УДК 004.855.5:004.932 **DOI:** 10.31673/2412-9070.2026.027603

Т. П. ДОВЖЕНКО, канд. техн. наук,
ORCID:0000-0002-0352-8391
Державний університет інформаційно-комунікаційних технологій, Київ

ТОПОЛОГІЧНЕ ЯКОРУВАННЯ ТА АДАПТИВНІ ШТРАФИ: АРХІТЕКТУРА HYBRID AWRED ДЛЯ РОЗПІЗНАВАННЯ ДЕФЕКТІВ У ЗАБРУДНЕНИХ ВІЗУАЛЬНИХ ДАНИХ

У сучасних системах комп'ютерного зору припущення про повну чистоту навчальних вибірок далеко не завжди відповідає практиці. Якщо дані містять значну кількість структурних спотворень, традиційні методи виявлення аномалій без учителя часто втрачають стійкість, оскільки дефектні зразки частково включаються до простору норми. У статті розглянуто архітектуру Hybrid AWRED v3, призначену для розпізнавання дефектів у забруднених візуальних даних. Запропонований підхід поєднує топологічне якорування, адаптивне зважування похибки реконструкції та зовнішній контур Байєсівської оптимізації гіперпараметрів.

Експериментальну перевірку виконано на модифікованому наборі MNIST-C у сценаріях із 5% і 20% забруднення. Показано, що за рівня забруднення 20% запропонована архітектура забезпечує найвище значення повноти виявлення (Recall = 0.880 ± 0.005) серед розглянутих моделей, перевищуючи, зокрема, DAGMM (0.397 ± 0.039) і Deep SVDD (0.584 ± 0.163). Водночас за інтегральною метрикою AUC-ROC найбільше середнє значення у цьому сценарії продемонстрував DAE (0.785 ± 0.008), що вказує на різну чутливість моделей до окремих аспектів якості виявлення.

З огляду на отримані результати Hybrid AWRED v3 доцільно розглядати як перспективний підхід для задач, у яких критично важливо зменшити кількість пропущених дефектів за умов значного забруднення навчальної вибірки. Разом із тим переваги методу варто інтерпретувати передусім через повноту виявлення та стійкість до data poisoning.

Ключові слова: виявлення аномалій, Data Poisoning, комп'ютерний зір, Hybrid AWRED, топологічне якорування, Байєсівська оптимізація, поетапне навчання.

Вступ

Сучасні системи інтелектуального аналізу зображень - від автоматичного розпізнавання символів (OCR) до засобів візуального контролю якості на складальних лініях Industry 4.0 – працюють в умовах, які суттєво відрізняються від лабораторних. У реальних візуальних потоках наявні дефекти, артефакти стиснення, імпульсний шум і різні типи структурних спотворень. За таких обставин класичне навчання з учителем часто виявляється економічно невигідним, оскільки потребує трудомісткого ручного маркування великих масивів піксельних даних. Саме тому зростає практичний інтерес до алгоритмів виявлення аномалій без учителя (Unsupervised Anomaly Detection).

Більшість існуючих моделей глибокого навчання, зокрема класичні автокодувальники або методи на основі оцінки щільності, спираються на припущення, що аномалії трапляються вкрай рідко. Зазвичай мається на увазі, що частка дефектних зразків у навчальній вибірці не перевищує 1–2%, тому їхній вплив на мінімізацію функції втрат залишається незначним. На практиці ця умова виконується далеко не завжди. Якщо рівень структурного шуму або частка дефектних символів зростає до 20% і більше, виникає ефект «отруєння даних» (Data Poisoning) [5]–[7].

За масованого забруднення візуальних вибірок навіть сучасні SOTA-архітектури демонструють відчутне зниження ефективності. Моделі однокласової класифікації, такі як Deep SVDD,

схильні до колапсу гіперсфери, намагаючись охопити забруднений простір одним центром. Складні ймовірнісні моделі, зокрема DAGMM (Deep Autoencoding Gaussian Mixture Model), натомість стикаються з ефектом «розмазування кластерів»: за високої концентрації аномалій алгоритм починає інтерпретувати їх як окремі легітимні компоненти гауссівської суміші. У підсумку модель пристосовується до дефектів, а повнота їх виявлення (Recall) різко знижується.

Для подолання цих обмежень у роботі запропоновано спеціалізовану архітектуру Hybrid AWRED (Adaptive Weighted Reconstruction with Regularized Energy and Dynamics), адаптовану до розпізнавання дефектів у забруднених візуальних даних. На відміну від попередніх версій методу, де використовувалася статична просторова регуляризація, запропонований варіант спирається на концепцію топологічного якорування, яку автор раніше розробляв для матричних даних у задачах кібербезпеки, а тепер адаптував до візуальних просторів високої розмірності.

У версії v3 реалізовано дві ключові зміни. Перша - механізм адаптивних штрафів, керований динамічною регуляризацією (Curriculum Dynamics), який дає змогу моделі спочатку м'яко пристосуватися до структури легітимних символів, а вже потім застосовувати жорстке квантильне відсікання аномалій. Друга - зовнішній автономний контур Байєсівської оптимізації, що зменшує потребу в ручному налаштуванні гіперпараметрів.

Метою цього дослідження є експериментальна перевірка припущення, що поєднання топологічного якорування з автономними адаптивними штрафами дає змогу послабити вплив «отруєння» нейромережі та зберегти стійкість виявлення графічних аномалій навіть тоді, коли рівень забруднення навчальної вибірки сягає 20%.

Аналіз останніх досліджень і публікацій

Аналіз споріднених робіт у цій статті охоплює три фундаментальні парадигми виявлення аномалій без учителя (реконструктивні методи, однокласову класифікацію та оцінку щільності розподілу), дослідження стратегій поетапного навчання (Curriculum Learning), а також ретроспективу еволюції власної базової архітектури Hybrid AWRED.

– Реконструктивні методи та автокодувальники. Відомим базовим підходом є автокодувальники (AE), в яких аномалії ідентифікуються за суттєвим зростанням помилки реконструкції (MSE). Шумоподавлювальні автокодувальники DAE [1] є стійкіші до піксельного шуму, але при структурному забрудненні понад 20% сприймають дефекти як різновид шуму для апроксимації, що критично знижує Recall.

– Однокласова класифікація та оцінка щільності. Мережевий алгоритм Deep SVDD [2] стискає дані у гіперсферу, але при отруєнні даними він зазнає виродження, а ймовірнісна модель DAGMM [3] оптимізує параметри суміші розподілів (GMM), проте при понад 15-20% аномалій виділяє під них окремі компоненти, «легалізуючи» їх, як нормальні. Нормалізаційні потоки [8] та методи на базі екстракції ознак (зокрема, алгоритм PatchCore) [9] також вразливі через обчислювальну складність і чутливість до чистоти базового простору.

– Поетапне навчання (Curriculum Learning). У підході Бенжіо та ін. [4] навчальні приклади ускладнюються поступово. Однак у задачах без учителя мережі стає складно автономно відрізнити «складний нормальний зразок» від аномалії.

– Еволюція Hybrid AWRED. Метод розвинувся від мультимодальних наборів зі статичною регуляризацією (v1) [14] та задач кібербезпеки з kNN-якоруванням (v2) [15]. У поточній версії (v3) для візуальних даних впроваджено автономне керування (Байєсівська оптимізація) та Curriculum Learning, що дозволяє мережі самостійно визначати поріг ігнорування шуму.

Методологія та архітектура HYBRID AWRED

Запропонована у цій роботі архітектура Hybrid AWRED v3 розроблена для вирішення проблеми «отруєння» нейромережі масованим візуальним шумом. На відміну від класичного навчання без учителя, де всі вхідні зразки мають однакову вагу під час градієнтного спуску, цей підхід базується на керованому перерозподілі внеску окремих зразків у процес навчання. Алгоритм працює завдяки поєднанню трьох механізмів: початкової метричної гібридизації

вхідного простору для просторової ізоляції дефектів, базового циклу оптимізації з адаптивно зваженою функцією втрат та зовнішнього контуру динамічної регуляризації, який автономно керує фазами навчання і Байєсівським пошуком гіперпараметрів.

Топологічне якорування візуальних даних

Відомо, що традиційні автокодувальники обробляють піксельні масиви як ізольовані вектори $x \in \mathrm{R}^D$ (де D=64 для згорнутих зображень формату 8x8 пікселів набору MNIST). В умовах сильного забруднення нейромережа може знайти хибні нелінійні кореляції між дефектними пікселями. Для запобігання цьому вводиться поняття гібридного метричного простору, де вхідний вектор доповнюється зовнішнім «топологічним якорем».

Гібридний вектор $x^* \in \mathrm{R}^{D+1}$ формується шляхом конкатенації оригінального піксельного масиву та скалярної ознаки локальної щільності $\phi(x)$:

$$x_i^* = \left[x_{i,1}, x_{i,2}, \dots, x_{i,D}, \phi(x_i)\right]. \quad (1)$$

Ознака $\phi(x)$ обчислюється до початку навчання нейромережі за допомогою алгоритму k найближчих сусідів (k-Nearest Neighbors, kNN) в оригінальному просторі ознак [10]. Математично вона виражається як нормалізована логарифмічна відстань:

$$\phi(x_i) = \sigma\left(\ln\left(1 + \frac{1}{k}\sum_{j \in \mathcal{N}_k(x_i)} \left\|x_i - x_j\right\|_2^2\right) - \mu_\phi\right), \quad (2)$$

де $\mathcal{N}_k(x_i)$ - множина з k найближчих сусідів для зразка x_i, μ_ϕ - середнє значення логарифмічної відстані по всій вибірці.

Ця 65-та ознака вводить додатковий просторовий орієнтир уже на першому прихованому шарі. Навіть якщо аномальні цифри візуально частково схожі на норму, через деформації вони матимуть інше значення ϕ(x), оскільки розташовані в більш розріджених ділянках багатовимірного простору. Завдяки цьому мережі складніше змішувати нормальні та дефектні зразки в одному латентному кластері.

Механізм адаптивного зважування (Adaptive Soft-Thresholding)

Основою моделі є здатність самостійно відрізняти складні, але легітимні символи від істинних дефектів. Це досягається шляхом модифікації функції втрат (Loss Function). Замість мінімізації середньоквадратичної помилки (MSE) для всього міні-пакету, ми використовуємо зважену помилку реконструкції (Robust Loss) [11]:

$$L(\theta, t) = \frac{1}{N}\sum_{i=1}^{N} w_i(t) \cdot \|x_i - \hat{x}_i\|^2 + \lambda_{\text{reg}} \|\theta\|_2^2, \quad (3)$$

де $\hat{x}_i$ - реконструйований вихід мережі, $w_i(t)$ - індивідуальна динамічна вага зразка у поточну епоху t. Вага обчислюється через диференційовну логістичну функцію (сигмоїду), що забезпечує м'яке квантильне відсікання (рис. 1):

$$w_i(t) = \frac{1}{1+\exp\left(G(t)\cdot\left(e_i - Q_q(E_{\text{batch}})\right)\right)}. \quad (4)$$

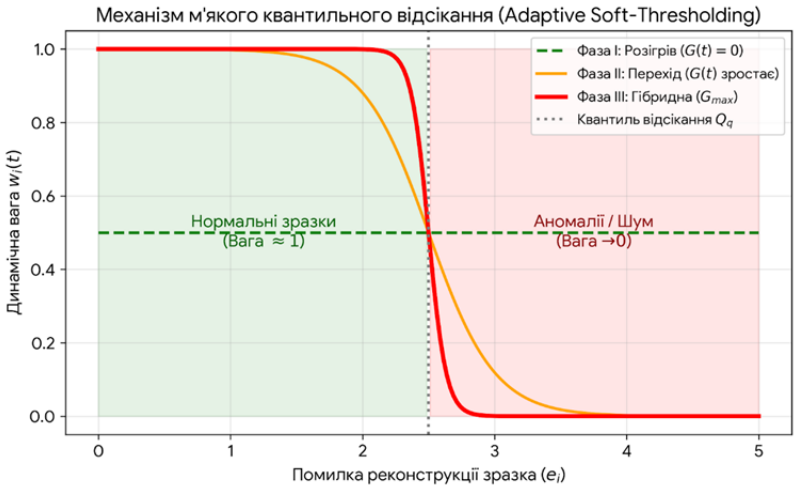

Рис. 1. Механізм адаптивного м'якого відсікання (Soft-Thresholding)

Графік демонструє еволюцію вагової функції $w_i(t)$ залежно від помилки реконструкції зразка e_i. На фазі I (розігрів) усі зразки мають максимальну вагу. На фазі III (гібридна), завдяки зростанню штрафу G_{max}, вага зразків, чия помилка перевищує поріг Q_q, стрімко наближається до нуля, виключаючи їх із процесу оновлення градієнтів.

Тут e_i - поточна помилка реконструкції зразка. Параметр $Q_q(E_{\text{batch}})$ позначає емпіричний

квантиль рівня q_{tail} від розподілу помилок у поточному міні-пакеті. Зразки, помилка яких перевищує поріг Q_q, отримують експоненціально малу вагу (штраф), завдяки чому їхні градієнти практично не оновлюють ваги нейромережі.

Динамічна регуляризація (Curriculum Dynamics)

Критичною проблемою ранніх ітерацій методів робастного навчання є передчасне застосування штрафів. Якщо застосувати жорстке відсікання з першої епохи, нейромережа може випадково відкинути нормальні зразки зі складним накресленням. Для вирішення цього в Hybrid AWRED v3 впроваджено розклад навчання за програмою (Curriculum Learning) з часозалежним параметром крутизни штрафу *G(t)*:

$$G(t) = \begin{cases} 0, & t \le T_{\text{warm}} \\ G_{\max} \cdot \frac{t - T_{\text{warm}}}{T_{\text{ramp}} - T_{\text{warm}}}, & T_{\text{warm}} < t \le T_{\text{ramp}} \\ G_{\max}, & t > T_{\text{ramp}} \end{cases}. \quad (5)$$

Даний процес розбито на три фази:

1. Фаза розігріву $(\text{Warm} - \text{up}, t \le T_{\text{warm}})$: $G(t) = 0$. Усі зразки мають вагу $w_i \approx 1$. Мережа вільно мінімізує глобальну енергію, формуючи базову структуру цифр.

2. Перехідна фаза $(\text{Ramp}, T_{\text{warm}} < t \le T_{\text{ramp}})$: Штраф $G(t)$ лінійно зростає. Модель починає плавно «виштовхувати» аномалії з легітимного простору, перелаштовуючи свої ваги без різких градієнтних стрибків.

3. Гібридна фаза $\left(t > T_{\text{ramp}}\right)$: Повний штраф $G_{\max}$ і відбувається формування чітких меж нормального кластера.

Автономний контур Байєсівської оптимізації

Ефективність описаного механізму критично залежить від двох гіперпараметрів: квантиля відсікання q_{tail} (яку частку даних вважати підозрілою) та сили штрафу $G_{\max}$ (наскільки жорстко їх відкидати). Оскільки рівень отруєння даних у реальних задачах заздалегідь невідомий, ручний підбір цих параметрів є неможливим.

Запропонована архітектура вирішує цю проблему за допомогою автономного контуру оптимізації на базі Гауссівських процесів (Gaussian Processes, GP) [12]. Формулюється пошук оптимальних гіперпараметрів $\text{p}^* = [q_{\text{tail}}, G_{max}]$ як задача максимізації невідомої цільової функції $\Psi(\text{p})$:

$$\text{p}^* = \arg\max_{\text{p}} \Psi(\text{p}), P(\Psi \mid \text{p}_{1:m}) \sim \mathcal{N}\left(\mu_m(\text{p}), \sigma_m^2(\text{p})\right). \quad (6)$$

Алгоритм формує апроксимуючу ймовірнісну модель простору параметрів і використовує функцію очікуваного покращення (Expected Improvement, EI) для вибору наступної точки перевірки:

$$EI(\text{p}) = \text{E}[max(0, \Psi(\text{p}) - \Psi_{\text{best}})]. \quad (7)$$

Цей механізм дозволяє моделі за обмежену кількість ітерацій автономно знаходити ідеальну робочу точку - глобальний мінімум простору втрат, де забезпечується максимальна здатність розрізняти легітимні візуальні структури та їхні дефектні аномалії, усуваючи необхідність втручання експерта.

Експериментальне дослідження

Дизайн експерименту. Для експерименту використано набір MNIST-C (Hard Mode) [13] зі спотвореннями. Вхідний вектор було розширено до R^{65} шляхом додавання ознаки локальної щільності до піксельного представлення. Тестування виконано у двох сценаріях: Low Contamination (5% аномалій) та High Contamination (20%). Для порівняння обрано базові моделі AE, DAE, Deep SVDD і DAGMM.

Аналіз ландшафту (рис. 2).

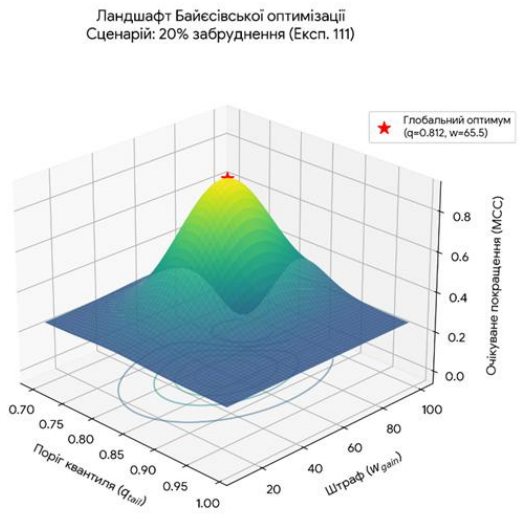

Рис. 2. Візуалізація апроксимуючої моделі (Gaussian Process) простору гіперпараметрів

Поверхня відображає прогнозований рівень MCC залежно від квантиля відсікання (q_{tail}) та сили штрафу (G_{max}). Червоною зіркою позначено знайдений глобальний оптимум. Аналіз ландшафту гіперпараметрів показав, що в просторі квантиля відсікання та сили штрафу формується локалізована область найкращих значень. У межах проведеного пошуку найвищі значення MCC спостерігалися за відносно агресивного відсікання ($q \approx 0.875$), що відповідає вилученню 12.5% найгірших зразків у міні-пакеті та за високого значення максимального штрафу ($G_{max} \approx 58.9$). Разом із тим надмірне посилення штрафу без достатнього квантильного запасу призводило до погіршення якості. Це вказує на необхідність узгодженого налаштування обох параметрів.

Динаміка навчання

Для перевірки ефективності стратегії Curriculum Learning було проведено порівняльний аналіз динаміки мінімізації функції втрат (Loss Curves). Щоб результати було зручніше інтерпретувати, графічне подання розділено на дві панелі: початкову збіжність для всіх моделей у межах перших 5 епох і повний трифазний цикл для Hybrid AWRED у межах 20 епох.

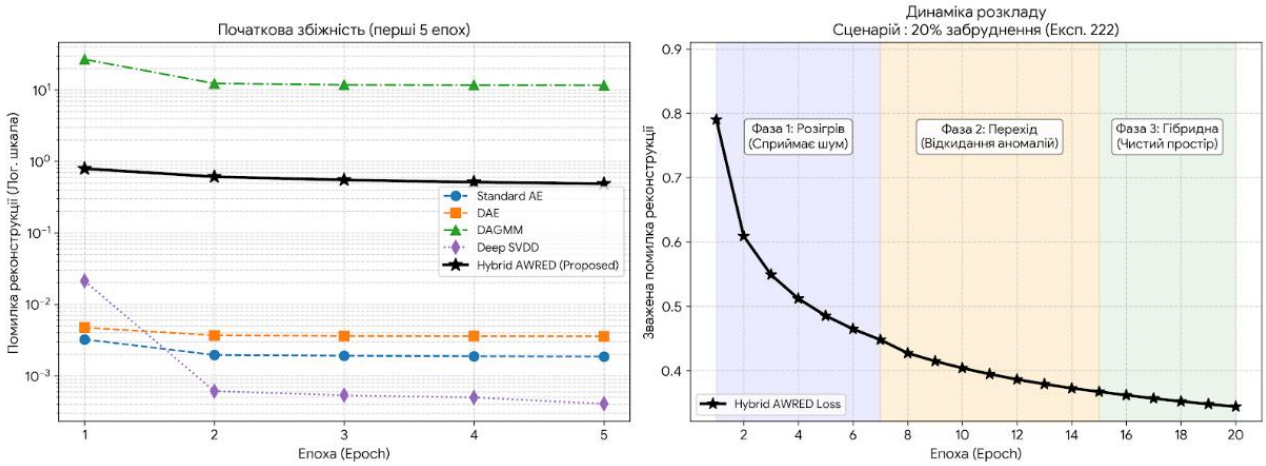

Рис. 3. Динаміка мінімізації функції втрат (сценарій 20% забруднення, експеримент 222)

Ліва панель: початкова збіжність базових моделей (перші 5 епох). Права панель: трифазний цикл навчання Hybrid AWRED із динамічною регуляризацією.

Аналіз кривих втрат (рис. 3) показав, що навчання моделі проходить у три послідовні фази, які відповідають розкладу Curriculum Dynamics. На етапі розігріву формується базове подання норми. Далі, у перехідній фазі, поступово посилюється вплив адаптивних штрафів. На завершальному етапі крива втрат набуває більш стабільного характеру. Така динаміка свідчить, що запропонований розклад дає змогу послабити вплив аномальних зразків на процес оптимізації без різких градієнтних коливань навіть за 20% забруднення вибірки.

Результати дослідження

Помірне забруднення (P = 0.05). У сценарії помірного забруднення Hybrid AWRED продемонстрував найвище значення AUC-ROC (0.874 ± 0.028) серед розглянутих моделей, перевищивши DAGMM (0.808 ± 0.013) та Deep SVDD (0.826 ± 0.020). Це свідчить, що за відносно невеликої частки аномалій запропонований підхід зберігає конкурентну якість ранжування дефектних зразків.

Масоване отруєння даних (P = 0.20). Аналізуємо таблицю 1.

Таблиця 1

Порівняльна ефективність моделей на наборі MNIST-C (рівень забруднення 5% і 20%)

Модель	AUC-ROC		F1-Score		MCC		Повнота	
	(5%)	(20%)	(5%)	(20%)	(5%)	(20%)	(5%)	(20%)
Hybrid AWRED	0.874 ± 0.028	0.722 ± 0.018	0.605 ± 0.055	0.594 ± 0.022	0.185 ± 0.005	0.217 ± 0.028	0.730 ± 0.021	0.880 ± 0.005
Standard AE	0.845 ± 0.012	0.728 ± 0.004	0.670 ± 0.000	0.603 ± 0.015	0.172 ± 0.001	0.239 ± 0.008	0.805 ± 0.045	0.706 ± 0.055
DAE	0.853 ± 0.004	0.785 ± 0.008	0.702 ± 0.006	0.717 ± 0.009	0.175 ± 0.003	0.282 ± 0.006	0.849 ± 0.031	0.786 ± 0.058
Deep SVDD	0.826 ± 0.020	0.754 ± 0.056	0.600 ± 0.010	0.517 ± 0.077	0.162 ± 0.003	0.243 ± 0.065	0.854 ± 0.045	0.584 ± 0.163
DAGMM	0.808 ± 0.013	0.732 ± 0.002	0.507 ± 0.026	0.465 ± 0.013	0.161 ± 0.005	0.240 ± 0.014	0.684 ± 0.010	0.397 ± 0.039

Аналіз результатів для сценарію 20% забруднення дозволяє виділити кілька важливих закономірностей. По-перше, базові моделі демонструють помітне погіршення показників повноти виявлення. Зокрема, у DAGMM значення Recall знижується до 0.397 ± 0.039, що може бути пов’язано з ефектом “розмазування кластерів”, тоді як у Deep SVDD спостерігається нестабільність результатів за окремими метриками (рис. 4).

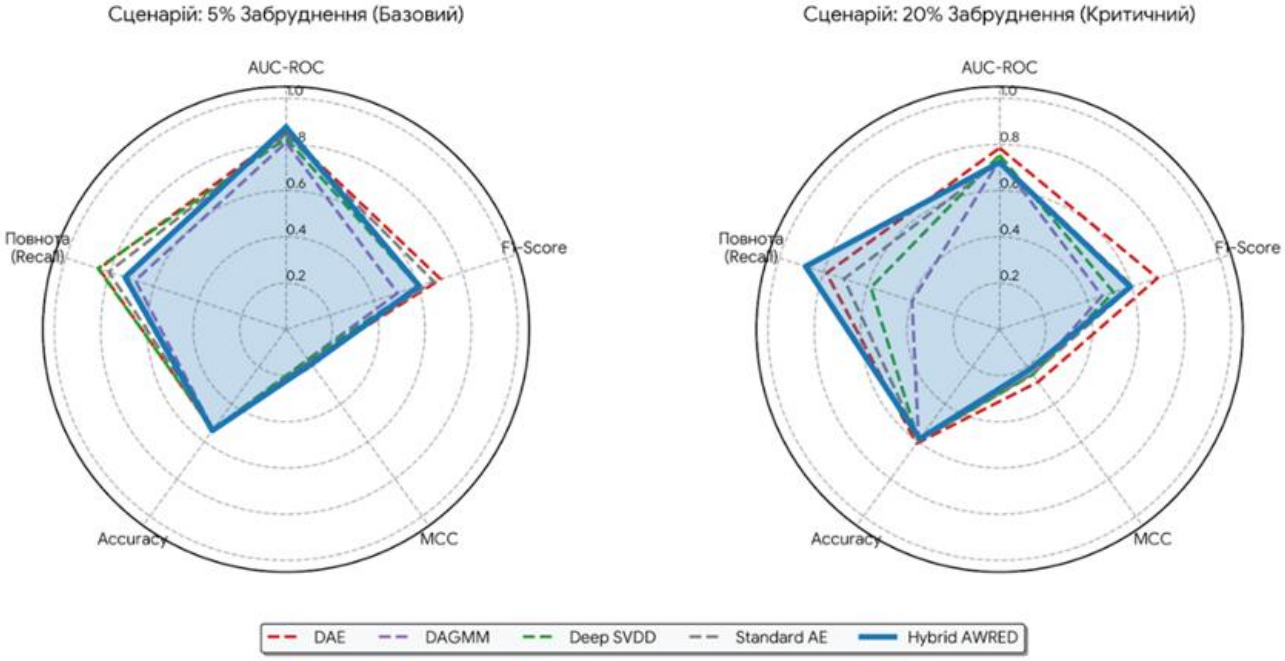

Рис. 4. Порівняльні пелюсткові діаграми (Radar Charts) комплексної ефективності моделей

На лівій панелі наведено результати для базового сценарію (5% забруднення вибірки), де більшість класичних архітектур демонструють прийнятну стабільність. На правій панелі візуалізовано значну деградацію метрик (зокрема Recall та F1-Score) для базових моделей в умовах екстремального отруєння (20%), на фоні яких запропонований метод Hybrid AWRED (суцільна лінія) демонструє більш збалансований профіль метрик. По-друге, Hybrid AWRED показує найвище значення Recall (0.880 ± 0.005) серед усіх розглянутих архітектур. Це означає, що в умовах масованого забруднення модель пропускає менше дефектних зразків, ніж альтернативні підходи. Для прикладних систем візуального контролю така властивість може бути особливо важливою, оскільки зменшення кількості пропущених дефектів часто має вищий пріоритет, ніж максимізація однієї узагальненої метрики.

По-третє, DAE у цьому сценарії демонструє найвище середнє значення AUC-ROC (0.785 ± 0.008). Це вказує, що його архітектура краще пристосована до певних типів штучних візуальних спотворень, представлених у MNIST-C. Проте, на відміну від DAE, адаптивний механізм Hybrid AWRED менш жорстко пов'язаний зі специфікою окремих візуальних артефактів, оскільки враховує порушення топології нормальних цифр. Це дає підстави розглядати його як перспективний підхід для підвищення повноти виявлення.

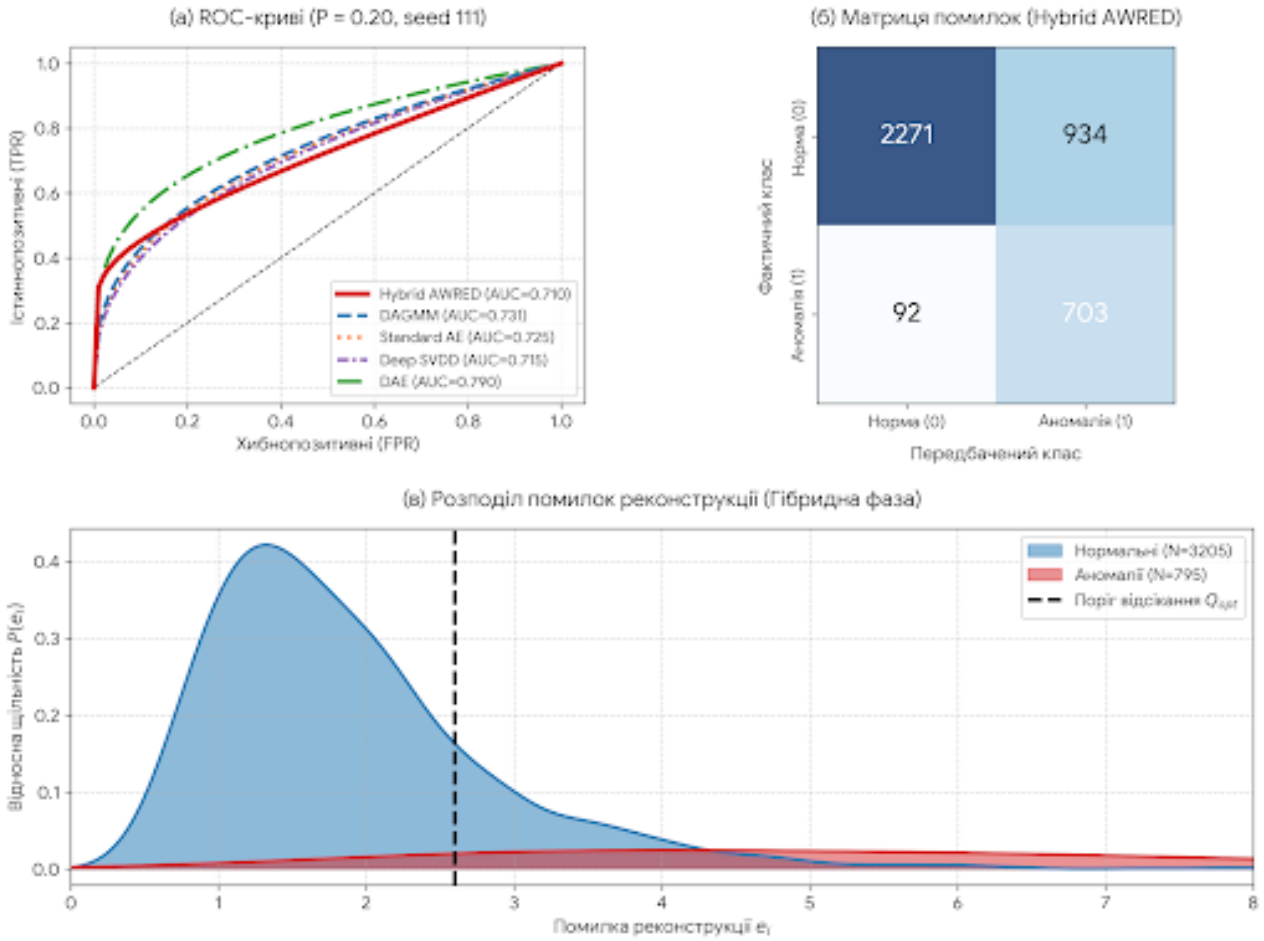

Рис. 5. Комплексна візуалізація ефективності моделі Hybrid AWRED (ілюстрація на базі конкретного запуску seed 111 при 20% забрудненні)

На рис. 5 подано ключові візуалізації результатів: ROC-криві (верхня ліва панель), матрицю помилок (верхня права панель) та графік щільності розподілу помилок (нижня панель). Вони показують, що запропонований алгоритм демонструє високу чутливість до дефектів, низьку кількість пропусків (Recall = 88.4%) і помітне розділення нормальних та аномальних зразків у просторі помилок. У цілому ці результати підтримують висновок про ефективність підходу за умов значного візуального шуму.

Висновки

У роботі досліджено стійкість систем машинного зору до масованого забруднення навчальних даних у задачі виявлення візуальних аномалій без учителя. Отримані результати показали, що за рівня структурних спотворень 20% традиційні підходи (AE, Deep SVDD, DAGMM) втрачають частину здатності до надійного виявлення дефектних зразків, що проявляється у зниженні окремих показників якості, насамперед повноти виявлення.

Для зменшення цього ефекту запропоновано архітектуру Hybrid AWRED v3, яка поєднує топологічне якорування, адаптивне зважування похибки реконструкції, динамічну регуляризацію та автономний контур Байєсівської оптимізації. У проведених експериментах така комбінація забезпечила найвище значення Recall у сценарії 20% забруднення (0.880 ± 0.005), що вказує на меншу кількість пропущених дефектів порівняно з іншими розглянутими моделями.

Разом із тим результати показали, що перевага запропонованого методу не є однаковою за всіма метриками. Зокрема, за AUC-ROC у сценарії сильного забруднення найвище середнє значення продемонстрував DAE. Тому, Hybrid AWRED v3 доцільно розглядати передусім як підхід, орієнтований на підвищення повноти виявлення та зниження ризику пропуску дефектів.

Висновки та перспективи подальших досліджень

Попри отримані обнадійливі результати для архітектури Hybrid AWRED v3, залишаються напрями її подальшого розвитку. По-перше, доцільним є дослідження можливості інтеграції механізмів просторової уваги (Spatial Attention) у ланцюг реконструкції. Такий підхід може підвищити інтерпретованість рішень моделі, дозволяючи не лише класифікувати візуальний масив як аномальний, а й формувати теплові карти, що відображатимуть найбільш підозрілі піксельні області. По-друге, перспективним напрямом є подальша оптимізація обчислювальної складності Байєсівського контуру та модуля kNN-якорування. Це може бути корисним для масштабування методу на багатоканальні зображення високої роздільної здатності, а також для його застосування в апаратних системах промислового візуального контролю, зокрема в режимах, наближених до реального часу.

Декларація про штучний інтелект

Автор не використовував штучний інтелект при створенні матеріалів статті.

Конфлікт інтересів

Автор заявляє про відсутність конфлікту інтересів та підтверджує, 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 використаної літератури

1. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research, 11(110), 3371-3408. URL: https://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf

2. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep One-Class Classification. Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80:4393-4402. URL: http://proceedings.mlr.press/v80/ruff18a.html

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. Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. Proceedings of the 26th Annual International Conference on Machine Learning (ICML), 41–48. DOI: 10.1145/1553374.1553380. URL: https://dl.acm.org/doi/10.1145/1553374.1553380

5. Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified Defenses for Data Poisoning Attacks. Advances in Neural Information Processing Systems (NeurIPS), 30. URL: https://papers.nips.cc/paper_files/paper/2017/hash/9d7311ba459f9e45ed746755a32dcd11-Abstract.html

6. Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Mądry, A., Li, B., & Goldstein, T. (2023). Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 1563-1580. DOI: 10.1109/TPAMI.2022.3162397. URL: https://ieeexplore.ieee.org/document/9743317

7. Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection: A Review. ACM Computing Surveys, 54(2), 1-38. URL: https://dl.acm.org/doi/10.1145/3439950

8. Rudolph, M., Wandt, B., & Rosenhahn, B. (2021). Same Same But DifferNet: Semi-Supervised Defect Detection with Normalizing Flows. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1907-1916. DOI: 10.1109/WACV48630.2021.00195. URL: https://openaccess.thecvf.com/content/WACV2021/html/Rudolph_Same_Same_but_DifferNet_Semi-Supervised_Defect_Detection_With_Normalizing_Flows_WACV_2021_paper.html

9. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. (2022). Towards Total Recall in Industrial Anomaly Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14318-14328. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html

10. Sun, Y., Guo, C., & Li, Y. (2022). Out-of-Distribution Detection with Deep Nearest Neighbors. International Conference on Machine Learning (ICML), 19812-19827. URL: https://proceedings.mlr.press/v162/sun22d.html

11. Barron, J. T. (2019). A General and Adaptive Robust Loss Function. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4331-4339. URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Barron_A_General_and_Adaptive_Robust_Loss_Function_CVPR_2019_paper.html

12. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems (NeurIPS), 25. URL: https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

13. Mu, N., & Gilmer, J. (2019). MNIST-C: A Robustness Benchmark for Computer Vision. ICML 2019 Workshop on Security and Privacy of Machine Learning. URL: https://arxiv.org/abs/1906.02337

14. Довженко, Т. П. (2026). Hybrid AWRED: Синергія адаптивної реконструкції та топологічної кластеризації для виявлення аномалій у мультимодальних даних.- Зв'язок. – 2026.- № 1 – с. 80-88. https://doi.org/10.31673/2412-9070.2026.017405

15. Довженко, Т. П., Зінченко, О. В. (2026). Стабільність моделей глибокого виявлення вторгнень в умовах масованих кібератак: стрес-тестування та архітектурні особливості Hybrid AWRED. - Телекомунікаційні та інформаційні технології.- 2026.- № 1 – с. 199 – 207. https://doi.org/10.31673/2412-4338.2026.019019

T. Dovzhenko

TOPOLOGICAL ANCHORING AND ADAPTIVE PENALTIES: THE HYBRID AWRED ARCHITECTURE FOR DEFECT RECOGNITION IN CONTAMINATED VISUAL DATA

In practical computer vision systems, the assumption of fully clean training data is often violated. Under structural contamination, anomalous visual samples can be partially absorbed into the latent space of normal patterns, which reduces the stability of unsupervised anomaly detection models and complicates reliable defect recognition. This issue becomes especially important when the concentration of visual distortions is high enough to affect the optimization process itself. To address this problem, the paper considers the Hybrid AWRED v3 methodology, representing the third generation of the proposed approach for defect recognition in contaminated visual data. The architecture combines topological anchoring, adaptive weighting of reconstruction errors, staged penalization, and an external Bayesian optimization loop for hyperparameter tuning.

The experimental study was conducted on the modified MNIST-C benchmark under two contamination regimes, 5% and 20%, and compared the proposed method with several baseline models, including standard autoencoders, denoising autoencoders, Deep SVDD, and DAGMM. In the moderate contamination scenario (P = 0.05), Hybrid AWRED v3 achieved the highest AUC-ROC = 0.874 ± 0.028, indicating competitive anomaly ranking quality under limited poisoning. In the severe conta-

mination scenario (P = 0.20), the proposed architecture obtained the highest Recall = 0.880 ± 0.005 among all compared methods, outperforming DAGMM (0.397 ± 0.039) and Deep SVDD (0.584 ± 0.163) in terms of detection completeness. At the same time, the highest mean AUC-ROC in this regime was observed for DAE (0.785 ± 0.008), which suggests that the compared models differ in their sensitivity to ranking quality and to the minimization of missed defects.

Overall, the obtained results indicate that Hybrid AWRED v3 is particularly effective in scenarios where reducing the number of missed anomalies is more important than maximizing a single aggregate metric. Its main advantage in the present study is therefore most directly supported by recall under severe contamination, while broader claims of superiority across all evaluation criteria would require additional evidence. This makes the method a promising candidate for machine vision applications operating in noisy or weakly controlled environments, where contamination of the training sample cannot be excluded in advance.

Keywords: anomaly detection, data poisoning, computer vision, Hybrid AWRED, topological anchoring, Bayesian optimization, staged learning (Curriculum Learning).

Надійшла до редакції: 06.03.2026
Прийнята до друку: 21.04.2026
Опубліковано: 27.04.2026