UDC: 004.056:[004.891:004.77]

А. В. Паламарчук

ЗАСТОСУВАННЯ МЕТОДІВ ПОЯСНЮВАНОГО ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ МОНІТОРИНГУ ТА ОЦІНКИ РИЗИКІВ У ГЕТЕРОГЕННИХ МЕРЕЖАХ

Вінницький національний технічний університет
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна

Анотація.

У роботі досліджено актуальну проблему інтеграції методів пояснюваного штучного інтелекту (XAI) у системи виявлення вторгнень (IDS) для сучасних гетерогенних середовищ, таких як IoT-мережі, хмарні застосунки та інтелектуальні транспортні системи (ICV). Проаналізовано ефективність алгоритму XGBoost у виявлении мережевих аномалій та роль методів інтерпретації (SHAP, LIME) у подоланні проблеми «чорної скриньки», що властива складним моделям машинного навчання. Визначено, що прозорість прийняття рішень ШІ є критичним фактором для верифікації загроз фахівцями з кібербезпеки та мінімізації ризиків помилкових спрацювань у критичній інфраструктурі. На основі аналізу сучасних фреймворків запропоновано комплексну стратегію захисту, яка поєднує високу точність класифікації з детальною аргументацією кожного виявленого інциденту.

Ключові слова: кібербезпека, пояснюваний штучний інтелект (XAI), виявлення вторгнень (IDS), IoT, XGBoost, хмарні технології, оцінка ризиків, інтерпретованість.

Abstract

The paper explores the current problem of integrating Explainable Artificial Intelligence (XAI) methods into Intrusion Detection Systems (IDS) for modern heterogeneous environments such as IoT networks, cloud applications, and Intelligent Connected Vehicles (ICV). The effectiveness of the XGBoost algorithm in detecting network anomalies and the role of interpretation methods (SHAP, LIME) in overcoming the "black box" problem inherent in complex machine learning models are analyzed. It is determined that the transparency of AI decision-making is a critical factor for threat verification by cybersecurity experts and minimizing the risks of false positives in critical infrastructure. Based on the analysis of modern frameworks, a comprehensive protection strategy is proposed that combines high classification accuracy with detailed reasoning for each detected incident.

Keywords: cybersecurity, explainable artificial intelligence (XAI), intrusion detection (IDS), IoT, XGBoost, cloud technologies, risk assessment, interpretability.

Вступ

Стрімка цифровізація та масове впровадження Інтернету речей (IoT), хмарних обчислень і автономних транспортних засобів докорінно змінили ландшафт кіберзагроз, зробивши класичні засоби захисту на основі сигнатур малоефективними проти нових типів атак. Впровадження моделей машинного навчання (ML) та глибокого навчання (DL) дозволило значно підвищити точність ідентифікації складних аномалій, проте виникла нова критична перешкода — відсутність прозорості в алгоритмах прийняття рішень, що часто називають ефектом «чорної скриньки». Актуальність теми зумовлена необхідністю переходу до концепції пояснюваного штучного інтелекту (XAI), який не лише виявляє шкідливу активність, а й надає фахівцям із безпеки чітке обґрунтування причин спрацювання системи. Метою даної роботи є аналіз сутності XAI-підходів у контексті кібербезпеки, дослідження ефективності градієнтного бустингу для моніторингу трафіку та визначення ролі інтерпретованості у забезпеченні стійкості хмарних і транспортних екосистем.

Результати дослідження

1. Ефективність градієнтного бустингу у виявленні вторгнень для IoT-мереж

Сучасні екосистеми Інтернету речей (IoT) характеризуються надзвичайною гетерогенністю пристроїв, високою динамічністю мережевого трафіку та обмеженими обчислювальними ресурсами кінцевих вузлів, що висуває жорсткі вимоги до архітектури систем виявлення вторгнень (IDS). Дослідження підтверджують, що використання алгоритмів на основі екстремального градієнтного бустингу (XGBoost) забезпечує значну перевагу у швидкості обробки великих масивів даних та

точності класифікації порівняно з традиційними методами, такими як випадкові ліси (Random Forest) або базові нейронні мережі. Завдяки вбудованим механізмам регуляризації (L1 та L2), паралелізації обчислень та здатності ефективно працювати з пропусками у даних, фреймворки на базі XGBoost дозволяють ідентифікувати складні та багатовекторні патерни атак, зокрема DDoS-напади, сканування портів та спроби перебору паролів (brute-force), навіть в умовах сильно незбалансованих наборів даних, що є типовим для реального мережевого середовища. Висока продуктивність алгоритму дозволяє розгортати його на рівні периферійних обчислень (edge computing), що мінімізує затримки при передачі даних до централізованих серверів обробки.

Проте, попри вражаючі показники метрик точності (Accuracy) та повноти (Recall), ключовою проблемою залишається рівень довіри до автоматизованих рішень. В умовах IoT, де помилкове блокування критичного датчика або виконавчого механізму може призвести до зупинки технологічного процесу, для адміністратора безпеки стає життєво необхідним розуміти внутрішню логіку моделі. Саме тут впровадження методів інтерпретованості (XAI) дозволяє трансформувати модель із «чорної скриньки» на прозорий інструмент аудиту. Аналіз інтерпретованості дає змогу визначити, які саме ознаки мережевого пакету — наприклад, середній час між прибуттями пакетів (IAT), специфічні прапорці TCP-заголовків або об'єм переданих байтів за певне вікно часу — відіграли вирішальну роль у винесенні вердикту про атаку. Це не лише полегшує процес верифікації загрози, а й дозволяє спеціалістам проводити тонке налаштування моделі, виключаючи нерелевантні чи зашумлені ознаки, що в результаті підвищує загальну стійкість системи до нових, раніше невідомих типів кіберзагроз [1].

2. Специфіка XAI у хмарних інфраструктурах та автономних транспортних системах

Застосування пояснюваного штучного інтелекту в Cloud-native застосунках та інтелектуальних транспортних системах (ICV) обумовлене критичною ціною помилкового рішення та надзвичайною складністю архітектурних взаємодій у цих середовищах. У хмарних інфраструктурах, що базуються на мікросервісній архітектурі та контейнеризації, вектори атак часто маскуються під легітимний міжсервісний трафік або легальні запити до API. Традиційні методи моніторингу часто не здатні виявити складні ланцюжки зловмисних дій, що розгорнуті в ефемерних контейнерах, тоді як моделі глибокого навчання (DL) можуть фіксувати ці аномалії, але не пояснювати їхню природу. Використання XAI дозволяє фахівцям проводити глибоку оцінку ризиків на рівні системних викликів та мережевих потоків, ідентифікуючи, наприклад, що конкретний контейнер був класифікований як скомпрометований через нетипову послідовність звернень до чутливих директорій хоста у поєднанні з аномальним вихідним трафіком. Це дозволяє адміністраторам не просто блокувати поди, а й розуміти першопричину вразливості, що є ключовим для забезпечення стійкості всієї хмарної екосистеми [3].

Аналогічно, у сфері підключених автомобілів та інтелектуального транспорту, де затримка у виявленні атаки на CAN-шину, сенсори або блок управління (ECU) може безпосередньо загрожувати життю пасажирів, методи XAI забезпечують необхідну прозорість процесів у реальному часі. Оскільки сучасні транспортні засоби інтегровані в мережі V2X (Vehicle-to-Everything), вони стають вразливими до дистанційного зламу та ін’єкцій шкідливих повідомлень. Впровадження таких інструментів інтерпретації, як SHAP (SHapley Additive exPlanations) або LIME (Local Interpretable Model-agnostic Explanations), дозволяє системі IDS не лише виявляти атаку, а й надавати контекстуальні пояснення: наприклад, вказувати, що активність класифікована як ін'єкція фальшивих даних через невідповідність фізичних показників руху та отриманих мережевих команд. Це створює надійну основу для реалізації стратегії «людина-в-циклі» (human-in-the-loop), де система безпеки здатна автоматично пропонувати заходи з пом’якшення наслідків (mitigation), базуючись на обґрунтованих доказах, що мінімізує ймовірність хибнопозитивних спрацювань, які могли б аварійно паралізувати роботу критичних систем автомобіля [2].

3. Інтеграція хаі у процеси моніторингу та вирішення проблем довіри

Ефективна протидія сучасним кіберзагрозам у складних мережевих інфраструктурах вимагає не лише високої швидкості автоматизації, а й створення умов для швидкої та безпомилкової верифікації інцидентів аналітиками безпеки. Впровадження методів пояснюваності в архітектуру IDS дозволяє розв’язати фундаментальну суперечність між складністю моделі (наприклад, ансамблевих методів чи глибоких нейромереж) та її інтерпретованістю. Використання аналізу важливості ознак (feature importance) на етапі експлуатації системи дає змогу виділити найбільш значущі параметри, що впливають на результат класифікації, тим самим відсікаючи інформаційний шум та суттєво знижуючи кількість хибнопозитивних спрацювань (False Positives). Це стає критично важливим для Cloud-native середовищ, де велика кількість легітимних запитів може бути помилково ідентифікована як аномальна через високу динаміку розгортання мікросервісів. Завдяки наданню візуальних або текстових пояснень, наприклад, через графіки локального внеску ознак за методом SHAP, XAI забезпечує прозорість, яка

дозволяє команді SOC (Security Operations Center) миттєво оцінити серйозність загрози, зрозуміти логіку зловмисника та обрати найбільш адекватний сценарій реагування — від ізоляції сегмента мережі до блокування конкретних API-ключів [2].

Окрім підвищення операційної ефективності, інтеграція XAI відіграє ключову роль у процесі безперервного навчання та вдосконалення самих моделей захисту. Аналізуючи пояснення, надані системою, інженери з кібербезпеки можуть виявляти "зміщення" (bias) у даних або випадки, коли модель приймає рішення на основі нерелевантних кореляцій, що часто трапляється в гетерогенних IoT-мережах. Таким чином, XAI виступає сполучною ланкою між потужними математичними алгоритмами та практичною експертизою людини, трансформуючи систему захисту зі статичного фільтра на динамічного інтелектуального асистента. Проте інтеграція таких методів вимагає ретельного балансування між деталізацією пояснень та обчислювальною потужністю, адже генерація інтерпретацій для кожної події в реальному часі створює додаткове навантаження на інфраструктуру. Оптимальним рішенням тут стає дворівнева архітектура, де XAI-шар активується лише для подій з високим рівнем критичності або для проведення поглибленого ретроспективного аналізу (форензики). Такий підхід створює екосистему «людина-в-циклі» (human-in-the-loop), де штучний інтелект виконує обробку колосальних масивів даних, а людина отримує готові, аргументовані інсайти для прийняття стратегічних рішень щодо захисту периметра та оптимізації політик безпеки [1, 2].

Висновки

У роботі було детально проаналізовано роль та перспективи застосування пояснюваного штучного інтелекту (XAI) у сфері кібербезпеки. Встановлено, що використання сучасних алгоритмів, таких як XGBoost, у поєднанні з методами інтерпретації (SHAP, LIME), дозволяє досягти високої точності виявлення аномалій у складних екосистемах IoT та хмарних інфраструктурах без втрати розуміння логіки прийняття рішень. Дослідження підтверджує, що головною перевагою XAI є подолання проблеми «форензичного розриву» та «чорної скриньки», оскільки система надає вичерпні докази для кожного зафіксованого інциденту, що є критично важливим для критичної інфраструктури та автономних систем. На основі проведеного аналізу визначено, що перехід до інтерпретованих моделей машинного навчання є необхідним кроком для підвищення рівня довіри до автоматизованих систем захисту. Подальший розвиток галузі має бути спрямований на оптимізацію обчислювальних витрат на XAI-шар, що дозволить впроваджувати ці методи в режимі реального часу навіть у найбільш обмежених ресурсами вузлах мережі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] An xgboost-based intrusion detection framework with interpretability analysis for iot networks [Electronic resource] / Yunwen Hu [et al.] // Applied sciences. – 2026. – Vol. 16, no. 2. – P. 980. – Mode of access: https://doi.org/10.3390/app16020980 (date of access: 02.03.2026). – Title from screen.

[2] Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: a review [Electronic resource] / Cosmas Ifeanyi Nwakanma [et al.] // Applied sciences. – 2023. – Vol. 13, no. 3. – P. 1252. – Mode of access: https://doi.org/10.3390/app13031252 (date of access: 02.03.2026). – Title from screen.

[3] Sowjanya Y. FBZX: A novel explainable AI based security model for iot healthcare systems / Yemineni Sowjanya, S. Gopalakrishnan, R. Diner Kumar // Third international conference on augmented intelligence and sustainable systems (ICAISS). – 2025. – No. 2025. – P. 106–110.

Паламарчук Андрій Володимирович – бакалавр, Вінницький Національний Університет 21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна,por314rop@gmail.com

Науковий керівник: Кирилашчук Тетяна Григорівна – асистент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця. kgt0998@gmail.com

Palamarchuk Andrii Volodymyrovyc – Bachelor Student, Vinnytsia National Technical University, 95 Khmelnytske Shose St., Vinnytsia, 21021, Ukraine, por314rop@gmail.com

Scientific Supervisor: Kyrylashchuk Tetiana Henadiivna – assistant of the Department of Information Protection, Vinnytsia National Technical University, Vinnytsia. kgt0998@gmail.com