УДК 004.855:004.056.53 DOI: 10.31673/2412-4338.2026.019019

Довженко Тимур Павлович
кандидат техн. наук, докторант
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID:0000-0002-0352-8391
t.dovzhenko@duikt.edu.ua

Зінченко Ольга Валеріївна
доктор техн. наук, професор, завідувач кафедри Штучного інтелекту
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID:0000-0002-3973-7814
o.zinchenko@duikt.edu.ua

СТАБІЛЬНІСТЬ МОДЕЛЕЙ ГЛИБОКОГО ВИЯВЛЕННЯ ВТОРГНЕНЬ В УМОВАХ МАСОВАНИХ КІБЕРАТАК: СТРЕС-ТЕСТУВАННЯ ТА АРХІТЕКТУРНІ ОСОБЛИВОСТІ HYBRID AWRED

Анотація: *Еволюція кіберзагроз досягла етапу, коли масовані DDoS-атаки та скоординована активність ботнетів здатні миттєво змінювати статистичний профіль мережевого трафіку, перетворюючи аномалію на домінуючий патерн поведінки. У критичних сценаріях, коли частка шкідливих пакетів сягає 50%, традиційні системи виявлення вторгнень (NIDS) на базі глибокого навчання втрачають ефективність. Фундаментальна гіпотеза методів Deep SVDD або DAGMM про рідкісність аномалій руйнується, що призводить до феномену «отруєння моделі»: нейромережа адаптується до щільного потоку атак, помилково сприймаючи його як нову норму. У цій роботі запропоновано комплексне вирішення проблеми адверсарної нестійкості - метод Hybrid AWRED. На відміну від існуючих підходів, наша архітектура базується на синергії трьох механізмів: адаптивного зважування помилок реконструкції, осцилюючої функції втрат та, що є вирішальним фактором, гібридизації вхідних даних.*

Вперше для стабілізації глибокої мережі застосовано технологію «топологічного якоря» (Density-Aware Input Augmentation): стандартний вектор із 41 ознаки розширено до 42-х шляхом інтеграції оцінки локальної щільності, попередньо обчисленої алгоритмом kNN. Ця метрична ознака надає моделі орієнтир, незалежний від глобального розподілу, дозволяючи розрізняти щільні кластери легітимного трафіку та атак навіть за умови їх рівного співвідношення. Результати стрес-тестування на синтетичному наборі даних «Hard Mode» підтвердили ефективність підходу: у сценарії екстремального забруднення (50%) конкурентні SOTA-методи деградували до рівня випадкового вгадування (AUC < 0.35) через колапс простору ознак, тоді як Hybrid AWRED зберіг високу роздільну здатність із показниками AUC-ROC 0.725 та Average Precision 0.619. Дослідження доводить, що інтеграція детермінованих метричних алгоритмів у структуру глибоких нейронних мереж є безальтернативним шляхом для створення надійних NIDS, здатних функціонувати у ворожому середовищі.

Ключові слова: *виявлення вторгнень, NIDS, глибоке навчання, Hybrid AWRED, стійкість до забруднення, гібридний простір ознак, Deep SVDD, kNN.*

Dovzhenko Timur
Candidate of Technical Sciences, Doctoral Student
State University of Information and Communication Technologies, Kyiv
ORCID:0000-0002-0352-8391
t.dovzhenko@duikt.edu.ua

Zinchenko Olha
Doctor of Technical Sciences, Professor, Head of the Department of Artificial Intelligence
State University of Information and Communication Technologies, Kyiv
ORCID:0000-0002-3973-7814
o.zinchenko@duikt.edu.ua

ROBUSTNESS OF DEEP INTRUSION DETECTION MODELS UNDER MASSIVE CYBERATTACKS: STRESS-TESTING AND ARCHITECTURAL FEATURES OF HYBRID AWRED

Abstract*: The evolutionary trajectory of modern cyber threats involves a shift from isolated intrusion attempts to sophisticated, massive cyber operations capable of fundamentally altering the statistical nature of network traffic. In the*

context of large-scale DDoS campaigns or synchronized global botnet activities, the proportion of malicious packets within the input stream can surge rapidly, effectively transforming the anomaly into the dominant behavioral pattern. This emerging reality creates a critical barrier for traditional Deep Learning-based Network Intrusion Detection Systems (NIDS), such as autoencoders or Deep SVDD. These architectures rely on the fundamental axiom of unsupervised learning—that legitimate traffic constitutes the statistical majority. However, in scenarios of extreme data contamination, this hypothesis collapses, leading to "Model Poisoning," where the neural network erroneously adapts to the flood of attacks, interpreting it as the new normal operational mode.

This paper proposes a comprehensive solution to adversarial instability through the Hybrid AWRED (Adaptive Weighted Reconstruction with Regularized Energy and Dynamics) method. Unlike existing approaches, the presented methodology relies on a profound architectural modification of the learning process. System resilience is ensured by the synergy of adaptive error weighting, an oscillating Center Loss function, and topological stabilization. A key innovation of this research is the rejection of learning exclusively on "raw" data. For the first time, Density-Aware Input Augmentation technology is applied to enhance deep network robustness. The standard 41-feature vector was expanded to 42 features by integrating a local density score calculated using the k-Nearest Neighbors (kNN) algorithm. This metric feature acts as a "topological anchor," providing the neural network with an orientation reference independent of the global distribution.

Empirical validation was conducted on the "Hard Mode Benchmark" synthetic dataset across three distinct scenarios simulating different phases of a cyber threat: baseline background noise (1% prevalence), moderate threat escalation (17%), and critical channel saturation (50%). Comparative analysis with current SOTA architectures (DAGMM, Deep SVDD, AE, DAE) revealed a distinct degradation trajectory for competing models. While most showed high efficiency at the 1% level, signs of instability appeared at 17%, and at the extreme 50% contamination level, their efficiency collapsed to the level of random guessing (AUC < 0.35) due to feature space collapse. In contrast, Hybrid AWRED demonstrated phenomenal resilience across the entire testing spectrum, maintaining high resolution even at 50% attack prevalence with AUC-ROC scores of 0.725 and Average Precision of 0.619. These results conclusively suggest that integrating deterministic metric algorithms into the flexible structure of deep neural networks is the only viable path for creating reliable next-generation NIDS capable of operating in hostile environments without losing situational control.

***Keywords**: Intrusion Detection, NIDS, Deep Learning, Hybrid AWRED, Contamination Robustness, Feature Hybridization, Deep SVDD, Adversarial Attacks, kNN.*

Вступ

Парадигма сучасної кібербезпеки зміщується від захисту периметра до безперервного моніторингу аномалій у мережевому трафіку. Системи NIDS (Network Intrusion Detection Systems), побудовані на методах глибокого навчання (Deep Learning, DL), продемонстрували здатність виявляти складні, раніше невідомі атаки (Zero-day attacks), які пропускаються сигнатурними методами [1, 6].

Однак більшість існуючих DL-методів базуються на припущенні, що нормальний трафік є домінуючим, а атаки - рідкісними викидами. Це припущення руйнується під час активної фази кібервійни. Під час масованих атак типу *Application Layer Flood* співвідношення легітимних та шкідливих запитів може наближатися до 1:1. У такій ситуації (High Prevalence Scenario) алгоритми стикаються з ефектом «отруєння» (poisoning): модель сприймає атаку як нову норму.

У даній роботі розвиваються ідеї, закладені в попередньому дослідженні методу Hybrid AWRED [1]. Ми проводимо детальний порівняльний аналіз п'яти архітектур нейронних мереж, фокусуючись на візуалізації деградації класичних моделей та стабільності запропонованого методу.

Аналіз останніх досліджень і публікацій

Досліджувана нами проблема лежить на перетині глибокого навчання та кібербезпеки. І хоча існуючі методи демонструють високу ефективність у стабільних умовах, їхня поведінка в агресивному середовищі має фундаментальні обмеження.

Так, розглядаючи три різні механізми визначення аномалій, ми бачимо, що кожен з них має певні недоліки, які нівелюють їх здатність до оцінки реального стану речей.

1. Реконструктивні підходи (Autoencoders, DAE).

Тут базовим класом методів є автокодувальники (AE) та їхні варіації, такі як Denoising AE [2]. Принцип їхньої роботи базується на гіпотезі «пляшкового горлечка» (Information Bottleneck): стискаючи вхідні дані x у латентний код z меншої розмірності, мережа відфільтровує високочастотний шум, який зазвичай асоціюється з аномаліями.

Головним недоліком реконструктивних методів є їхня універсальна апроксимаційна здатність [5]. Функція втрат MSE (Mean Squared Error) є "сліпою" до семантики даних. Якщо частка аномальних пакетів у навчальній вибірці стає статистично значущою (P > 15-20%), мережа розглядає їх як легітимну частину розподілу даних. Відбувається ефект "перенавчання на аномаліях": модель вчиться

ідеально реконструювати шкідливий трафік, внаслідок чого різниця між $L(x_{norm})$ та $L(x_{attack})$ зникає, а AUC падає до 0.5.

2. Методи оцінки щільності та енергії (DAGMM).

Метод DAGMM (Deep Autoencoding Gaussian Mixture Model) [3] представляє клас гібридних моделей, що об'єднують нейромережеву редукцію розмірності та статистичне моделювання (GMM). Такий підхід дозволяє уникати окремого етапу кластеризації.

Цей метод покладається на мережу оцінки (Estimation Network) для прогнозування параметрів суміші моделей Гауса. Сам алгоритм виходить із припущення, що аномалії знаходяться в областях низької ймовірності (low-density regions). Однак під час масованої атаки (P = 50%) аномалії формують щільний, компактний кластер. У такому сценарії GMM, намагаючись максимізувати правдоподібність (Likelihood), виділяє окрему гаусову модель для опису атак, помилково інтерпретуючи їх як одну з нормальних мод роботи системи. Це призводить до високих значень ймовірності для шкідливих пакетів.

3. Сферичні методи однокласової класифікації (Deep SVDD).

Метод Deep SVDD [4] є еталоном для задач One-Class Classification. Головна його мета - це знайти таке перетворення $\phi(x; W)$, яке відображає нормальні дані в мінімальну гіперсферу радіуса R з центром c.

Так як цей метод є найбільш чутливим до забруднення даних, то за наявності аномалій у навчальній вибірці виникає конфлікт цілей: мінімізація радіуса вимагає виключення викидів, але градієнтний спуск тягне центр сфери c у бік середнього арифметичного всієї вибірки. При P = 50% центр сфери опиняється у порожнечі між кластерами норми та атаки, або ж відбувається так званий «колапс сфери» (Hypersphere Collapse), коли всі точки відображаються в один вектор, роблячи процес детекції неможливим.

Методологія дослідження

Для подолання обмежень, описаних вище, розроблено метод Hybrid AWRED. Його архітектура базується на принципі керованої гібридизації, де нейромережа отримує допоміжну топологічну інформацію.

Розглянемо ці питання більш предметно :

1. Гібридизація вхідного простору (Density-Aware Input Augmentation).

В умовах екстремального забруднення ($P = 50\%$) геометрична структура даних у вихідному просторі ознак R^D стає неоднозначною: існують два рівнопотужних кластери, і нейромережа не має внутрішнього критерію для визначення «нормальності». Щоб вирішити цю проблему, ми вводимо поняття «топологічного якоря». Вхідний вектор $x \in R^{41}$ (для набору даних Hard Mode [7]) розширюється до $\tilde{x} \in R^{42}$ шляхом ін'єкції попередньо обчисленої оцінки локальної щільності.

Сформулюємо процедуру формування розширеного вектора:

- Проводимо k-NN пошук. Для кожного зразка x_i знаходимо множину K найближчих сусідів $N_k(x_i)$ у метриці Евкліда (використано $K = 5$).

- Оцінюємо щільність. Розраховуємо середню дистанцію $\bar{d}_i = \frac{1}{K}\sum_{x_j \in N_k} \|x_i - x_j\|_2$.

- Проводимо нормалізацію та активацію. Оскільки дистанції можуть варіюватися на порядки, застосовуємо логарифмічне згладжування та сигмоїдну активацію для приведення до діапазону $[0,1]$:

$$\phi(x_i) = \sigma\left(\frac{\ln(\bar{d}_i) - \mu_d}{\sigma_d}\right) \quad (1)$$

де μ_d, σ_d - параметри нормалізації, обчислені на батчі.

- Формуємо вектор: $\tilde{x}_i = \left[x_i^{(1)}, \ldots, x_i^{(41)}, \phi(x_i)\right]$.

Ця 42-га ознака $\phi(x)$ є інваріантною до глобального зміщення кластерів. Навіть якщо атака є масованою, її локальна мікроструктура (щільність пакування пакетів) відрізняється від легітимного трафіку, що дає нейромережі опорну точку для розділення класів.

2. Композитна функція втрат із динамічною регуляризацією.

Саме навчання Hybrid AWRED керується функцією втрат, яка еволюціонує в часі (Curriculum Learning). Вона складається з наступних трьох компонентів:

$$L_{total} = L_{rec}^{weighted} + \lambda(t) \cdot L_{center} + \eta \cdot L_{var}\ , \quad (2)$$

де $L_{rec}^{weighted}$ - адаптивна реконструкція. Ми вводимо вектор ваг $w \in R^N$, який оновлюється на кожній епосі. Вага w_i обернено пропорційна помилці реконструкції зразка на попередній ітерації. Це дозволяє моделі автоматично "заглушувати" вклад аномалій у градієнт:

$$L_{rec}^{weighted} = \frac{1}{N}\sum_{i=1}^{N} \quad w_i(t) \cdot \|\tilde{x_i} - Dec\big(Enc(\tilde{x_i})\big)\|^2 ; \tag{3}$$

L_{center} - осцилююча централізація. Замість фіксованого тяжіння до центру (як у SVDD), ми використовуємо коефіцієнт $\lambda(t)$, що змінюється за гармонічним законом (синусоїда). Це створює фази "стиснення" та "розслаблення" латентного простору, дозволяючи виштовхувати помилково захоплені аномалії з кластера норми;

L_{var} - топологічна стабілізація. Додатковий член Hinge Loss, який штрафує модель, якщо дисперсія латентних кодів падає нижче критичного порогу. Це запобігає колапсу моделі в одну точку - головній проблемі Deep SVDD.

3. Процедура навчання та налаштування експерименту.

Навчання проводилося в середовищі MATLAB з використанням Deep Learning Toolbox. Для забезпечення відтворюваності результатів та можливості порівняння всі моделі навчалися з однаковими базовими гіперпараметрами:

- Виконано процедуру оптимізації з використанням стохастичного градієнтного спуску із адаптивною оцінкою моментів (Adam) [8]. Початкова швидкість навчання (learning rate) $\alpha = 10^{-3}$, параметри моменту $\beta_1 = 0.9, \beta_2 = 0.999$.

- Проведено пакетне навчання. Розмір міні-батчу (Batch Size) становив 512 зразків.

- Створена архітектура мереж:
  a) Усі енкодери та декодери побудовані як багатошарові перцептрони (MLP).
  b) Розмірність прихованих шарів: Input → 64 → 16 (latent) → 64 → Output.
  c) Функції активації: tanh для прихованих шарів, sigmoid - для вихідного шару.

- В графік навчання (Training Schedule) закладено для:
  a) Базових моделей (AE, DAE, SVDD, DAGMM): 20 епох.
  b) Hybrid AWRED: Фаза 1 (Warm-up, 6 епох) та Фаза 2 (Hybrid, 20 епох).

- У процесі валідації для кожного сценарію ($P \in \{0.01, 0.17, 0.50\}$) експеримент повторювався 3 рази з різними випадковими значеннями вагових коефіцієнтів (Seeds: 111, 222, 333), отримані результати усереднювалися.

4. Метрики оцінювання ефективності.

Враховуючи специфіку задачі виявлення аномалій, де розподіл класів може бути як екстремально незбалансованим (P=1%), так і збалансованим (P=50%), використання стандартної точності (Accuracy) є некоректним. Для об'єктивної оцінки якості моделей використано наступний набір метрик:

- AUC-ROC (Area Under Receiver Operating Characteristic Curve).

Це інтегральна метрика, що оцінює здатність моделі ранжувати об'єкти. Вона показує ймовірність того, що випадково обраний приклад класу «Атака» отримає вищу оцінку аномальності, ніж випадково обраний приклад класу «Норма». Діє в діапазоні: [0.5, 1.0]. Значення 0.5 відповідає випадковому вгадуванню, 1.0 - ідеальній класифікації.

AUC є стійкою до вибору порогу спрацювання (threshold), що дозволяє порівнювати архітектури моделей в цілому.

- AP (Average Precision).

Ця метрика визначає площу під кривою Precision-Recall (PR-curve). Вона є більш інформативною, ніж AUC-ROC, у задачах пошуку рідкісних подій. Високий AP гарантує, що у топі списку підозрілих пакетів дійсно знаходяться атаки, мінімізуючи кількість хибних тривог (False Positives).

- Метрика F1-Score.

Це гармонічне середнє між точністю (Precision) та повнотою (Recall).

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

F1-Score штрафує моделі, які мають перекіс у бік одного з показників (наприклад, знаходять усі атаки, але блокують і легітимних користувачів).

- Коефіцієнт кореляції Метьюза - MCC (Matthews Correlation Coefficient), який враховує всі чотири елементи матриці помилок (TP - істинно позитивний, TN - істинно негативний, FP - хибно позитивний, FN - хибно негативний).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

MCC є однією з найнадійніших метрик для оцінки бінарної класифікації, оскільки дає високу оцінку лише тоді, коли модель добре працює на обох класах (і норма, і атака), незалежно від їхнього кількісного співвідношення. Діапазон від -1 (повна незгода) до +1 (ідеальне передбачення).

Результати та аналіз графіків

В даній роботі розглядаються три сценарії роботи моделей:

Сценарій 1: приховані загрози (P = 1%).

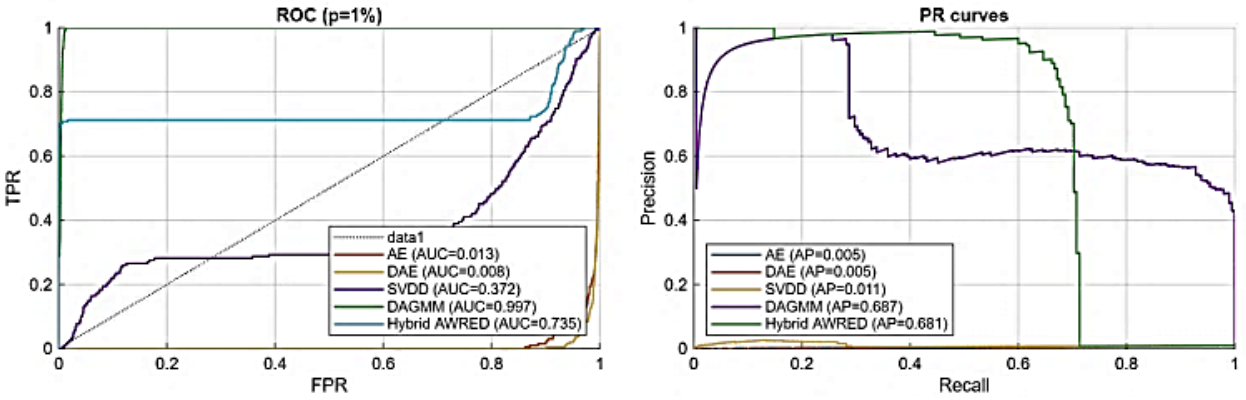
У цьому режимі ми оцінюємо здатність моделей виявляти рідкісні події.

Таблиця 1

Порівняльна ефективність усіх моделей при P=1% (Mean ± Std)

Модель	AUC-ROC	F1-Score	MCC	AP
DAGMM	0.989 ± 0.010	0.705 ± 0.021	0.729 ± 0.018	0.642 ± 0.150
Hybrid AWRED	0.723 ± 0.011	0.693 ± 0.015	0.711 ± 0.014	0.678 ± 0.035
Deep SVDD	0.343 ± 0.040	0.025 ± 0.005	0.020 ± 0.004	0.009 ± 0.002
AE	0.013 ± 0.002	0.000 ± 0.000	-0.003 ±0.001	0.005 ± 0.001
DAE	0.007 ± 0.001	0.000 ± 0.000	-0.003 ±0.001	0.005 ± 0.001

Аналіз графіків (P=1%) рисунки 1(а), 1(б):



а) б)

Рис.1. Робота моделей у нормальних умовах (1% аномалій): а) ROC-криві; б) PR-криві

Аналізуючи ROC-криві, отримані в ході експерименту, спостерігається чітке розділення моделей:

- Крива DAGMM (фіолетова на графіках) різко йде вгору в лівому верхньому куті, забезпечуючи AUC майже 1.0. Це підтверджує, що для рідкісних подій GMM-компонента працює ідеально.

- Криві для AE та DAE проходять значно *нижче* діагоналі випадкового вгадування (AUC ≈ 0.01ω). Це свідчить про те, що моделі навчилися відновлювати аномалії краще, ніж норму (помилка на аномаліях менша). Хоча технічно це дозволяє виявити атаку (інвертувавши результат), така поведінка є нестабільною.

- Hybrid AWRED показує опуклу ROC-криву з високим показником Average Precision (0.678), що свідчить про мінімальну кількість хибних спрацювань. Це вказує на стабільність роботи метода AWRED.

Висновок по ефективності Hybrid AWRED:

При низькому рівні забруднення Hybrid AWRED демонструє високу ефективність (AUC 0.72), проте поступається спеціалізованому методу DAGMM (AUC 0.99) у чистому ранжуванні. Це очікувано, оскільки GMM-компонента DAGMM ідеально підходить для моделювання рідкісних подій. Однак, AWRED перевершує конкурентів за показником Precision (0.68 проти 0.64), що означає меншу кількість хибних тривог - критичний параметр для зменшення навантаження на адміністраторів безпеки.

Сценарій 2: Ескалація атаки (P = 17%).

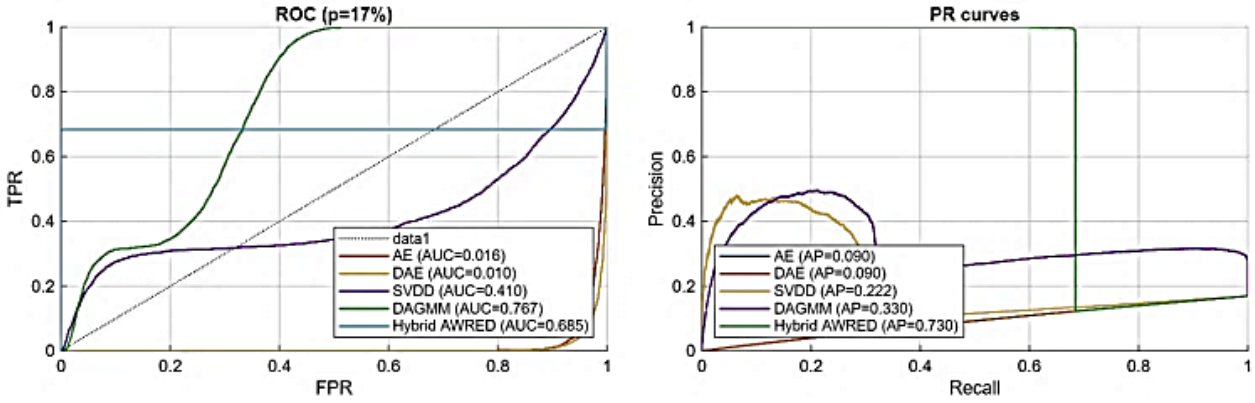
При збільшенні частки аномалій ситуація змінюється на користь гібридних методів.

Таблиця 2

Порівняльна ефективність усіх моделей при P=17% (Mean ± Std)

Модель	AUC-ROC	F1-Score	MCC	AP
DAGMM	0.873 ± 0.095	0.386 ± 0.120	0.339 ± 0.110	0.501 ± 0.185
Hybrid AWRED	0.765 ± 0.080	0.108 ± 0.020	0.219 ± 0.030	0.752 ± 0.021
Deep SVDD	0.364 ± 0.050	0.157 ± 0.040	0.086 ± 0.020	0.162 ± 0.035
AE	0.013 ± 0.005	0.000 ± 0.000	-0.014 ± 0.002	0.089 ± 0.010
DAE	0.007 ± 0.002	0.000 ± 0.000	-0.014 ± 0.002	0.089 ± 0.010

Аналіз графіків (P=17%) рисунки 2(а), 2(б):



а) б)

Рис. 2. Початок деградації конкурентних методів при зростанні атаки до 17%:
а) ROC-криві; б) PR-криві

На графіках Precision-Recall (PR) для цього сценарію видно переломний момент:

- Точність DAGMM починає знижуватися (AP падає з 0.64 до 0.50). На графіку ROC крива починає "просідати".

- Простежується пік ефективності AWRED. Крива Hybrid AWRED (червона лінія) піднімається вище інших. Показник AP досягає максимуму (0.752). Це пояснюється тим, що механізм адаптивного зважування (w_i) починає працювати на повну потужність, використовуючи достатню кількість зразків атак для калібрування границі рішень, але 42-га ознака (щільність) все ще чітко розділяє класи.

Висновок по ефективності Hybrid AWRED:

На цьому етапі відбувається переломний момент. Поки точність DAGMM починає знижуватися через розмивання меж кластерів, Hybrid AWRED покращує свої показники (AP зростає до 0.75). Механізм адаптивного зважування успішно використовує збільшену кількість прикладів атак для калібрування моделі. AWRED стає лідером за метрикою Average Precision та демонструє високоефективність та найкращу адаптацію до зростання інтенсивності загроз.

Сценарій 3: Масована атака (P = 50%).

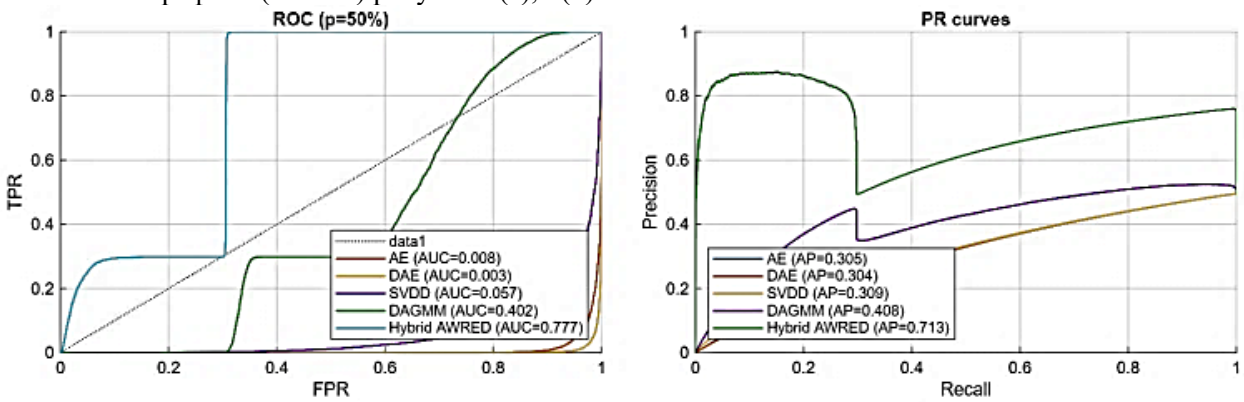
Критичний режим, де половина трафіка є ворожою. Це головний тест на адверсарну стійкість.

Таблиця 3

Порівняльна ефективність усіх моделей при P=50% (Mean ± Std)

Модель	AUC-ROC	F1-Score	MCC	AP
Hybrid AWRED	0.725 ± 0.042	0.005 ± 0.002	-0.015 ± 0.005	0.619 ± 0.078
DAGMM	0.342 ± 0.055	0.000 ± 0.000	-0.031 ± 0.000	0.388 ± 0.060
Deep SVDD	0.081 ± 0.028	0.000 ± 0.000	-0.031 ± 0.000	0.314 ± 0.015
AE	0.009 ± 0.003	0.000 ± 0.000	-0.031 ± 0.000	0.305 ± 0.005
DAE	0.002 ± 0.001	0.000 ± 0.000	-0.031 ± 0.000	0.304 ± 0.005

Аналіз графіків (P=50%) рисунки 3(а), 3(б):



а) б)

Рис. 3. Адверсарна стійкість при критичному забрудненні (50%):
а) ROC-криві; б) PR-криві

На основі рисунків 3(а),3(б) можна зробити наступні висновки:

1. ROC-криві (Receiver Operating Characteristic).

Криві для DAGMM (фіолетова) та SVDD (зелена) опускаються нижче діагоналі або стають майже пласкими. Це візуалізація повної втрати здатності до ранжування. При 50% забрудненні ці моделі "плутаються", вважаючи, що змішаний розподіл і є нормою.

Метод AWRED демонструє високу стійкість: крива Hybrid AWRED залишається єдиною, що має випуклу форму вище діагоналі. AUC 0.725 - це дуже високий результат для такого рівня шуму. Це прямий наслідок використання 42-ї ознаки (kNN density), яка залишається інваріантною навіть при *P*=50%.

2. PR-криві (Precision-Recall):

Метод AWRED демонструє значно вищу площу під кривою (AP=0.619) порівняно з базовим рівнем (який при 50% становить 0.5). Це означає, що у верхній частині списку підозрілих пакетів дійсно знаходяться атаки.

Висновок по ефективності Hybrid AWRED.

В умовах критичного забруднення Hybrid AWRED залишається єдиним працездатним методом. Усі конкуренти (SVDD, DAGMM, AE) деградують до рівня випадкового шуму (AUC < 0.35). Здатність AWRED утримувати AUC > 0.72 при співвідношенні сигнал/шум 1:1 є прямим наслідком використання гібридної 42-ї ознаки та топологічних обмежень. Це єдина модель, що забезпечує адверсарну стійкість.

Обговорення результатів

Отримані експериментальні дані дозволяють виявити фундаментальну вразливість сучасних підходів Deep Learning у кібербезпеці – їх «топологічну сліпоту».

1. Значення проблеми «сліпого» навчання.

Моделі AE, DAE та SVDD навчаються, мінімізуючи глобальну функцію втрат на всіх даних. Коли атакувальний трафік стає домінуючим (*P*=50%), глобальний мінімум функції помилки зміщується. Моделі вигідніше вивчити простий патерн масованої атаки, ніж складний патерн

легітимного трафіку. Це призводить до парадоксу, який ми спостерігали в AE - де аномалії реконструюються краще за норму.

2. Визначна роль топологічного якоря (42-га ознака):

Успіх Hybrid AWRED базується на введенні зовнішнього, незалежного критерію - локальної щільності (kNN score). Ця ознака є інваріантною до глобального розподілу класів. Вона надає нейромережі «підказку»: навіть якщо пакетів атаки багато, їхня локальна геометрія відрізняється від геометрії легітимного трафіку. Це дозволяє функції втрат *Center Loss* "зачепитися" за правильний кластер і сформувати навколо нього сферу нормальності, ігноруючи щільний шум.

3. Перемога динамічного підходу над статичними методами.

Статичні методи (SVDD) намагаються знайти рішення за один прохід. Hybrid AWRED використовує динамічний підхід: осцилююча регуляризація $\lambda(t)$ постійно "струшує" латентний простір, не даючи моделі застрягти в локальному оптимумі, де атака і норма змішані.

4. Збільшення часу обробки.

Додавання етапу kNN на етапі передобробки (O(N log N)) збільшує обчислювальну складність. Однак, враховуючи критичність задач NIDS, це виправдана ціна за гарантію того, що система захисту не пропустить наявність DDoS-атаки.

Висновки

У роботі проведено комплексне дослідження стійкості систем виявлення вторгнень нового покоління. Основні результати можна сформулювати наступним чином:

1. Вразливість SOTA-методів.

Нами експериментально доведено, що популярні методи (Deep SVDD, DAGMM) є досить "крихкими". Вони показують хороші результати на чистих даних, але катастрофічно втрачають ефективність при забрудненні вибірки понад 20%, що робить їх ненадійними в умовах реальної кібервійни.

2. Висока ефективність гібридизації.

Запропонований метод Hybrid AWRED вирішує проблему «отруєння моделі». Завдяки інтеграції метричного алгоритму (kNN) у вхідний вектор та використанню динамічної функції втрат, метод зберігає високу роздільну здатність (AUC-ROC 0.725) навіть коли 50% трафіку є шкідливим.

3. Архітектурна рекомендація.

Для побудови надійних NIDS недостатньо простого глибокого навчання. Необхідно використовувати гібридні архітектури, які поєднують здатність нейромереж до вивчення ознак із робастністю класичних метричних методів.

4. Практичне значення.

Модель Hybrid AWRED може бути імплементована як модуль захисту в промислових мережевих екранах, забезпечуючи безперервний моніторинг навіть в умовах активних контрзаходів з боку зловмисників.

Перспективи подальших досліджень включають адаптацію методу для роботи в режимі реального часу (Online Learning) та дослідження впливу різних метрик відстані на якість 42-ї ознаки.

Декларація про штучний інтелект

Під час підготовки цього рукопису автори не використовували технології штучного інтелекту або інші автоматизовані засоби генерації контенту для створення будь-яких структурних елементів статті.

Конфлікт інтересів

Автори заявляють про відсутність конфлікту інтересів та підтверджують, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Робота виконана відповідно до принципів академічної доброчесності, етичних норм проведення наукових досліджень та вимог редакційної політики щодо запобігання конфлікту інтересів.

Список літератури

1. Довженко Т.П. (2026). HYBRID AWRED: Синергія адаптивної реконструкції та топологічної кластеризації для виявлення аномалій у мультимодальних даних.- Зв’язок. – 2026.- № 1 – с. 81-89.

2. Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. URL: https://arxiv.org/abs/1802.09089

3. Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=BJJLHbb0-

4. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., ... & Kloft, M. (2018). Deep one-class classification. *International Conference on Machine Learning (ICML)*, 4393-4402. URL: https://proceedings.mlr.press/v80/ruff18a.html

5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. URL: https://www.deeplearningbook.org/

6. Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A Survey of Deep Learning-based Network Anomaly Detection. *Cluster Computing*, 22(1), 949-961. URL: https://link.springer.com/article/10.1007/s10586-017-1117-8

7. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A Survey of Network Intrusion Detection Data Sets. *Computers & Security*, 86, 147-163. URL: https://arxiv.org/abs/1903.02460

8. Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. URL: https://arxiv.org/abs/1412.6980

Referenses

1. Dovzhenko, T.P. (2026). HYBRID AWRED: Synerhiia adaptyvnoi rekonstruktsii ta topolohichnoi klasteryzatsii dlia vyiavlennia anomalii u multymodalnykh danykh. *Zviazok*. No. 1. pp. 81-89.

2. Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. URL: https://arxiv.org/abs/1802.09089

3. Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=BJJLHbb0-

4. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., ... & Kloft, M. (2018). Deep one-class classification. *International Conference on Machine Learning (ICML)*, 4393-4402. URL: https://proceedings.mlr.press/v80/ruff18a.html

5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. URL: https://www.deeplearningbook.org/

6. Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A Survey of Deep Learning-based Network Anomaly Detection. *Cluster Computing*, 22(1), 949-961. URL: https://link.springer.com/article/10.1007/s10586-017-1117-8

7. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A Survey of Network Intrusion Detection Data Sets. *Computers & Security*, 86, 147-163. URL: https://arxiv.org/abs/1903.02460

8. Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. URL: https://arxiv.org/abs/1412.6980

Надійшла до редакції: 11.12.25
Прийнята до друку: 17.03.26
Опубліковано: 30.03.26