УДК 004.056.5 doi: 10.20998/2413-4295.2026.02.02

АНАЛІЗ ІНТЕЛЕКТУАЛЬНИХ МЕТОДІВ ВИЯВЛЕННЯ КІБЕРІНЦИДЕНТІВ У АТОМНІЙ ЕНЕРГЕТИЦІ НА ОСНОВІ ОДНОКЛАСОВОГО НАВЧАННЯ

С.С. ЛИС [1*], О.М. ЛИС[1], І.О. ДЗЮБА[2]

[1]*Інститут комп'ютерних технологій, автоматики та метрології, Національний університет «Львівська політехніка», Львів, УКРАЇНА*
[2]*Інституту енергетики та систем керування, Національний університет «Львівська політехніка», Львів, УКРАЇНА*
[*]*e-mail: Lysss@ukr.net*

АНОТАЦІЯ *У роботі представлено концептуальний зсув у підходах до забезпечення кібербезпеки та технічної надійності атомних електричних станцій (АЕС). Замість традиційної реактивної класифікації відомих атак запропоновано проактивне моделювання «нормального стану» об'єкта (Normal State Recognition), яке дозволяє виявляти будь-які відхилення від фізично обґрунтованої поведінки системи. Це є актуальним з огляду на критичний дефіцит емпіричних даних про кіберінциденти в атомній енергетиці та обмеженість сигнатурних методів у протидії атакам «нульового дня». Обґрунтовано застосування методів однокласового навчання (One-Class Classification, OCC), які навчають алгоритми розпізнавати адекватну поведінку системи без потреби у великих масивах аварійних даних. Для формалізації нормального стану використано чотирьох елементну схему аналізу «режим-стан-об'єкт-взаємозв'язок», що забезпечує структуроване представлення багатовимірного простору даних. Досліджено внутрішні принципи функціонування сучасних архітектур – автоенкодерів (AE/TSAE) та Isolation Forest (iForest), здатних ідентифікувати приховані закономірності та мікроаномалії на ранніх етапах, до виникнення критичних станів.*
Ключові слова*: кіберінцидент, однокласове навчання, виявлення аномалій, пояснюваний штучний інтелект, атомна енергетика, нормальний стан роботи об'єкта, критична інфраструктура.*

ANALYSIS OF INTELLIGENT METHODS FOR DETECTING CYBER INCIDENTS IN NUCLEAR POWER ENGINEERING BASED ON ONE-CLASS LEARNING

S. LYS[1], O. LYS[1], I. DZYUBA[2]

[1]*Institute of Computer Technologies, Automation and Metrology, Lviv Polytechnic National University, Lviv, Ukraine*
[2]*Institute of Power Engineering and Control Systems, Lviv Polytechnic National University, Lviv, Ukraine*

ABSTRACT *The paper presents a conceptual shift in approaches to ensuring cybersecurity and technical reliability of nuclear power plants (NPPs). Instead of the traditional reactive classification of known attacks, a proactive modeling of the "normal state" of an object (Normal State Recognition) is proposed, which makes it possible to detect any deviations from the physically justified behavior of the system. This is particularly relevant given the critical lack of empirical data on cyber incidents in nuclear power engineering and the limitations of signature-based methods in countering zero-day attacks. The application of one-class classification (OCC) methods is substantiated, as they train algorithms to recognize the proper system behavior without the need for large datasets of аварійних (incident/failure) data. To formalize the normal state, a four-element analysis scheme "mode–state–object–relationship" is used, providing a structured representation of a multidimensional data space. The internal principles of operation of modern architectures – autoencoders (AE/TSAE) and Isolation Forest (iForest) – are studied, as they are capable of identifying hidden patterns and micro-anomalies at early stages, before critical conditions arise.*
Keywords: *cyber incident, one-class learning, anomaly detection, explainable artificial intelligence, nuclear power engineering, normal operating state of an object, critical infrastructure.*

Вступ

Сучасна атомна енергетика перебуває на етапі інтенсивної цифрової трансформації, що супроводжується переходом від аналогових до повністю цифрових систем контролю та управління (I&C). Поєднання операційних технологій (OT) та інформаційних систем (IT) створює нові вектори загроз, включаючи ін'єкції хибних даних (FDI) та атаки типу «відмова в обслуговуванні» (DoS), які здатні маскуватися під фізичні несправності [1-7]. Традиційні системи виявлення вторгнень, засновані на сигнатурах, демонструють обмеженість у протидії атакам «нульового дня», що вимагає розробки методів, здатних ідентифікувати аномалії без попередніх знань про їхній тип. Структурна складність сучасних АЕС та ієрархія інформаційних потоків між різними рівнями управління (рис. 1) обумовлюють необхідність впровадження інтелектуальних засобів моніторингу на кожному етапі обробки даних [1-7].

У цьому контексті особливої актуальності набувають підходи, орієнтовані не на виявлення відомих шаблонів атак, а на формування цілісного уявлення про нормальне функціонування технологічного об'єкта. Така парадигма передбачає

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

побудову моделей, здатних відображати фізично узгоджену поведінку системи в умовах штатної експлуатації, що дозволяє фіксувати навіть незначні відхилення, які можуть бути індикаторами як кібернетичних впливів, так і прихованих технічних несправностей.

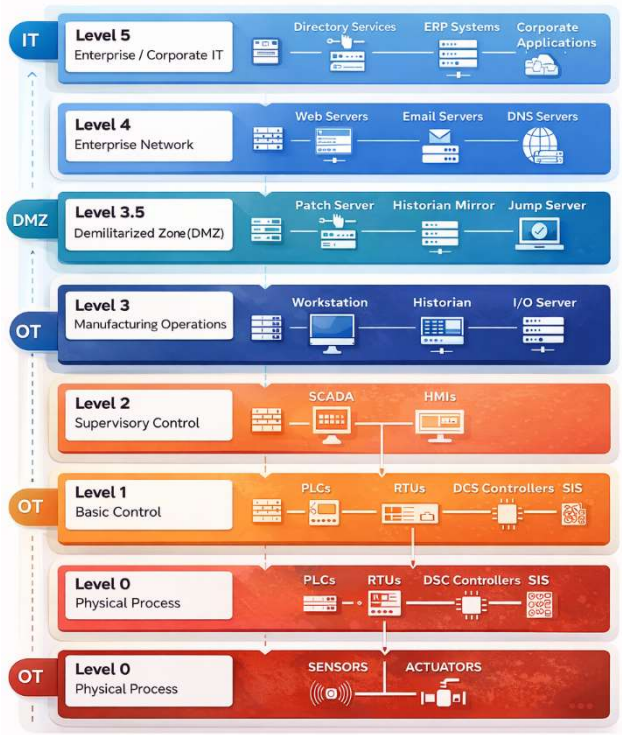

Рис. 1 – Архітектура кіберфізичної системи за моделлю Пердью.

Важливою особливістю кіберфізичних систем атомної енергетики є тісний взаємозв'язок між інформаційними сигналами та фізичними процесами. Це означає, що будь-яке втручання в інформаційний контур потенційно відображається на параметрах технологічного процесу, однак ці зміни можуть бути латентними або компенсованими системами автоматичного регулювання. У зв'язку з цим виникає потреба у методах, здатних виявляти не лише явні аномалії, але й приховані, слабовиражені відхилення у багатовимірному просторі параметрів.

Крім того, зростає значення пояснюваності прийнятих рішень, оскільки в умовах критичної інфраструктури будь-яке автоматизоване втручання має бути обґрунтованим та зрозумілим для оперативного персоналу. Це обумовлює інтеграцію підходів пояснюваного штучного інтелекту, які дозволяють інтерпретувати результати роботи моделей та встановлювати причинно-наслідкові зв'язки між виявленими аномаліями та станом системи.

Таким чином, виникає науково-практична задача розробки інтелектуальних методів виявлення кіберінцидентів, що базуються на аналізі нормального стану об'єкта, враховують багаторівневу структуру АЕС та забезпечують ранню діагностику відхилень у режимі реального часу.

Аналіз літературних джерел та постановка проблеми дослідження

Аналіз сучасних досліджень у галузі кібербезпеки на об'єктах атомної енергетики демонструє зростаючу увагу до методів машинного навчання, орієнтованих на виявлення аномалій в умовах дефіциту позначених даних.

Liu et al. [8] запропонували метод захисту ICS від адверсарних атак на основі LSTM-ED, який ефективно генерує протокольно-валідні зразки та використовує захисний механізм LSTM-FWED без попереднього знання про тип атаки, однак залишається у парадигмі реконструкційних помилок і не використовує переваг однокласового навчання, що обмежує застосовність в умовах критичного дисбалансу класів (30:1), характерного для ядерних енергетичних установок (ЯЕУ). Soomro et al. [9] надали вичерпний огляд застосування supervised learning (CNN, SVM, ANN) для діагностики корозії трубопроводів АЕС, визначивши дефіцит даних як критичну проблему та запропонувавши SMOTE, GAN і transfer learning для її вирішення, проте не розглядали методи OCC, які є природним рішенням саме у випадку відсутності аварійних даних. Dahm et al. [10] розробили XAI-фреймворк на основі RNN та модифікованого SHAP для виявлення FDI-атак у сигналах PUR-1 з точністю понад 93%, застосувавши адаптивне віконування для аналізу залишків, однак використання RNN обмежує масштабованість на надвеликих вибірках, тоді як iForest завдяки лінійній складності ефективно обробляє такі обсяги. Liu et al. [11] запропонували метод на основі DDPM із стратегією "noise-to-noise" для підвищення стійкості моделі до шуму датчиків, продемонструвавши перевагу над AE, VAE та GAN, проте не розглянули питання пояснюваності рішень, що є критичною вимогою NUREG-2261 [3].

Caixeta et al. [12] досягли точності 99,94% у прогнозуванні TRIP на основі LSTM/Transformer із 10-річними даними APS реактора Angra 2, однак їхній supervised підхід вимагає повної розмітки аварійних послідовностей, на противагу цьому OCC-метод навчається виключно на нормі та виявляє раніше невідомі відхилення. Rivas et al. [13] представили інтегровану SDP-систему (LSTM-AE + CNN + LSTM-D) для прогресивних реакторів, що виявляє аномалії за 20 секунд та прогнозує RUL за 720 секунд до порушення меж безпеки, проте модульна архітектура ускладнює масштабування, а відсутність XAI для CNN-класифікатора обмежує довіру операторів. Park et al. [14] розробили RIDA із GRU-AE, LightGBM та SHAP, застосувавши концепцію різноманітності (*diversity*) для підвищення надійності та rule-based систему для оцінки симптомів АОР, однак їхній підхід базується на supervised класифікації 16 попередньо відомих сценаріїв, тоді як iForest є unsupervised і виявляє раніше невідомі типи відхилень. Li et al. [15] підтвердили життєздатність VAE + iForest для виявлення аномалій АЕС у реальному часі (~3 мс), але відсутність XAI-інструментів обмежує можливість

операторів верифікувати фізичну природу аномалій. Натомість сучасна парадигма безпеки обґрунтовує доцільність інтеграції SHAP для кількісної атрибуції внеску сенсорів, забезпечуючи відповідність регуляторним вимогам.

Отже, виникає потреба у комплексному аналізі та систематизації зміни парадигми, замість класифікації конкретних атак система повинна навчаєтися досконало розуміти норму об’єкта, що забезпечує універсальність виявлення. Такий підхід дозволяє знаходити відхилення, спричинені як хакерським втручанням, так і природною деградацією обладнання, забезпечуючи високу надійність критичної інфраструктури.

Формалізація «нормального стану» об’єкта

Формалізація нормального стану є фундаментальним етапом у створенні надійної системи виявлення аномалій для об’єктів атомної енергетики. У роботі використано спеціалізовану чотирьохелементну схему аналізу для структурування багатовимірного простору даних:

- Режим (або mode) визначає глобальний операційний статус реактора, такий як пуск, робота на потужності або зупинка, встановлюючи фізичні межі для очікуваної поведінки сигналів.
- У межах кожного режиму система може проходити через множину валідних станів, які відповідають конкретним сукупностям умов усіх системних змінних у певний момент часу.
- Об’єкти визначаються як окремі фізичні або віртуальні одиниці системи, починаючи від датчиків нейтронного потоку і закінчуючи пакетами даних у мережевих потоках.
- Взаємозв’язок описує логічні або фізичні взаємодії між цими об’єктами, наприклад, протокольний зв’язок між програмованим логічним контролером (ПЛК) та віддаленим терміналом (див. табл. 1). Для моделювання зв'язків між об'єктами використовуються промислові протоколи, зокрема Modbus та Ethernet, які є стандартом для SCADA-систем, що використовуються на об'єктах атомної енергетики [1, 2].

Таблиця 1 – Елементи системного формалізму характеристики станів об'єктів атомної енергетики

Компонент	Фізична реалізація	Роль
Режим	Глобальний статус (пуск, робота, зупинка)	Встановлює контекст для порогів
Стан	Набір умов об'єктів та зв'язків у момент *t*	Визначає точку в просторі ознак
Об'єкт	Датчики, прилади, ПЛК, віртуальні дані	Суб'єкт прямого моніторингу
Взаємозв'язок	Ethernet, Modbus	Індикатор цілісності комунікацій

Аномалії виникають, коли внутрішня або зовнішня подія змушує систему змінити властивості об’єкта або порушити встановлені зв’язки між ними. Шляхом представлення цих компонентів у вигляді високовимірного вектора ознак, нормальний стан формує щільне скупчення даних у математичному просторі. Визначення відхилення базується на тому, чи потрапляє поточний вектор стану в межі цього заздалегідь вивченого кластера «адекватної» роботи. Представлена логіка є особливо ефективною для застосувань на об'єктах атомної енергетики, де фізичні закони, зокрема нейтронна кінетика, суворо обмежують можливі переходи між станами. Таким чином, модель вивчає не просто набір цифр, а динамічні залежності, як-от співвідношення між потужністю реактора та температурою теплоносія. Це забезпечує відповідність нормального стану фізичній реальності об’єкта та мінімізує вплив електронного шуму.

Механізм та сутність навчання алгоритмів

Основна задача дослідження – перехід до концепції навчання на основі одного класу (OCC), який дозволяє моделі ідентифікувати аномалії за принципом «не-норма», що є критично важливим для атомної енергетики, де дані про аварійні режими та кібератаки є надзвичайно дефіцитними або недоступними через міркування безпеки. Використання методів однокласового навчання забезпечує стійкість системи до раніше невідомих типів атак та поступової фізичної деградації компонентів. Механізм навчання спрямований на вилучення глибинних статистичних та фізичних кореляцій, які визначають стабільну поведінку реактора. У результаті система стає експертом у розпізнаванні здорового стану, автоматично маркуючи будь-яку невідповідність як підозрілу подію. Цей процес вимагає суворого математичного визначення меж прийнятної поведінки в багатовимірному просторі сигналів.

Побудова опису меж норми. У межах парадигми OCC навчання алгоритмів фокусується на створенні компактної описової моделі, яка оточує скупчення нормальних даних у багатовимірному просторі. На відміну від стандартних класифікаторів, які намагаються розділити два класи гіперплощиною, OCC будує замкнену межу навколо здорових точок. Процес навчання мінімізує об’єм цієї межі, одночасно гарантуючи, що максимальна кількість прикладів нормальної експлуатації потрапляє всередину. Алгоритм вивчає не лише абсолютні значення параметрів (потік нейтронів, тиск), а й складні часові залежності між ними, що формує «цифровий відбиток» норми. Якщо вектор стану системи S_i виходить за межі цієї оболонки, система констатує відхилення без потреби знати причину його виникнення. Використання ансамблевих методів, зокрема Isolation Forest (iForest) [12, 13], дозволяє цій межі бути адаптивною до різних режимів роботи –

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

аномальний бал (*anomaly score*) визначається глибиною ізоляції точки в дереві розбиттів: нормальні точки ізолюються значно глибше, ніж аномальні. Такий механізм забезпечує нульовий рівень помилкових спрацьовувань (False Positives) при правильному налаштуванні гіперпараметрів на валідаційній вибірці. Важливим аспектом є стійкість цієї межі до природних коливань інтенсивності нейтронів, які не повинні сприйматися як аномалії. Впровадження інструментів стабільності навчання гарантує, що модель не перенавчається на випадкових шумах, зберігаючи здатність до узагальнення.

Математична логіка виявлення відхилень через реконструкцію. Фундаментальним механізмом виявлення в некерованих нейронних архітектурах є аналіз похибки реконструкції (Reconstruction Error). Автоенкодери [11, 15] навчаються стискати вхідний вектор ознак X, що складається з 67 сигналів ОТ та 11 ІТ, у латентне представлення низької розмірності $h = f(X)$, а потім відновлювати його до початкового стану $\hat{X} = g(h)$. Оскільки нейронна мережа в ході навчання засвоїла виключно приклади нормальної експлуатації, вона оптимізується для ідеального відновлення саме цих патернів. У разі появи аномалії – ін'єкції хибних даних або технічної несправності – мережа не здатна ефективно реконструювати вхідний сигнал, що призводить до математичного зростання похибки E_{rec} , що розраховується як квадрат евклідової відстані між оригіналом та відновленим вектором:

$$E_{rec} = ||X - \hat{X}||^2$$

Якщо значення перевищує заздалегідь визначений статистичний поріг τ, система ініціює сигнал тривоги. Він встановлюється на основі валідаційних даних «норми» з урахуванням допустимого рівня помилкових тривог і фактично відображає верхню межу природних коливань сигналів у штатних режимах роботи. На практиці такий поріг часто визначається через статистичні характеристики похибки на нормальних даних, зокрема:

$$\tau = \mu_E + k * \sigma_E,$$

де μ_E позначає математичне сподівання похибки реконструкції, σ_E – її середньоквадратичне відхилення, а коефіцієнт k визначає поріг чутливості детектора. Використання такої статистичної оцінки дозволяє алгоритму гнучко адаптуватися до природної варіативності технологічних процесів, ефективно мінімізуючи ймовірність хибних спрацьовувань.

З програмної точки зору, принцип дії автоенкодера ґрунтується на зниженні розмірності вхідних даних із подальшим їх відновленням, що дозволяє сформувати базовий профіль нормальної роботи системи. Відповідно, похибка реконструкції виконує функцію метрики віддаленості поточного вектора стану від множини нормальних режимів. За умов штатної експлуатації значення цієї похибки прямує до мінімуму. Однак у разі виникнення аномалій порушуються приховані кореляційні зв'язки між сигналами, що спричиняє різке зростання помилки відновлення. Концептуально це можна інтерпретувати як оцінку відхилення від нормальних станів M:

$$E_{rec}(X) \approx dist(X, M)^2$$

Навчена модель неявно фіксує фізично припустимі межі функціонування ЯЕУ, детерміновані законами термодинаміки та нейтронної кінетики. Її головна перевага у здатності ідентифікувати порушення складних багатовимірних взаємозв'язків, не зосереджуючись лише амплітудні відхилення ізольованих параметрів. Що нижчим є значення E_{rec}, то вищою є ймовірність належності стану до номінального експлуатаційного профілю. Фундаментальний принцип методу візуалізовано на рисунку 2. Процес відновлення нормальних сигналів відбувається з високою точністю, тоді як поява аномалій супроводжується різким зростанням похибки, що і слугує їх тригером.

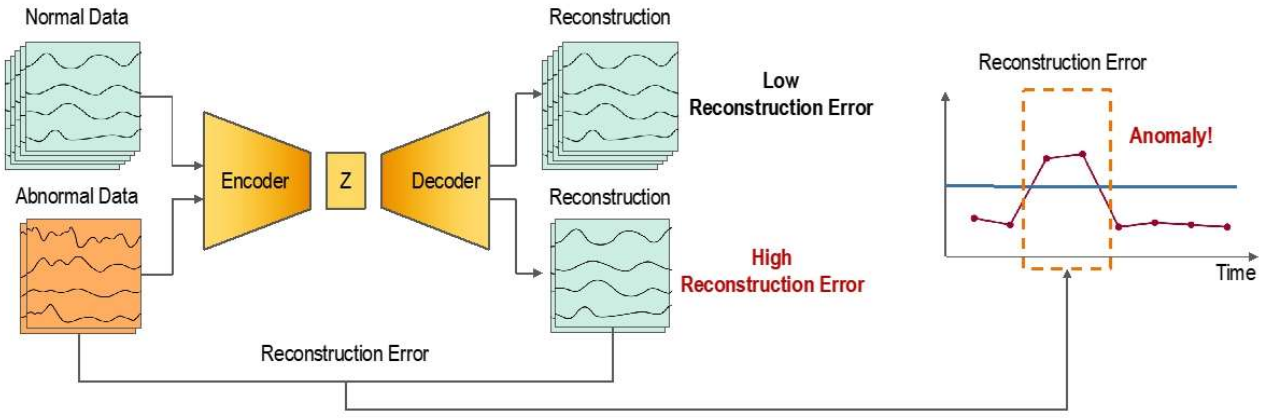

Рис. 2 – Інтерпретація моделі виявлення аномалій на основі автоенкодера.

Використання двоетапних автоенкодерів (TSAE) [7] дозволяє розділити навчання на аналіз динаміки часових рядів та аналіз відхилень, що значно підвищує чутливість до мікро-аномалій. Дана логіка дозволяє виявляти приховані тренди, які ще не досягли критичних порогів спрацювання традиційних систем захисту, забезпечуючи оператору додатковий час для прийняття рішень.

Аналіз методологій досліджень та збору даних

Проведено аналіз дослідження, виконаних на базі цифрового дослідницького реактора PUR-1 (Університет Пердью) [6, 10], які передбачали повний цикл збору даних у реальному часі, їхню попередню обробку та структурування для потреб алгоритмів машинного навчання. Було розроблено сценарій використання, що охоплює 14 різних станів системи, включаючи нормальну роботу та 13 аномальних станів різної природи. Такий підхід дозволив підтвердити здатність ШІ розпізнавати норму в умовах, максимально наближених до реальної експлуатації АЕС. Особлива увага приділялася фізичній достовірності даних та їхній відповідності законам нейтронної кінетики та термодинаміки. Зібрані дані є одними із найбільших у галузі відкритих досліджень кібербезпеки АЕС.

Параметризація та характеристика наборів даних. Для побудови репрезентативного профілю «нормального стану» було зібрано 67 багатовимірних сигналів операційних технологій (ОТ) та 11 сигналів інформаційних технологій (ІТ). ОТ-дані включають критичні фізичні показники, такі як потік нейтронів (*n*-flux), положення регулюючих стрижнів (RR position), температуру теплоносія першого контуру та напругу на магнітах приводів. ІТ-дані відображають мережеву активність між рівнями 3 та 4 архітектури, включаючи кількість пакетів на секунду, використання центрального процесора (CPU load) та мережеву затримку (Latency). Перелік сигналів і приклади даних наведено в табл. 2.

Таблиця 2 – Склад багатовимірних сигналів для навчання моделі ШІ

Параметр	Категорія (ОТ/ІТ)	Походження	Приклад даних
Потік нейтронів	ОТ (Process)	Level 0	Фізичні імпульси $(n/cm^2 \cdot s)$
Положення стрижнів	ОТ (Control)	Level 2	Дискретні кроки приводу
Пакети на секунду	ІТ (Comm)	Level 4	Мережевий трафік TCP/IP
Завантаження CPU	ІТ (Host)	Level 4	Ресурси робочої станції

Загальний обсяг «чистої норми» для навчання склав понад 13,4 мільйона точок даних для ОТ та 638,000 для ІТ, що охоплюють період експлуатації з серпня 2022 по червень 2023 року [9, 11]. Такий масштаб вибірки дозволив алгоритмам ШІ вивчити всі можливі стаціонарні коливання та допустимі перехідні процеси реактора. Дані були розподілені на навчальну, валідаційну та тестову вибірки у пропорції 60/20/20 відповідно, що є академічним стандартом для запобігання перенавчанню. Співвідношення балансу класів (*Balance Ratio*) у тестах підтримувалося на рівні 30:1 (норма до аномалії), що відображає реальний експлуатаційний дисбаланс подій. Вибір саме цих сигналів ґрунтувався на знаннях домену, відсікаючи нерелевантні параметри температуру в приміщенні чи стан системи HVAC.

Попередня обробка та робота з артефактами даних. Реальні дані ЯЕУ містять значну кількість артефактів, які за відсутності належної обробки можуть призвести до високого рівня хибнопозитивних тривог. У ході дослідження було виявлено, що випадкові викиди (*outliers*), спричинені електронним шумом датчиків [9, 11], або фізичними процесами (бульбашки на термопарі), становлять приблизно 0,0482% від загального обсягу. Окремою проблемою стали нульові значення (0,78% даних), які з'являються кластерами в режимі вимкнення через специфіку протоколів Modbus та UDP. Методологія обробки включала етап очищення від NaN-значень та нормалізацію за методом Standard Scaling для приведення всіх сигналів до єдиного масштабу. Для захоплення часових залежностей було застосовано метод «Sliding Window» з оптимальною довжиною 20 секунд та кроком в 1 секунду. Таке вікно дозволяє ШІ зафіксувати розвиток перехідного процесу, відрізняючи його від миттєвого шуму (рис. 3). Крім того, враховувалися «артефакти оператора» – варіації в сигналах, спричинені різними підходами персоналу до управління потужністю, які модель повинні сприймати як частину нормальної поведінки. Використання методів зниження розмірності через кодування в автоенкодерах дозволило зменшити вплив випадкових шумів, фокусуючи увагу алгоритму на ключових фізичних параметрах.

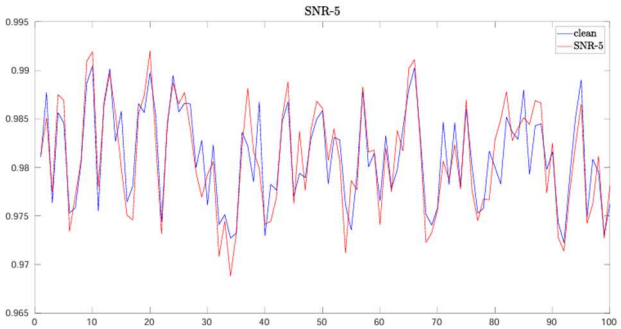

Рис. 3 – Приклади коливань даних та артефактів.

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

Поданий графік ілюструє приклад сигналу, змодельованого за допомогою S3K, до якого було додано шум із співвідношенням сигнал/шум (SNR-5). Синя крива представляє еталонний сигнал без артефактів, тоді як червона демонструє, як випадкові коливання впливають на форму сигналу, ускладнюючи його аналіз. Такий підхід візуалізації важливий для дослідження систем, що використовуються на об'єктах атомної енергетики, адже навіть незначний рівень шуму може спричинити помилкові спрацьовування алгоритмів виявлення аномалій. Це підкреслює необхідність застосування методів попередньої обробки даних та зниження розмірності, що допомагають виділити фізично значущі характеристики, усуваючи випадкові перешкоди.

Обговорення результатів дослідження

За результатами порівняльного аналізу чотирьох алгоритмів одноклasового навчання, представлених у таблиці 3, модель Isolation Forest продемонструвала найвищу ефективність у розпізнаванні нормального стану з показником F1-score на рівні 0,94-0,97. Ця перевага зумовлена принциповою відповідністю механізму ізоляції парадигмі OCC: алгоритм будує ансамбль бінарних дерев розбиттів виключно на нормальних даних, а аномальний бал (*anomaly score*) визначається глибиною ізоляції точки – без будь-яких міток аварійних станів. На відміну від LOF, iForest ефективно обробляє високорозмірний простір 78 сигналів із мінімальною деградацією продуктивності при варіюванні довжини вікна навчання.

Алгоритм COPOD, попри свою пояснюваність через хвостові ймовірності, демонструє знижену ефективність в умовах сильних фізичних кореляцій між параметрами реактора (нейтронний потік – температура теплоносія), оскільки спирається на припущення про незалежність маргінальних розподілів. SVDD, у свою чергу, стикається з квадратичною обчислювальною складністю при роботі з масивом понад 13,4 мільйона точок та суттєво деградує в умовах дисбалансу класів 30:1, характерного для реальної експлуатації АЕС. Важливою перевагою iForest стала його відносна пояснюваність у поєднанні з інструментами SHAP [7, 10], що дозволяє кількісно визначити внесок конкретних сенсорів у формування аномального балу та верифікувати фізичну природу відхилення. Згідно з вимогами NRC (NUREG-2261) [2] та МАГАТЕ [4, 5], така прозорість є обов'язковою умовою для впровадження ШІ у критичні функції безпеки АЕС.

Таблиця 3 – Порівняльна ефективність алгоритмів у виявленні відхилень

Алгоритм	Показник F1	Рівень XAI	Стійкість до шуму
Isolation Forest (iForest)	0.94 -0.97	Високий	Висока
LOF (Local Outlier Factor)	0.81 -0.88	Низький	Низька
COPOD (Copula-Based OD)	0.84 -0.90	Середній	Середня
SVDD (Support Vector Data Description)	0.79 -0.86	Низький	Середня

Для інтерпретації відхилень у розглянутих дослідженнях інтегровано інструменти SHAP, які кількісно оцінюють внесок кожного сенсора у формування аномального балу, та 1D Grad-CAM [14] для аналізу частотних спектрів у реконструкційних архітектурах. Ці методи забезпечують інженерний контроль над автоматизованою системою, дозволяючи персоналу підтвердити фізичну природу виявленого відхилення. Таким чином, інтегрована архітектура не лише констатує аномалію, а й надає доказову базу для подальшого прийняття рішень.

Висновок

Навчання алгоритмів виключно на нормальних станах об'єкта визначено як найбільш життєздатну стратегію для забезпечення безпеки інфраструктури об'єктів атомної енергетики. Оскільки передбачити всі можливі конфігурації майбутніх кібератак та фізичних відмов неможливо, ШІ має виступати як експерт із розпізнавання «здорової» поведінки системи. Підхід дозволяє ідентифікувати загрози «нульового дня» та складні маніпуляції з даними за фактом їхньої статистичної невідповідності еталонному розподілу. Це нівелює потребу в позначених аварійних даних, які практично відсутні в реальних умовах експлуатації АЕС.

Встановлено, що використання реконструкційних моделей, зокрема архітектури TSAE, забезпечує можливість ранньої ідентифікації аномальних трендів ще до активації традиційних систем захисту. Показано, що модель здатна виявляти ознаки аномалій за 38 хвилин до спрацювання автоматичного аварійного відключення реактора (*scram*). Це вікно часу надає операторам стратегічну перевагу для проведення керованої зупинки, що є критичним для мінімізації економічних втрат та запобігання пошкодженню дороговартісного обладнання.

Доведено, що інтеграція інструментів прозорості, таких як SHAP та 1D Grad-CAM, є обов’язковим фактором для успішного ліцензування систем ШІ, що використовуються на об'єктах атомної енергетики. Відповідно до стратегічного плану NRC США (NUREG-2261), оператор повинен чітко розуміти логіку, за якою система ініціювала сигнал тривоги. Прозорість прийняття рішень забезпечує довіру

персоналу до рекомендацій ШІ та дозволяє швидко локалізувати джерело проблеми. Відсутність XAI залишається основним бар'єром для впровадження «чорних скриньок» глибокого навчання у критичні цикли управління.

Підтверджено спроможність запропонованих алгоритмів функціонувати в умовах насиченого інформаційного фону реактора, що містить електронний шум та систематичні похибки приладів. Застосування методу «Sliding Window» тривалістю 20 секунд у поєднанні з попередньою нормалізацією дозволяє моделі ігнорувати випадкові сплески без втрати чутливості до реальних аномалій, що гарантує високу точність виявлення навіть при втраті частини мережевих пакетів під час DoS-атак.

Список літератури

1. Purdue Model For ICS Security. *Purdue Model enhances ICS security through network segmentation and defense-in-depth to safeguard critical infrastructure*. 2025.
2. Cybersecurity of Digital I&C Systems | Nuclear Regulatory Commission. *Nuclear Regulatory Commission*, January 12, 2026.
3. U.S. Nuclear Regulatory Commission. *Artificial Intelligence Strategic Plan: Fiscal Years 2023-2027* (NUREG-2261). Washington, DC, 2023.
4. International Atomic Energy Agency. Computer Security for Nuclear Security. IAEA Nuclear Security Series No. 42-G, Implementing Guide. Vienna: IAEA, 2021.
5. International Atomic Energy Agency. New Research Project on Computer Security For Nuclear AI. International Atomic Energy Agency | Atoms for Peace and Development. Paulina Rosol-Barrass**,** IAEA Department of Nuclear Safety and Security, 20 October 2025.
6. Purdue University / U.S. NRC. Characterization of Nuclear Cyber Security States Using Artificial Intelligence and Machine Learning. Technical Letter Report TLR-RES/DE-2024-03e. 2024.
7. Jiapeng Yang, Zuhua Jiang, TSAE: A teacher-student transformer autoencoder for restoration of fNIRS signals with channel contribution weights analysis, Advanced Engineering Informatics, Volume 72, 2026, https://doi.org/10.1016/j.aei.2026.104450.
8. Yaru Liu, Lijuan Xu, Shumian Yang, Dawei Zhao, Xin Li, Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems, Computers & Security, Volume 140, 2024, https://doi.org/10.1016/j.cose.2024.103750.
9. Afzal Ahmed Soomro, Osman K. Siddiqui, Afaque Shams, Belal Almomani, Machine learning applications in nuclear power plant piping inspection: A review of methods, data, and future trends, Annals of Nuclear Energy, Volume 225, 2026, https://doi.org/10.1016/j.anucene.2025.111760.
10. Zachery Dahm, Vasileios Theos, Konstantinos Vasili, William Richards, Konstantinos Gkouliaras, Stylianos Chatzidakis, A one-class explainable AI framework for identification of non-stationary concurrent false data injections in nuclear reactor signals, Nuclear Engineering and Design, Volume 444, 2025, https://doi.org/10.1016/j.nucengdes.2025.114359.
11. Shiqiao Liu, Zifei Zhu, Xinwen Zhao, Yangguang Wang, Xiang Sun, Lei Yu, Unsupervised anomaly detection for Nuclear Power Plants based on Denoising Diffusion Probabilistic Models, Progress in Nuclear Energy, Volume 178, 2025, https://doi.org/10.1016/j.pnucene.2024.105521.
12. Bernardo M. Caixeta, Marcelo C. Santos, Alan M.M. de Lima, Victor H.C. Pinheiro, Roberto Schirru, LSTM and transformer-based approach for nuclear reactor event sequence forecasting and TRIP detection, Progress in Nuclear Energy, Volume 196, 2026, https://doi.org/10.1016/j.pnucene.2026.106354.
13. Andy Rivas, Gregory Kyriakos Delipei, Ian Davis, Satyan Bhongale, Jason Hou, A system diagnostic and prognostic framework based on deep learning for advanced reactors, Progress in Nuclear Energy, Volume 170, 2024, https://doi.org/10.1016/j.pnucene.2024.105114.
14. Ji Hun Park, Hye Seon Jo, Sang Hyun Lee, Sang Won Oh, Man Gyun Na, A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP, Nuclear Engineering and Technology, Volume 54, Issue 4, 2022, Pages 1271-1287,https://doi.org/10.1016/j.net.2021.10.024.
15. Xiangyu Li, Tao Huang, Kun Cheng, Zhifang Qiu, Tan Sichao, Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model, Annals of Nuclear Energy, Volume 167, 2022, https://doi.org/10.1016/j.anucene.2021.108785.

References

1. Purdue Model For ICS Security. *Purdue Model enhances ICS security through network segmentation and defense-in-depth to safeguard critical infrastructure*. 2025.
2. Cybersecurity of Digital I&C Systems | Nuclear Regulatory Commission. *Nuclear Regulatory Commission*, January 12, 2026.
3. U.S. Nuclear Regulatory Commission. *Artificial Intelligence Strategic Plan: Fiscal Years 2023-2027* (NUREG-2261). Washington, DC, 2023.
4. International Atomic Energy Agency. Computer Security for Nuclear Security. IAEA Nuclear Security Series No. 42-G, Implementing Guide. Vienna: IAEA, 2021.
5. International Atomic Energy Agency. New Research Project on Computer Security For Nuclear AI. International Atomic Energy Agency | Atoms for Peace and Development. Paulina Rosol-Barrass**,** IAEA Department of Nuclear Safety and Security, 20 October 2025.
6. Purdue University / U.S. NRC. Characterization of Nuclear Cyber Security States Using Artificial Intelligence and Machine Learning. Technical Letter Report TLR-RES/DE-2024-03e. 2024.
7. Jiapeng Yang, Zuhua Jiang, TSAE: A teacher-student transformer autoencoder for restoration of fNIRS signals with channel contribution weights analysis, Advanced Engineering Informatics, Volume 72, 2026, https://doi.org/10.1016/j.aei.2026.104450.
8. Yaru Liu, Lijuan Xu, Shumian Yang, Dawei Zhao, Xin Li, Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems, Computers & Security, Volume 140, 2024, https://doi.org/10.1016/j.cose.2024.103750.
9. Afzal Ahmed Soomro, Osman K. Siddiqui, Afaque Shams, Belal Almomani, Machine learning applications in nuclear power plant piping inspection: A review of methods, data, and future trends, Annals of Nuclear Energy, Volume 225, 2026, https://doi.org/10.1016/j.anucene.2025.111760.
10. Zachery Dahm, Vasileios Theos, Konstantinos Vasili, William Richards, Konstantinos Gkouliaras, Stylianos Chatzidakis, A one-class explainable AI framework for identification of non-stationary concurrent false data

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

injections in nuclear reactor signals, Nuclear Engineering and Design, Volume 444, 2025, https://doi.org/10.1016/j.nucengdes.2025.114359.
11. Shiqiao Liu, Zifei Zhu, Xinwen Zhao, Yangguang Wang, Xiang Sun, Lei Yu, Unsupervised anomaly detection for Nuclear Power Plants based on Denoising Diffusion Probabilistic Models, Progress in Nuclear Energy, Volume 178, 2025, https://doi.org/10.1016/j.pnucene.2024.105521.
12. Bernardo M. Caixeta, Marcelo C. Santos, Alan M.M. de Lima, Victor H.C. Pinheiro, Roberto Schirru, LSTM and transformer-based approach for nuclear reactor event sequence forecasting and TRIP detection, Progress in Nuclear Energy, Volume 196, 2026, https://doi.org/10.1016/j.pnucene.2026.106354.
13. Andy Rivas, Gregory Kyriakos Delipei, Ian Davis, Satyan Bhongale, Jason Hou, A system diagnostic and prognostic framework based on deep learning for advanced reactors, Progress in Nuclear Energy, Volume 170, 2024, https://doi.org/10.1016/j.pnucene.2024.105114.
14. Ji Hun Park, Hye Seon Jo, Sang Hyun Lee, Sang Won Oh, Man Gyun Na, A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP, Nuclear Engineering and Technology, Volume 54, Issue 4, 2022, Pages 1271-1287,https://doi.org/10.1016/j.net.2021.10.024.
15. Xiangyu Li, Tao Huang, Kun Cheng, Zhifang Qiu, Tan Sichao, Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model, Annals of Nuclear Energy, Volume 167, 2022, https://doi.org/10.1016/j.anucene.2021.108785.

Відомості про авторів / About the Authors

Лис Степан Степанович – кандидат технічних наук, доцент; Інститут комп'ютерних технологій, автоматики та метрології, Національний університет «Львівська політехніка», вул. С. Бандери, 12, м. Львів, Україна, 79013; e-mail: lysss@ukr.net, тел.: (032) 258-23-15; ORCID: 0000-0002-7359-1177.

Stepan Lys – Assoc. Prof., Ph.D., Institute of Computer Technologies, Automation and Metrology, Lviv Polytechnic National University, 12 S. Bandery St., Lviv, 79013, Ukraine, Tel. 0038 032 258 25 15; Email: lysss@ukr.net; ORCID: 0000-0002-7359-1177.

Лис Ольга Михайлівна – студентка; Інститут комп'ютерних технологій, автоматики та метрології, Національний
університет «Львівська політехніка», вул. С. Бандери, 12, м. Львів, Україна, 79013; e-mail: olha.lys.kb.2024@lpnu.ua, тел.: (032) 258-23-15.

Olha Lys – student, Institute of Computer Technologies, Automation and Metrology, Lviv Polytechnic National University, 12 S. Bandery St., Lviv, 79013, Ukraine, Tel. 0038 032 258 25 15; Email: olha.lys.kb.2024@lpnu.ua.

Дзюба Ігор Орестович – аспірант; Інституту енергетики та систем керування, Національний університет «Львівська політехніка», вул. С. Бандери, 12, м. Львів, Україна, 79013; e-mail: ihor.o.dziuba@lpnu.ua, тел.: (032) 258-26-20.

Ihor Dzyuba – postgraduate student, Institute of Power Engineering and Control Systems, Lviv Polytechnic National University, 12 S. Bandery St., Lviv, 79013, Ukraine, Tel. 0038 032 258 26 20; Email: ihor.o.dziuba@lpnu.ua.

Будь ласка, посилайтесь на цю статтю наступним чином:

Лис С. С., Лис О. М., Дзюба І. О. Аналіз інтелектуальних методів виявлення кіберінцидентів у атомній енергетиці на основі однокласового навчання. *Вісник Національного технічного університету «ХПІ». Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». 2026. № 2 (28). С. 15-22. doi: 10.20998/2413-4295.2026.02.02

Please cite this article as:

Lys S., Lys O., Dzyuba I. Analysis of intelligent methods for detecting cyber incidents in nuclear power engineering based on one-class learning. *Bulletin of the National Technical University "KhPI". Series: New solutions in modern technology*. – Kharkiv: NTU "KhPI", 2025, no. 2(28), pp. 15–22, doi: 10.20998/2413-4295.2026.02.02.

Надійшла (received) 23.04.2026
Прийнята (accepted) 07.05.2026
Опублікована (published) 05.06.2026