КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.31.1049
УДК 004.056.5:004.852:004.891.2:004.272

Тетяна Миколаївна Фесенко
к.т.н., доцент, доцент кафедри комп'ютерних та інформаційних технологій і систем
Національний університет «Полтавська політехніка імені Юрія Кондратюка», м. Полтава, Україна
ORCID: 0009-0006-1698-3795
tanifesenko@gmail.com

Юлія Вадимівна Калашнікова
асистент кафедри комп'ютерних та інформаційних технологій і систем
Національний університет «Полтавська політехніка імені Юрія Кондратюка», м. Полтава, Україна
ORCID: 0000-0001-9899-4784
kalashjulia74@gmail.com

ФЕДЕРАТИВНА GNN-XAI МОДЕЛЬ ПРОГНОЗУ КОМПРОМЕТАЦІЇ ОБЛІКОВИХ ЗАПИСІВ У ZERO TRUST-СЕРЕДОВИЩІ

Анотація. У статті представлено методологічний підхід до розробки інтелектуальної системи прогнозування компрометації облікових записів користувачів у корпоративних інформаційних середовищах. Запропонована система інтегрує федеративне навчання, графові нейронні мережі та пояснювальний штучний інтелект у рамках концепції Zero Trust, що забезпечує підвищений рівень безпеки процесів автентифікації та управління доступом шляхом децентралізованої обробки даних та збереження конфіденційності при спільному навчанні моделей. Однією з ключових особливостей є локальне навчання моделей без передавання первинних даних до центрального сховища, що виключає можливість їхнього перехоплення або несанкціонованого доступу. Агрегація локальних оновлень здійснюється за допомогою механізмів федеративної оптимізації, які враховують гетерогенність даних різних корпоративних доменів. Графовий модуль формалізує міжкористувацькі та міжсистемні взаємозв'язки у вигляді орієнтованого графа, що дозволяє ідентифікувати латентні поведінкові залежності та потенційні ризики компрометації на рівні окремих зв'язків між об'єктами автентифікації. Інтеграція компонента пояснювального штучного інтелекту забезпечує прозорість процесу прийняття рішень, формалізує обґрунтування прогнозів та знижує частоту хибних спрацьовувань. Система реалізована у парадигмі Zero Trust, що передбачає постійну верифікацію дій користувачів і пристроїв незалежно від рівня довіри, забезпечуючи адаптивне реагування на аномалії у режимі реального часу. Отримані результати демонструють підвищення точності прогнозування компрометації облікових записів порівняно з традиційними централізованими моделями машинного навчання та зменшення частоти хибнопозитивних спрацьовувань. Запропонований підхід може бути використаний для побудови адаптивних систем моніторингу безпеки у критичних інформаційних інфраструктурах, які функціонують у високодинамічних та розподілених середовищах.

Ключові слова: федеративне навчання; графові нейронні мережі; пояснювальний штучний інтелект; компрометація; кібербезпека; поведінкове моделювання; SIEM; Zero Trust.

ВСТУП

Постановка проблеми. Компрометація облікових записів користувачів у корпоративних інформаційних системах залишається однією з ключових загроз сучасного кіберпростору, оскільки забезпечує зловмисникам доступ до критичних ресурсів та даних організацій. Незважаючи на широке впровадження систем управління

подіями безпеки (SIEM), традиційні підходи до детекції загроз, що базуються на централізованому зборі та аналізі логів, мають низку принципових обмежень. По-перше, централізація даних створює ризики порушення конфіденційності та політик безпеки, оскільки передача великих обсягів критичної інформації може суперечити нормативним вимогам, таким як GDPR або внутрішні корпоративні регламенти [1]. По-друге, гетерогенність і розподіленість даних у різних підсистемах ускладнює створення уніфікованих моделей та знижує їхню точність, а централізоване навчання моделей зазвичай не забезпечує належної адаптивності до нових або раптових аномалій [2]. Крім того, існуючі моделі часто функціонують у режимі пакетної обробки даних, що не дозволяє виявляти ризики компрометації у реальному часі, а відсутність механізмів пояснюваності робить рішення системи непрозорими для аналітиків безпеки, що знижує довіру до автоматизованих рішень [3, 4].

Для подолання цих обмежень актуальним стає застосування концепції Zero Trust, яка передбачає постійну перевірку всіх дій користувачів і пристроїв незалежно від їхнього рівня довіри, а також інтеграцію передових алгоритмічних підходів до прогнозування ризиків. Сучасні дослідження демонструють перспективність використання графових нейронних мереж (GNN) для моделювання складних поведінкових та міжкористувацьких взаємозв’язків, що дозволяє виявляти латентні аномалії та залежності, які неможливо детектувати традиційними методами [5, 6]. Однак централізоване застосування GNN обмежене потребою у збиранні великих обсягів чутливих даних, що суперечить вимогам конфіденційності та безпеки, особливо у розподілених корпоративних середовищах.

Вирішення цієї проблеми можливо через впровадження федеративного навчання (Federated Learning, FL), яке забезпечує децентралізоване навчання моделей без необхідності передачі сирих даних до центрального вузла. Так кожен вузол навчає локальну модель на власних даних, після чого оновлення параметрів агрегуються для формування глобальної моделі, що враховує знання від усіх учасників [7, 8]. Попри значний потенціал, поєднання FL із графовими моделями та механізмами пояснюваності (XAI) у задачах прогнозування компрометації облікових записів залишається недостатньо вивченим, що обмежує можливості практичного застосування таких підходів у реальних умовах корпоративної безпеки.

Таким чином, наукова і практична проблема полягає у створені інтегрованої, адаптивної та конфіденційної моделі прогнозування компрометації облікових записів, яка поєднує можливості GNN для виявлення міжкористувацьких і поведінкових залежностей, механізми FL для збереження приватності даних, а також інструменти пояснюваності для підвищення довіри аналітиків до автоматизованих рішень. Вирішення цієї проблеми дозволяє розробити адаптивні системи моніторингу безпеки, здатні функціонувати у динамічних умовах Zero Trust, підвищує точність прогнозування та зменшує ризик хибних спрацьовувань, що є критично важливим для захисту корпоративних інформаційних ресурсів [9, 10].

Аналіз останніх досліджень і публікацій. У сучасних наукових дослідженнях з кібербезпеки значна увага приділяється застосуванню графових нейронних мереж для моделювання складних міжкористувацьких і поведінкових взаємозв’язків. Оглядові праці, такі як [11], демонструють ефективність GNN у виявленні латентних аномалій та потенційних загроз, проте підкреслюють обмеження централізованих підходів, що вимагають збирання великих обсягів чутливих даних, що суперечить вимогам корпоративної безпеки та нормативним актам щодо приватності. Додатково, систематичні огляди, такі як [12], відзначають, що GNN здатні забезпечувати потужне

представлення зв'язків між об'єктами у складних інформаційних середовищах, проте їхнє використання обмежене централізованими архітектурами та відсутністю розподілених механізмів навчання.

Натомість FL надає можливість колективного навчання моделей без передачі сирих даних у центральний репозиторій, що забезпечує збереження конфіденційності і дозволяє інтегрувати знання з різних підсистем. Дослідження [13] та [14] висвітлюють потенціал FL для побудови розподілених систем виявлення аномалій, звертаючи увагу на ключові виклики, такі як гетерогенність даних, масштабованість алгоритмів та стандартизація критеріїв оцінки. Проте більшість існуючих досліджень зосереджено на табличних або мережевих даних, ігноруючи складні графові структури міжкористувацьких взаємодій, що обмежує ефективність прогнозування ризиків компрометації облікових записів.

Окремий напрямок сучасних досліджень присвячений пояснюваному штучному інтелекту (Explainable AI, XAI), який забезпечує прозорість і інтерпретованість рішень автоматизованих систем безпеки. Роботи, такі як [15], підкреслюють важливість пояснюваності для підвищення довіри аналітиків та зниження ризику хибних спрацьовувань. Разом з тим інтеграція XAI з графовими та федеративними моделями у практичних системах прогнозування компрометації облікових записів залишається недостатньо дослідженою.

Таким чином, проведений аналіз літератури дозволяє виокремити низку невирішених аспектів, які формують наукову нішу для подальших досліджень. По-перше, централізоване застосування GNN не забезпечує приватність даних та не враховує вимог розподілених корпоративних систем. По-друге, існуючі рішення FL не інтегрують графові моделі для відображення міжкористувацьких зв'язків і контекстних залежностей, що критично важливо для прогнозування компрометації облікових записів. По-третє, механізми XAI, хоча і підвищують прозорість, здебільшого не враховують специфіку графових структур і федеративної децентралізації.

В цілому, наведені невирішені аспекти вказують на необхідність розробки інтегрованого підходу, що поєднує графові нейронні мережі, федеративне навчання та пояснювальний інтелект для побудови адаптивної та конфіденційної системи прогнозування компрометації облікових записів у середовищі Zero Trust.

Метою статті є розробка інтегрованої адаптивної моделі прогнозування компрометації облікових записів користувачів у корпоративних інформаційних системах**,** що функціонує в рамках архітектури Zero Trust і поєднує три ключові технологічні компоненти: GNN, FL та XAI.

Реалізація цієї мети дозволяє забезпечити наукову новизну дослідження за рахунок поєднання графового представлення, FL та XAI у контексті прогнозування компрометації облікових записів, що раніше залишалося недостатньо вивченим. Практична значущість полягає у підвищенні кіберстійкості корпоративних інформаційних систем, зменшенні ризику несанкціонованого доступу та підвищенні ефективності прийняття рішень аналітиками безпеки в умовах Zero Trust.

Таким чином, мета статті формує основу для подальшої розробки алгоритмічних, структурних та програмних рішень, що забезпечують комплексний підхід до прогнозування компрометації облікових записів у сучасних інформаційних середовищах.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Одним із ключових завдань дослідження є розробка методики формалізації динамічних взаємозв'язків між користувачами та інформаційними ресурсами корпоративної системи у вигляді графових структур, що дозволяє виявляти латентні аномалії та потенційні ризики компрометації облікових записів. Така формалізація є критичною для побудови моделей прогнозування, оскільки дозволяє структурувати поведінкові та контекстні залежності у формат, придатний для обробки GNN, що ефективно виявляють складні зв'язки у великих даних [11, 12].

У запропонованій методиці корпоративне інформаційне середовище моделюється як динамічний орієнтований граф $G_t = (V, E_t, X_t)$, де множина вузлів V об'єднує сутності користувачів, ресурсів і служб, а множина ребер E_t відображує факти взаємодії між цими сутностями у момент часу t. Матриця ознак X_t поєднує векторні представлення вузлів та ребер, що включають поведінкові, часові й контекстні характеристики. Така формалізація дозволяє динамічно відображати структуру системи, що змінюється з часом, і готувати дані для подальшого оброблення GNN, які вже довели свою здатність до виявлення складних структурних залежностей у різних доменах [16].

Для квантифікації рівня взаємодії між вузлами вводиться вагове представлення ребер де кожне ребро (v_i, v_j) на час t отримує вагу $w_{ij}(t)$, яка є агрегатом таких компонент, як частота доступів, середня тривалість сесій, критичність ресурсу, історичні аномальні ознаки та інші контекстні змінні. Враховуючи зазначене формула може набувати вигляду:

$$w_{ij}(t) = \alpha \cdot \text{freq}_{ij} + \beta \cdot \text{dur}_{ij} + \gamma \cdot \text{crit}_j + \delta \cdot \text{anom}_{ij}, \quad (1)$$

де $\alpha, \beta, \gamma, \delta$ – налаштовуванні коефіцієнти, а anom_{ij} може бути отримане, наприклад, з моделі попередньої аномальної оцінки. Таке зважене представлення дозволяє GNN-методам взяти до уваги не лише структурні, але й інтенсивні сигнали взаємодій.

Матриця суміжності $A_t \in \mathbb{R}^{|V|\times|V|}$ формується на основі цих ваг:

$$(A_t)_{ij} = \begin{cases} w_{ij}(t), & \text{якщо } (v_i, v_j) \in E_t \\ 0, & \text{інакше.} \end{cases} \quad (2)$$

Крім того, матриці ознак вузлів X_t^V і ознак ребер X_t^V формуються на підставі метаданих, що супроводжують кожну взаємодію (наприклад, тип дії, часовий штамп, роль, рівень доступу).

Оскільки граф динамічний, щоразу при надходженні нової події (логін, доступ, зміна прав) необхідно оновлювати структуру графа. При цьому нехай подія $e_k = (v_i, v_j, a_k, t_k)$. Якщо ребро (v_i, v_j) вже існує в $E_{t_{k-1}}$, його вага оновлюється:

$$w_{ij}(t_k) = w_{ij}(t_{k-1}) + f(a_k), \quad (3)$$

де функція $f(a_k)$ визначає приріст інтенсивності взаємодії згідно з типом дії a_k. Якщо ребро не існувало раніше, додається нове ребро з початковою вагою $w_{ij}(t_k) = f(a_k)$. Одночасно оновлюються відповідні записи в X^E і X^V відповідно до поточного контексту події.

Отриманий динамічний граф подається на вхід GNN-моделі, яка здійснює обмін повідомленнями між вузлами. Зокрема, типовий шар агрегування можна представити у такій формі:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathrm{N}(i)} \frac{w_{ij}(t)}{\sum_{j \in \mathrm{N}(i)} w_{ik}(t)} W^{(l)} h_j^{(l)} \right), \quad (4)$$

де $h_i^{(l)}$ – ознака вузла v_i на шарі l, $\mathrm{N}(i)$ – сусіди вузла i, $W^{(l)}$ – параметри шару, а σ – функція активації (ReLU, tanh тощо). Після останнього шару модель видає ризик-компрометації:

$$r_i(t) = \mathrm{sigmoid}\left(h_i^{(L)} \cdot w_r \right) \quad (5)$$

де w_r – параметр проєкції на ризиковий простір. Значення $r_i(t) \approx 1$ свідчить про високий ризик компрометації.

Для виявлення латентних аномалій використовується порівняння прогнозного ризику з історичною базовою лінією:

$$\mathrm{AnomScore}(v_i, t) = \left| r_i(t) - \bar{r}_i(t - \Delta) \right| \quad (6)$$

де $\bar{r}_i(t - \Delta)$ – середнє прогнозне значення за попередній проміжок часу Δ. Поріг спрацьовування може бути визначений через аналіз ROC-кривих або адаптивні алгоритми контролю аномалій.

Подібна динамічна обробка графів та використання GNN досліджується в контексті виявлення аномалій у змінних графових потоках. Так, модель SAD (Semi-Supervised Anomaly Detection on Dynamic Graphs) використовує пам’ять із часовими ознаками та псевдо-лейбли для виявлення аномалій у еволюційних графах [17]. Ще один підхід – TADDY, який застосовує трансформери для виявлення аномалій у динамічних графах, поєднуючи просторово-часові представлення [18]. Також в галузі федеративних графових моделей запропоновано FedCLGN, контрастивний підхід до федеративного виявлення аномалій, що зберігає приватність даних клієнтів [19]. У мережевій безпеці AD FG реалізує міждоменний федеративний підхід до графового подання даних для виявлення аномальних мережевих шаблонів, за якого передається лише структурна інформація, тоді як чутливі атрибути зберігаються локально. [20].

У підсумку, математично формалізована методика дозволяє крок за кроком трансформувати поведінкові й контекстні події системи у динамічний, ваговий графовий формат, підтримувати його оновлення в реальному часі, подавати до GNN-моделі й отримувати інтерпретовану оцінку ризику компрометації з можливістю виявлення латентних аномалій на основі порівняння з історичною базовою лінією. Такий підхід створює технічну основу для подальшої інтеграції з FL та XAI в межах комплексної системи прогнозування компрометації облікових записів у середовищі Zero Trust.

У розподілених корпоративних системах із високим рівнем взаємозалежності компонентів централізована схема навчання моделей машинного навчання є не лише неефективною, а й небезпечною, оскільки концентрація журналів безпеки на єдиному вузлі створює ризики порушення політик конфіденційності (GDPR, NIST SP 800-207) і знижує стійкість системи до перевантажень. У цьому контексті FL набуває практичної значущості яка дозволяє тренувати моделі на локальних вузлах без передачі сирих даних до централізованого сховища, що створює умови для розподіленого навчання з збереженням приватності [21].

Архітектуру системи спроєктовано на основі еталонної моделі [22], що визначає структурні принципи побудови клієнт-серверних компонентів, забезпечує безпечну агрегацію даних, контроль оновлень та стійкість до відмов. У представленому рішенні архітектуру (Рис.1) умовно поділено на чотири рівні: джерела даних, периферійні агенти (edge clients), сервер-агрегатор (orchestrator) та рівень аналітики/пояснюваності (XAI/SOC).

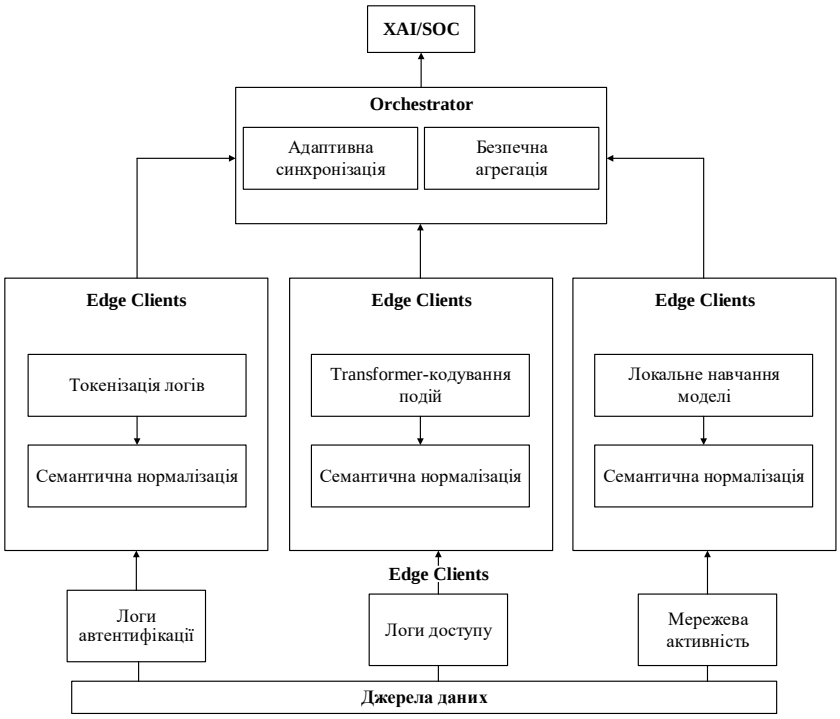

Рис.1. Архітектура федеративної системи прогнозування компрометації облікових записів користувачів

Семантична нормалізація та модуль підготовки локальних даних.

Оскільки корпоративні журнали мають гетерогенну структуру та різнорідні формати, зокрема журнали автентифікації, доступу, мережевого трафіку та службового аудиту, то перед їх передачею до локального агента необхідним є етап попередньої семантичної нормалізації. Тому враховуючи зазначене запропоновано модуль, що реалізує лог-токенізацію та рансформерне кодування подій:

$$\mathrm{h}_i = \mathrm{Proj}\Big(\mathrm{TransformerEnc}\big(\mathrm{Tokenize}\big(e_i\big)\big)\Big),\ \mathrm{h}_i \in \square^{\,d} \qquad (7)$$

де подія журналу e_i розглядається як вхідна одиниця спостереження, результат обробки h_i якої передається локальному модулю побудови графа. Подібні підходи передбачають застосування трансформер-архітектур для обробки лог-даних та уніфікації їхніх семантичних представлень.

Програмна реалізація (псевдокод) обробки подій журналу через трансформерну модель:

```
tokens = tokenizer(event_text)
emb    = transformer_encoder(tokens)  # [L, d_model]
h      = pooler(emb)                  # [d_model]
h_norm= projection_layer(h)           # [d]
```

Цей модуль забезпечує стандартизацію векторних представлень подій, що дозволяє подальшу агрегацію навчальних зразків у графовій нейронній моделі й забезпечує стійкість до різноформатних джерел.

Архітектура федеративного навчання з урахуванням гетерогенності і асинхронності. На рівні федерації клієнти $C_k\ (k=1,...,K)$ навчають локальні моделі $w_k^{(t)}$ на своїх наборах даних D_k і передають лише оновлення параметрів, а не дані. Агрегація центральним сервером відбувається за класичною схемою:

$$w^{(t+1)} = \sum_{k\text{-}1}^{K} \frac{n_k}{n} w_k^{(t)}, \quad n = \sum_{k\text{-}1}^{K} n_k \tag{8}$$

Проте через гетерогенність даних та ресурсів клієнтів, а також мережеві затримки, застосовано адаптивну асинхронну стратегію:

$$w^{(t+1)} = w^{(t)} + \sum_{k \in S_t} \eta_k \Delta w_k, \quad \eta_k = \frac{1}{1+\alpha\tau_k} \tag{9}$$

де τ_k – затримка від клієнта k; α – коефіцієнт адаптації. Такий підхід враховується в оглядах, які досліджують аспекти гетерогенності та особливості системного проєктування FL-архітектур [23].

Для забезпечення безпечної агрегації у федеративних системах застосовується модуль безпечної агрегації [24], що підтримує секретний розподіл і гомоморфне шифрування. Це дозволяє клієнтам передавати замасковані оновлення без розкриття локальних градієнтів, забезпечуючи конфіденційність даних у процесі навчання.

Протоколи комунікації та центральний контроль. Транспортний рівень реалізовано за допомогою протоколів gRPC або HTTPS із підтримкою TLS 1.3 та взаємної автентифікації (mTLS) [25]. Так, повідомлення клієнта містить метадані (наприклад, n_k, затримку τ_k, *dp-epsilon*) та цифровий підпис ED25519. При цьому сервер забезпечує вибір клієнтів, контроль версій моделей, робастну атестацію оновлень (медіанна агрегація, Krum) і журналювання раундів для аудиту.

Інтеграція з XAI та SOC-процесами. Результати роботи глобальної моделі надходять до модуля пояснюваності (XAI), який здійснює інтеграцію локальних і глобальних атрибуцій, формуючи узагальнені представлення типу «why-вузол» та «why-граф». На основі цих даних формується інтерактивна аналітична панель, що відображає причинно-наслідкові зв'язки, рівень довіри до результатів та контрфактичні інтерпретації. Механізм сповіщень інтегрується з платформами класу SIEM і процесами SOC, що забезпечує автоматизоване реагування на події безпеки у режимі реального часу.

Використання в корпоративному середовищі. Архітектура системи реалізована з урахуванням вимог масштабованості, що забезпечує підтримку до кількох тисяч клієнтів і можливість розподілу за доменними областями. Для керування життєвим циклом моделей застосовується механізм версіонування на базі MLflow/S3, а безперервна інтеграція та розгортання реалізуються за допомогою CI/CD-пакетів (Helm-chart). Контейнеризація компонентів у середовищі Docker підвищує портативність і керованість сервісів. Концепція Zero Trust покладена в основу політики безпеки де кожен клієнт і

сервер функціонують із мінімально необхідними привілеями, усі канали комунікації шифруються, а журнали подій ведуться з метою аудиту та контролю доступу.

Підсумовуючи, слід зазначити, що інтеграція механізмів пояснюваного штучного інтелекту (XAI) у федеративну GNN-модель прогнозування компрометації облікових записів є критично важливим напрямом для підвищення довіри, контрольованості та операційної ефективності корпоративних систем кіберзахисту. Така інтеграція забезпечує аналітикам SOC не лише автоматизовану ідентифікацію аномалій, але й можливість інтерпретації причинно-наслідкових зв'язків, які призводять до конкретного рішення моделі, що підвищує прозорість і передбачуваність поведінки системи.

Водночас у сучасних SIEM системах зберігається проблема «чорної скриньки» моделей глибинного навчання, коли результати класифікації компрометацій формуються без видимих пояснень. Це ускладнює проведення аудиту та знижує рівень довіри з боку аналітиків. За спостереженнями експертів, застосування XAI у контексті GNN дозволяє локалізувати критично важливі області графа, включно з вузлами, ребрами та їх атрибутами, що мають визначальний вплив на прогнозні результати. Таким чином, XAI виконує роль шару «семантичної верифікації», пояснюючи, чому певна поведінка користувача або взаємодія між системами класифікується як потенційно компрометована.

У представленій архітектурі інтеграція XAI реалізується на двох взаємопов'язаних рівнях, що забезпечує як локальну адаптацію моделей, так і координацію прогнозів на глобальному рівні.

Локальний рівень (Edge XAI) передбачає обробку журналів подій, побудову підграфів взаємодій користувачів, обчислення прогнозу ризику $r_i(t)$ та локальну генерацію пояснень $E_k = f_{XAI}(G_k, \theta_k)$, де G_k – локальний граф подій безпеки, а θ_k – параметри моделі клієнта. Пояснення формуються засобами GNNExplainer або PGExplainer, які визначають впливові вузли, ребра та ознаки, що визначили рішення моделі.

Глобальний рівень (Federated XAI Aggregator) реалізує узагальнення локальних пояснень без передачі сирих даних, забезпечуючи конфіденційність окремих організаційних даних. Введемо агрегатор:

$$\Phi^t = \mathrm{Agg}\left(\left\{\Phi_k^t\right\}_{k-1}^{K}\right) \tag{10}$$

де Φ^t – глобальний вектор пояснень, який інтегрує семантичні патерни локальних GNN-моделей у межах федерації. Цей механізм дозволяє зберегти як цілісність прогнозів, так і приватність окремих організаційних даних.

Для узгодження пояснень між гетерогенними джерелами застосовується модуль семантичної нормалізації, який забезпечує уніфікацію атрибутів різних доменів без передачі первинних даних. Це досягається шляхом побудови функції відповідності $\varphi : (V_k, E_k) \to \Omega$, яка відображає локальні структурні ознаки до єдиного простору ознак Ω. Тоді семантична відповідність між вузлами різних організацій визначається за допомогою косинусної подібності ембеддингів:

$$S(v_i, v_j) = \frac{\langle \mathrm{e}_i, \mathrm{e}_j \rangle}{\|\mathrm{e}_i\| \|\mathrm{e}_j\|} \tag{11}$$

Це дозволяє адаптивно синхронізувати локальні вузли в умовах непостійного каналу або часткової участі клієнтів у циклі навчання.

ISSN 2663 - 4023

В цілому FL архітектура реалізує динамічну синхронізацію вузлів, де клієнти з нестабільним з'єднанням передають не повні моделі, а градієнтні патчі з метаданими XAI, що зменшує трафік і прискорює глобальну конвергенцію.

Інтеграційна архітектура XAI реалізується через багатоступеневий обчислювальний конвеєр (Рис. 2), що гарантує обчислювальну ізоляцію, безпечний обмін даними та прозорість результатів.

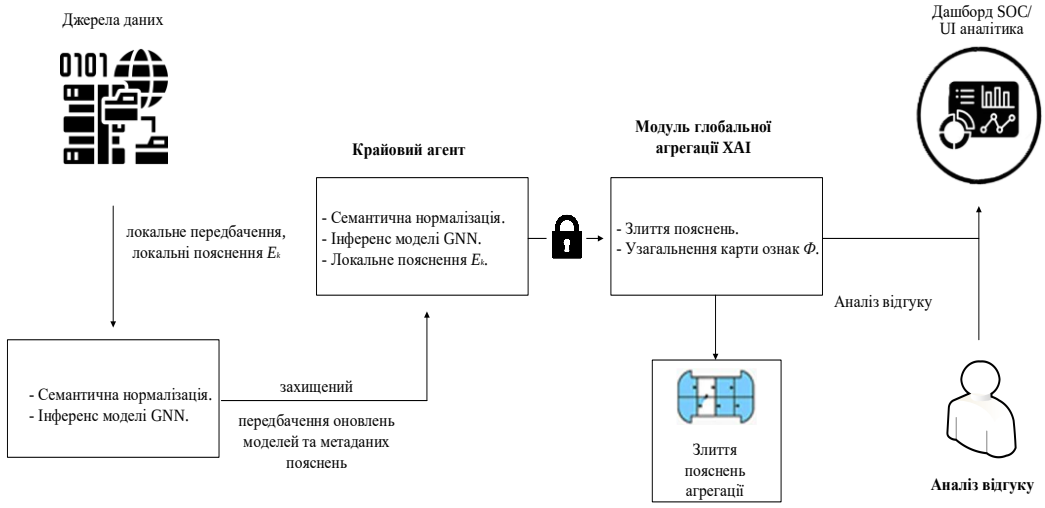

Рис. 2. Архітектура інтеграції XAI у федеративну GNN

Технічно локальний вузол генерує пояснення виключно для виявлених аномалій, нормалізує їх у форматі JSON або Protobuf та передає через захищений канал комунікації. Локальні пояснення агрегуються оркестратором у безпечний спосіб, після чого об'єднувальний модуль XAI обчислює глобальний вектор важливості, що відображає ключові чинники прийняття рішень моделі на рівні всієї федеративної системи.

Отримані пояснення візуалізуються на інтерактивній панелі SOC Dashboard у вигляді субграфів ризику, ранжованих атрибутів та контрфактичних рекомендацій.

За оцінками SOC-аналітиків, інтеграція XAI-процесів у корпоративні системи моніторингу забезпечує скорочення кількості хибних спрацьовувань на рівні 40-50% у мережах з високим ступенем складності.

В архітектурі модуль XAI реалізовано у вигляді автономного мікросервісу з інтерфейсом прикладного програмування, який взаємодіє з оркестратором через шину повідомлень. Локальна обробка даних та формування пояснень здійснюються за кілька етапів, що можуть бути представлені у вигляді псевдокоду для ілюстрації логіки обчислень:

```
explainer = GNNExplainer(local_gnn)
node_imp, edge_imp = explainer.explain_graph(graph.x, graph.edge_index)
payload = compress_and_encode(node_imp, edge_imp)
secure_send(payload)
```

Завдяки такій архітектурі забезпечується повна сумісність із середовищами з обмеженими обчислювальними ресурсами, підтримується автономність обчислень локальних вузлів і зберігається конфіденційність організаційних даних.

Таким чином, інтеграція XAI у федеративну GNN-модель формує новий тип кіберзахисної аналітики, де автоматизовані прогнози не лише точні, але й пояснювані.

Загалом поєднання GNN, XAI і федеративного навчання створює архітектуру довіри для Zero Trust-середовищ, де кожен прогноз може бути прозоро верифікований, а процес ухвалення рішень залишається під контролем людини.

Тому, саме цей підхід формує основу для розроблення адаптивних, етичних та нормативно узгоджених систем моніторингу безпеки корпоративного класу.

У рамках федеративного GNN-проєкту, спрямованого на прогнозування компрометації облікових записів, критично важливим завданням є оперативне та ефективне оновлення параметрів моделі без необхідності повного перенавчання. Реалізація такого механізму дозволяє системі своєчасно реагувати на зміну поведінкових патернів користувачів та адаптуватися до нових загроз у режимі реального часу, одночасно мінімізуючи затримки й оптимізуючи обчислювальні та комунікаційні ресурси.

Для формалізації цього процесу пропонується математично-алгоритмічна модель, адаптована під завдання кібербезпеки, що включає імовірнісні та адаптивні компоненти.

1. Постановка локального оновлення з вибіркою важливих параметрів. Нехай клієнт k має локальні параметри моделі $\mathbf{W}_t^{(k)} \in \mathbb{R}^d$. Замість повного оновлення всього вектору ваг застосовується селективне оновлення лише тих компонентів, які демонструють суттєву зміну чутливості або градієнту.

Спочатку обчислюється локальний градієнт:

$$\mathbf{g}_t^{(k)} = \nabla_w L_t^{(k)}\left(\mathbf{W}_t^{(k)}\right) \tag{12}$$

де $L_t^{(k)}$ – локальна функція втрат на нещодавніх подіях.

Одночасно накопичується експоненційне середнє квадратів градієнтів:

$$s_{i,t}^{(k)} = \beta s_{i,t-1}^{(k)} + (1-\beta)\left(g_{i,t}^{(k)}\right)^2, \quad \forall_i \in \{1,\ldots,d\}. \tag{13}$$

Цей механізм подібний до адаптивних оптимізаторів (RMSProp / Adam).

Далі обчислюється локальна важливість параметрів на базі зміни цих середніх:

$$\Delta F_{i,t}^{(k)} = S_{i,t}^{(k)} - S_{i,t-T_\Delta}^{(k)}, \tag{14}$$

де T_Δ – часовий інтервал для оцінки змін. При цьому обирається підмножина індексів:

$$S_t^{(k)} = \left\{ i : \left|\Delta F_{i,t}^{(k)}\right| > \tau_F \right\}, \quad \left|S_t^{(k)}\right| \leq d \tag{15}$$

де τ_F – поріг чутливості.

Отже, оновлення форми є:

$$\Delta \mathbf{W}_t^{(k)}[i] = \begin{cases} -\eta_{i,t} \cdot g_{i,t}^{(k)}, & \text{якщо } i \in S_t^{(k)}, \\ 0, & \text{інакше,} \end{cases} \tag{16}$$

де адаптивний крок:

$$\eta_{i,t} = \frac{\eta_0}{\sqrt{s_{i,t}^{(k)}} + \varepsilon}. \tag{17}$$

Таким чином запропонований селективний підхід забезпечує суттєве скорочення обсягу переданих оновлень при збереженні адаптивності.

2. Компресія та безпечна передача зміщень. Стиснення оновлень виконується через представлення лише індексів $i \in S_t^{(k)}$ та їхніх значень, наприклад, у форматі sparse CSR. Для додаткового зменшення трафіку можна виконати низькорангову апроксимацію:

$$\Delta \mathbf{W}_t^{(k)} \approx U_t^{(k)} \left(V_t^{(k)} \right)^{\mathsf{T}} \tag{18}$$

де $U,V \in \square^{d \times r}, r \square d$.

В подальшому клієнт надсилає повідомлення:

```
{ client_id, indices S, values Δw[S], metadata (n_k, timestamp), signature, dp_noise_info }
```

через захищений канал (TLS + mTLS) із реалізацією безпечного агрегування [26], завдяки якому сервер не має доступу до окремих оновлень клієнтів.

3. Серверна адаптивна агрегація з урахуванням запізнення. Нехай зібрано оновлення $\Delta \mathbf{W}_t^{(k)}$ від клієнтів з множини $\mathcal{S}(\mathrm{t})$. Сервер виконує вагову агрегацію:

$$\Delta \mathbf{W}_t^{(G)} = \frac{\sum_{k \in \mathcal{S}_t} \omega_k \phi(\tau_k) \Delta \mathbf{W}_t^{(k)}}{\sum_{k \in \mathcal{S}_t} \omega_k \phi(\tau_k)}, \tag{19}$$

де: ω_k – довіра до клієнта (на основі якісних метрик contribution/history); $\phi(\tau_k) = \exp(-\gamma \tau_k)$ – функція згортання за затримкою τk\tau_kτk (старі оновлення мають менший вплив).

Глобальне оновлення:

$$\Delta \mathbf{W}_{t+1}^{(G)} = \Delta \mathbf{W}_t^{(G)} + \lambda_G \cdot \Delta \mathbf{W}_t^{(G)}. \tag{20}$$

4. Регуляризація збереження знань (EWC-подібна схема). Щоб уникнути катастрофічного забування раніше вивчених патернів, вводимо регуляризацію між локальними оновленнями:

$$L_{\text{reg}} = \sum_{i=1}^{d} \frac{1}{2} F_i^{(G)} \left(w_i^{(k)} - w_{i,ref}^{(G)} \right)^2, \tag{21}$$

де $F_i^{(G)}$ – глобальна оцінка важливості параметра (усереднений Fisher) і $w_{i,ref}^{(G)}$ – контрольна точка глобальних ваг. Цей термін включається у локальну функцію втрат перед оновленням, що дозволяє балансувати локальну адаптацію і стабільність глобального знання.

5. Псевдокод: оперативне оновлення на стороні клієнта та процес агрегації.

Клієнт k:

```
g = compute_gradients(w_local, recent_events)
s = beta * s + (1-beta) * (g**2)
F_hat = rho*F_hat + (1-rho)*(g**2)
deltaF = s - s_window
S = select_topK_indices(abs(deltaF), K)eta = eta0 / (sqrt(s) + eps)
eta = eta0 / (sqrt(s) + eps)
eta_tilde = eta * sigmoid(kappa * deltaF)
delta_w = zeros_like(w_local)
delta_w[S] = -eta[S] * g[S]
payload = compress(delta_w)
secure_send(payload, meta={n_k, last_update, dp_eps})
```

На сервері:

```
updates = collect_updates()
numer = 0; denom = 0
for upd in updates:
    w_k = upd.client_id
    weight = omega[w_k] * exp(-gamma * upd.delay)
    numer += weight * upd.delta_w
    denom += weight
Delta_G = numer / denom
w_global += lambda_G * Delta_G
```

Підсумовуючи, слід зазначити, що розроблений підхід до адаптивного оновлення вагових коефіцієнтів у федеративній GNN-моделі забезпечує комплекс ключових технічних та методологічних переваг, що підвищують ефективність моделі та її придатність для систем прогнозування компрометації облікових записів у реальному часі.

По-перше, досягається мінімізація затримок адаптації завдяки модифікації лише найбільш інформативних параметрів моделі, тоді як незначущі коефіцієнти залишаються незмінними. Така вибіркова оновлювана стратегія зменшує обчислювальні витрати та скорочує час навчання, що критично для режиму безперервного прогнозування.

По-друге, підвищується ефективність комунікаційного обміну між клієнтськими вузлами та центральним оркестратором за рахунок передачі лише вагових оновлень з ненульовими значеннями, тоді як інші параметри ігноруються у процесі синхронізації. Це значно зменшує мережевий трафік у федеративній інфраструктурі та забезпечує масштабованість для корпоративних систем із великою кількістю вузлів.

По-третє, архітектура демонструє стійкість до запізнілих оновлень за рахунок застосування компенсаційної функції синхронізації $\psi(t-\Delta t)$, яка дозволяє коректно враховувати часові зсуви під час агрегації локальних моделей. Такий підхід підвищує узгодженість глобальної моделі навіть в умовах нестабільних мережевих з'єднань та неоднорідної частоти оновлень вузлів.

По-четверте, підтримується баланс між адаптивністю та стабільністю навчання завдяки використанню механізму регуляризації EWC (Elastic Weight Consolidation) [27], який мінімізує катастрофічне забування попередніх станів моделі при динамічному оновленні параметрів. Це дозволяє системі зберігати набуті знання й водночас реагувати на нові патерни поведінки користувачів.

Нарешті, запропонований підхід забезпечує сумісність із вимогами конфіденційності, підтримуючи механізми безпечної агрегації [24] та опціональне додавання диференційного шуму [28]. Це гарантує неможливість відновлення первинних даних користувачів під час спільного навчання, що критично для корпоративних середовищ із високими вимогами до приватності.

Таким чином, інтеграція запропонованого алгоритмічного підходу формує оптимальний компроміс між продуктивністю, безпекою, масштабованістю та надійністю федеративної системи адаптивного прогнозування. Це створює фундамент для побудови інтелектуальних платформ Zero Trust нового покоління, здатних ефективно функціонувати у складних корпоративних мережах із мінімальними затримками та високим рівнем контролю за даними.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У межах дослідження розроблено інтегровану федеративну GNN-XAI модель прогнозування компрометації облікових записів у Zero Trust-середовищі. Модель поєднує GNN для виявлення латентних аномалій, FL для збереження конфіденційності даних та XAI для інтерпретації рішень і зниження частоти хибних спрацьовувань.

Слід зазначити, що застосування графової формалізації корпоративних логів дозволило відобразити складні взаємозв'язки між користувачами, подіями та ресурсами, що підвищило точність прогнозування потенційних компрометацій. Впровадження федеративної архітектури забезпечило розподілене навчання без передачі сирих даних, що відповідає сучасним вимогам до кіберконфіденційності.

Необхідно наголосити що інтеграція механізмів пояснюваного ШІ підвищила довіру аналітиків до автоматизованих прогнозів, надаючи прозорі інтерпретації рішень моделі. Алгоритми адаптивного оновлення вагових коефіцієнтів із використанням регуляризації EWC забезпечили швидку реакцію на поведінкові зміни користувачів без ризику катастрофічного забування.

Запропонований підхід продемонстрував ефективність, масштабованість і стійкість у розподілених інформаційних системах, поєднуючи продуктивність із дотриманням принципів безпеки та приватності.

Подальші дослідження доцільно спрямувати на розробку онлайн-алгоритмів адаптації вагових коефіцієнтів у потокових середовищах для забезпечення безперервного навчання моделі в реальному часі. Перспективним є поєднання FL з підкріпленням (Federated RL) для автоматизації політик реагування на інциденти без централізованого контролю. Важливим напрямом залишається створення енергоефективних схем навчання для периферійних вузлів та розширення модулів XAI із урахуванням експертних знань фахівців з безпеки. Додатково слід інтегрувати квантово-стійкі протоколи secure aggregation для підвищення криптостійкості та збереження приватності даних.

Отже, отримані результати формують основу для розвитку інтелектуальних кіберзахисних систем нового покоління, орієнтованих на довіру, автономність і безперервну адаптацію до динамічних загроз.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. González-Granadillo, G., González-Zarzosa, S., & Diaz, R. (2021). *Security Information and Event Management (SIEM): Analysis, trends, and usage in critical infrastructures*. Sensors, 21(14), 4759. https://doi.org/10.3390/s21144759.
2. Kindervag, J. (2010). *Build security into your network's DNA: The Zero Trust network architecture*. Forrester Research. Retrieved from https://media.paloaltonetworks.com/documents/Forrester-Build-Security-Into-Your-Network.pdf.
3. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207.
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778.
5. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). *A comprehensive survey on graph neural networks*. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.

6. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The graph neural network model*. IEEE Transactions on Neural Networks, 20(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.
7. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The graph neural network model*. IEEE Transactions on Neural Networks, 20(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.
8. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). *Communication-efficient learning of deep networks from decentralized data*. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017)* (Vol. 54, pp. 1273–1282). PMLR. Retrieved from https://proceedings.mlr.press/v54/mcmahan17a.html
9. Kairouz, P., McMahan, H. B., Avent, B., ... & Zhao, S. (2021). *Advances and open problems in federated learning*. *Foundations and Trends® in Machine Learning*, *14*(1–2), 1–210. https://doi.org/10.1561/2200000083.
10. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Lulu.com.ISBN 978-0-244-76652-2. https://christophm.github.io/interpretable-ml-book/.
11. Guan, F., Zhu, T., Zhou, W. et al. Graph neural networks: a survey on the links between privacy and security. Artif Intell Rev 57, 40 (2024). https://doi.org/10.1007/s10462-023-10656-4.
12. Alshehri, S. M., Sharaf, S. A., & Molla, R. A. (2025). Systematic Review of Graph Neural Network for Malicious Attack Detection. *Information*, *16*(6), 470. https://doi.org/10.3390/info16060470.
13. Ansam Khraisat, Ammar Alazab, Sarabjot Singh, Tony Jan, Alfredo Jr. Gomez, Survey on Federated Learning for Intrusion Detection System: Concept, Architectures, Aggregation Strategies, Challenges and Future Directions. ACM Computing Surveys, Volume 57, Issue 1, Article No.: 7, Pages 1 - 38, https://doi.org/10.1145/3687124.
14. Lim, L.-H., Ong, L.-Y., & Leow, M.-C. (2025). Federated Learning for Anomaly Detection: A Systematic Review on Scalability, Adaptability, and Benchmarking Framework. *Future Internet*, *17*(8), 375. https://doi.org/10.3390/fi17080375.
15. Gupta, L., & Misra, D. C. (2025). Cybersecurity Threat Detection Through Explainable Artificial Intelligence (XAI): A Data-Driven Framework. International Research Journal of MMC, 6(2), 119–131. https://doi.org/10.3126/irjmmc.v6i2.80687.
16. Ekle, O. A., & Eberle, W. (2024). Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, *18*(8), 1-44. https://arxiv.org/abs/2406.00134.
17. Tian, S., Dong, J., Li, J., Zhao, W., Xu, X., Song, B., ... & Chen, L. (2023). Sad: Semi-supervised anomaly detection on dynamic graphs. *arXiv preprint arXiv:2305.13573*.
18. Liu, Y., Pan, S., Wang, Y. G., Xiong, F., Wang, L., Chen, Q., & Lee, V. C. (2021). Anomaly detection in dynamic graphs via transformer. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12081-12094.10.
19. Wu, Nannan, Zhao, Yazheng, Dong, Hongdou, Xi, Keao, Yu, Wei, & Wang, Wenjun. Federated Graph Anomaly Detection Through Contrastive Learning with Global Negative Pairs. *IEEE Transactions on Information Forensics and Security*, 2024, Vol. 19, pp. 4583–4597.
20. Zhao, Y., Liu, Z., & Pang, J. (2025). Anomaly Detection in Network Traffic via Cross-Domain Federated Graph Representation Learning. *Applied Sciences*, *15*(11), 6258. https://doi.org/10.3390/app15116258.
21. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2019). *Federated learning: Challenges, methods, and future directions.* arXiv. https://arxiv.org/abs/1908.07873.
22. Lo, S. K., Lu, Q., Paik, H. Y., & Zhu, L. (2021, August). FLRA: A reference architecture for federated learning systems. In *European Conference on Software Architecture* (pp. 83–98). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-78831-5_6.
23. Zhan, S., Huang, L., Luo, G., Zheng, S., Gao, Z., & Chao, H.-C. (2025). A review on federated learning architectures for privacy-preserving AI: Lightweight and secure cloud–edge–end collaboration. *Electronics*, *14*(13), 2512. https://doi.org/10.3390/electronics14132512.
24. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017, October). *Practical Secure Aggregation for Privacy-Preserving Machine Learning.* In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191). ACM. https://doi.org/10.1145/3133956.3133982.
25. Shanmugam, L., Tillu, R., & Tomar, M. (2023). *Federated learning architecture: Design, implementation, and challenges in distributed AI systems*. *Journal of Knowledge Learning and Science Technology*, *2*(2), 384–396. https://doi.org/10.60087/jklst.vol2.n2.p384.
26. Bonawitz, K., Salehi, F., Konečný, J., McMahan, B., & Gruteser, M. (2019). *Federated Learning with Autotuned Communication-Efficient Secure Aggregation*. arXiv. https://arxiv.org/abs/1912.00131.

27. Kirkpatrick, J., Pascanu, R., Rabinowitz, … & Hadsell, R. (2017). *Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114.
28. Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy.* Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023

Tatiana Fesenko
candidate of technical sciences, associate professor, associate professor of the department of computer and information technologies and systems
National University «Poltava Polytechnic named after Yury Kondratyuk», Poltava, Ukraine
ORCID: 0009-0006-1698-3795
tanifesenko@gmail.com

Yuliya Kalashnicova
assistant of the department of computer and information technologies and systems
National University «Poltava Polytechnic named after Yury Kondratyuk», Poltava, Ukraine
ORCID: 0000-0001-9899-4784
kalashjulia74@gmail.com

FEDERATIVE GNN-XAI MODEL FORPREDICTING COMPROMISE OF ACCOUNT RECORDS IN ZERO TRUST ENVIRONMENT

Abstract. The article presents a methodological approach to developing an intelligent system for predicting user account compromise in corporate information environments. The proposed system integrates federated learning, graph neural networks, and explainable artificial intelligence within the Zero Trust concept, ensuring an enhanced level of security for authentication and access management processes through decentralized data processing and privacy preservation during collaborative model training. One of the key features is local model training without transferring primary data to a central repository, which eliminates the possibility of interception or unauthorized access. Aggregation of local updates is performed using federated optimization mechanisms that account for the heterogeneity of data from various corporate domains. The graph module formalizes inter-user and inter-system relationships as a directed graph, allowing for the identification of latent behavioral dependencies and potential compromise risks at the level of individual connections between authentication objects. The integration of an explainable artificial intelligence component ensures transparency in the decision-making process, formalizes the justification of predictions, and reduces the frequency of false positives. The system is implemented in a Zero Trust paradigm, which involves continuous verification of user and device actions regardless of the trust level, ensuring adaptive real-time anomaly response. The obtained results demonstrate an increase in the accuracy of predicting account compromise compared to traditional centralized machine learning models and a reduction in the frequency of false positives. The proposed approach can be used to build adaptive security monitoring systems in critical information infrastructures operating in highly dynamic and distributed environments.

Keywords: federated learning; graph neural networks; explainable artificial intelligence; compromise; cybersecurity; behavioral modeling; SIEM; Zero Trust.

REFERENCES

1. González-Granadillo, G., González-Zarzosa, S., & Diaz, R. (2021). *Security Information and Event Management (SIEM): Analysis, trends, and usage in critical infrastructures*. Sensors, 21(14), 4759. https://doi.org/10.3390/s21144759.
2. Kindervag, J. (2010). *Build security into your network's DNA: The Zero Trust network architecture*. Forrester Research. Retrieved from https://media.paloaltonetworks.com/documents/Forrester-Build-Security-Into-Your-Network.pdf.
3. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207.
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778.

5. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). *A comprehensive survey on graph neural networks*. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.
6. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The graph neural network model*. IEEE Transactions on Neural Networks, 20(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.
7. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). *The graph neural network model*. IEEE Transactions on Neural Networks, 20(1), 61–80. https://doi.org/10.1109/TNN.2008.2005605.
8. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). *Communication-efficient learning of deep networks from decentralized data*. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017)* (Vol. 54, pp. 1273–1282). PMLR. Retrieved from https://proceedings.mlr.press/v54/mcmahan17a.html
9. Kairouz, P., McMahan, H. B., Avent, B., ... & Zhao, S. (2021). *Advances and open problems in federated learning*. *Foundations and Trends® in Machine Learning*, *14*(1–2), 1–210. https://doi.org/10.1561/2200000083.
10. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Lulu.com.ISBN 978-0-244-76652-2. https://christophm.github.io/interpretable-ml-book/.
11. Guan, F., Zhu, T., Zhou, W. et al. Graph neural networks: a survey on the links between privacy and security. Artif Intell Rev 57, 40 (2024). https://doi.org/10.1007/s10462-023-10656-4.
12. Alshehri, S. M., Sharaf, S. A., & Molla, R. A. (2025). Systematic Review of Graph Neural Network for Malicious Attack Detection. *Information*, *16*(6), 470. https://doi.org/10.3390/info16060470.
13. Ansam Khraisat, Ammar Alazab, Sarabjot Singh, Tony Jan, Alfredo Jr. Gomez, Survey on Federated Learning for Intrusion Detection System: Concept, Architectures, Aggregation Strategies, Challenges and Future Directions. ACM Computing Surveys, Volume 57, Issue 1, Article No.: 7, Pages 1 - 38, https://doi.org/10.1145/3687124.
14. Lim, L.-H., Ong, L.-Y., & Leow, M.-C. (2025). Federated Learning for Anomaly Detection: A Systematic Review on Scalability, Adaptability, and Benchmarking Framework. *Future Internet*, *17*(8), 375. https://doi.org/10.3390/fi17080375.
15. Gupta, L., & Misra, D. C. (2025). Cybersecurity Threat Detection Through Explainable Artificial Intelligence (XAI): A Data-Driven Framework. International Research Journal of MMC, 6(2), 119–131. https://doi.org/10.3126/irjmmc.v6i2.80687.
16. Ekle, O. A., & Eberle, W. (2024). Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, *18*(8), 1-44. https://arxiv.org/abs/2406.00134.
17. Tian, S., Dong, J., Li, J., Zhao, W., Xu, X., Song, B., ... & Chen, L. (2023). Sad: Semi-supervised anomaly detection on dynamic graphs. *arXiv preprint arXiv:2305.13573*.
18. Liu, Y., Pan, S., Wang, Y. G., Xiong, F., Wang, L., Chen, Q., & Lee, V. C. (2021). Anomaly detection in dynamic graphs via transformer. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12081-12094.10.
19. Wu, Nannan, Zhao, Yazheng, Dong, Hongdou, Xi, Keao, Yu, Wei, & Wang, Wenjun. Federated Graph Anomaly Detection Through Contrastive Learning with Global Negative Pairs. *IEEE Transactions on Information Forensics and Security*, 2024, Vol. 19, pp. 4583–4597.
20. Zhao, Y., Liu, Z., & Pang, J. (2025). Anomaly Detection in Network Traffic via Cross-Domain Federated Graph Representation Learning. *Applied Sciences*, *15*(11), 6258. https://doi.org/10.3390/app15116258.
21. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2019). *Federated learning: Challenges, methods, and future directions.* arXiv. https://arxiv.org/abs/1908.07873.
22. Lo, S. K., Lu, Q., Paik, H. Y., & Zhu, L. (2021, August). FLRA: A reference architecture for federated learning systems. In *European Conference on Software Architecture* (pp. 83–98). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-78831-5_6.
23. Zhan, S., Huang, L., Luo, G., Zheng, S., Gao, Z., & Chao, H.-C. (2025). A review on federated learning architectures for privacy-preserving AI: Lightweight and secure cloud–edge–end collaboration. *Electronics*, *14*(13), 2512. https://doi.org/10.3390/electronics14132512.
24. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017, October). *Practical Secure Aggregation for Privacy-Preserving Machine Learning.* In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191). ACM. https://doi.org/10.1145/3133956.3133982.
25. Shanmugam, L., Tillu, R., & Tomar, M. (2023). *Federated learning architecture: Design, implementation, and challenges in distributed AI systems*. *Journal of Knowledge Learning and Science Technology*, *2*(2), 384–396. https://doi.org/10.60087/jklst.vol2.n2.p384.

26. Bonawitz, K., Salehi, F., Konečný, J., McMahan, B., & Gruteser, M. (2019). *Federated Learning with Autotuned Communication-Efficient Secure Aggregation*. arXiv. https://arxiv.org/abs/1912.00131.
27. Kirkpatrick, J., Pascanu, R., Rabinowitz, … & Hadsell, R. (2017). *Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114.
28. Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy.* Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042.