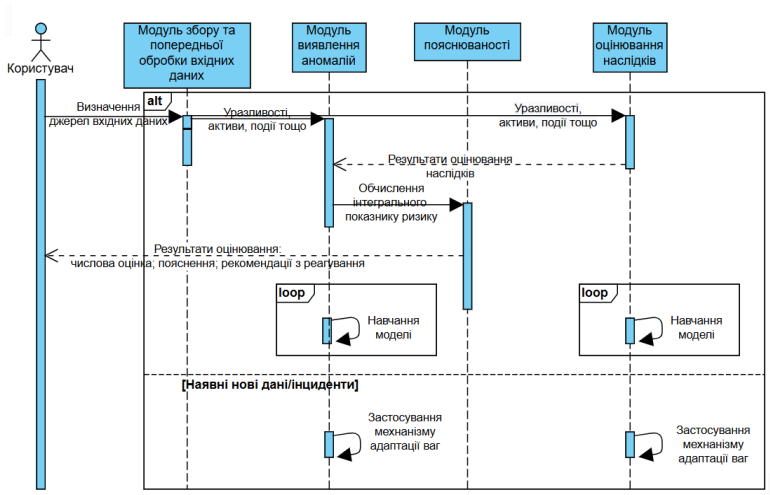

Рис.4. Діаграма послідовності одного циклу роботи пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки

Збір вхідних даних відбувається з різних джерел інформаційної безпеки, якими є параметри мережевого трафіку, журнали подій, результати сканування уразливостей, дані SIEM-систем, індикатори компрометації (IoC), а також характеристики поведінки користувачів. Зібрані дані піддаються первинній обробці, яка полягає в очищенні, нормалізації, усуненні дублікатів, зведенні даних про події.

Після обробки формується n-елементна множина параметрів ризику $X = \{x_1, x_2, \dots, x_n\}$, де через x_i позначено параметр стану системи. До основних параметрів зазвичай відносять кількість аномальних подій; інтенсивність мережевого трафіку; частоту помилок автентифікації; рівень критичності уразливостей; відхилення у поведінці користувачів.

Елементи множини X подаються на вхід функції f ШІ-моделі для виявлення аномалій та загроз. За таку ШІ-модель доцільно взяти модель машинного навчання типу автокодувальника (autoencoder), яка дозволяє визначати аномальні стани системи на основі відхилень від нормальної поведінки. В цьому випадку вихідний вектор відновлених параметрів $\tilde{X}$ визначається наступним чином:

$$\tilde{X} = f(X)$$

де X – вхідний вектор параметрів ризику, $f(X)$ – функція ШІ-моделі.

Якщо похибка відновлення $|X - \tilde{X}|$ не перевищує встановленої межі, подія класифікується як нормальна, в противному випадку – як аномальна.

На основі аналізу історичних даних та поточного стану системи ШІ-модель оцінює ймовірність реалізації загрози.

З урахуванням ймовірності реалізації загрози, рівня уразливості, критичності активу та ступеня аномальності поведінки розраховується інтегральний показник ризику R за формулою:

$$R = P \cdot V \cdot C \cdot A\ ,$$

де P – ймовірність реалізації загрози; V – рівень уразливості; C – критичність активу; A – коефіцієнт аномальності.

Нормалізований показник ризику R_{norm} розраховується за формулою

$$R_{norm} = \frac{R - R_{min}}{R_{max} - R_{min}}$$

і використовується для класифікації рівнів небезпеки від низького до критичного.

Виходячи з отриманого рівня ризику формуються рекомендації щодо реагування, спрямовані на зниження ймовірності реалізації загроз та мінімізацію їхнього впливу. Заходами реагування є блокування підозрілої активності, посилення контролю доступу, застосування додаткових механізмів автентифікації, ізоляція окремих вузлів мережі, а також ініціювання аудиту безпеки для комплексної перевірки стану системи.

У процесі оцінювання ризиків доцільним є застосування моделі пояснюваного ШІ (SHAP або LIME), використання якої допомагає з'ясувати причини формування інтегральної оцінки ризику, визначити параметри, які найбільше вплинули на результат і загалом підвищити довіру до ШІ-моделі.

Практичне та освітнє застосування пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки.

З практичної точки зору впровадження пояснюваного інтелектуального методу оцінювання у сферу управління ризиками інформаційної безпеки дозволяє вирішити з найкращими результатами наступні критичні завдання:

- здійснити валідацію та верифікацію рішень, тобто забезпечити можливість виявлення випадкових або хибних кореляцій у даних і підтвердження коректності результатів;
- оптимізувати витрати на захист за рахунок пріоритезації заходів шляхом кількісного ранжування чинників ризику;
- забезпечити комплаєнс шляхом обґрунтування рішень для аудиту та відповідності нормативним вимогам;
- збільшити стійкість до змагальних атак завдяки виявленню аномалій у логіці прийняття рішень та підвищення захищеності від отруєння даних.

Пояснюваний інтелектуальний метод може бути використаний не лише у практичних системах кібербезпеки, але й у навчальному процесі підготовки фахівців з кібербезпеки.

Великою цінністю пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки є його освітнє застосування. Воно передбачає:

- аналіз поведінкових даних з метою вивчення патернів нормальних та аномальних дій;
- дослідження аномальної активності у контрольованому середовищі;
- моделювання кіберінцидентів для відпрацювання сценаріїв реагування;
- навчання роботі з ШІ-моделями, включно з інтерпретацією результатів за допомогою пояснюваного ШІ;
- формування навичок оцінювання ризиків та прийняття обґрунтованих рішень.

Введення методу в освітній процес дозволить поєднати теоретичну підготовку з потужною практичною складовою, сформувати міждисциплінарні компетентності ШІ та кібербезпеки. Набуття практичних навичок аналізу, оцінювання, інтерпретації та прийняття обґрунтованих рішень щодо ризиків інформаційної безпеки значно підвищать якість підготовки майбутніх фахівців із інформаційної безпеки.

Планування сценаріїв експериментальних досліджень. Для визначення напрямів подальшої апробації запропонованого пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки доцільно сформувати набір типових сценаріїв його застосування. Такі сценарії охоплюють основні аспекти функціонування методу, зокрема оцінювання ризиків, виявлення аномалій, аналіз впливу факторів ризику, забезпечення стійкості до спотворення даних та використання в освітньому процесі і можуть бути використані іншими науковцями як основа для подальших експериментальних досліджень.

Сценарій 1. Оцінювання ризиків на основі багатоджерельних даних.

Мета сценарію – перевірка працездатності всіх етапів запропонованого пояснюваного інтелектуального методу оцінювання ризиків інформаційної безпеки в умовах використання різнорідних джерел даних.

Вхідні дані – параметри мережевого трафіку, журнали подій, результати сканування уразливостей, індикатори компрометації та повідомлення систем виявлення вторгнень.

Як інструмент виявлення аномалій пропонується використовувати автокодувальник, який формує оцінку реконструкційної помилки як показник відхилення поточного стану системи від нормального. Отримані значення разом із характеристиками уразливостей, критичністю активів та ймовірністю реалізації загроз використовуються моделлю оцінювання ризику. Для пояснення результатів

пропонується застосовувати методи SHAP або LIME, які дозволяють визначити внесок окремих факторів у підсумкову оцінку.

Вихідні дані сценарію – інтегральний показник ризику, категорія ризику (низький, середній, високий, критичний) та пояснення причин отриманої оцінки із зазначенням найбільш впливових факторів.

Успішність виконання сценарію пропонується оцінювати за такими критеріями:

– точність класифікації рівнів ризику відносно експертних оцінок або еталонних даних (Accuracy, Precision, Recall, F1-score), при цьому цільовими значеннями можуть вважатися Accuracy ≥ 0,90 та F1-score ≥ 0,85;

– середнє відхилення інтегральної оцінки ризику при додаванні нових джерел інформації (цільове значення не повинно перевищувати 5–10 % за відсутності суттєвих змін у стані системи);

– частка пояснень моделі, підтверджених експертами як коректні та релевантні (цільове значення на рівні не менше 80 %);

– рівень узгодженості пояснень з експертними висновками, визначений за коефіцієнтом узгодженості (наприклад, коефіцієнтом Каппа Коена, для якого бажаним є значення $k \geq 0{,}70$, що відповідає високому рівню узгодженості);

– частка коректно виявлених ризикових станів, відсутніх у навчальній вибірці (True Positive Rate для нових або раніше невідомих сценаріїв, цільове значення повинно становити не менше 80 %).

Крім того, доцільно оцінювати частку хибнопозитивних спрацювань (False Positive Rate), цільове значення якої не повинно перевищувати 10 %, що дозволить забезпечити практичну придатність методу для використання в реальних системах управління ризиками інформаційної безпеки.

Сценарій 2. Виявлення аномальних станів інформаційної системи.

Мета сценарію – перевірка здатності запропонованого пояснюваного інтелектуального методу своєчасно виявляти аномальні стани інформаційної системи та відокремлювати їх від нормальної активності.

Вхідні дані формуються у вигляді часових послідовностей показників функціонування інформаційної системи, що характеризують динаміку її роботи протягом визначеного періоду спостереження. Особливістю сценарію є акцент не на інтеграції різнорідних джерел даних, а на аналізі відхилень від усталених закономірностей функціонування системи. Для цього передбачається моделювання різних типів аномалій, зокрема різких змін інтенсивності подій, нетипових послідовностей дій користувачів, аномальних навантажень на окремі компоненти системи та інших відхилень від сформованого профілю нормальної поведінки.

Виявлення аномалій пропонується здійснювати за допомогою автокодувальника, який навчається на вибірці нормальних станів та визначає ступінь відхилення поточного стану від сформованої моделі штатного функціонування. Отримана оцінка аномальності може бути використана для коригування інтегрального показника ризику та підвищення чутливості методу до потенційно небезпечних змін у роботі системи. Для інтерпретації результатів доцільно застосовувати методи SHAP або LIME, які дозволяють визначити характеристики стану системи, що найбільше вплинули на формування висновку про наявність аномалії.

Вихідні дані сценарію – оцінка аномальності, інтегральний показник ризику та пояснення причин виявленого відхилення.

Успішність виконання сценарію пропонується оцінювати за такими критеріями:

- точність виявлення аномальних станів (Accuracy, цільове значення не менше 90 %);
- повнота виявлення аномалій (Recall, цільове значення не менше 85 %);
- частка хибнопозитивних спрацьовувань (False Positive Rate, цільове значення не більше 10 %;);
- F1-міра як узагальнений показник якості класифікації аномальних та нормальних станів (цільове значення не менше 0,85);
- середній час виявлення аномалії від моменту її виникнення до формування попередження (цільове значення визначатиметься експериментально залежно від характеристик інформаційної системи, обсягу даних та обчислювальних ресурсів);
- частка коректно виявлених аномалій нових типів, відсутніх у навчальній вибірці (цільове значення не менше 75 %).

Сценарій 3. Оцінювання впливу факторів ризику.

Мета сценарію – визначення ступеня впливу окремих факторів ризику на формування інтегральної оцінки ризику інформаційної безпеки та перевірка узгодженості пояснень, сформованих моделлю, з експертними оцінками фахівців з кібербезпеки.

Вхідні дані – набір параметрів, що характеризують стан інформаційної системи, зокрема показники мережевої активності, характеристики уразливостей, поведінкові дані користувачів, журнали подій безпеки та результати роботи засобів моніторингу й виявлення загроз. Для кожного набору даних модель формує інтегральний показник ризику.

Як інструменти дослідження пропонується використовувати розроблений пояснюваний інтелектуальний метод оцінювання ризиків, модель машинного навчання для прогнозування рівня ризику, а також методи пояснюваного штучного інтелекту SHAP та/або LIME для визначення внеску окремих ознак у кінцевий результат.

Вихідні дані сценарію – інтегральна оцінка ризику, ранжований перелік факторів ризику за ступенем їх впливу на результат, а також візуалізації та пояснення, що відображають внесок кожного параметра у процес прийняття рішення моделлю.

Успішність виконання сценарію пропонується оцінювати за такими критеріями:

- частка факторів ризику, визначених моделлю як найбільш значущі та підтверджених експертами (цільове значення не менше 80 %);
- коефіцієнт рангової кореляції Спірмена між ранжуванням факторів моделлю та експертним ранжуванням (цільове значення не менше 0,7);
- середня зміна позиції фактора у рейтингу при повторних запусках моделі (цільове значення не більше 10 % від загальної кількості факторів);
- частка пояснень, які експерти визнали достатніми для обґрунтування прийнятих рішень (цільове значення не менше 80 %).

Сценарій 4. Аналіз стійкості до спотворення вхідних даних.

Метою сценарію є оцінювання стійкості пояснюваного інтелектуального методу до неповних, зашумлених або навмисно спотворених даних.

Вхідними даними виступають журнали подій інформаційної безпеки, мережевий трафік, поведінкові характеристики користувачів та інші параметри, що використовуються для формування оцінки ризику.

Для проведення дослідження до набору даних вносяться контрольовані зміни у вигляді випадкового шуму, пропусків значень та модифікації окремих ознак. Як інструменти аналізу використовуються механізми оцінювання ризику, засоби пояснюваного штучного інтелекту (SHAP/LIME) та методи аналізу чутливості моделі.

Критеріями успішності є збереження стабільності інтегральної оцінки ризику в допустимих межах, незначне зниження точності моделі та узгодженість пояснень щодо найбільш впливових факторів ризику після внесення спотворень у дані:

- відносна зміна інтегральної оцінки ризику не більше 10 % при внесенні до 15 % шумових або спотворених даних;
- зниження показників якості моделі (Accuracy, F1-score або ROC-AUC) не більше ніж на 5 % порівняно з базовим станом без спотворень;
- збереження не менше 80 % ознак у переліку п’яти найбільш впливових факторів ризику за результатами SHAP/LIME-аналізу до та після внесення спотворень;
- коефіцієнт рангової кореляції Спірмена між вагами факторів ризику до та після модифікації даних не нижче 0,8;
- відсутність критичних змін категорії ризику (низький, середній, високий) більш ніж у 90 % досліджуваних випадків.

Досягнення зазначених цільових значень свідчитиме про достатню стійкість методу до неповноти, зашумлення та навмисного спотворення вхідних даних.

Сценарій 5. Навчально-демонстраційний сценарій для освітнього процесу.

Метою сценарію є формування когнітивних навичок інтерпретації результатів моделі та прийняття обґрунтованих рішень щодо реагування на інциденти. Сценарій орієнтований на використання методу у підготовці фахівців з кібербезпеки та передбачає аналіз типових інцидентів інформаційної безпеки у вигляді навчальних кейсів. До таких кейсів можуть належати:

- виявлення фішингової атаки на корпоративну електронну пошту,
- аналіз спроб несанкціонованого доступу до інформаційних ресурсів,
- дослідження аномальної мережевої активності, пов’язаної з поширенням шкідливого програмного забезпечення,
- виявлення внутрішніх загроз з боку користувачів із підвищеними привілеями,
- оцінювання ризиків експлуатації відомих уразливостей у програмному забезпеченні.

У межах кожного кейсу здобувачі отримують набір вхідних даних, результати оцінювання ризику та пояснення моделі щодо впливу окремих факторів на прийняте рішення. Успішність виконання сценарію доцільно оцінювати за такими показниками як точність визначення рівня ризику порівняно з еталонним рішенням експертів, правильність і повнота інтерпретації пояснень моделі, обґрунтованість запропонованих заходів реагування, а також час, необхідний здобувачеві для аналізу кейсу та прийняття рішення.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Проведений аналіз показав, що розвиток інтелектуальних методів оцінювання ризиків інформаційної безпеки нерозривно пов'язаний із використанням технологій штучного інтелекту та машинного навчання, які забезпечують можливість оброблення великих обсягів різнорідних даних і підтримки процесів прийняття рішень. Разом з цим існує фундаментальна суперечність у таких системах: чим складніша модель ШІ, що застосовується, тим менше людина-оператор здатна зрозуміти логіку її висновків та рекомендацій. Пояснюваний інтелектуальний метод оцінювання ризиків долає цю суперечність за рахунок додавання рівня інтерпретації висновків ШІ. Такий підхід дозволяє у процесі оцінювання ризиків перейти від сліпої довіри рішенням ШІ до усвідомленого прийняття рішень. Впровадження механізмів пояснень в інтелектуальні методи оцінювання ризиків дозволяє підвищити прозорість роботи моделей ШІ та спростити аналіз отриманих результатів. Завдяки цьому фахівець може не лише отримати оцінку ризику, а й зрозуміти, які фактори найбільше вплинули на її формування, що сприяє прийняттю більш обґрунтованих рішень у сфері кібербезпеки.

У роботі також запропоновано систему сценаріїв застосування методу, яка охоплює задачі оцінювання ризиків на основі багатоджерельних даних, виявлення аномальних станів, аналізу впливу факторів ризику, дослідження стійкості до спотворення даних та використання методу в освітньому процесі. Запропоновані сценарії формують методичну основу для подальших експериментальних досліджень і дозволяють комплексно оцінити точність, адаптивність, стійкість та інтерпретованість розробленого підходу.

Подальші дослідження спрямовані на підтвердження ефективності запропонованого методу шляхом порівняння його результатів із класичними підходами оцінювання ризиків (кількісними, якісними та гібридними) за показниками точності, адаптивності та інтерпретованості.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Mirtaheric, S. L., et al. (2025). Cybersecurity in the age of generative AI: A systematic taxonomy of AI-powered vulnerability assessment and risk management. *Future Generation Computer Systems*. https://iris.unipa.it/retrieve/9073d615-0bc5-405b-a240-e043431d85fc/cyber_compressed.pdf
2. Uddin, M., Irshad, M. S., Kandhro, I. A., Alanazi, F., Ahmed, F., & Ullah, S. S. (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. *Artificial Intelligence Review, 58*, Article 236. https://doi.org/10.1007/s10462-025-11219-5
3. Hamid, I., & Rahman, M. M. H. (2025). AI, machine learning and deep learning in cyber risk management: A review. *Discover Sustainability, 6*(1), Article 112. https://doi.org/10.1007/s43621-025-01012-3
4. Razavi, H., Franco, M. F., Ouaissa, M., Ouaissa, M., & Srivastava, G. (Eds.). (2026). *AI-driven cyber risk management* (1st ed.). River Publishers. https://doi.org/10.1201/9788743808077
5. Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems, 67*, 6969-7055. https://doi.org/10.1007/s10115-025-02429-y
6. Ali, S. M., Razzaque, A., Abbass, H., & Yousaf, M. (2025). A novel AI-based integrated cybersecurity risk assessment framework and resilience of national critical infrastructure. *IEEE Access, 13*, 12427-12446. https://doi.org/10.1109/ACCESS.2024.3524884
7. Zeijlemaker, S., Lemiesa, Y. K., Schröer, S. L., Abhishta, A., & Siegel, M. (2025). How does AI transform cyber risk management? *Systems, 13*(10), 835. https://doi.org/10.3390/systems13100835
8. Aydin, Y. (2025). *CIA+TA risk assessment framework for AI reasoning vulnerabilities*. arXiv. https://arxiv.org/abs/2508.15839
9. Shapira, B., et al. (2025). *FRAME: A risk assessment framework for adversarial machine learning systems*. arXiv. https://arxiv.org/abs/2508.17405
10. Tian, J. (2025). Integrating artificial intelligence into the cybersecurity curriculum in higher education: A systematic literature review. *Education Sciences, 15*(11), 1540. https://doi.org/10.3390/educsci15111540

11. Lysetskyi, Y. M. (2025). Artificial intelligence in cybersecurity. *Military Strategy and Technology*, *3*(3), 94-99. https://doi.org/10.63978/3083-6476.2025.3.3.08
12. Islam, S., et al. (2026). Hybrid AI-based dynamic risk assessment framework with explainable AI for cybersecurity applications. *International Journal of Information Security*. Advance online publication. https://doi.org/10.1007/s10207-026-01218-0
13. Sukailo, I., & Korshun, N. (2022). The impact of NLU and generative AI on the development of cyber defense systems. *Cybersecurity: Education, Science, Technique, 2*(18), 187–196. https://doi.org/10.28925/2663-4023.2022.18.187196
14. Ilienko, A., Kryvokulska, O., Yakovenko, O., & Teliushchenko, V. (2026). Intelligent technologies in cybersecurity: Analysis of the potential and challenges of artificial intelligence applications. *Cybersecurity: Education, Science, Technique, 4*(32), 711-723. https://doi.org/10.28925/2663-4023.2026.32.1139
15. Dakov, S., Mankovskyi, D., & Bilokon, I. (2024). Artificial intelligence systems in cybersecurity and their capabilities. *Security of Information Systems and Technologies, 2*(8), 42–48. https://doi.org/10.17721/ISTS.2024.8.42-48
16. Wisakanto, R., et al. (2025). *Adapting probabilistic risk assessment for AI systems: Concepts and applications*. arXiv. https://arxiv.org/abs/2504.18536
17. Okdem, S., & Okdem, S. (2024). Artificial intelligence in cybersecurity: A review and a case study. *Applied Sciences*, *14*(22), 10487. https://doi.org/10.3390/app142210487
18. Ivanchenko, Y., Averichev, I., & Ryzhakov, M. (2025). Generalized model for forecasting and detecting cybersecurity anomalies based on artificial intelligence. *Cybersecurity: Education, Science, Technique, 2*(28), 493-510. https://doi.org/10.28925/2663-4023.2025.28.823
19. Melko, T., & Kotsun, V. (2025). Theoretical and technical aspects of machine learning applications in cybersecurity. *Cybersecurity: Education, Science, Technique, 4*(28), 162-175. https://doi.org/10.28925/2663-4023.2025.28.774
20. Haidur, H. I., Hakhov, S. O., & Skybun, O. Z. (2025). Artificial intelligence in critical infrastructure cybersecurity. *Modern Information Protection, 4*(64), 24-37. https://doi.org/10.31673/2409-7292.2025.041203
21. Zavrazhnyi, K. Y., & Kulyk, A. K. (2024). Modern challenges of business cybersecurity and the role of artificial intelligence. *Economic Bulletin of NTUU KPI, 30*, 81-86. https://doi.org/10.20535/2307-5651.30.2024.313042
22. Zavrazhnyi, K. Y., & Kulyk, A. K. (2024). Methodological foundations for assessing the impact of artificial intelligence on information security of enterprise management systems. *Kyiv Economic Scientific Journal*, *7*, 71–78. https://doi.org/10.32782/2786-765X/2024-7-10
23. Skitsko, O., Skladannyi, P., Shyrshov, R., Humeniuk, M., & Vorokhob, M. (2023). Threats and risks of artificial intelligence use. *Cybersecurity: Education, Science, Technique, 2*(22), 6-18. https://doi.org/10.28925/2663-4023.2023.22.618
24. Kret, T., & Martseniuk, Y. (2025). Integrated approach to threat modeling in artificial intelligence systems. *Cybersecurity: Education, Science, Technique, 2*(30), 555–567. https://doi.org/10.28925/2663-4023.2025.30.993-
25. Tkach, Y., Odnokolov, V., & Petrenko, T. (2026). Risks of artificial intelligence implementation: Security, legal, and socio-economic aspects. *Technical Sciences and Technologies, 1*(43), 90-104. https://doi.org/10.25140/2411-5363-2026-1(43)-90-104
26. Elkhodr, M., & Gide, E. (2025). *Integrating generative AI in cybersecurity education: Case study insights on pedagogical strategies, critical thinking, and responsible AI use*. arXiv. https://doi.org/10.48550/arXiv.2502.15357
27. Hurevych, R., Konoshevskyi, L., Konoshevskyi, O., Voievoda, A., & Liulchak, S. (2024). Integration of artificial intelligence into education: Problems, challenges, threats, and prospects. *Modern Information Technologies and Innovative Teaching Methods in Training Specialists: Methodology, Theory, Experience, Problems, 72*, 170-186. https://doi.org/10.31652/2412-1142-2024-72-170-186
28. Grover, S., Broll, B., & Babb, D. (2023). Cybersecurity education in the age of AI: Integrating AI learning into cybersecurity high school curricula. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education (SIGCSE 2023)* (pp. 980–986). ACM. https://doi.org/10.1145/3545945.3569750
29. Shevchenko, H., Shevchenko, S., Zhdanova, Y., Spasiteleva, S., & Negodenko, O. (2021). Information security risk analysis SWOT. In *Cybersecurity Providing in Information and Telecommunication Systems* (Vol. 2923, pp. 309-317). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2923/paper34.pdf

30. Shevchenko, S., Zhdanova, Y., Storozhenko, V., Rashevska, V., & Horbach, V. (2026). Integrated information security risk assessment based on Bayesian networks and maturity auditing. *Cybersecurity: Education, Science, Technique, 4*(32), 892-907. https://doi.org/10.28925/2663-4023.2026.32.1203
31. Shevchenko, S., Zhdanova, Y., Kryvytska, O., Shevchenko, H., & Spasiteleva, S. (2024). Fuzzy cognitive mapping as a scenario approach for information security risk analysis. In *Cybersecurity Providing in Information and Telecommunication Systems* (Vol. 3826, pp. 356–362). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3826/short28.pdf
32. Mohamed, N. (2025). A comprehensive framework for cyber threat detection: Leveraging AI, NLP, and malware analysis. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-025-02466-4
33. Palko, D., Vialkova, V., & Babenko, T. (2019). Intellectual models for cyber security risk assessment. In *Processing, Transmission and Security of Information* (Vol. 2, pp. 284-288). Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej w Bielsku-Białej.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (33), 2026
ISSN 2663 – 4023

Yuliia Zhdanova
PhD, Associate Professor,
Associate Professor of the Department of Information and Cybersecurity
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: 0000-0002-9277-4972
y.zhdanova@kubg.edu.ua

Svitlana Shevchenko
PhD, Associate Professor,
Associate Professor of the Department of Information and Cybersecurity
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: 0000-0002-9736-8623
s.shevchenko@kubg.edu.ua

Oksana Zolotukhina
PhD, Associate Professor,
Associate Professor of the Department of Intellectual Technologies
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0000-0002-3314-417X
oksana.zolotukhina@knu.ua

Olena Nehodenko
PhD, Associate Professor,
Associate Professor of the Department of Computer Science
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID: 0000-0001-6645-1566
o.nehodenko@kubg.edu.ua

APPLICATION OF EXPLAINED ARTIFICIAL INTELLIGENCE TO ASSESS INFORMATION SECURITY RISKS

Abstract. The article considers modern approaches to the analysis and assessment of information security risks with an emphasis on the application of artificial intelligence (AI) methods.
A systematic review of classical qualitative, quantitative and hybrid risk management methods is conducted in the context of increasing complexity of cyber threats, dynamic changes in attack vectors and rapid adaptation of attackers. Based on the analysis of scientific literature, the need to transition from static procedures to adaptive models that rely on objective data and analytics and provide continuous monitoring, self-learning and operational updating of risk assessments is substantiated. Classes of intelligent methods are considered - expert intelligent systems, probabilistic-statistical models, neural network approaches, hybrid AI systems and behavioral analysis systems - their advantages, limitations and areas of application in the tasks of anomaly detection, incident forecasting and response automation. Special attention is paid to the role of explainable AI (XAI) in increasing the transparency of decision-making, the possibility of auditing models, and trust from users and regulators. Specific risks associated with the use of AI in cybersecurity are analyzed, in particular the vulnerability of intelligent systems themselves to specialized attacks, and directions for their mitigation through combined technical and organizational measures are proposed. Recommendations are given for the integration of AI components into educational programs for training cybersecurity specialists, which include the formation of competencies in the field of machine learning, interpretability of models, and safe deployment practices. Based on a comparative analysis, conceptual provisions are proposed for building an adaptive risk assessment method that combines automated threat detection, probabilistic assessment of consequences, and mechanisms for explaining results for making informed management decisions. To verify the method, five application scenarios were developed that allow testing the functional capability of the method in identifying hidden threats, ranking impact factors, and using it in training cases. The practical significance of the work lies in the formation of a methodological basis for the implementation of intelligent risk management

systems in critical information infrastructures and organizations of various levels, as well as in determining the priorities of further research in the field of safe and transparent use of AI in cyberspace.

Keywords: information security; cybersecurity; cognitive modeling; artificial intelligence; explainable AI; risk assessment; intelligent risk analysis; adaptive models.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Mirtaheric, S. L., et al. (2025). Cybersecurity in the age of generative AI: A systematic taxonomy of AI-powered vulnerability assessment and risk management. *Future Generation Computer Systems*. https://iris.unipa.it/retrieve/9073d615-0bc5-405b-a240-e043431d85fc/cyber_compressed.pdf
2. Uddin, M., Irshad, M. S., Kandhro, I. A., Alanazi, F., Ahmed, F., & Ullah, S. S. (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. *Artificial Intelligence Review, 58*, Article 236. https://doi.org/10.1007/s10462-025-11219-5
3. Hamid, I., & Rahman, M. M. H. (2025). AI, machine learning and deep learning in cyber risk management: A review. *Discover Sustainability, 6*(1), Article 112. https://doi.org/10.1007/s43621-025-01012-3
4. Razavi, H., Franco, M. F., Ouaissa, M., Ouaissa, M., & Srivastava, G. (Eds.). (2026). *AI-driven cyber risk management* (1st ed.). River Publishers. https://doi.org/10.1201/9788743808077
5. Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems, 67*, 6969-7055. https://doi.org/10.1007/s10115-025-02429-y
6. Ali, S. M., Razzaque, A., Abbass, H., & Yousaf, M. (2025). A novel AI-based integrated cybersecurity risk assessment framework and resilience of national critical infrastructure. *IEEE Access, 13*, 12427-12446. https://doi.org/10.1109/ACCESS.2024.3524884
7. Zeijlemaker, S., Lemiesa, Y. K., Schröer, S. L., Abhishta, A., & Siegel, M. (2025). How does AI transform cyber risk management? *Systems, 13*(10), 835. https://doi.org/10.3390/systems13100835
8. Aydin, Y. (2025). *CIA+TA risk assessment framework for AI reasoning vulnerabilities*. arXiv. https://arxiv.org/abs/2508.15839
9. Shapira, B., et al. (2025). *FRAME: A risk assessment framework for adversarial machine learning systems*. arXiv. https://arxiv.org/abs/2508.17405
10. Tian, J. (2025). Integrating artificial intelligence into the cybersecurity curriculum in higher education: A systematic literature review. *Education Sciences, 15*(11), 1540. https://doi.org/10.3390/educsci15111540
11. Lysetskyi, Y. M. (2025). Artificial intelligence in cybersecurity. *Military Strategy and Technology, 3*(3), 94-99. https://doi.org/10.63978/3083-6476.2025.3.3.08
12. Islam, S., et al. (2026). Hybrid AI-based dynamic risk assessment framework with explainable AI for cybersecurity applications. *International Journal of Information Security*. Advance online publication. https://doi.org/10.1007/s10207-026-01218-0
13. Sukailo, I., & Korshun, N. (2022). The impact of NLU and generative AI on the development of cyber defense systems. *Cybersecurity: Education, Science, Technique, 2*(18), 187–196. https://doi.org/10.28925/2663-4023.2022.18.187196
14. Ilienko, A., Kryvokulska, O., Yakovenko, O., & Teliushchenko, V. (2026). Intelligent technologies in cybersecurity: Analysis of the potential and challenges of artificial intelligence applications. *Cybersecurity: Education, Science, Technique, 4*(32), 711-723. https://doi.org/10.28925/2663-4023.2026.32.1139
15. Dakov, S., Mankovskyi, D., & Bilokon, I. (2024). Artificial intelligence systems in cybersecurity and their capabilities. *Security of Information Systems and Technologies, 2*(8), 42–48. https://doi.org/10.17721/ISTS.2024.8.42-48
16. Wisakanto, R., et al. (2025). *Adapting probabilistic risk assessment for AI systems: Concepts and applications*. arXiv. https://arxiv.org/abs/2504.18536
17. Okdem, S., & Okdem, S. (2024). Artificial intelligence in cybersecurity: A review and a case study. *Applied Sciences, 14*(22), 10487. https://doi.org/10.3390/app142210487
18. Ivanchenko, Y., Averichev, I., & Ryzhakov, M. (2025). Generalized model for forecasting and detecting cybersecurity anomalies based on artificial intelligence. *Cybersecurity: Education, Science, Technique, 2*(28), 493-510. https://doi.org/10.28925/2663-4023.2025.28.823

19. Melko, T., & Kotsun, V. (2025). Theoretical and technical aspects of machine learning applications in cybersecurity. *Cybersecurity: Education, Science, Technique, 4*(28), 162-175. https://doi.org/10.28925/2663-4023.2025.28.774
20. Haidur, H. I., Hakhov, S. O., & Skybun, O. Z. (2025). Artificial intelligence in critical infrastructure cybersecurity. *Modern Information Protection, 4*(64), 24-37. https://doi.org/10.31673/2409-7292.2025.041203
21. Zavrazhnyi, K. Y., & Kulyk, A. K. (2024). Modern challenges of business cybersecurity and the role of artificial intelligence. *Economic Bulletin of NTUU KPI, 30*, 81-86. https://doi.org/10.20535/2307-5651.30.2024.313042
22. Zavrazhnyi, K. Y., & Kulyk, A. K. (2024). Methodological foundations for assessing the impact of artificial intelligence on information security of enterprise management systems. *Kyiv Economic Scientific Journal, 7*, 71–78. https://doi.org/10.32782/2786-765X/2024-7-10
23. Skitsko, O., Skladannyi, P., Shyrshov, R., Humeniuk, M., & Vorokhob, M. (2023). Threats and risks of artificial intelligence use. *Cybersecurity: Education, Science, Technique, 2*(22), 6-18. https://doi.org/10.28925/2663-4023.2023.22.618
24. Kret, T., & Martseniuk, Y. (2025). Integrated approach to threat modeling in artificial intelligence systems. *Cybersecurity: Education, Science, Technique, 2*(30), 555–567. https://doi.org/10.28925/2663-4023.2025.30.993-
25. Tkach, Y., Odnokolov, V., & Petrenko, T. (2026). Risks of artificial intelligence implementation: Security, legal, and socio-economic aspects. *Technical Sciences and Technologies, 1*(43), 90-104. https://doi.org/10.25140/2411-5363-2026-1(43)-90-104
26. Elkhodr, M., & Gide, E. (2025). *Integrating generative AI in cybersecurity education: Case study insights on pedagogical strategies, critical thinking, and responsible AI use*. arXiv. https://doi.org/10.48550/arXiv.2502.15357
27. Hurevych, R., Konoshevskyi, L., Konoshevskyi, O., Voievoda, A., & Liulchak, S. (2024). Integration of artificial intelligence into education: Problems, challenges, threats, and prospects. *Modern Information Technologies and Innovative Teaching Methods in Training Specialists: Methodology, Theory, Experience, Problems, 72*, 170-186. https://doi.org/10.31652/2412-1142-2024-72-170-186
28. Grover, S., Broll, B., & Babb, D. (2023). Cybersecurity education in the age of AI: Integrating AI learning into cybersecurity high school curricula. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education (SIGCSE 2023)* (pp. 980–986). ACM. https://doi.org/10.1145/3545945.3569750
29. Shevchenko, H., Shevchenko, S., Zhdanova, Y., Spasiteleva, S., & Negodenko, O. (2021). Information security risk analysis SWOT. In *Cybersecurity Providing in Information and Telecommunication Systems* (Vol. 2923, pp. 309-317). CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2923/paper34.pdf
30. Shevchenko, S., Zhdanova, Y., Storozhenko, V., Rashevska, V., & Horbach, V. (2026). Integrated information security risk assessment based on Bayesian networks and maturity auditing. *Cybersecurity: Education, Science, Technique, 4*(32), 892-907. https://doi.org/10.28925/2663-4023.2026.32.1203
31. Shevchenko, S., Zhdanova, Y., Kryvytska, O., Shevchenko, H., & Spasiteleva, S. (2024). Fuzzy cognitive mapping as a scenario approach for information security risk analysis. In *Cybersecurity Providing in Information and Telecommunication Systems* (Vol. 3826, pp. 356–362). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3826/short28.pdf
32. Mohamed, N. (2025). A comprehensive framework for cyber threat detection: Leveraging AI, NLP, and malware analysis. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-025-02466-4
33. Palko, D., Vialkova, V., & Babenko, T. (2019). Intellectual models for cyber security risk assessment. In *Processing, Transmission and Security of Information* (Vol. 2, pp. 284-288). Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej w Bielsku-Białej.

Отримано редакцією журналу / Received: 28.02.26
Прорецензовано / Revised: 06.03.26
Схвалено до друку / Accepted: 25.06.26