УДК 620.9:004.056:004.89 DOI: 10.18372/2073-4751.86.21276

Ковилін А.В.,
orcid.org/0009-0001-6844-8931,
e-mail: anton.v.kovylin@gmail.com

ОЦІНКА ЗДАТНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ВИЯВЛЯТИ НЕВІДОМІ КІБЕРАТАКИ НА ЦИФРОВІ ПІДСТАНЦІЇ

Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України

Вступ

Цифрові підстанції є важливими елементами сучасної енергетичної інфраструктури, у яких функції моніторингу, керування, автоматизації та захисту значною мірою залежать від мережевої взаємодії між інтелектуальними електронними пристроями, серверами, комунікаційними шлюзами та системами диспетчерського керування. Використання стандартизованих протоколів, запозичених з IT-галузі, підвищує ефективність обміну даними, але одночасно збільшує ризики потенційних кібератак. Взагалі, інтеграція інформаційно-комунікаційних технологій в операційні процеси електричних підстанцій створює нові виклики для кіберзахисту.

Одним із перспективних напрямів підвищення кіберстійкості цифрових підстанцій є застосування систем виявлення вторгнень (СВВ), побудованих на основі методів машинного навчання. Такі системи здатні аналізувати параметри мережевого трафіку, виявляти відхилення від нормальної поведінки та класифікувати потенційно шкідливу активність. На сьогодні вже з'являються спеціалізовані набори даних (датасети), такі як SANDI-2024, які призначені для покращення навчання та оцінки таких систем [1-3].

У сучасних дослідженнях СВВ на основі машинного навчання розглядаються як важливий компонент захисту інтелектуальних енергетичних мереж та середовищ цифрових підстанцій [4]. Разом із тим значна частина експериментальних досліджень оцінює моделі за умов, коли приклади всіх типів атак уже присутні у навчальній вибірці. У таких сценаріях моделі машинного навчання з учителем можуть демонструвати високі значення точності, однак це не завжди означає їхню здатність виявляти нові або модифіковані атаки. Для об'єктів критичної інфраструктури ця проблема є особливо важливою, оскільки реальні кібератаки можуть відрізнятися від сценаріїв, на яких модель навчалася. У сучасних роботах із виявлення невідомих атак для цифрових підстанцій підкреслюється, що узагальнення моделей на раніше невідомі атаки або невідомі для моделей атаки залишається складним дослідницьким завданням [5, 6]. У цій роботі основну увагу приділено сценарію виявлення типів атак, не представлених під час навчання моделі, у якому певний тип атаки повністю виключається з навчальної вибірки та використовується лише на етапі тестування. Така постановка дозволяє моделювати ситуацію появи нового типу кібератаки в мережевому трафіку цифрової підстанції та перевірити, чи здатна модель виявити її без попереднього навчання на відповідних прикладах.

Аналіз останніх досліджень і публікацій

Питання використання машинного навчання для систем виявлення вторгнень у смарт-грід та цифрових підстанціях активно досліджується в останні роки. В роботі [4] систематизовано підходи до механізмів виявлення вторгнень на основі машинного навчання в інтелектуальних

енергетичних мережах та підкреслено, що складність енергетичних кіберфізичних систем вимагає поєднання ефективних моделей, якісних даних і коректних сценаріїв оцінювання. Згаданий вище датасет SANDI-2024 (представлений в журналі Data in Brief та опублікований у Zenodo) став важливим кроком для відтворюваного експериментального аналізу [1, 2]. Його цінність полягає в орієнтації саме на підстанційні середовища та наявності попередньо оброблених CSV-даних, придатних для машинного аналізу, придатних для навчання і тестування моделей машинного навчання. У GitHub-репозиторії авторів наведено опис конвеєра попередньої обробки даних, формування CSV-файлів і додавання поля Label, яке визначає тип атаки або її відсутність [3]. Окремий напрям сучасних досліджень спрямований на виявлення атак нульового дня (невідомих атак) та типів атак, не представлених у навчальній вибірці, у середовищах цифрових підстанцій. В роботі [5] запропоновано підхід до виявлення атак нульового дня в середовищах цифрових підстанцій на базі стандарту МЕК 61850 із використанням навчання в контексті (in-context learning); автори підкреслюють, що традиційні методи на основі машинного навчання часто мають обмежену здатність узагальнюватися на атаки, які не були представлені у навчальних даних. Суміжні дослідження також розглядають СВВ для комунікацій за протоколом GOOSE стандарту МЕК 61850, зокрема підходи до виявлення аномалій у GOOSE-трафіку на основі машинного навчання [7], набір даних GOOSE Secure, призначений для аналізу атак підміни повідомлень у GOOSE-трафіку [8], виявлення вторгнень на основі RNN для МЕК 61850 [9], гібридні техніки навчання для смарт-грід СВВ [10], ансамблеве моделювання (ensemble modelling) [11], вибір функцій (feature selection) для СВВ на основі МЕК 61850 [12], методи виявлення аномалій у GOOSE-мережах на основі навчання без учителя та часових моделей [13], метод виявлення аномалій на основі автоенкодерів із можливістю пояснення результатів [14], а також формування реалістичних датасетів систем СВВ для протоколу МЕК 61850 [15]. Незважаючи на наявність значної кількості робіт, відкритим залишається питання оцінювання моделей у сценаріях, де певні атаки повністю відсутні у навчальній вибірці. Саме така постановка є ближчою до практичного використання СВВ у критичній інфраструктурі, де повний перелік майбутніх атакувальних сценаріїв заздалегідь невідомий.

Мета та постановка задачі дослідження

Метою статті є оцінка здатності моделей машинного навчання з учителем та підходів, заснованих на аномаліях (тобто на основі виявленні аномалій), до виявлення раніше невідомих типів кібератак у попередньо обробленій частині набору даних SANDI-2024, що стосується протоколу МЕК 104, що має відношення до використання промислового протоколу для зв’язку SCADA-систем із підстанціями та енергетичним обладнанням через TCP/IP-мережі.

Для досягнення поставленої мети проаналізовано структуру набору даних і розподіл класів, сформовано базовий сценарій оцінювання моделей на відомих атаках, реалізовано сценарій leave-one-attack-out (за якого один тип атаки вилучається з навчальної вибірки та використовується лише на етапі тестування як раніше невідома атака), здійснено порівняння моделей машинного навчання з учителем різних класів, оцінено підходи до виявлення аномалій, навчені лише на нормальному трафіку, а також досліджено вплив порогу прийняття рішення на ефективність виявлення найскладнішої невідомої атаки. Об’єктом дослідження є мережевий трафік цифрових підстанцій у попередньо обробленій МЕК 104 частині датасету SANDI-2024. Предметом дослідження є моделі машинного навчання з учителем і

моделі, засновані на аномаліях, для виявлення типів кібератак, не представлених у навчальній вибірці (невідомих атак), у мережевому трафіку цифрових підстанцій. Наукова новизна роботи полягає у застосуванні сценарію leave-one-attack-out для оцінювання здатності моделей машинного навчання виявляти раніше не представлені у навчальній вибірці типи атак у попередньо обробленій МЕК 104 частині трафіку цифрових підстанцій, а також у порівнянні моделей машинного навчання з учителем та підходів, заснованих на аномаліях з урахуванням Recall, F1-score, Balanced Accuracy, False Positive Rate і впливу порогу класифікації.

Матеріали та методика експерименту

Для експериментального дослідження використано датасет SANDI-2024, призначений для навчання та оцінювання систем виявлення вторгнень в електричних підстанціях. Згідно з описом авторів, датасет містить сирі та попередньо оброблені мережеві захоплення, охоплює протоколи МЕК 61850 (міжнародний стандарт зв'язку для цифрових підстанцій), МЕК 104, NTP (Network Time Protocol) і PTP (Precision Time Protocol), а також поєднує реальний безпечний трафік із лабораторними сценаріями атак [1, 2].

У цій роботі використано попередньо оброблену МЕК 104 частину датасету, оскільки вона містить табличні CSV-файли з flow-based ознаками, придатними для безпосереднього застосування моделей машинного навчання. Відповідно до опису GitHub-репозиторію авторів, для протоколу МЕК 104 ознаки формуються з використанням CICFlowMeter (інструмент для перетворення мережевого трафіку з PCAP-файлів у табличні ознаки потоків), а після вилучення ознак до кожного CSV-файлу додається поле Label, яке визначає тип атаки або її відсутність [3]. Нормальний трафік представлений класом attackfree, а атакувальний трафік – сімома типами атак. Файл capture104-ntpddosattack.csv у межах експериментів позначався як клас NTP DDoS (ntpddosattack); у подальшому використовується скорочене позначення NTP DDoS.

Результати початкового аналізу (Табл. 1) показали суттєвий дисбаланс класів: найбільшу кількість записів має dosattack, тоді як attackfree, floodattack і mitmattack представлені значно меншою кількістю прикладів. Тому для оцінювання моделей використовувалися Precision, Recall, F1-score, Balanced Accuracy, False Positive Rate та ROC-AUC.

Таблиця 1. Розподіл класів у попередньо обробленій МЕК 104 частині SANDI-2024

Клас	Кількість записів
dosattack	304627
portscanattack	9710
NTP DDoS (ntpddosattack)	2278
iec104starvationattack	2028
fuzzyattack	939
attackfree	255
floodattack	108
mitmattack	26

На етапі попередньої обробки всі CSV-файли були об'єднані в єдину таблицю. Для задачі бінарного виявлення атак клас attackfree позначався як нормальний трафік, а всі інші класи – як атакувальний трафік. З набору ознак було вилучено службові та ідентифікаційні поля, які можуть призводити до витоку інформації або не мають узагальнювальної цінності для моделі, зокрема Flow ID, IP-адреси, часові мітки, назву вихідного файлу та цільові мітки. Числові ознаки використовувалися без зміни для деревоподібних моделей, а для моделей, чутливих до масштабу ознак, застосовувалася стандартизація. Категоріальні ознаки, якщо вони були наявні, перетворювалися за допомогою one-hot encoding. Оскільки клас dosattack суттєво домінував у вихідному наборі, для

експериментів застосовано обмеження максимальної кількості прикладів атакувальних класів до 255 записів. Для класів, що містили менше 255 прикладів, використовувалися всі доступні записи. Першим етапом експерименту було формування базового сценарію виявлення відомих атак, у якому навчальна та тестова вибірки містили приклади всіх типів атак. Основним сценарієм дослідження був leave-one-attack-out: один тип атаки повністю виключався з навчальної вибірки та використовувався лише на етапі тестування. Ця процедура повторювалася для кожного атакувального класу.

До групи моделей машинного навчання з учителем було включено Logistic Regression, K-Nearest Neighbors, Linear SVM, Gaussian Naive Bayes, Random Forest, Extra Trees та HistGradientBoosting. До групи моделей, заснованих на аномаліях, було включено Isolation Forest, One-Class SVM та Local Outlier Factor. Моделі машинного навчання з учителем навчалися з використанням як нормального, так і атакувального трафіку, тоді як моделі, засновані на аномаліях, навчалися лише на нормальному трафіку, позначеному як attackfree.

Результати експериментів та їх обговорення

У базовому сценарії виявлення відомих атак моделі машинного навчання з учителем продемонстрували майже ідеальні результати. Random Forest та HistGradientBoosting досягли значень Accuracy, Precision, Recall, F1-score і Balanced Accuracy на рівні 1.000 за FPR = 0.000. Logistic Regression показала Accuracy = 0.998, Precision = 1.000, Recall = 0.998, F1-score = 0.999, Balanced Accuracy = 0.999 та FPR = 0.000. Такі результати підтверджують, що стандартний benchmark-сценарій може переоцінювати практичну ефективність СВВ, оскільки в реальних умовах не можна припускати наявність усіх майбутніх типів атак у навчальній вибірці. Основним етапом дослідження стало оцінювання моделей у сценарії leave-one-attack-out. Серед моделей машинного навчання з учителем найвищі середні показники ефективності продемонструвала модель Extra Trees, яка досягла Recall = 0.942, F1-score = 0.964 та Balanced Accuracy = 0.971 за FPR = 0.000. Random Forest, HistGradientBoosting, Linear SVM та Logistic Regression показали близькі результати за Balanced Accuracy, однак поступилися Extra Trees за Recall і F1-score. KNN мав порівняно високий Recall, але створював більше хибних спрацювань, тоді як Gaussian Naive Bayes продемонстрував найнижче середнє значення повноти (Recall) серед моделей машинного навчання з учителем.

Окремий аналіз за типами виключених атак (Табл. 2) показав, що більшість атак добре детектувалася навіть за умов, коли відповідний тип атаки був повністю відсутній у навчальній вибірці. Водночас атака NTP DDoS виявилася найскладнішою для виявлення серед усіх атак у сценарії оцінювання моделей машинного навчання з учителем. Для неї середній Recall становив 0.425, Precision – 0.993, F1-score – 0.587, Balanced Accuracy – 0.707, FPR – 0.011, ROC-AUC – 0.869. Для інших атак Recall був близьким до 1.000: mitmattack – 0.984, portscanattack – 0.996, iec104starvationattack – 0.999, dosattack – 1.000, floodattack – 1.000, fuzzyattack – 1.000. Це означає, що високі результати у стандартному benchmark-сценарії не гарантують однаково ефективного виявлення всіх типів невідомих атак.

Оскільки NTP DDoS виявилася найскладнішою атакою, для неї додатково проведено аналіз впливу порогу класифікації на прикладі Random Forest.

Результати показали, що при стандартному порозі 0.50 модель Random Forest виявляла лише 111 із 255 атакувальних прикладів NTP DDoS. Зниження порогу до 0.25-0.30 дозволило виявити всі 255 прикладів атаки без створення хибних спрацювань на нормальному трафіку (Табл. 3). Отже, для окремих раніше невідомих атак

ефективність СВВ залежить не лише від вибору моделі, а й від механізму прийняття рішення. Серед моделей виявлення аномалій найкращий баланс між виявленням атак і кількістю хибних спрацювань продемонструвала Local Outlier Factor. Вона досягла Recall = 1.000, F1-score = 0.946 та Balanced Accuracy = 0.914 за FPR = 0.171. One-Class SVM також забезпечила повне виявлення атак, однак мала дещо вищий рівень хибних спрацювань. Isolation Forest показала нижчий FPR, але поступилася за Recall, що означає пропуск частини атакувальних прикладів. Порівняння моделей машинного навчання з учителем та моделей, заснованих на аномаліях, засвідчило, що ці групи підходів характеризуються різними сильними сторонами.

Таблиця 2. Узагальнені результати ключових експериментів

Сценарій / група моделей	Найважливіший результат	Recall	F1-score	Balanced Accuracy	FPR	Основний висновок
Відомі атаки	Random Forest / HistGradientBoosting	1	1	1	0	Стандартний сценарій дає майже ідеальні метрики
Leave-one-attack-out, моделі машинного навчання з учителем	Extra Trees	0.942	0.964	0.971	0	Найкращий середній результат серед моделей машинного навчання з учителем
Найскладніша для виявлення невідома атака	NTP DDoS, середнє значення для моделей машинного навчання з учителем	0.425	0.587	0.707	0.011	Стандартний поріг пропускає значну частину атаки
Підхід на основі аномалій	Local Outlier Factor	1	0.946	0.914	0.171	Повне виявлення атак, але вища наявність хибних тривог
Підхід на основі аномалій	One-Class SVM	1	0.942	0.908	0.184	Повне виявлення атак із вищим FPR
Підхід на основі аномалій	Isolation Forest	0.868	0.868	0.862	0.145	Нижчий FPR, але наявний пропуск частини атак

Таблиця 3. Вплив порогу класифікації Random Forest на виявлення NTP DDoS

Поріг	Precision	Recall	F1-score	Balanced Accuracy	FPR	TP	FN	FP	TN
0.5	1	0.435	0.607	0.718	0	111	144	0	76
0.35	1	0.773	0.872	0.886	0	197	58	0	76
0.3	1	1	1	1	0	255	0	0	76
0.25	1	1	1	1	0	255	0	0	76
0.2	0.996	1	0.998	0.993	0.013	255	0	1	75

Моделі машинного навчання з учителем, особливо Extra Trees, забезпечили високі середні метрики та нульовий рівень хибних спрацювань у сценарії leave-one-attack-out. Водночас моделі виявлення аномалій, зокрема Local Outlier Factor та One-Class SVM, продемонстрували повне виявлення атак у

тестових сценаріях, але ціною більшої кількості хибних спрацювань. З практичного погляду це вказує на доцільність комбінованої архітектури СВВ для цифрових підстанцій, у якій модель машинного навчання з учителем виконує роль основного класифікатора, а компонент виявлення аномалій використовується як додатковий рівень контролю для виявлення нетипової активності.

Фінансування

Статтю підготовлено за матеріалами дослідження, яке фінансується Національною академією наук України в рамках виконання науково-дослідної роботи "Розвиток методів і засобів підвищення рівня кіберзахищеності цифрових підстанцій (шифр: МОД-К)", номер державної реєстрації: 0124U002384.

Висновки

У статті розглянуто задачу оцінювання здатності моделей машинного навчання до виявлення невідомих кібератак у цифрових підстанціях на основі попередньо обробленої МЕК 104 частини датасету SANDI-2024. На відміну від стандартного сценарію виявлення відомих атак, основну увагу приділено сценарію leave-one-attack-out, який дозволяє моделювати появу раніше не представленого в навчальних даних типу атаки. Проведений аналіз показав суттєвий дисбаланс класів у попередньо обробленій МЕК 104 частині SANDI-2024. Найбільшу кількість записів має клас dosattack, тоді як нормальний трафік attackfree та окремі типи атак представлені значно меншою кількістю прикладів. Це зумовило необхідність використання метрик, стійких до дисбалансу класів, зокрема Recall, F1-score, Balanced Accuracy та False Positive Rate. У базовому сценарії виявлення відомих атак моделі машинного навчання з учителем продемонстрували майже ідеальні результати: Random Forest та HistGradientBoosting досягли значень основних метрик на рівні 1.000, а Logistic Regression показала Accuracy = 0.998. Водночас сценарій leave-one-attack-out виявив складніші закономірності. Більшість атак добре детектувалася навіть за відсутності відповідних прикладів у навчальній вибірці, однак атака NTP DDoS виявилася найскладнішою для моделей машинного навчання з учителем з погляду ефективності виявлення: середній Recall для неї становив лише 0.425. Серед моделей машинного навчання з учителем найкращий середній результат у сценарії leave-one-attack-out показала Extra Trees, яка досягла Recall = 0.942, F1-score = 0.964 та Balanced Accuracy = 0.971 за False Positive Rate = 0.000. Додатковий аналіз впливу порогу класифікації для NTP DDoS показав, що зниження порогу Random Forest з 0.50 до 0.25-0.30 дозволило підвищити Recall з 0.435 до 1.000 без збільшення False Positive Rate. Це свідчить про важливість адаптивного або сценарно залежного налаштування порогів у системах кіберзахисту критичної інфраструктури. Серед моделей, заснованих на аномаліях, найкращий баланс між повнотою виявлення атак і кількістю хибних спрацювань показала Local Outlier Factor, яка досягла Recall = 1.000, F1-score = 0.946 та Balanced Accuracy = 0.914 за False Positive Rate = 0.171. Отримані результати засвідчують, що моделі машинного навчання з учителем і моделі, засновані на аномаліях, мають різні практичні переваги: перші забезпечують нижчий рівень хибнопозитивних спрацювань, тоді як другі демонструють вищу чутливість до атак і здатність виявляти відхилення від нормального трафіку. Тому для практичного застосування в цифрових підстанціях доцільним є комбінований підхід, у якому модель машинного навчання з учителем виконує роль основного класифікатора, а компонент, заснований на аномаліях, використовується як додатковий рівень контролю для виявлення нетипової активності та потенційно невідомих атак.

Перспективами подальших досліджень є розширення експериментів на МЕК 61850 частину датасету SANDI-

2024, перевірка міжпротокольної переносимості моделей, аналіз впливу різних способів вилучення ознак із сирого PCAP-трафіку, а також дослідження методів адаптивного налаштування порогів СВВ для зменшення ризику пропуску нових кібератак.

Література

1. A dataset to train intrusion detection systems based on machine learning models for electrical substations / E. D. G. Mlot et al. Data in brief. 2024. P. 111153. URL: https://doi.org/10.1016/j.dib.2024.111153 (date of access: 10.05.2026).
2. Dataset to train intrusion detection systems based on machine learning models for electrical substations / G. M. E. Damian et al. Zenodo. URL: https://zenodo.org/records/15487636 (date of access: 10.05.2026).
3. GitHub – esguti / cybersecurity-datasets: tools to process network captures in PCAP format from IEC61850 or IEC60870-5-104 (also known as IEC104). GitHub. URL: https://github.com/esguti/cybersecurity-datasets (date of access: 10.05.2026).
4. Machine learning-based intrusion detection for smart grid computing: a survey / N. Sahani et al. ACM transactions on cyber-physical systems. 2023. URL: https://doi.org/10.1145/3578366 (date of access: 10.05.2026).
5. Zero-day attack detection in digital substations using in-context learning / F. Manzoor et al. 2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2024. P. 220-225. URL: https://doi.org/10.1109/SmartGridComm60555.2024.10738025 (date of access: 10.05.2026).
6. Detecting zero-day attacks in digital substations via in-context learning. arXiv.org. URL: https://arxiv.org/abs/2501.16453 (date of access: 10.05.2026).
7. Machine-Learning-Based anomaly detection for GOOSE in digital substations / H. Nhung-Nguyen et al. Energies. 2024. Vol. 17, no. 15. P. 3745. URL: https://doi.org/10.3390/en17153745 (date of access: 10.05.2026).
8. GOOSE secure: A comprehensive dataset for in-depth analysis of GOOSE spoofing attacks in digital substations / O. A. Tobar-Rosero et al. Energies. 2024. Vol. 17, no. 23. P. 6098. URL: https://doi.org/10.3390/en17236098 (date of access: 10.05.2026).
9. Alves de Oliveira J. A., Pereira dos Santos A. F. P., Salles R. M. RNN for intrusion detection in digital substations based on the IEC 61850. Journal of information security and applications. 2025. Vol. 94. P. 104197. URL: https://doi.org/10.1016/j.jisa.2025.104197 (date of access: 10.05.2026).
10. Hamdi N. A hybrid learning technique for intrusion detection system for smart grid. Sustainable computing: informatics and systems. 2025. P. 101102. URL: https://doi.org/10.1016/j.suscom.2025.101102 (date of access: 10.05.2026).
11. Intrusion detection in smart grCBB using artificial intelligence-based ensemble modelling / A. Alsirhani et al. Cluster computing. 2025. Vol. 28, no. 4. URL: https://doi.org/10.1007/s10586-024-04964-9 (date of access: 10.05.2026).
12. Research on intrusion detection of IEC 61850 protocol based on feature selection and triadic concept analysis / H.-M. Wang et al. Cybersecurity. 2025. Vol. 8, no. 1. URL: https://doi.org/10.1186/s42400-025-00463-5 (date of access: 10.05.2026).
13. Anomaly detection in IEC-61850 GOOSE networks: evaluating unsupervised and temporal learning for real-time intrusion detection. arXiv.org. URL: https://arxiv.org/abs/2604.14233 (date of access: 10.05.2026).
14. Explainable autoencoder-based anomaly detection in IEC 61850 GOOSE networks. arXiv.org. URL: https://arxiv.org/abs/2601.09287 (date of access: 10.05.2026).
15. ERENO: A framework for generating realistic IEC-61850 intrusion detection datasets for smart grCBB / S. E. Quincozes et al. IEEE transactions on dependable and secure computing. 2023. P. 1-15. URL: https://doi.org/10.1109/tdsc.2023.3336857 (date of access: 10.05.2026).

Ковилін А.В.

ОЦІНКА ЗДАТНОСТІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ВИЯВЛЯТИ НЕВІДОМІ КІБЕРАТАКИ НА ЦИФРОВІ ПІДСТАНЦІЇ

У статті розглянуто задачу оцінювання здатності моделей машинного навчання до виявлення невідомих кібератак у мережевому трафіку цифрових підстанцій. Актуальність дослідження зумовлена тим, що цифрові підстанції є складовою критичної інфраструктури, а їх функціонування залежить від захищеної мережевої взаємодії між пристроями автоматизації, системами моніторингу та промисловими протоколами. Для експериментального дослідження використано попередньо оброблену МЕК 104 частину набору даних SANDI-2024, призначену для навчання та оцінювання систем виявлення вторгнень в електричних підстанціях. Реалізовано сценарій leave-one-attack-out, у межах якого один тип атаки повністю вилучається з навчальної вибірки та використовується лише на етапі тестування. Проведено порівняння моделей машинного навчання з учителем та моделей виявлення аномалій. Результати показали, що у стандартному сценарії виявлення відомих атак моделі досягають майже ідеальних метрик, однак у сценарії невідомих атак їхня ефективність залежить від типу атаки та порогу прийняття рішення. Найкращий середній результат серед моделей машинного навчання з учителем показала модель Extra Trees, а серед моделей виявлення аномалій – Local Outlier Factor. Отримані результати підтверджують доцільність використання сценаріїв виявлення невідомих атак для реалістичного оцінювання систем виявлення вторгнень у цифрових підстанціях.

Ключові слова: *цифрова підстанція, кібербезпека, критична інфраструктура, система виявлення вторгнень, машинне навчання, невідомі атаки, SANDI-2024, МЕК 104.*

Kovylin A.V.

EVALUATING THE CAPABILITY OF MACHINE LEARNING MODELS TO DETECT UNKNOWN CYBERATTACKS AT DIGITAL SUBSTATIONS

The article addresses the problem of evaluating the capability of machine learning models to detect unknown cyberattacks in network traffic of digital substations. The relevance of the study is determined by the fact that digital substations are part of critical infrastructure, and their operation depends on secure network interaction between automation devices, monitoring systems and industrial communication protocols. The experimental study is based on the processed IEC 104 part of the SANDI-2024 dataset, which was designed for training and evaluating intrusion detection systems for electrical substations. The paper proposes a leave-one-attack-out evaluation scenario, in which one attack type is completely excluded from the training set and used only during testing. This approach makes it possible to evaluate the ability of models to detect attack types that were not represented in the training data. The study compares supervised machine learning models and anomaly-based approaches. The supervised group includes Logistic Regression, K-Nearest Neighbors, Linear SVM, Gaussian Naive Bayes, Random Forest, Extra Trees and HistGradientBoosting. The anomaly-based group includes Isolation Forest, One-Class SVM and Local Outlier Factor. The experimental results show that in the standard known-attack scenario, the models achieve nearly perfect classification metrics. However, in the leave-one-attack-out scenario, model performance depends significantly on the type of unseen attack and on the decision threshold. Extra Trees achieved the best average result among supervised models, while Local Outlier Factor demonstrated the best balance among anomaly-based models. Additional threshold analysis showed that the detection of the most difficult unseen attack can be significantly improved by adjusting the classification threshold. The obtained results confirm the importance of using unseen attack detection scenarios for a more realistic evaluation of IDS models in digital substations.

Keywords: *digital substation, cybersecurity, critical infrastructure, intrusion detection system, machine learning, unseen attack detection, SANDI-2024, IEC 104.*

Стаття подана до редакції: 13/05/2026
Стаття прийнята до опублікування:15/05/2026
Стаття опублікована: 30/05/2026