КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (34), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.34.1333
УДК 004.056.5:004.89

**Партика Ольга Олександрівна**
кандидат фізико-математичних наук, доцент, доцент кафедри захисту інформації
Національний університет «Львівська політехніка», Львів, Україна
ORCID:0000-0002-3086-3160
*olha.o.mykhailova@lpnu.ua*

# РИЗИК-ОРІЄНТОВАНЕ ВИЯВЛЕННЯ КІБЕРАТАК У БЕЗПРОВОДОВИХ МЕРЕЖАХ З УРАХУВАННЯМ СТІЙКОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

**Анотація.** Детектори вторгнень для безпроводових мереж зазвичай оцінюють на незмінених тестових даних. Така перевірка не враховує ситуації, коли противник цілеспрямовано змінює доступні ознаки мережевого запису або використовує пояснення моделі, щоб знайти чутливий напрям впливу. За цих умов висока точність на тестовій вибірці ще не означає, що детектор є стійким. У статті розглянуто ризик-орієнтовану схему, до якої входять базовий класифікатор із градієнтним підсиленням, SHAP-пояснення, логістичний метакласифікатор ризику та передавання частини подій аналітику. Ризик оцінюється за трьома сигналами: імовірністю атаки, стабільністю локального пояснення і відстанню Махаланобіса до навчального розподілу. Завдяки такому поєднанню враховується не тільки прогноз, а й те, наскільки звичною для моделі є логіка рішення та сам запис. На ручну перевірку передавали фіксовані 5 % подій, відібраних за нестабільністю пояснення, невпевненістю прогнозу або предиктивною ентропією. Експерименти проведено на синтетичному наборі та NSL-KDD; перевірено однокрокове безградієнтне SHAP-sign збурення і перенесену PGD-атаку. Результати залежать і від набору даних, і від способу впливу. На синтетичних даних кращими були правила, побудовані за невпевненістю та ентропією. На чистій вибірці NSL-KDD і після SHAP-sign впливу кориснішим виявився відбір за стабільністю SHAP. Проте за перенесеної PGD-атаки селективна перевірка не дала виграшу: macro-F1 для C3 становила 0,699, для C4 – 0,696, а для C4-conf/ent – 0,694. Отже, правило ручної перевірки, вибране лише за чистою валідаційною вибіркою, може не спрацювати після зміни розподілу або типу атаки. Селективний режим допомагає керувати ризиком на автоматично опрацьованій частині потоку, але сам по собі не забезпечує стійкості до змагальних впливів. Для практичного використання потрібна перевірка на сучасному трафіку Wi-Fi, IoT, 5G і 6G, зокрема в умовах, коли противнику відома логіка метакласифікатора.
**Ключові слова:** кібербезпека; безпроводові мережі; система виявлення вторгнень; машинне навчання; змагальна атака; SHAP; селективна класифікація; оцінювання ризику.

## ВСТУП

Постановка проблеми. Машинне навчання вже стало звичним інструментом у системах виявлення вторгнень для стільникових мереж, Wi-Fi та Інтернету речей. Такі моделі знаходять у мережевих з'єднаннях залежності, які важко задати наперед у вигляді сигнатур. Проте результат на відомій тестовій вибірці мало говорить про поведінку детектора після розгортання. Робочий трафік змінюється, і разом із ним можуть втрачати каліброваність оцінки, одержані під час навчання. Додатковий ризик створює навмисна зміна керованих ознак запису, спрямована на зниження ймовірності розпізнавання атаки. Протокольні вимоги та властивості фізичного каналу звужують можливості такого впливу в безпроводовій мережі, але не усувають їх повністю [1, 2].

Пояснюваність у цьому контексті має не лише захисну цінність. Для аналітика SHAP або LIME показує, які ознаки вплинули на віднесення запису до атаки. Та сама інформація може бути корисною й противнику, оскільки вказує на найчутливіші місця моделі. У праці [3], наприклад, локальні пояснення використано для зондування мережевого детектора, доступного лише як «чорна скринька». Тому пояснювальний модуль варто розглядати одночасно як засіб аналізу і як можливу частину поверхні атаки.

Є й суто операційне обмеження: центр моніторингу не може вручну перевіряти кожне сумнівне сповіщення. Якщо таких подій надто багато, черга швидко перевищує реальну спроможність аналітиків.

Селективна класифікація розв'язує це питання частково — система утримується від автоматичного рішення лише для певної частини записів, а її роботу оцінюють через співвідношення ризику і покриття [4, 5]. Водночас невпевненість прогнозу, нестабільність пояснення та віддаленість запису від навчального розподілу не однаково реагують на цілеспрямовані збурення. Саме цю різницю досліджено далі.

Аналіз останніх досліджень і публікацій. У попередніх роботах можна виокремити дві близькі, але здебільшого розділені лінії досліджень. Перша стосується пояснюваного виявлення вторгнень. В. Мохале та І. Обагбува [6], а також О. Аррече, Т. Гунтур, Дж. Робертс і М. Абдалла [7] застосували SHAP для тлумачення рішень мережевих детекторів, не перевіряючи окремо стійкість пояснень до цілеспрямованих збурень. Для безпроводових сенсорних мереж С. Бірахім та співавтори [8] поєднали ансамблеву модель, оптимізацію і пояснення. Це підвищує прозорість рішення, однак не відповідає на практичне питання: які саме події слід першими передати аналітику, коли ресурс ручної перевірки обмежений. Друга лінія охоплює змагальне машинне навчання. Д. Адесіна, К.-К. Сє, Є. Сагдую та Л. Цянь [1] систематизували атаки на класифікатори радіосигналів, а К. Хе, Д. Кім і М. Асгар [2] – загрози для мережевих систем виявлення вторгнень. Можливість перенесення збурень між моделями досліджено у праці [9]. А. Мадрі та співавтори [10] показали, чому захисний механізм потрібно випробовувати за участю достатньо сильного противника. Сертифіковані підходи, зокрема рандомізоване згладжування [11], дають гарантії для наперед заданої норми збурення та конкретної моделі загроз, але не визначають, як розподілити сповіщення між автоматичною системою й людиною. Рамкова настанова NIST AI RMF [20] описує загальні принципи керування ризиком, тоді як вибір робочого сигналу для селективної перевірки залишається завданням власника конкретної системи.

Табл. 1 стисло зіставляє ці підходи. Помітна прогалина полягає в тому, що базовий детектор, пояснювальний модуль і правило передавання подій аналітику зазвичай оцінюють окремо. Так само незрозуміло, чи зберігається перевага одного сигналу ризику над іншим після зміни способу атаки.

*Таблиця 1*

**Порівняння підходів до пояснюваного та стійкого виявлення атак**

| Джерело | Контекст | Основний результат | Обмеження відносно цієї роботи |
|---|---|---|---|
| [3] | Мережевий IDS; SHAP/LIME | Пояснення використано для зондування за моделлю «чорної скриньки» | Не досліджено селективну перевірку та безпроводові обмеження |
| [6] | Мережевий IDS; XAI | Оцінено прозорість рішень детектора | Основний акцент на чистих даних |
| [8] | Безпроводова сенсорна мережа | Ансамбль із SHAP та оптимізацією | Немає оцінки адаптивних атак і невизначеності |
| [1] | Радіочастотні дані; огляд AML | Систематизовано поверхні й обмеження атак | Не визначено правило ручної перевірки |
| [11] | Класифікація; randomized smoothing | Гарантія для заданої норми збурення | Не охоплює ризик-орієнтовану маршрутизацію подій |
| [5] | Селективна класифікація | Формалізовано компроміс ризик-покриття | Не враховано атаковані пояснення та IDS |

Мета статті – розробити й експериментально перевірити ризик-орієнтовану схему виявлення кібератак, у якій прогноз базового детектора, стабільність SHAP-пояснення та відстань до навчального розподілу використовуються для відбору обмеженої частини подій на ручну перевірку. Для цього сформовано модель загроз, задано конфігурації конвеєра і правила відбору, проведено порівняння на чистих та атакованих даних і визначено межі, в яких одержані результати можна тлумачити.

## ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Розглядається двокласовий детектор, який за вектором мережевих ознак оцінює ймовірність належності запису до класу атак. Порогове значення для базового рішення дорівнює 0,5. Противник взаємодіє з моделлю як із «чорною скринькою» або має лише обмежені відомості про її внутрішні параметри; змінювати він може тільки числові ознаки в допустимих межах.

Це припущення потребує важливого застереження. Вектор може задовольняти математичні обмеження і водночас бути нереалізовним у живій мережі. Ознаки, пов'язані з протоколом, часом або агрегуванням потоку, повинні залишатися семантично узгодженими. Тому результати нижче

характеризують стійкість обчислювального конвеєра до заданих збурень, але не підтверджують фізичної здійсненності кожного сформованого запису.

Для кожного запису TreeSHAP формує вектор локальних атрибуцій [13]. На відміну від LIME [14], де локальне рішення апроксимується окремою моделлю, тут використано точний для деревного ансамблю алгоритм. Стабільність пояснення визначається косинусною подібністю цього вектора до центроїда пояснень відповідного класу, обчисленого за навчальними спостереженнями. Низька подібність означає, що прогноз спирається на незвичне поєднання ознак. Інший сигнал – відстань Махаланобіса; вона враховує коваріаційну структуру даних і зростає для записів, віддалених від навчального розподілу.

$$\mathrm{c}_y = \frac{1}{N_y} \sum_{i:\, y_i = y} \varphi\,(\mathrm{x}_i) \tag{1}$$

$$s_{\mathrm{SHAP}}(\mathrm{x}) = \frac{\varphi(\mathrm{x})^{\mathrm{T}} \mathrm{c}_{\hat{y}}}{\| \varphi(\mathrm{x}) \|_2 \, \| \mathrm{c}_{\hat{y}} \|_2} \tag{2}$$

$$d_{\mathrm{M}}(\mathrm{x}) = [(\mathrm{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathrm{x} - \mu)]^{1/2} \tag{3}$$

На вхід метакласифікатора подаються базова ймовірність атаки, стабільність SHAP і відстань Махаланобіса. Логістичну регресію обрано через її прозорість: внесок трьох сигналів можна відокремити від роботи складнішого базового ансамблю, а невелика кількість ознак обмежує ризик прихованого перенавчання. Вихід конфігурації C3 позначено r(x). За ним у формулах (5) і (6) визначаються невпевненість u(x) та предиктивна ентропія H(x). Далі записи ранжуються за одним із трьох сигналів g(x), наведених у формулі (7), і 5 % із найвищим пріоритетом передаються аналітику. Важливо, що r(x) є оцінкою C3, а не окремим правилом відбору. Показники селективного рішення обчислюються для решти 95 % потоку, де система зберігає автоматичне рішення.

$$r(\mathrm{x}) = \sigma\big(\beta_0 + \beta_1 p(\mathrm{x}) + \beta_2 s_{\mathrm{SHAP}}(\mathrm{x}) + \beta_3 d_{\mathrm{M}}(\mathrm{x})\big) \tag{4}$$

$$u(\mathrm{x}) = 1 - 2|r(\mathrm{x}) - 0{,}5| \tag{5}$$

$$H(\mathrm{x}) = -r(\mathrm{x}) \ln r(\mathrm{x}) - [1 - r(\mathrm{x})] \ln [1 - r(\mathrm{x})] \tag{6}$$

$$g(\mathrm{x}) \in \{1 - s_{\mathrm{SHAP}}(\mathrm{x}),\, u(\mathrm{x}),\, H(\mathrm{x})\} \tag{7}$$

$$R_\rho^{(g)} = \{\mathrm{x}: g(\mathrm{x}) \geq q_{1-\rho}^{(g)}\}, \quad \rho = 0{,}05 \tag{8}$$

$$A_\rho^{(g)} = \mathcal{X}_{\mathrm{test}} \setminus R_\rho^{(g)} \tag{9}$$

$$\mathrm{coverage}(\rho, g) = \frac{\left|A_\rho^{(g)}\right|}{N} = 1 - \rho \tag{10}$$

$$\mathrm{risk}_{\mathrm{sel}}(\rho, g) = \frac{1}{\left|A_\rho^{(g)}\right|} \sum_{\mathbf{x}_i \in A_\rho^{(g)}} \ell\,(\hat{y}_i, y_i) \tag{11}$$

У конфігурації C4 сигнал g(x)=1−sSHAP(x) виділяє записи з нетиповою логікою рішення. C4-conf використовує невпевненість u(x), тобто найменше абсолютне відхилення ймовірності C3 від 0,5, а C4-ent – предиктивну ентропію H(x). У двокласовому завданні u(x) і H(x) дають однаковий порядок записів, тому ці два варіанти мають той самий склад ручної перевірки й надалі подаються разом. Більше g(x) завжди відповідає вищому пріоритету для аналітика. Множини ручної та автоматичної обробки задано у формулах (8) і (9), покриття та селективний ризик – у формулах (10) і (11). Послідовність оброблення нової події показано на рисунку 1.

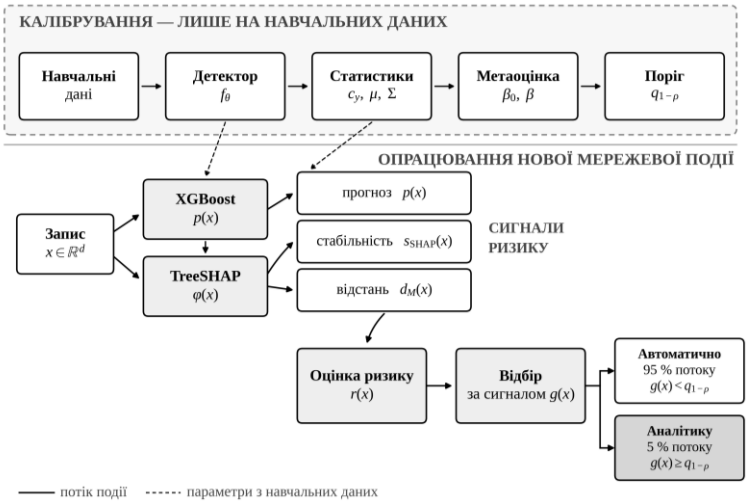


*Рис. 1. Архітектура ризик-орієнтованого конвеєра виявлення кібератак із використанням базового прогнозу, стабільності SHAP-пояснення та відстані Махаланобіса*

## МЕТОДИКА ДОСЛІДЖЕННЯ

Експеримент проведено у двох частинах. Синтетичний набір містив 10 000 двокласових записів із 20 ознаками: 30 % атак і 70 % доброякісних спостережень. Для нього використано п'ятикратну стратифіковану крос-валідацію, повторену двічі, тобто загалом 10 оцінювальних розбиттів. Реальну частину побудовано на офіційних вибірках KDDTrain+ і KDDTest+ набору NSL-KDD. Словник категорій формували без міток, а параметри масштабування числових ознак оцінювали лише за KDDTrain+. Після кодування кожен запис мав 122 ознаки, з яких 38 були числовими; збурення вносили тільки у стандартизовані числові ознаки записів атак. Для NSL-KDD подано середні значення та стандартні відхилення за 10 bootstrap-вибірками з фіксованої KDDTest+ (seed 42).

Базовим детектором був XGBoost із 200 деревами максимальної глибини 4, локальні атрибуції обчислювали методом TreeSHAP [13], а метакласифікатором слугувала логістична регресія за трьома сигналами ризику. Конфігурації конвеєра наведено в табл. 2. C2 є лише постфактумним пояснювальним шаром і повністю відтворює прогнози C1. Тому окремих показників для C2 немає: вони збігаються з C1. Для селективних конфігурацій усі метрики розраховано після передавання 5 % записів аналітику, тобто лише на автоматично опрацьованій частині потоку.

На NSL-KDD базовий детектор і сурогатну модель навчали за всією тренувальною вибіркою. Для SHAP-центроїдів і логістичного метакласифікатора використано фіксовану підвибірку з 20 000 тренувальних записів (seed 1). Натомість середнє та коваріаційну матрицю, потрібні для відстані Махаланобіса, оцінювали за повною KDDTrain+. Після навчання параметри не змінювали. Тому bootstrap тут відображає мінливість повторного відбору з однієї тестової вибірки, а не розкид між окремими навчаннями моделі.

Обчислення виконано з такими версіями бібліотек: NumPy 1.26, pandas 2.2, SciPy 1.13, scikit-learn 1.5.2, XGBoost 2.1.1 і SHAP 0.46.0.

*Таблиця 2*

**Конфігурації експериментального конвеєра**

| ID | Конфігурація | Зміст |
|---|---|---|
| C1 | Базовий класифікатор | XGBoost; рішення без пояснення та селективної перевірки |
| C2 | C1 + SHAP | Прогноз C1 не змінюється; додаються локальні атрибуції |
| C3 | Метакласифікатор ризику | Логістична регресія за ймовірністю, стабільністю SHAP та відстанню Махаланобіса |
| C4 | C3 + перевірка за стабільністю | На перевірку передаються 5 % записів із найнижчою стабільністю пояснення |
| C4-conf | C3 + перевірка за невпевненістю | На перевірку передаються 5 % найменш упевнених прогнозів |
| C4-ent | C3 + перевірка за ентропією | На перевірку передаються 5 % записів із найбільшою ентропією; для двох класів збігається з C4-conf |

У межах експерименту перевірено два сценарії. Перший – однокрокове безградієнтне SHAP-sign збурення, де напрям зміни числових ознак задають локальні атрибуції. У другому сценарії PGD-збурення будували на диференційовній сурогатній моделі, а потім переносили на деревний детектор [9]. В обох випадках змінювали лише записи атак і тільки ті числові ознаки, для яких збурення було допустимим. Адаптивну атаку на пояснення зі збереженням прогнозованого класу не моделювали; її перевірка на повному конвеєрі залишається окремим завданням.

$$\begin{aligned} \mathrm{v}(\mathrm{x}) &= \eta\,\mathrm{sign}\big(\varphi(\mathrm{x})\big), \quad \eta \in \{-1,+1\}, \\ \mathrm{x}_{\mathrm{adv}} &= \Pi_{\mathcal{X}\cap B_\infty(\mathrm{x},\varepsilon)}\big(\mathrm{x}+\varepsilon\mathrm{v}(\mathrm{x})\big), \quad \varepsilon = 0{,}10. \end{aligned} \tag{12}$$

У формулі (12) v(x) задає напрям SHAP-sign збурення, а η – його глобальну орієнтацію. Спочатку орієнтацію перевірили скінченнорізницевою пробою; у звітних запусках використано η=+1. Для SHAP-sign задано ε=0,10 у стандартизованих одиницях і виконано один проєктований крок. Знаки SHAP-атрибуцій переважно не збігалися зі знаками скінченнорізницевих похідних імовірності атаки. Тому правило знака не переносили механічно з градієнтної атаки: застосували напрям, попередньо перевірений щодо потрібної зміни прогнозу. Саме це пояснює результативність SHAP-sign атаки за η=+1 і відсутність суперечності між спостережуваними знаками.

Перенесену PGD-атаку виконували за T=10 і кроку α=ε/4=0,025, проєктуючи запис після кожної ітерації. Категоріальні ознаки в обох сценаріях залишалися незмінними. Схему однокрокового SHAP-sign збурення разом із перевіркою орієнтації наведено на рисунку 2.


*Рис. 2. Однокрокова процедура безградієнтного SHAP-sign збурення з перевіреною орієнтацією знака та проєкцією на множину допустимих значень*

Якість оцінювали за macro-F1, precision, recall, частотою хибнопозитивних рішень (FPR), частотою хибнонегативних рішень (FNR) та площею під ROC-кривою (AUC). Для селективних конфігурацій окремо розглядали помилку на покритій, тобто автоматично опрацьованій, частині. Частка ручної перевірки в усіх варіантах однакова. Тому різниця між C4 і C4-conf/ent показує якість сигналу ранжування, а не ефект від вилучення більшої кількості записів. Правило формування автоматичної частини за фіксованого покриття показано на рисунку 3.


*Рис. 3. Ранжування мережевих подій за узагальненим сигналом g(x) і формування автоматично опрацьованої частини потоку за фіксованої частки ручної перевірки ρ=0,05*

## РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати синтетичного експерименту подано в табл. 3. На чистих даних macro-F1 зросла з 0,912 ± 0,007 для C1 до 0,916 ± 0,006 для C3. Після вилучення однакових 5 % записів C4 дала 0,925 ± 0,007, а

C4-conf/ent – 0,938 ± 0,005. Порядок конфігурацій не змінився ні після SHAP-sign збурення, ні після перенесеної PGD-атаки. У випадку PGD одержано 0,882 ± 0,008 для C1, 0,889 ± 0,007 для C3, 0,898 ± 0,008 для C4 і 0,908 ± 0,008 для C4-conf/ent. Для цього контрольованого розподілу близькість прогнозу до межі рішення виявилася кориснішою для відбору, ніж нетиповість локального пояснення.

*Таблиця 3*

**Результати на синтетичному наборі за фіксованої частки ручної перевірки 5 %**

| Умова | Конфігурація | Macro-F1 | Precision | Recall | FPR | FNR | AUC |
|---|---|---|---|---|---|---|---|
| Чисті | C1 | 0,912± 0,007 | 0,923± 0,008 | 0,832± 0,017 | 0,030± 0,003 | 0,168± 0,017 | 0,968± 0,004 |
| Чисті | C3 | 0,916± 0,006 | 0,907± 0,008 | 0,857± 0,016 | 0,038± 0,003 | 0,143± 0,016 | 0,967± 0,004 |
| Чисті | C4* | 0,925± 0,007 | 0,927± 0,010 | 0,859± 0,017 | 0,028± 0,004 | 0,141± 0,017 | 0,965± 0,004 |
| Чисті | C4-conf/ent* | 0,938± 0,005 | 0,935± 0,010 | 0,888± 0,012 | 0,026± 0,004 | 0,112± 0,012 | 0,967± 0,004 |
| SHAP-sign | C1 | 0,907± 0,008 | 0,922± 0,008 | 0,820± 0,019 | 0,030± 0,003 | 0,180± 0,019 | 0,967± 0,003 |
| SHAP-sign | C3 | 0,912± 0,006 | 0,906± 0,008 | 0,847± 0,015 | 0,038± 0,003 | 0,153± 0,015 | 0,966± 0,003 |
| SHAP-sign | C4* | 0,921± 0,007 | 0,927± 0,009 | 0,850± 0,015 | 0,028± 0,004 | 0,150± 0,015 | 0,964± 0,003 |
| SHAP-sign | C4-conf/ent* | 0,931± 0,006 | 0,933± 0,011 | 0,873± 0,011 | 0,026± 0,004 | 0,127± 0,011 | 0,966± 0,003 |
| PGD | C1 | 0,882± 0,008 | 0,916± 0,009 | 0,759± 0,017 | 0,030± 0,003 | 0,241± 0,017 | 0,953± 0,004 |
| PGD | C3 | 0,889± 0,007 | 0,900± 0,008 | 0,789± 0,016 | 0,038± 0,003 | 0,211± 0,016 | 0,952± 0,004 |
| PGD | C4* | 0,898± 0,008 | 0,921± 0,009 | 0,792± 0,018 | 0,028± 0,004 | 0,208± 0,018 | 0,950± 0,004 |
| PGD | C4-conf/ent* | 0,908± 0,008 | 0,924± 0,010 | 0,815± 0,020 | 0,027± 0,004 | 0,185± 0,020 | 0,952± 0,004 |

Примітка. C2 має ті самі прогнози й показники, що й C1, тому окремо не наведена. Зірочкою позначено селективні конфігурації: показники обчислено на автоматично опрацьованих 95 % записів (покриття 0,95).

Показники селективних конфігурацій мають сенс лише разом із покриттям 0,95. Вищий macro-F1 не означає, що система правильно розпізнала всі вилучені події: для цих записів автоматичне рішення взагалі не формувалося, їх передали аналітику. Такий режим придатний для центру моніторингу лише тоді, коли 5 % вхідного потоку відповідають реальній спроможності команди. Інакше виграш у метриках може обернутися чергою сповіщень і довшим часом реагування.

На NSL-KDD базовий детектор мав macro-F1 0,791 ± 0,003 на чистій вибірці. Однокрокове SHAP-sign збурення знизило цей показник до 0,579 ± 0,004, перенесена PGD-атака – до 0,637 ± 0,002. У прийнятій моделі загроз SHAP-керований вплив на деревний ансамбль, таким чином, був сильнішим за перенесення градієнтної атаки із сурогатної моделі. Результат стосується перевіреної орієнтації $\eta=+1$. Оскільки знаки атрибуцій переважно були протилежними знакам скінченнорізницевих похідних, неперевірене правило знака спрямувало б збурення у протилежний бік.

На NSL-KDD порядок правил відбору вже не повторював синтетичний експеримент. Для чистої вибірки та SHAP-sign впливу стабільність пояснення краще відокремлювала ризикові записи. Перенесена PGD-атака дала негативний результат: macro-F1 для C3 становила 0,699 ± 0,002, для C4 – 0,696 ± 0,003, а для C4-conf/ent – 0,694 ± 0,002. Отже, жоден із двох сигналів не виділив підмножину, вилучення якої поліпшило б якість на решті 95 % записів. Водночас перехід від C1 до C3 підвищив повноту виявлення атак із 0,408 ± 0,003 до 0,503 ± 0,004. На чистих даних macro-F1 C3 була лише трохи вищою за C1 (0,792 проти 0,791), тоді як AUC знизилася з 0,968 до 0,954. Це не суперечність: метакласифікатор дещо поліпшив рішення за фіксованого порога, але гірше впорядкував прогнози в усьому діапазоні порогів. Стабільність SHAP і відстань до навчального розподілу, отже, додають інформацію до базової ймовірності, але не утворюють універсального правила ранжування. Повні результати наведено в табл. 4.

*Таблиця 4*

**Результати на NSL-KDD за фіксованої частки ручної перевірки 5 %**

| Умова | Конфігурація | Macro-F1 | Precision | Recall | FPR | FNR | AUC |
|---|---|---|---|---|---|---|---|
| Чисті | C1 | 0,791± 0,003 | 0,968± 0,002 | 0,655± 0,003 | 0,029± 0,002 | 0,345± 0,003 | 0,968± 0,001 |
| Чисті | C3 | 0,792± 0,003 | 0,968± 0,002 | 0,656± 0,003 | 0,029± 0,002 | 0,344± 0,003 | 0,954± 0,001 |
| Чисті | C4* | 0,821± 0,003 | 0,968± 0,002 | 0,698± 0,003 | 0,028± 0,002 | 0,302± 0,003 | 0,957± 0,001 |
| Чисті | C4-conf/ent* | 0,809± 0,003 | 0,968± 0,002 | 0,676± 0,003 | 0,027± 0,002 | 0,324± 0,003 | 0,953± 0,001 |
| SHAP-sign | C1 | 0,579± 0,004 | 0,937± 0,005 | 0,324± 0,004 | 0,029± 0,002 | 0,676± 0,004 | 0,963± 0,002 |
| SHAP-sign | C3 | 0,578± 0,004 | 0,936± 0,005 | 0,322± 0,004 | 0,029± 0,002 | 0,678± 0,004 | 0,958± 0,002 |
| SHAP-sign | C4* | 0,599± 0,004 | 0,934± 0,005 | 0,339± 0,005 | 0,029± 0,002 | 0,661± 0,005 | 0,960± 0,002 |
| SHAP-sign | C4-conf/ent* | 0,569± 0,004 | 0,925± 0,005 | 0,296± 0,005 | 0,029± 0,002 | 0,704± 0,005 | 0,956± 0,002 |
| PGD | C1 | 0,637± 0,002 | 0,947± 0,003 | 0,408± 0,003 | 0,030± 0,002 | 0,592± 0,003 | 0,959± 0,002 |
| PGD | C3 | 0,699± 0,002 | 0,957± 0,003 | 0,503± 0,004 | 0,030± 0,002 | 0,497± 0,004 | 0,976± 0,002 |
| PGD | C4* | 0,696± 0,003 | 0,953± 0,003 | 0,486± 0,005 | 0,029± 0,002 | 0,514± 0,005 | 0,977± 0,002 |
| PGD | C4-conf/ent* | 0,694± 0,002 | 0,951± 0,003 | 0,483± 0,004 | 0,030± 0,002 | 0,517± 0,004 | 0,977± 0,002 |

Примітка. Позначення C2 і зірочки – як у табл. 3. Наведено середнє ± стандартне відхилення за 10 bootstrap-вибірками. Пооб'єктні прогнози вихідного запуску не збереглися, тому значення precision відновлено з агрегованих показників матриці помилок (macro-F1, recall і FPR); стандартне відхилення precision оцінено дельта-методом.

У всіх конфігураціях збільшення допустимого збурення погіршувало результати базового детектора, метакласифікатора і селективних варіантів. Передавання аналітику 5 % записів із найбільшим g(x) зменшувало частоту помилок серед тих подій, для яких система все ще формувала автоматичне рішення, але загальної деградації не зупиняло. Селективна перевірка, таким чином, перерозподіляє операційний ризик, а не усуває вразливість детектора. Для самої стійкості потрібні окремі механізми захисту та виявлення атакованих вхідних даних.

Для практичного впровадження правило відбору слід навчати й перевіряти на часово відокремлених даних конкретної мережі. У таку валідацію варто включати зміну обладнання, версій протоколів, профілів користувачів і частоти атак. Для радіосигналів потрібно враховувати обмеження каналу та фізичну здійсненність збурення, як у працях [17, 18], а для Wi-Fi – структуру реального набору атак [19]. Залишається й адаптивний сценарій: знаючи правило g(x), противник може одночасно оптимізувати прогноз детектора та сигнал відбору. Повної білоскринькової атаки такого типу наведений експеримент не охоплює.

Числові результати мають кілька суттєвих обмежень. NSL-KDD застарів і не відтворює робочий трафік сучасної безпроводової мережі. Самі збурення перевірено в межах математичної моделі, без окремої реалізації на протокольному або фізичному рівні. Крім того, експеримент охоплює одне сімейство базових моделей і лише одну частку ручної перевірки – 5 %. Тому наведені значення придатні для порівняння конфігурацій у відтворюваних умовах, але їх не слід читати як прогноз ефективності в мережах 5G чи 6G.

## ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Запропонована схема поєднує прогноз класифікатора з градієнтним підсиленням, стабільність SHAP-пояснення та відстань Махаланобіса. Після прогнозу C3 система передає аналітику 5 % записів із найбільшим значенням вибраного сигналу g(x). Експерименти показали, що одного найкращого правила для всіх умов немає. На синтетичних даних кориснішими були невпевненість прогнозу й еквівалентна їй

у двокласовому випадку ентропія. Для NSL-KDD стабільність пояснення краще працювала на чистих даних і після SHAP-sign впливу, але за перенесеної PGD-атаки жодна селективна конфігурація не перевищила C3 за macro-F1. Політика ручної перевірки, налаштована лише на чистій валідаційній вибірці, тому може втратити користь після зміни трафіку або способу атаки.

Практична користь селективної перевірки полягає в тому, що вона узгоджує потік сповіщень із ресурсом аналітиків. Її не можна оцінювати окремо від покриття, адже поліпшення метрик стосується тільки записів, для яких залишено автоматичне рішення. Подальша робота має охопити сучасні набори трафіку Wi-Fi, IoT, 5G і 6G, семантично допустимі збурення та різні частки ручної перевірки. Окремо потрібно перевірити адаптивні атаки на весь конвеєр, включно з цілеспрямованим спотворенням пояснення без зміни прогнозованого класу, а також калібрування ризику за умов часового дрейфу трафіку.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Adesina, D., Hsieh, C.-C., Sagduyu, Y. E., & Qian, L. (2023). Adversarial machine learning in wireless communications using RF data: A review. IEEE Communications Surveys & Tutorials, 25(1), 77–100. https://doi.org/10.1109/COMST.2022.3205184
2. He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE* Communications Surveys & Tutorials, *25*(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793
3. Senevirathna, T., Siniarski, B., Liyanage, M., & Wang, S. (2024). Deceiving post-hoc explainable AI (XAI) methods in network intrusion detection. In 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC) (pp. 107-112). IEEE. https://doi.org/10.1109/CCNC51664.2024.10454633
4. Chow, C. K. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41-46. https://doi.org/10.1109/TIT.1970.1054406
5. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4878-4887). NeurIPS proceedings
6. Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. Frontiers in Computer Science, 7, Article 1520741. https://doi.org/10.3389/fcomp.2025.1520741
7. Arreche, O., Guntur, T. R., Roberts, J. W., & Abdallah, M. (2024). E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection. IEEE Access, 12, 23954-23988. https://doi.org/10.1109/ACCESS.2024.3365140
8. Birahim, S. A., Paul, A., Rahman, F., Islam, Y., Roy, T., Hasan, M. A., Haque, F., & Chowdhury, M. E. H. (2025). Intrusion detection for wireless sensor network using particle swarm optimization based explainable ensemble machine learning approach. IEEE Access, *13*, 13711-13730. https://doi.org/10.1109/ACCESS.2025.3528341
9. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: From phenomena to black-box attacks using adversarial samples [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1605.07277
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR).
11. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 1310-1320). PMLR. PMLR proceedings
12. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1-6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774). NeurIPS proceedings
14. 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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778
15. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 274-283). PMLR. PMLR proceedings

16. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, *84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023
17. Kim, B., Sagduyu, Y. E., Davaslioglu, K., Erpek, T., & Ulukus, S. (2022). Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. IEEE Transactions on Wireless Communications, 21(6), 3868-3880. https://doi.org/10.1109/TWC.2021.3124855
18. Zhang, W., Krunz, M., & Ditzler, G. (2024). Stealthy adversarial attacks on machine learning-based classifiers of wireless signals. IEEE Transactions on Machine Learning in Communications and Networking, 2, 261-279. https://doi.org/10.1109/TMLCN.2024.3366161
19. Kolias, C., Kambourakis, G., Stavrou, A., & Gritzalis, S. (2016). Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset. IEEE Communications Surveys & Tutorials, 18*(*1), 184-208. https://doi.org/10.1109/COMST.2015.2402161
20. Tabassi, E. (2023*).* Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

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

**Olha Partyka**
Candidate of Physical and Mathematical Sciences (PhD), Associate Professor, Associate Professor at the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID:0000-0002-3086-3160
*olha.o.mykhailova@lpnu.ua*

## RISK-AWARE CYBERATTACK DETECTION IN WIRELESS NETWORKS CONSIDERING THE ROBUSTNESS OF MACHINE LEARNING MODELS

**Abstract.** Intrusion detectors for wireless networks are still commonly assessed on unchanged test sets. Such evaluation does not cover a deliberate modification of controllable traffic features or the use of model explanations to identify a vulnerable direction. A high test score therefore says little about detector behaviour under an active attack. This paper examines a risk-aware pipeline that combines a gradient-boosted base classifier, SHAP explanations, a logistic risk meta-classifier, and selective referral to a human analyst. Risk is described by three signals: the base attack probability, cosine stability of a local SHAP explanation relative to a class-specific training centroid, and Mahalanobis distance from the training distribution. This combination is intended to capture not only the predicted class, but also whether the reasoning pattern and the record itself are familiar to the model. A fixed five percent of records is referred for review according to explanation instability, predictive uncertainty, or predictive entropy. The experiments use a controlled synthetic dataset and NSL-KDD. Two attacks are evaluated: a one-step, gradient-free SHAP-sign perturbation and a transferred projected-gradient attack. The results are not uniform across datasets. On the synthetic benchmark, uncertainty and entropy provide the best selective results; under transferred PGD, macro-F1 on the retained 95 percent rises from 0.882 ± 0.008 for the base detector to 0.908 ± 0.008. NSL-KDD gives a different ordering. Explanation stability is more useful on clean and SHAP-sign-perturbed records, whereas under transferred PGD neither selective configuration improves on C3: macro-F1 is 0.699 for C3, 0.696 for C4, and 0.694 for C4-conf/ent. The base detector also deteriorates markedly, from 0.791 ± 0.003 on clean NSL-KDD to 0.579 ± 0.004 under SHAP-sign and 0.637 ± 0.002 under transferred PGD. These results do not support choosing a review rule from clean validation data alone. Selective referral can manage the accuracy–coverage trade-off, but it is not an adversarial defence. Before deployment, the pipeline should be tested on current Wi-Fi, IoT, 5G, and 6G traffic, with semantically valid perturbations, adaptive attacks, and review rates matched to analyst capacity.

**Keywords:** cybersecurity; wireless networks; intrusion detection system; machine learning; adversarial attack; SHAP; selective classification; risk assessment.

### REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Adesina, D., Hsieh, C.-C., Sagduyu, Y. E., & Qian, L. (2023). Adversarial machine learning in wireless communications using RF data: A review. IEEE Communications Surveys & Tutorials, 25(1), 77–100. https://doi.org/10.1109/COMST.2022.3205184
2. He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE* Communications Surveys & Tutorials, *25*(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793
3. Senevirathna, T., Siniarski, B., Liyanage, M., & Wang, S. (2024). Deceiving post-hoc explainable AI (XAI) methods in network intrusion detection. In 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC) (pp. 107-112). IEEE. https://doi.org/10.1109/CCNC51664.2024.10454633
4. Chow, C. K. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41-46. https://doi.org/10.1109/TIT.1970.1054406
5. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4878-4887). NeurIPS proceedings
6. Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. Frontiers in Computer Science, 7, Article 1520741. https://doi.org/10.3389/fcomp.2025.1520741

7. Arreche, O., Guntur, T. R., Roberts, J. W., & Abdallah, M. (2024). E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection. IEEE Access, 12, 23954-23988. https://doi.org/10.1109/ACCESS.2024.3365140
8. Birahim, S. A., Paul, A., Rahman, F., Islam, Y., Roy, T., Hasan, M. A., Haque, F., & Chowdhury, M. E. H. (2025). Intrusion detection for wireless sensor network using particle swarm optimization based explainable ensemble machine learning approach. IEEE Access, *13*, 13711-13730. https://doi.org/10.1109/ACCESS.2025.3528341
9. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: From phenomena to black-box attacks using adversarial samples [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1605.07277
10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR).
11. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 1310-1320). PMLR. PMLR proceedings
12. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1-6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774). NeurIPS proceedings
14. 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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778
15. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 274-283). PMLR. PMLR proceedings
16. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, *84*, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023
17. Kim, B., Sagduyu, Y. E., Davaslioglu, K., Erpek, T., & Ulukus, S. (2022). Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. IEEE Transactions on Wireless Communications, 21(6), 3868-3880. https://doi.org/10.1109/TWC.2021.3124855
18. Zhang, W., Krunz, M., & Ditzler, G. (2024). Stealthy adversarial attacks on machine learning-based classifiers of wireless signals. IEEE Transactions on Machine Learning in Communications and Networking, 2, 261-279. https://doi.org/10.1109/TMLCN.2024.3366161
19. Kolias, C., Kambourakis, G., Stavrou, A., & Gritzalis, S. (2016). Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset. IEEE Communications Surveys & Tutorials, 18*(*1), 184-208. https://doi.org/10.1109/COMST.2015.2402161
20. Tabassi, E. (2023*)*. Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Отримано редакцією журналу / Received: 29.01.26
Прорецензовано / Revised: 08.02.26
Схвалено до друку / Accepted: 24.09.26