КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1362
УДК 004.056:519.226

Радівілова Тамара Анатоліївна
д. т. н., проф, професор кафедри інфокомунікаційної інженерії ім.В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID: 0000-0001-5975-0269
tamara.radivilova@nure.ua

Пантєлєєв Вадим Олегович
аспірант кафедри інфокомунікаційної інженерії ім.В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID: 0009-0008-7824-8782
vadym.pantielieiev@nure.ua

МЕТОД ТЕМПОРАЛЬНОГО РАННЬОГО ПРОГНОЗУВАННЯ ВНУТРІШНІХ ІНЦИДЕНТІВ З КАЛІБРУВАННЯМ ЙМОВІРНОСТЕЙ РИЗИКУ

Анотація. У роботі запропоновано метод темпорального раннього прогнозування внутрішніх (інсайдерських) інцидентів, який здійснює перехід від бінарної класифікації «інцидент / не інцидент» до оцінювання динаміки ризику в часі. Часові ряди поведінкової телеметрії та розріджених сигналів розвідки з відкритих джерел (OSINT) агрегуються ієрархічною байєсівською моделлю простору станів, у якій латентний ризик користувача еволюціонує як авторегресійний процес, а спостереження різної природи входять через власні функції правдоподібності. Структура організаційного графа враховується не гомофілійним усередненням сусідів, а гетерофілійно-свідомим контрастом «вузол проти околу» та структурним пріором, що відповідає емпірично встановленій властивості інсайдерських зв'язків: аномальні ребра з'єднують порушника переважно з нормальними носіями чутливих ресурсів. Виходом методу є калібрована апостеріорна ймовірність настання інциденту в межах горизонту прогнозу, а правило тривоги будується за керованим бюджетом хибних спрацювань із вимогою стійкості сигналу протягом кількох діб. Обчислювальний експеримент на темпоральному наборі, параметризованому за описом CERT (тисяча користувачів, 420 діб, п'ять повторень), показав, що за фіксованого навантаження 0,1 хибної тривоги на користувача сама лише телеметрія забезпечує виявлення 0,11 інсайдерів, додавання OSINT-сигналів підвищує цей показник до 0,62, а повний метод – до 0,66 за майже вдвічі меншої частки хибних тривог; площа під ROC-кривою становить 0,957, а температурне калібрування знижує оцінку Браєра з 0,037 до 0,004 та очікувану похибку калібрування з 0,158 до 0,002. Окремо досліджено змагальну стійкість: приховування та підроблення відкритого цифрового сліду знижують виявлення до 0,01–0,06, тобто раннє попередження, побудоване на OSINT, є принципово вразливим до маніпуляції з боку самого порушника, тоді як ранжування користувачів частково зберігається завдяки структурному складнику. Показано, що довірче злиття та масковане донавчання компенсують деградацію лише частково і ціною чутливості за чистих умов.

Ключові слова: інсайдерські загрози; раннє прогнозування; часове випередження; калібрування ймовірностей; моделювання; граф; трансформер; змагальна стійкість; OSINT.

ВСТУП ТА ПОСТАНОВКА ПРОБЛЕМИ

Внутрішні інциденти залишаються категорією загроз, для якої запізнення виявлення дорівнює втраті керованості. Інсайдер діє в межах наданих повноважень, тому момент, коли його активність стає технічно аномальною, здебільшого збігається з моментом заподіяння шкоди. Практичну цінність має не стільки констатація факту, скільки завчасність: інтервал між першим достовірним сигналом і подією визначає, чи залишається в розпорядженні організації нетехнічний інструментарій – розмова з керівником, перегляд прав доступу, програми підтримки персоналу.

Найчастіше використовуваною постановкою задачі є бінарна класифікація користувача або доби активності на «інцидент/не-інцидент» [1]. Така постановка має три суттєві вади. По-перше, вона позбавлена часової семантики: модель відповідає на питання «чи є аномалія», але не на питання «наскільки близько подія» [2]. По-друге, вихід класифікатора здебільшого не є ймовірністю: бали ранжування придатні для сортування, але непридатні для встановлення порогів втручання, що стосуються конкретних людей [3]. По-третє, для вкрай незбалансованого класу навіть висока площа під ROC-кривою сумісна з неприйнятним навантаженням хибними тривогами, а саме воно визначає операційну придатність системи [1].

Аналіз операціоналізації OSINT-ознак показав, що відкриті сигнали підвищують прогнозну здатність, але водночас збільшують частку хибних спрацювань і потребують калібрування та контролю порогів. Дослідження гібридної моделі з гетерофілійно-стійкими графовими мережами встановило, що структурний контекст поліпшує ранжування, проте модель залишається статичною за часом і вразливою до маніпуляції відкритим слідом [4]. Обидва результати вказують на той самий незакритий напрям: потрібен метод, який оцінює ризик як функцію часу, видає калібровану ймовірність і має явно керований рівень хибних тривог.

Мета роботи – розробити метод темпорального раннього прогнозування внутрішніх інцидентів із калібруванням ймовірностей ризику та оцінити його якість, часове випередження і стійкість до змагальної маніпуляції відкритим цифровим слідом. Завдання: сформулювати задачу як оцінювання динаміки ризику в часі; обґрунтувати вибір ймовірнісної моделі простору станів порівняно з архітектурами на самоувазі; побудувати метод із калібруванням і правилом тривоги за бюджетом; валідувати його та кількісно оцінити деградацію під атаками приховування й підроблення сигналів.

АНАЛІЗ НАУКОВИХ ПУБЛІКАЦІЙ

Огляд глибокого навчання для виявлення інсайдерів [5] систематизував перехід від ознакових класифікаторів до послідовних архітектур і водночас зафіксував брак робіт, орієнтованих на випередження події. Практичні реалізації цього напряму – виявлення аномалій на основі мереж довгої короткочасної пам’яті [1], ансамблі без учителя [3], темпоральна агрегація ознак із механізмом уваги над журналами активності [7] та поєднання статистичного і послідовнісного аналізу [2], [6]. Ці роботи демонструють, що врахування порядку подій підвищує чутливість, однак цільова змінна в них залишається бінарною позначкою доби або сесії, а показник завчасності не вимірюється. Поведінкові фреймворки на основі глибокого навчання [8] та багатошарові інтелектуальні схеми [9] підтверджують ту саму закономірність.

Систематичний огляд графового виявлення інсайдерів [10] обґрунтовує перевагу графового подання інтранету над табличним і окремо називає дисбаланс класів та брак реальних розмічених даних серед відкритих проблем напряму. Прикладні моделі охоплюють згорткові мережі над графами журналів [11], дводоменну графову згортку та навчання на графі знань із керуванням оцінкою ризику [12]. Паралельно сформувався клас архітектур, що вносять структуру безпосередньо в механізм самоуваги: узагальнення трансформера на графи [13], структурні кодування у вигляді адитивного зсуву уваги [14] та гібридні схеми, що поєднують передавання повідомлень із глобальною увагою [15]. Саме порівняння з Graph Transformers було заявлене як напрям подальших досліджень у попередній публікації, і в цій роботі воно виконується.

Критично важливим для постановки є висновок про гетерофілію. Роботи з виявлення шахрайства [16] показали, що аномальні вузли рідкісні й з’єднані переважно з нормальними, через що гомофілійне усереднення сусідів шкодить, а не допомагає; для гетерофілійних графів запропоновано розділення его- та сусідських подань [17] і навчувані ваги узагальненого PageRank [18]. Ця властивість прямо переноситься на інсайдерів, чиї аномальні зв’язки ведуть до нормальних носіїв чутливих ресурсів.

Питання відповідності прогнозованих ймовірностей спостережуваним частотам залишалося поза увагою літератури з інсайдерських загроз, хоча в машинному навчанні воно опрацьоване докладно: показано системну надмірну впевненість сучасних нейромереж і запропоновано температурне масштабування як просте й дієве посткалібрування [19]. Розвиток напряму дав методи контролю ризику з формальними гарантіями – конформне прогнозування та конформний контроль ризику [20], адаптивні схеми за умов зсуву розподілу [21] і конформізовані моделі аналізу виживання [22], придатні для задач із цензурованим часом до події. Застосування цих інструментів до прогнозування інсайдерських інцидентів у переглянутій літературі не виявлено.

Вразливість графових моделей до цілеспрямованих збурень структури та ознак встановлено у класичних роботах [23] і систематизовано в оглядах стійкості [24]. Специфіка інсайдерської задачі полягає в тому, що частину вхідних сигналів – публічний цифровий слід – контролює сам порушник, тому сценарії приховування й підроблення є не гіпотетичними, а базовими. Емпіричною основою досліджень

залишаються синтетичний набір CERT [25], зібраний у гейміфікованому експерименті TWOS [26] та корпус ділового листування Enron [27].

Таким чином, існує нашальна потреба в розробці методу, який одночасно оцінює ризик як динаміку в часі з вимірюваним випередженням, видає калібровану ймовірність, керовано обмежує навантаження хибними тривогами та кількісно перевіряється на стійкість до маніпуляції відкритим слідом. Усунення цієї сукупної прогалини є предметом цієї роботи.

ОБҐРУНТУВАННЯ ВИБОРУ МОДЕЛЕЙ ТА ПОБУДОВА МЕТОДУ

Замість бінарної позначки користувача вводимо цільову величину, визначену для кожної доби спостереження: подія настання інциденту в межах горизонту H. Для користувача i та доби t цільова змінна та шукана величина задаються співвідношенням (1), де τ_i – момент інциденту, а $\mathcal{F}_t$ – уся інформація, доступна до моменту t включно.

$$y_{i,t} = 1\{\tau_i \in (t, t+H]\}, \pi_{i,t} = P\big(y_{i,t} = 1|\mathcal{F}_t\big). \quad (1)$$

Така постановка змінює семантику виходу: модель оцінює не наявність аномалії, а близькість події, що робить осмисленими показники часового випередження та калібрування.

В ході роботи було розглянуто два класи кандидатів. Перший: архітектури на самоувазі, зокрема Graph Transformer зі структурним зсувом уваги [14], [15], які здатні вловлювати довгі залежності та поєднувати часовий і структурний контексти. Другий: ієрархічні ймовірнісні моделі простору станів із послідовним виведенням. Вибір на користь другого класу як основи методу зроблено з чотирьох міркувань, кожне з яких випливає з умов задачі.

1) Розрідженість і різнорідність спостережень. OSINT-сигнали надходять нерегулярно, у нашій конфігурації приблизно в третині діб, тоді як телеметрія доступна щодня. Модель простору станів природно опрацьовує пропуски: у добу без спостереження виконується лише крок прогнозу, а функції правдоподібності для лічильникових і неперервних сигналів задаються окремо.

2) Потреба в ймовірності, а не в балі. Апостеріорний розподіл латентного стану дає ймовірнісний вихід за побудовою, тоді як вихід дискримінативної мережі потребує посткалібрування і, як показано нижче, залишається сильно надмірно впевненим до його застосування.

3) Дефіцит позитивних прикладів. Частка інсайдерів становить одиниці відсотків, а розмічених діб – частки відсотка. Модель із явно заданою структурою (авторегресійна динаміка, відомі канали спостережень) потребує на порядки менше прикладів, ніж навчання уваги з нуля.

4) Інтерпретовність для кадрових рішень. Траєкторія латентного ризику та внески окремих каналів пояснюються безпековому й кадровому підрозділу, що є передумовою правомірного застосування.

Graph Transformer збережено в дослідженні як конкурентну базову лінію (варіант GT-lite: одноголова самоувага по вікну діб зі структурним зсувом від сусідського агрегату), що безпосередньо реалізує заявлений раніше напрям порівняння.

Ієрархічна байєсівська модель динаміки ризику

Кожному користувачу зіставляється латентний скалярний стан $r_{i,t}$ у лог-шкалі, який еволюціонує як авторегресійний процес першого порядку зі структурним доданком:

$$r_{i,t} = a \cdot r_{i,t-1} + u_{i,t} + \varepsilon_{i,t},\ \varepsilon_{i,t} \sim N(0, q), \quad (2)$$

де $u_{i,t}$ – структурний доданок, який є відомим екзогенним вхідним сигналом, керуючою змінною або зміщенням дрейфу; a – авторегресійний параметр персистентності ($|a| < 1$ забезпечує стаціонарність); $\varepsilon_{i,t}$ – гаусівський шум переходу стану з дисперсією q.

Спостережуваний ризик визначається логістичним перетворенням стану (нелінійне відображення активації) згідно з (3), де $s_{i,t} \in (0,1)$ це логістичне сигмоїдне перетворення, яке відображає необмежений латентний стан $r_{i,t}$ на обмежений інтервал (який часто представляє рівень активації, ймовірність або показник насиченої спроможності).

$$s_{i,t} = \sigma\big(r_{i,t}\big) = \frac{1}{1+e^{-r_{i,t}}}. \quad (3)$$

Основний канал спостереження (телеметричний), після логарифмування та z-нормалізації відображено гаусовою правдоподібністю (4), а OSINT-канали – правдоподібністю (5).

$$p(\tilde{y}_{i,t}|r_{i,t}) \propto exp\left(-\frac{(\tilde{y}_{i,t}-\beta_T \cdot s_{i,t})^2}{2\sigma_T^2}\right), \quad (4)$$

де ймовірність – це гаусівська модель спостереження $\tilde{y}_{i,t} \mid r_{i,t} \sim \mathcal{N}(\beta_T s_{i,t}, \sigma_T^2)$; β_T – коефіцієнт масштабування, що відображає активацію $s_{i,t}$ на шкалу спостережуваних вимірювань; σ_T^2 – дисперсія шуму вимірювання для спостереження об'єкта.

$$p(o_{i,t}|r_{i,t})^{m_{i,t}} \propto exp\left(-\frac{m_{i,t}(o_{i,t}-\beta_S \cdot s_{i,t})^2}{2\sigma_S^2 \cdot \lambda^{-1}}\right), \quad (5)$$

де $o_{i,t}$ – допоміжне або сурогатне спостереження (наприклад, OSINT, самозвіти, поведінкові сигнали); $m_{i,t} \in \{0,1\}$ – індикатор бінарної маски: якщо $m_{i,t} = 1$, це спостереження є доступним і вносить внесок $\mathcal{N}\left(\beta_S s_{i,t}, \frac{\sigma_S^2}{\lambda}\right)$ у сукупну логарифмічну ймовірність; якщо $m_{i,t} = 0$, дані спостереження відсутні, тому показник степеня встановлено на 0, а значення ймовірності – на 1 (оновлення не відбувається); β_S – параметр масштабування викидів для допоміжного каналу; $\sigma_S^2 \lambda^{-1}$ – дисперсія допоміжного каналу, де $\lambda > 0$ є коефіцієнтом довіри до відкритих джерел: його зменшення розширює дисперсію правдоподібності OSINT і послаблює вплив цього каналу. Він відіграє роль механізму компенсації в умовах змагальної маніпуляції.

Гетерофілійно-свідомий структурний складник

Структурний доданок $u_{i,t}$ у (2) є ключовою відмінністю методу. Пряме усереднення апостеріорного ризику сусідів у цій задачі шкідливе, оскільки аномальні ребра ведуть від інсайдера до нормальних носіїв чутливих ресурсів.

У рівняннях (6) і (7) зовнішній чинник, що впливає на систему, визначається як поєднання впливу соціальних мереж (тиск оточення/поширення) та статичної структурної центральності мережі (перетин меж/посередництво).

Вплив мережі та формулювання зовнішнього впливу:

$$u_{i,t} = \gamma\left(\bar{s}_{i,t-1} - \frac{1}{N(i)}\sum_{j \in N(i)} \bar{s}_{j,t-1}\right) + \kappa c_i, \quad (6)$$

де $N(i)$ – множина сусідів вузла i у графі $G = (V, E)$, ступінь якого дорівнює $|N(i)|$; $\bar{s}_{i,t-1}$ – агрегований або згладжений історичний стан активації вузла i станом на момент часу $t-1$; $\frac{1}{|N(i)|}\sum_{j \in N(i)} \bar{s}_{j,t-1}$ – середнє локальне значення активації серед усіх прямих сусідів вузла i; $\left(\bar{s}_{i,t-1} - \frac{1}{|N(i)|}\sum_{j \in N(i)} \bar{s}_{j,t-1}\right)$ – відносний розрив в активації між вузлом i та його локальною соціальною групою: якщо $\gamma < 0$ – діє як притягання/гомофілія/консенсус, що тягне до середнього значення сусідів (негативний зворотний зв'язок, що спонукає вузол i до однорідності; якщо $\gamma > 0$ – представляє диференціацію або конкурентний вплив (позитивний зворотний зв'язок, що віддаляє від середнього значення групи); $\kappa \cdot c_i$ – статичний індивідуальний зсув, зважений на κ, що визначається топологічним положенням вузла i на межі c_i.

Оцінка перетину меж / оцінка посередництва розраховується так:

$$c_i = z\left(\frac{|\{(i,j) \in E : \text{comm}(i) \neq \text{comm}(j)\}|}{|N(i)|}\right), \quad (7)$$

де $\text{comm}(i)$ – спільнота, кластер або модульний сегмент, до якого належить вузол i; $|\{(i,j) \in E : \text{comm}(i) \neq \text{comm}(j)\}|$ – кількість міжспільнотних або міжмодульних ребер, пов'язаних із вузлом i; $\frac{|\ldots|}{|N(i)|}$ – коефіцієнт участі або коефіцієнт зовнішнього ступеня (співвідношення зв'язків вузла i, які з'єднують різні спільноти, а не залишаються в межах його локального модуля: близько 0 – вузол i є внутрішнім провінційним вузлом або повністю локалізований у межах своєї спільноти, близько 1 – вузол i є посередником між спільнотами / сполучною ланкою між спільнотами / брокер, що з'єднує різні групи);

$z(\cdot)$ – стандартне нормалізаційне перетворення (наприклад, стандартизація за z-балом для всіх вузлів V: $z(x) = \frac{x-\mu}{\sigma}$, або відображення активації).

Переходимо до зв'язаної системи, підставляючи $u_{i,t}$ у рівняння переходу (2):

$$r_{i,t} = a \cdot r_{i,t-1} + \gamma \left(\bar{s}_{i,t-1} - \frac{1}{|N(i)|} \sum_{j \in N(i)} \bar{s}_{j,t-1} \right) + \kappa \cdot c_i + \varepsilon_{i,t} \tag{8}$$

Повна модель утворює мережеву нелінійну авторегресійну систему стану-простору, де: прихований стан $r_{i,t}$ еволюціонує завдяки часовій стійкості (a); механізми дифузії/консенсусу поширюються через топологію мережі за допомогою відмінностей, подібних до лапласіана графа; топологічні посередники (c_i) отримують вищий або нижчий базовий рівень прихованого збудження, що регулюється κ; активації станів $s_{i,t} = \sigma(r_{i,t})$ спостерігаються через багатошвидкісні подвійні гаусові канали випромінювання ($\tilde{y}_{i,t}$ та частково спостережувані $o_{i,t}$).

Апостеріорний розподіл наближується бутстреп-фільтром частинок: для кожного користувача підтримується ансамбль із N частинок, які на кожному кроці поширюються згідно з (2), зважуються добутком правдоподібностей (4)–(5) та піддаються систематичному ресемплінгу. Оцінка ризику обчислюється як апостеріорне середнє:

$$\bar{s}_{i,t} = \sum_{k=1}^{N} w_{i,t}^{(k)} \, \sigma\left(r_{i,t}^{(k)}\right), \; w_{i,t}^{(k)} \propto p\left(\tilde{y}_{i,t} | r^{(k)}\right) p\left(o_{i,t} | r^{(k)}\right)^{m_{i,t}} \tag{9}$$

Визначимо рівень калібрування ймовірності, оцінки та виявлення тривог раннього попередження, побудований на основі латентної моделі простору станів. Апостеріорне середнє є монотонним показником ризику, але не ймовірністю події в межах горизонту. Тому виконується посткалібрування двопараметричним перетворенням на лог-шкалі (10), параметри якого оцінюються мінімізацією від'ємної логарифмічної правдоподібності на валідаційній вибірці користувачів, що не перетинається з тестовою

$$\hat{\pi}_{i,t} = \sigma\left(\frac{\text{logit}(\bar{s}_{i,t})}{T} + b \right), \tag{10}$$

де $\hat{\pi}_{i,t} \in (0,1)$ – відкалібрована ймовірність прогнозу, що об'єкт i зазнає події, яка цікавить, у момент часу t (або напередодні); $\text{logit}(\bar{s}_{i,t}) = \ln\left(\frac{\bar{s}_{i,t}}{1-\bar{s}_{i,t}}\right)$ – інвертує сигмоїдну активацію $\bar{s}_{i,t}$ назад до необмеженого латентного простору логарифмічних шансів; $T > 0$ – параметр температури ($T > 1$ зм'якшує/згладжує надмірно впевнені ймовірності в бік 0,5; $T < 1$ – посилює недостатньо впевнені ймовірності в бік 0 або 1); b – параметр зміщення калібрувального зсуву.

Якість калібрування для точності і чіткості прогнозів контролюється оцінкою Браєра (11) та очікуваною похибкою калібрування (12):

$$BS = \frac{1}{M} \frac{1}{M} \sum_{m=1}^{M} \left(\hat{\pi}_{i,t} - y_{i,t} \right)^2, \tag{11}$$

le $y_{i,t} \in \{0,1\}$ – справжній бінарний еталонний результат події для вибірки/моменту часу $m = (i,t)$; M – загальна кількість оцінених випадків просторово-часового прогнозування; BS $\in [0,1]$ – суворо правильне правило оцінювання, що вимірює середньоквадратичну похибку ймовірнісних прогнозів (значення 0 означає ідеальне калібрування та дискримінацію). Справжній бінарний результат події, що відповідає еталонному значенню, для вибірки/моменту часу m=(i,t).

Очікувана похибка калібрування (Expected Calibration Error**)** вимірює середнє відхилення між суб'єктивною впевненістю та спостережуваною емпіричною ймовірністю по всьому ймовірнісному розподілі:

$$\text{ECE} = \sum_{\beta=1}^{B} \frac{|B_\beta|}{M} \left| \text{freq}(B_\beta) - \text{conf}(B_\beta) \right|, \tag{12}$$

де прогнози поділяються на B діапазони впевненості $B_1, \ldots, B_B \subset [0,1]$; $\text{conf}(B_\beta) = \frac{1}{|B_\beta|} \sum_{m \in B_\beta} \hat{\pi}_m$ – середня прогнозована впевненість у діапазоні β; $\text{freq}(B_\beta) = \frac{1}{|B_\beta|} \sum_{m \in B_\beta} y_m$ – емпірична точність (фактичний відсоток позитивних результатів) у діапазоні β.

Правило обмеженої тривоги (фільтр стійкості послідовності) усуває артефакт, за якого шумна модель отримує формально велике випередження за рахунок випадкових одноденних перетинів порога:

$$\theta = \min\{\theta' : \mathrm{FA}(\theta') \le \Phi\}, \quad \text{тривога}(i,t) \Leftrightarrow \hat{\pi}_{i,t-k+1} \ge \theta \quad \forall k \in \{1,\dots,K\}, \tag{13}$$

де $\mathrm{FA}(\theta')$ – частота помилкових спрацьовувань (або кількість помилкових позитивних результатів) при заданому пороговому значенні θ'; Ф – верхня межа допустимої частоти помилкових спрацьовувань (наприклад, робоче обмеження Неймана-Пірсона); поріг θ обирається як найменший, за якого середня кількість стійких тривог на одного некритичного користувача за період спостереження не перевищує заданого рівня Ф, причому тривога вважається стійкою лише за K послідовних діб перевищення ; тривога (i,t) (умова спрацьовування сигналу тривоги) – сигнал тривоги спрацьовує в момент часу t для об'єкта i тоді й тільки тоді, коли відкалібрована ймовірність $\hat{\pi}_{i,\tau}$ безперервно залишається вище θ протягом K послідовних часових кроків $(t-K+1,\dots,t)$. Цей фільтр стійкості пригнічує однокрокові шумові сплески та будується не за фіксованим значенням ймовірності, а за операційним бюджетом.

Час випередження (горизонт раннього попередження) визначається як різниця між моментом інциденту та початком першої стійкої тривоги

$$\mathrm{LT}_i = \tau_i - \min\{t{:}\text{тривога}(i,t)\}, \tag{14}$$

де τ_i – точна часова мітка, коли для об'єкта i відбувається справжня несприятлива подія; $\min\{t{:}\text{тривога}(i,t)\}$ – час першої справжньої тривоги; LT_i – час випередження (вікно реагування, що надається системою раннього попередження). Більше позитивне значення LT_i дає фахівцям більше часу для втручання до настання події τ_i.

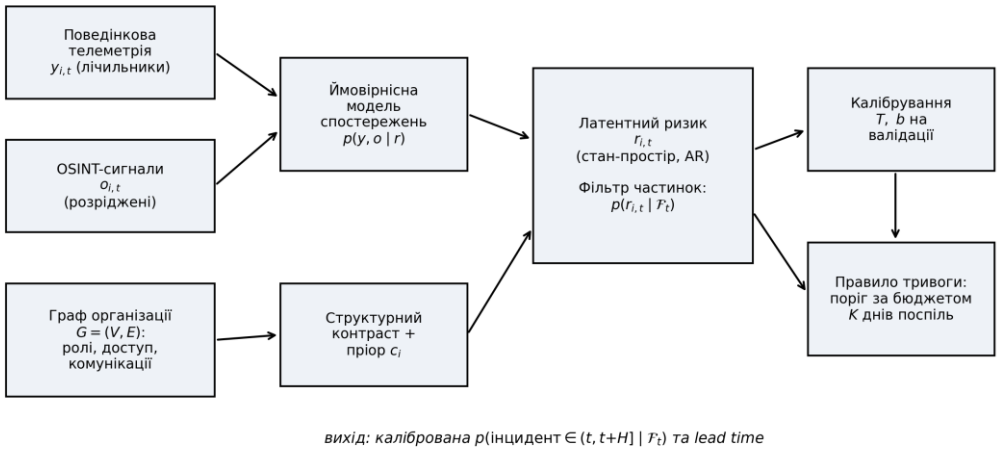

Рис. 1. Структурна схема методу темпорального раннього прогнозування

Таким чином, в роботі уперше задачу прогнозування інсайдерських інцидентів сформульовано як оцінювання каліброваної апостеріорної ймовірності настання події в межах горизонту, а не як бінарну класифікацію, що дає змогу вимірювати часове випередження та встановлювати пороги втручання за операційним бюджетом хибних тривог, а не за довільним значенням бала. Удосконалено спосіб урахування структури організаційного графа в темпоральній моделі: замість гомофілійного усереднення сусідів застосовано гетерофілійно-свідомий контраст «вузол проти околу» разом зі структурним пріором міжспільнотних зв'язків, що відповідає емпірично встановленій природі аномальних ребер інсайдера. Уперше для цього класу задач кількісно оцінено змагальну стійкість темпорального раннього прогнозування до приховування та підроблення відкритого цифрового сліду і показано принципову асиметрію: ранжування користувачів частково зберігається завдяки структурному складнику, тоді як завчасність втрачається майже повністю; введено коефіцієнт довіри до відкритих джерел як механізм часткової компенсації з явно вимірюваною ціною у вигляді зниження чутливості за чистих умов.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ ТА ЗМАГАЛЬНА СТІЙКІСТЬ

Валідацію виконано на темпоральному наборі, параметризованому за опублікованим описом CERT [25]. Набір відтворює організацію з 1000 користувачів упродовж 420 діб. Щоденна телеметрія подана п'ятьма лічильниковими каналами з тижневою періодичністю; OSINT-сигнали – трьома неперервними ознаками, доступними в середньому в 35 % діб. Частка інсайдерів становить близько 5 %; для кожного з них момент інциденту та два моменти появи передвісників задано окремо: OSINT-передвісник виникає за 20–45 діб і швидко виходить на плато, що відповідає природі поведінкових сигналів на кшталт пошуку роботи чи публічного невдоволення, тоді як телеметричний слід наростає за 5–15 діб до події. У 12 % нормальних користувачів змодельовано «життєві обставини» – тривалі підвищення OSINT-ознак без зв'язку з ризиком, що є джерелом хибних спрацювань. Граф організації побудовано за допомогою стохастичної блокової моделі з вісьмома спільнотами; аномальні ребра з'єднують інсайдерів із кастодіанами чутливих ресурсів з інших спільнот. Користувачів стратифіковано розподілено на навчальну, валідаційну та тестову частини у співвідношенні 60:20:20; результати усереднено за п'ятьма незалежними повтореннями. Горизонт прогнозу H = 30 діб, вимога стійкості K = 3 доби, бюджет Ф = 0,1 стійкої хибної тривоги на користувача.

Для порівняння реалізовано п'ять базових ліній: поденний персептрон без пам'яті (Static-MLP), експоненційне згладжування бала аномальності (EWMA), фільтр частинок лише на телеметрії (PF-T), фільтр частинок на телеметрії та OSINT без структурного складника (PF-TS) і Graph Transformer-lite із самоувагою по вікну діб та структурним зсувом (GT-lite). Пропонований метод позначено TERC.

Зведені результати наведено в табл. 1. Пропонований метод забезпечує найвищу площу під ROC-кривою (0,957) та найвищу площу під кривою «точність – повнота» (0,275), що є визначальним показником за частки позитивного класу нижче одного відсотка. Значення Brier і ECE в таблиці наведено після калібрування (до калібрування Brier і ECE становлять відповідно 0,037 та 0,158 для TERC). Найважливішим є порівняння джерел сигналу за операційною робочою точкою: за однакового бюджету хибних тривог сама лише телеметрія виявляє 0,11 інсайдерів, додавання OSINT підвищує виявлення до 0,62, а структурний складник – до 0,66 за нижчого фактичного навантаження тривогами (0,037 проти 0,053), як зазначено на рис.2.

Таблиця 1

Якість прогнозу, калібрування та операційні показники (п'ять повторень)

Модель	AUC-ROC	PR-AUC	Brier	ECE	Детек.	Lead, діб	Хибні тривоги
Static-MLP	0,836±0,035	0,069	0,0045	0,0079	0,12	7,2	0,028
EWMA	0,872±0,043	0,174	0,0046	0,0162	0,53	5,6	0,058
PF-T	0,548±0,074	0,019	0,0045	0,0018	0,11	28,0	0,049
PF-TS	0,841±0,033	0,204	0,0041	0,0021	0,62	6,7	0,053
TERC	0,957±0,013	0,275	0,0041	0,0020	0,66	7,8	0,037
GT-lite	0,867±0,048	0,113	0,0045	0,0136	0,15	10,3	0,041

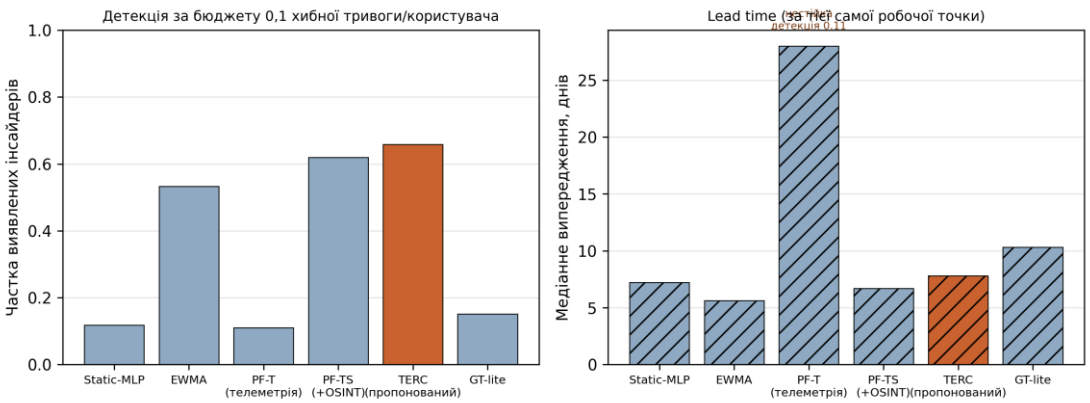

Рис. 2. Виявлення та випередження за фіксованого бюджету хибних тривог

Слід зауважити, що формально показник фільтра лише на телеметрії демонструє найбільше медіанне випередження (28 діб), проте за виявлення лише 0,11: його бал майже не несе сигналу (AUC 0,548), тому поодинокі спрацювання виникають випадково задовго до події. Це ілюструє, чому випередження не можна інтерпретувати у відриві від робочої точки, і виправдовує вимогу стійкості тривоги, введену в (13).

Ефект калібрування показано на рис. 3. До калібрування сирий апостеріорний бал є різко надмірно впевненим: значенням близько 0,3–0,6 відповідає спостережувана частота подій на два порядки нижча. Двопараметричне перетворення (10) знижує оцінку Браєра з 0,037 до 0,004, а очікувану похибку калібрування – з 0,158 до 0,002, приводячи прогнози до масштабу базової частоти. Водночас у верхньому децилі зберігається залишкова недостатня впевненість, що є прямим наслідком украй малої кількості позитивних прикладів у цій області.

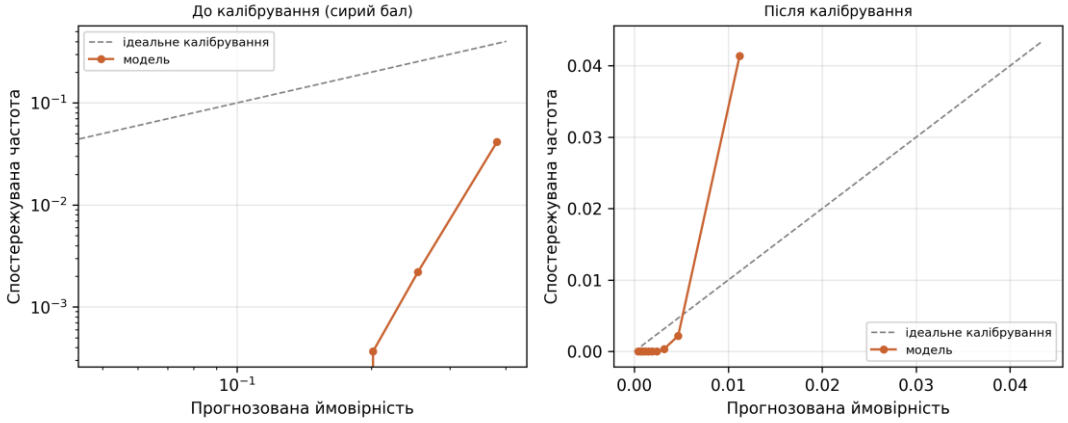

Рис. 3. Діаграма надійності прогнозу до та після калібрування

Стійкість до маніпуляції відкритим цифровим слідом. Досліджено чотири сценарії, що діють лише на тестових інсайдерів і відтворюють дії обізнаного порушника: A1 – приховування OSINT-сліду (заміщення відкритих ознак медіанними значеннями норми); A2 – підроблення, або камуфляж (імітація типового спокійного профілю зі зниженою дисперсією та рідшими оновленнями); A3 – приховування 60 % аномальних міжспільнотних зв'язків; A4 – комбінація A1, A2 і A3. Результати наведено в табл. 2 і на рис. 4.

Таблиця 2

Деградація виявлення та ранжування під атаками (п'ять повторень)

Модель	**Без атаки**	**A1**	**A2**	**A3**	**A4**	**AUC при A4**
Static-MLP	0,12	0,00	0,00	0,12	0,00	0,53
EWMA	0,53	0,00	0,00	0,53	0,00	0,55
PF-TS	0,62	0,03	0,00	0,62	0,00	0,49
TERC	0,66	0,06	0,01	0,43	0,01	0,74
GT-lite	0,15	0,00	0,00	0,13	0,00	0,53

Результат є однозначним і незручним для практики. Приховування або підроблення відкритого сліду знижує виявлення пропонованого методу з 0,66 до 0,06 і 0,01 відповідно, а комбінована атака – до 0,01; усі моделі, що спираються на OSINT, поводяться так само. Натомість приховування зв'язків (A3) є істотно менш дієвим: TERC зберігає 0,43, а PF-TS – 0,62, оскільки часові сигнали лишаються недоторканими. Отже, домінантним вектором є саме маніпуляція публічним слідом, який контролює сам порушник, а не структурою корпоративного графа, яку контролює організація.

Водночас виявлено асиметрію між ранжуванням і завчасністю. Під комбінованою атакою площа під ROC-кривою пропонованого методу знижується лише до 0,74, тоді як у решти моделей – до 0,49–0,55, тобто до рівня випадкового вгадування. Це пояснюється структурним пріором (8), який не залежить від

відкритих сигналів: метод продовжує коректно ранжувати користувачів за ризиком, але втрачає здатність вказати момент. Практичний висновок полягає в тому, що після втрати OSINT-каналу система деградує з інструмента раннього прогнозування до інструмента пріоритизації.

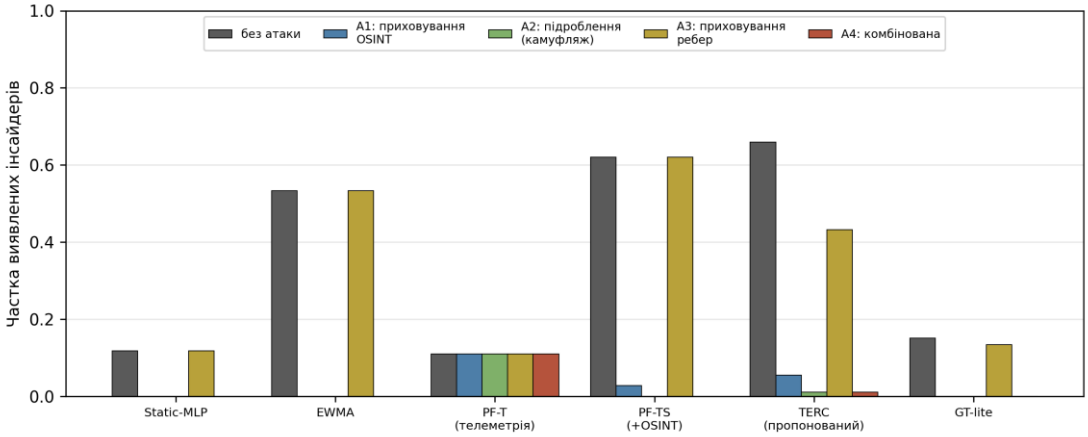

Рис. 4. Деградація виявлення під сценаріями маніпуляції

В ході експерименту перевірено два механізми: 1) довірче злиття: зниження коефіцієнта λ у (5), що послаблює вплив відкритих джерел на користь телеметрії та структури; 2) масковане донавчання дискримінативної моделі, за якого до навчальної вибірки додаються копії прикладів із обнуленим блоком OSINT-ознак. Результати наведено в табл. 3.

Таблиця 3

Ефект механізмів компенсації під комбінованою атакою A4

Конфігурація	**AUC, чисто**	**Детек., чисто**	**AUC при A4**	**Детек. при A4**
TERC (без захисту)	0,957	0,66	0,736	0,01
TERC + довірче злиття	0,932	0,41	0,757	0,04
MLP + масковане донавчання	0,833	0,15	0,703	0,00

Довірче злиття підвищує площу під ROC-кривою під атакою з 0,736 до 0,757, проте ціною зниження виявлення за чистих умов з 0,66 до 0,41. При цьому слід зауважити , що масковане донавчання не дає переваги взагалі. Отже, жоден із перевірених механізмів не відновлює раннє прогнозування: втрата інформативного каналу не компенсується перерозподілом ваг між рештою каналів, якщо ці канали за побудовою реагують пізніше. Це негативний результат, який необхідно зауважити, оскільки він визначає межі застосовності всього класу OSINT-орієнтованих методів раннього прогнозування.

ДИСКУСІЯ ТА ОБМЕЖЕННЯ

В роботі поза межами конкретної реалізації є три результати, які потрібно окремо зауважити. Перший: перехід до темпоральної постановки виправданий не стільки приростом площі під ROC-кривою, скільки появою вимірюваних операційних величин (випередження та навантаження тривогами), без яких порівняння методів залишається абстрактним. Другий: калібрування є не косметичною процедурою, а передумовою правомірності, оскільки саме калібрована ймовірність, а не бал ранжування, може бути покладена в основу порогу втручання щодо конкретної людини. Третій: цінність відкритих сигналів і їх вразливість є двома проявами однієї властивості – контрольованості цих сигналів суб'єктом спостереження; будь-яка оцінка ефективності OSINT-орієнтованого прогнозу без сценаріїв протидії є завищеною.

Порівняння з Graph Transformer дало результат, протилежний очікуваному: варіант GT-lite поступився ймовірнісній моделі за всіма операційними показниками (виявлення 0,15 проти 0,66) і виявився погано каліброваним до посткалібрування (ECE 0,402). Але не потрібно інтерпретувати це як загальну перевагу моделей простору станів над архітектурами на самоувазі. Радше це наслідок умов задачі

(дефіциту позитивних прикладів і розрідженості спостережень), за яких модель із заданою структурою має суттєву перевагу перед архітектурою, що має вивчати залежності з даних. За наявності великих реальних даних співвідношення може змінитися на протилежне.

Також існують обмеження запропонованих підходів:

– Синтетичність емпіричної бази. Усі оцінки одержано на даних, у які випереджальність OSINT-сигналів закладено конструкцією генератора. Числа характеризують здатність методу реалізувати цей потенціал, а не сам факт його існування в реальних організаціях.

– Принципова вразливість раннього прогнозування. Метод спирається на канал, який контролює порушник. Досліджені механізми компенсації відновлюють ранжування, але не завчасність; проти обізнаного та мотивованого інсайдера завчасність не гарантується.

– Спрощення моделі динаміки. Латентний ризик подано скалярним авторегресійним процесом із гаусовим шумом і стаціонарними параметрами. Реальна поведінка є нестаціонарною, а різні сценарії інциденту (саботаж, витік, змова) мають різну динаміку, що потребує багатовимірного або режимно-перемикального стану.

– Обмеження структурного складника. Структурний пріор (8) є статичним і відображає радше схильність користувача до ризику, ніж момент події: він поліпшує ранжування, майже не впливаючи на завчасність. Динаміка самого графа не моделюється.

– Незалежність користувачів у виведенні. Фільтрація виконується за кожним користувачем зі зв’язком через доданок (7) із затримкою на крок; змова кількох осіб як спільна подія не моделюється.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Розроблено метод темпорального раннього прогнозування внутрішніх інцидентів, який замінює бінарну класифікацію оцінюванням каліброваної ймовірності настання події в межах горизонту прогнозу. Часові ряди поведінкової телеметрії та розріджених OSINT-сигналів агрегуються ієрархічною байєсівською моделлю простору станів із послідовним виведенням фільтром частинок, а структура організаційного графа враховується гетерофілійно-свідомим контрастом і структурним пріором замість гомофілійного усереднення сусідів.

Обчислювальний експеримент показав, що за фіксованого бюджету 0,1 хибної тривоги на користувача виявлення зростає з 0,11 для самої лише телеметрії до 0,62 після додавання відкритих сигналів і до 0,66 після врахування структури, причому останнє досягається за нижчого фактичного навантаження тривогами; площа під ROC-кривою становить 0,957, під кривою «точність – повнота» – 0,275, медіанне випередження – близько восьми діб. Посткалібрування знижує оцінку Браєра з 0,037 до 0,004 та очікувану похибку калібрування з 0,158 до 0,002, роблячи вихід придатним для встановлення порогів втручання.

Уперше для цього класу задач кількісно оцінено змагальну стійкість темпорального раннього прогнозування. Установлено, що приховування та підроблення відкритого цифрового сліду знижують виявлення до 0,01–0,06, тоді як приховування зв’язків є значно менш дієвим; при цьому ранжування частково зберігається (площа під ROC-кривою 0,74 проти 0,49–0,55 у базових моделей) завдяки структурному складнику. Перевірені механізми компенсації (довірче злиття та масковане донавчання) відновлюють стійкість лише частково і ціною чутливості за чистих умов, що визначає межі застосовності всього класу OSINT-орієнтованих методів.

Усі оцінки одержано на синтетичних даних, і статус верхньої межі не знято через недоступність наборів CERT, TWOS та Enron у середовищі виконання; наведено детальний протокол їх валідації. Перспективу подальших досліджень становлять експериментальне дослідження на реальних даних, заміна посткалібрування на конформний контроль ризику з формальними гарантіями рівня хибних спрацювань, перехід до режимно-перемикальної моделі латентного стану для розрізнення сценаріїв інциденту, а також дослідження адаптивного змагальника, який ураховує наявність захисних механізмів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kirichenko, L., Radivilova, T., & Bulakh, V. (2019). Machine learning in classification time series with fractal properties. *Data, 4*(1), Article 5, 1–13. https://doi.org/10.3390/data4010005
2. Radivilova, T., Kirichenko, L., Alghawli, A. S., Ilkov, A., Tawalbeh, M., & Zinchenko, P. (2020). The complex method of intrusion detection based on anomaly detection and misuse detection. In *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 133–137). IEEE. https://doi.org/10.1109/DESSERT50317.2020.9125051

3. Kirichenko, L., Pichugina, O., Radivilova, T., & Pavlenko, K. (2023). Application of wavelet transform for machine learning classification of time series. In S. Babichev & V. Lytvynenko (Eds.), *Lecture notes in data engineering, computational intelligence, and decision making. ISDMCI 2022* (Lecture Notes on Data Engineering and Communications Technologies, Vol. 149, pp. 445–458). Springer, Cham. https://doi.org/10.1007/978-3-031-16203-9_31
4. Пантєлєєв, В., Радівілова, Т., Добринін, І., Фісенко, Д., Мазепа, А., & Білодід, В. (2025). Аналіз методів прогнозування внутрішніх загроз на основі аналізу даних соціальної мережі TWITTER. *Кібербезпека: освіта, наука, техніка*, *4*(28), 478–489. https://doi.org/10.28925/2663-4023.2025.28.818
5. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, Article 102221. https://doi.org/10.1016/j.cose.2021.102221
6. Daradkeh, Y. I., Kirichenko, L., & Radivilova, T. (2018). Development of QoS methods in the information networks with fractal traffic. *International Journal of Electronics and Telecommunications*, *64*(1), 27–32. https://doi.org/10.24425/118142
7. Radivilova, T., Kirichenko, L., Panteleev, V., & Mazepa, A. (2024). Analysis of authentication methods for full-stack applications and implementation of a web application with an integrated authentication system. *Innovative Technologies and Scientific Solutions for Industries*, (3), 76–90. https://doi.org/10.30837/ITSSI.2024.3.076
8. Nasir, R., Afzal, M., Latif, R., & Iqbal, W. (2021). Behavioral based insider threat detection using deep learning. *IEEE Access*, *9*, 143266–143274. https://doi.org/10.1109/ACCESS.2021.3118297
9. Al-Mhiqani, M. N., Ahmad, R., Abidin, Z. Z., Abdulkareem, K. H., Mohammed, M. A., Gupta, D., & Shankar, K. (2022). A new intelligent multilayer framework for insider threat detection. *Computers & Electrical Engineering*, *97*, Article 107597. https://doi.org/10.1016/j.compeleceng.2021.107597
10. Gong, Y., Cui, S., Liu, S., Jiang, B., Dong, C., & Lu, Z. (2024). Graph-based insider threat detection: A survey. *Computer Networks*, *254*, Article 110757. https://doi.org/10.1016/j.comnet.2024.110757
11. Fei, K., Zhou, J., Su, L., Wang, W., & Chen, Y. (2025). Log2Graph: A graph convolution neural network based method for insider threat detection. *Journal of Computer Security*, *33*(2). https://doi.org/10.3233/JCS-230092
12. Thi, V. D., Duc, T. T., Nguyen, T. V., & Luong, H.-M. P. (2026). Insider threat detection using knowledge graphs and RiskScore-guided graph neural networks. *Engineering, Technology & Applied Science Research*, *16*(2), 33712–33721. https://doi.org/10.48084/etasr.17187
13. Dwivedi, V. P., & Bresson, X. (2021). *A generalization of transformer networks to graphs*. arXiv. https://doi.org/10.48550/arXiv.2012.09699
14. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., & Liu, T.-Y. (2021). Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 28877–28888). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2106.05234
15. Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., & Beaini, D. (2022). Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 14501–14515). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2205.12454
16. Xu, F., Wang, N., Wu, H., Wen, X., Zhao, X., & Wan, H. (2024). Revisiting graph-based fraud detection in sight of heterophily and spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 8, pp. 9214–9222). AAAI Press. https://doi.org/10.1609/aaai.v38i8.28773
17. Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., & Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 7793–7804). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2006.11468
18. Chien, E., Peng, J., Li, P., & Milenkovic, O. (2021). *Adaptive universal generalized PageRank graph neural network*. arXiv. https://doi.org/10.48550/arXiv.2006.07988
19. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (pp. 1321–1330). PMLR. https://doi.org/10.48550/arXiv.1706.04599
20. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, *16*(4), 494–591. https://doi.org/10.1561/2200000101
21. Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 1660–1672). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2106.00170
22. Candès, E., Lei, L., & Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B*, *85*(1), 24–45. https://doi.org/10.1093/jrsssb/qkac004

23. Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2847–2856). ACM. https://doi.org/10.1145/3219819.3220078
24. Freitas, S., Yang, D., Kumar, S., Tong, H., & Chau, D. H. (2022). Graph vulnerability and robustness: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 5915–5934. https://doi.org/10.1109/TKDE.2022.3163672
25. Glasser, J., & Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops* (pp. 98–104). IEEE. https://doi.org/10.1109/SPW.2013.37
26. Harilal, A., Toffalini, F., Castellanos, J., Guarnizo, J., Homoliak, I., & Ochoa, M. (2017). TWOS: A dataset of malicious insider threat behavior based on a gamified competition. In *Proceedings of the 2017 International Workshop on Managing Insider Security Threats* (pp. 45–56). ACM. https://doi.org/10.1145/3139923.3139929
27. Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Machine Learning: ECML 2004* (Lecture Notes in Computer Science, Vol. 3201, pp. 217–226). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_22

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023

Tamara Radivilova
Dr. eng. sc., prof., professor of V.V. Popovskyy department of infocommunication engineering,
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID: 0000-0001-5975-0269
tamara.radivilova@gmail.com

Pantelieiev Vadym
Postgraduate student of the V.V. Popovskyy department of infocommunication engineering,
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID: 0009-0008-7824-8782
vadym.pantielieiev@nure.ua

A METHOD FOR TEMPORAL EARLY FORECASTING OF INSIDER INCIDENTS WITH RISK PROBABILITY CALIBRATION

Abstract. This paper proposes a temporal early-warning forecasting method for insider incidents that transitions from binary "incident / non-incident" classification to estimating risk dynamics over time. Time series of behavioural telemetry and sparse open-source intelligence (OSINT) signals are aggregated using a hierarchical Bayesian state-space model, in which the user's latent risk evolves as an autoregressive process and observations from different modalities are incorporated via their respective likelihood functions. The organizational graph structure is accounted for not by homophilic neighbour averaging, but rather through a heterophily-aware "node-versus-neighbourhood" contrast and a structural prior aligned with the empirically established property of insider connections: anomalous edges predominantly connect the malicious insider to benign custodians of sensitive resources. The output of the method is a calibrated posterior probability of an incident occurring within the forecast horizon, while the alarm rule is constructed under a controlled false-alarm budget that requires signal persistence over multiple days. A computational experiment on a temporal dataset parameterized according to the CERT specification (1,000 users, 420 days, five runs) demonstrated that at a fixed load of 0.1 false alarms per user, telemetry alone detects 0.11 of insiders; incorporating OSINT signals raises this metric to 0.62, and the full proposed method reaches 0.66 with nearly half the false-alarm rate; the area under the ROC curve reaches 0.957, while temperature calibration reduces the Brier score from 0.037 to 0.004 and the expected calibration error from 0.158 to 0.002. Adversarial robustness is evaluated separately: concealment and spoofing of the public digital footprint drops detection down to 0.01–0.06, indicating that OSINT-based early warning is fundamentally vulnerable to evasion by the perpetrator, whereas user ranking is partially preserved due to the structural component. It is shown that credible fusion and masked fine-tuning compensate for this degradation only partially and at the expense of sensitivity under clean conditions.

Keywords: insider threats; early forecasting; lead time; probability calibration; modelling; graph; transformer; adversarial robustness; OSINT.

REFERENCES

1. Kirichenko, L., Radivilova, T., & Bulakh, V. (2019). Machine learning in classification time series with fractal properties. *Data*, *4*(1), Article 5, 1–13. https://doi.org/10.3390/data4010005
2. Radivilova, T., Kirichenko, L., Alghawli, A. S., Ilkov, A., Tawalbeh, M., & Zinchenko, P. (2020). The complex method of intrusion detection based on anomaly detection and misuse detection. In *2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 133–137). IEEE. https://doi.org/10.1109/DESSERT50317.2020.9125051
3. Kirichenko, L., Pichugina, O., Radivilova, T., & Pavlenko, K. (2023). Application of wavelet transform for machine learning classification of time series. In S. Babichev & V. Lytvynenko (Eds.), *Lecture notes in data engineering, computational intelligence, and decision making. ISDMCI 2022* (Lecture Notes on Data Engineering and Communications Technologies, Vol. 149, pp. 445–458). Springer, Cham. https://doi.org/10.1007/978-3-031-16203-9_31
4. Panteleyev, V., Radivilova, T., Dobrynin, I., Fisenko, D., Mazepa, A., & Bilodid, V. (2025). Analiz metodiv prohnozuvannya vnutrishnikh zahroz na osnovi analizu danykh sotsial′noyi merezhi TWITTER [Analysis of methods for predicting internal threats based on data analysis of the social network TWITTER].

Cybersecurity: education, science, technology, *4*(28), 478–489. https://doi.org/10.28925/2663-4023.2025.28.818

5. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, Article 102221. https://doi.org/10.1016/j.cose.2021.102221
6. Daradkeh, Y. I., Kirichenko, L., & Radivilova, T. (2018). Development of QoS methods in the information networks with fractal traffic. *International Journal of Electronics and Telecommunications*, *64*(1), 27–32. https://doi.org/10.24425/118142
7. Radivilova, T., Kirichenko, L., Panteleev, V., & Mazepa, A. (2024). Analysis of authentication methods for full-stack applications and implementation of a web application with an integrated authentication system. *Innovative Technologies and Scientific Solutions for Industries*, (3), 76–90. https://doi.org/10.30837/ITSSI.2024.3.076
8. Nasir, R., Afzal, M., Latif, R., & Iqbal, W. (2021). Behavioral based insider threat detection using deep learning. *IEEE Access*, *9*, 143266–143274. https://doi.org/10.1109/ACCESS.2021.3118297
9. Al-Mhiqani, M. N., Ahmad, R., Abidin, Z. Z., Abdulkareem, K. H., Mohammed, M. A., Gupta, D., & Shankar, K. (2022). A new intelligent multilayer framework for insider threat detection. *Computers & Electrical Engineering*, *97*, Article 107597. https://doi.org/10.1016/j.compeleceng.2021.107597
10. Gong, Y., Cui, S., Liu, S., Jiang, B., Dong, C., & Lu, Z. (2024). Graph-based insider threat detection: A survey. *Computer Networks*, *254*, Article 110757. https://doi.org/10.1016/j.comnet.2024.110757
11. Fei, K., Zhou, J., Su, L., Wang, W., & Chen, Y. (2025). Log2Graph: A graph convolution neural network based method for insider threat detection. *Journal of Computer Security*, *33*(2). https://doi.org/10.3233/JCS-230092
12. Thi, V. D., Duc, T. T., Nguyen, T. V., & Luong, H.-M. P. (2026). Insider threat detection using knowledge graphs and RiskScore-guided graph neural networks. *Engineering, Technology & Applied Science Research*, *16*(2), 33712–33721. https://doi.org/10.48084/etasr.17187
13. Dwivedi, V. P., & Bresson, X. (2021). *A generalization of transformer networks to graphs*. arXiv. https://doi.org/10.48550/arXiv.2012.09699
14. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., & Liu, T.-Y. (2021). Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 28877–28888). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2106.05234
15. Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., & Beaini, D. (2022). Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 14501–14515). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2205.12454
16. Xu, F., Wang, N., Wu, H., Wen, X., Zhao, X., & Wan, H. (2024). Revisiting graph-based fraud detection in sight of heterophily and spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 8, pp. 9214–9222). AAAI Press. https://doi.org/10.1609/aaai.v38i8.28773
17. Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., & Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 7793–7804). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2006.11468
18. Chien, E., Peng, J., Li, P., & Milenkovic, O. (2021). *Adaptive universal generalized PageRank graph neural network*. arXiv. https://doi.org/10.48550/arXiv.2006.07988
19. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (pp. 1321–1330). PMLR. https://doi.org/10.48550/arXiv.1706.04599
20. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, *16*(4), 494–591. https://doi.org/10.1561/2200000101
21. Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 1660–1672). Curran Associates, Inc. https://doi.org/10.48550/arXiv.2106.00170
22. Candès, E., Lei, L., & Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B*, *85*(1), 24–45. https://doi.org/10.1093/jrsssb/qkac004
23. Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2847–2856). ACM. https://doi.org/10.1145/3219819.3220078
24. Freitas, S., Yang, D., Kumar, S., Tong, H., & Chau, D. H. (2022). Graph vulnerability and robustness: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *35*(6), 5915–5934. https://doi.org/10.1109/TKDE.2022.3163672

25. Glasser, J., & Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops* (pp. 98–104). IEEE. https://doi.org/10.1109/SPW.2013.37
26. Harilal, A., Toffalini, F., Castellanos, J., Guarnizo, J., Homoliak, I., & Ochoa, M. (2017). TWOS: A dataset of malicious insider threat behavior based on a gamified competition. In *Proceedings of the 2017 International Workshop on Managing Insider Security Threats* (pp. 45–56). ACM. https://doi.org/10.1145/3139923.3139929
27. Klimt, B., & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Machine Learning: ECML 2004* (Lecture Notes in Computer Science, Vol. 3201, pp. 217–226). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30115-8_22

Отримано редакцією журналу / Received: 19.02.26
Прорецензовано / Revised: 26.02.26
Схвалено до друку / Accepted: 24.09.26