Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи збільшення даних для покращення контрольованого навчання при виявленні кібератак

(тема)

Виконав:
здобувач 2 року навчання,
групи СПзм-23-1
Ілля БАЄВ
(власне ім'я, прізвище)

Спеціальність
123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма
Системне програмування
(повна назва освітньої програми)

Керівник: ст. викл. Василь ЗНАЙДЮК
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет навчально-науковий центр заочної форми навчання

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Баєву Іллі Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи збільшення даних для покращення контрольованого навчання при виявленні кібератак

затверджена наказом по університету від " 07 " квітня 2025 р. № 53 Стз

2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025 р.

3. Вхідні дані до роботи доповнення даних, SMOTE, ADASYN, виявлення аномалій, кібербезпека, ансамблеве навчання, інтерпретація моделей.

4. Перелік питань, що потрібно опрацювати у роботі

Теоретичні основи дослідження

Основні положення

Тестування метод

Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій 12

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання теми кваліфікаційної роботи	07.04	
2	Аналіз літератури	08.04-21.04	
3	Побудова методів	22.04-15.05	
4	Тестування системи та отримання результатів	16.05-30.05	
5	Формування пояснювальної записки	31.05-04.06	
6	Перевірка на плагіат	05.06-06.06	
7	Рецензування роботи	07.06-10.06	
8	Подача роботи в ЕК	12.06	

Дата видачі завдання " 07 " квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ ст. викл. Василь ЗНАЙДЮК
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 59 с., 1 рис., 5 табл., 2 дод., 2 джерел.

ДОПОВНЕННЯ ДАНИХ, SMOTE, ADASYN, ВИЯВЛЕННЯ АНОМАЛІЙ, КІБЕРБЕЗПЕКА, АНСАМБЛЕВЕ НАВЧАННЯ, ІНТЕРПРЕТАЦІЯ МОДЕЛЕЙ.

Метою кваліфікаційної роботи є вивчення та оцінка різних методів доповнення даних для підвищення ефективності моделей контрольованого навчання у виявленні кібератак.

У ході виконання кваліфікаційної роботи розроблено та апробовано гібридний підхід до синтетичного розширення незбалансованих вибірок із використанням методів SMOTE, ADASYN та Tomek Links. Такий підхід дозволив зменшити ризик перенавчання та суттєво підвищити здатність моделей до узагальнення, особливо в умовах мінливих кіберзагроз. Також запропоновано стратегії безперервного та ансамблевого навчання, інтеграцію зворотного зв’язку з центру операцій безпеки (SOC) та поєднання методів навчання з учителем і без учителя. У роботі обґрунтовано необхідність використання механізмів пояснення (LIME, SHAP) для підвищення прозорості моделей. Отримані результати підтверджують перспективність застосованих методів для адаптивного та надійного виявлення аномалій у кібербезпеці.

ABSTRACT

Master’s thesis: 59 pages, 1 figures, 5 tables, 1 appendices, 22 sources.

DATA AUGMENTATION, SMOTE, ADASYN, ANOMALY DETECTION, CYBERSECURITY, ENSEMBLE LEARNING, MODEL INTERPRETABILITY.

The major goal of this thesis is to study and evaluate various data augmentation methods to enhance the performance of supervised learning models in cyberattack detection.

In this research, a hybrid approach to synthetic augmentation of imbalanced datasets was developed and tested using SMOTE, ADASYN, and Tomek Links methods. This approach significantly reduced the risk of overfitting and improved the generalization ability of models, especially in dynamic cybersecurity environments. Additionally, the study proposed strategies for continuous and ensemble learning, integration of feedback from the Security Operations Center (SOC), and a combination of supervised and unsupervised learning methods. The importance of model interpretability was also addressed through the incorporation of explanation mechanisms such as LIME and SHAP. The results confirm the effectiveness of the proposed techniques for building adaptive and reliable anomaly detection systems in cybersecurity.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

ML – Machine Learning

DL – Deep Learning

IDS – Intrusion Detection System

SMOTE – Synthetic Minority Over-sampling Technique

ADASYN – Adaptive Synthetic Sampling

SOC – Security Operations Center

AUC – Area Under the Curve

SHAP – SHapley Additive exPlanations

ВСТУП

Зростаюча складність кібератак вимагає розробки передових систем виявлення, здатних точно виявляти та пом'якшувати потенційні загрози. Це дослідження вирішує критичну проблему виявлення кібератак, використовуючи комплексний підхід, який включає створення реалістичного, але незбалансованого набору даних, що імітує різні типи кібератак. Визнаючи притаманні обмеження, що виникають через незбалансовані дані, ми дослідили кілька методів доповнення даних, щоб підвищити ефективність навчання моделі та забезпечити надійну продуктивність у різних сценаріях атак. По-перше, ми створили детальний набір даних, що відображає реальні умови мережевих вторгнень, імітуючи ряд типів кібератак, гарантуючи, що він втілює типові дисбаланси, що спостерігаються в реальних загрозах кібербезпеці. Згодом ми застосували кілька методів доповнення даних, включаючи SMOTE та ADASYN, щоб усунути перекіс у розподілі класів, тим самим забезпечуючи більш збалансований набір даних для навчання моделей машинного навчання з учителем. Наша оцінка цих методів у різних моделях, таких як випадкові ліси та нейронні мережі, демонструє значне покращення можливостей виявлення. Крім того, аналіз також поширюється на дослідження важливості ознак, надаючи критичне розуміння того, які атрибути найбільше впливають на прогностичні результати моделей. Це не лише покращує інтерпретацію моделей, але й допомагає в удосконаленні інженерії ознак та процесів вибору для оптимізації продуктивності.

1 ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

1.1 Цілі дослідження

Національний інститут стандартів і технологій надає кілька визначень кібербезпеки, які можна об’єднати в один принцип: кібербезпека передбачає впровадження захисних заходів та засобів контролю, спрямованих на запобігання шкоді та захист інформації, що зберігається в комп’ютерних системах або передається через комунікаційні мережі. Ця практика має першорядне значення для забезпечення доступності, цілісності та конфіденційності інформації.

У звіті Агентства Європейського Союзу з кібербезпеки підкреслюється значне збільшення кількості кібератак наприкінці 2022 року та в першій половині 2023 року. Тривале російське вторгнення в Україну вважається основним каталізатором цього зростання. Крім того, у звіті визначено вісім основних груп загроз, підкреслюючи динамічний та еволюціонуючий характер кіберзагроз.

Загрози даним поділяються на витоки даних та порушення безпеки даних, які відрізняються головним чином метою розкриття інформації. Витік даних зазвичай є ненавмисним розкриттям інформації через людську помилку або системні вразливості, тоді як витік даних – це навмисна атака, спрямована на крадіжку інформації.

Загрози доступності включають атаки типу «відмова в обслуговуванні» (DoS), які порушують нормальну роботу, перевантажуючи системи або мережі надмірним трафіком, зазвичай з кількох джерел. Крім того, інтернет-загрози включають низку навмисних або випадкових перебоїв в електронному зв'язку, що призводять до перебоїв, відключень, вимкнень або цензури, спричинених різними факторами, включаючи дії уряду, стихійні лиха або кібератаки. Маніпулювання інформацією включає дії, які впливають

на цінності, процедури та політичні процеси, часто з маніпулятивним порядком денним. Хоча ці дії не завжди є незаконними, вони становлять значні загрози, потенційно підриваючи демократичні принципи, дестабілізуючи суспільства та руйнуючи інституційну довіру. Атаки на ланцюги поставок являють собою складні кібератаки, спрямовані на взаємопов'язані відносини між організаціями та їхніми постачальниками. Ці атаки можуть включати шкідливий код в оновленнях програмного забезпечення або апаратних компонентах, скомпрометовані облікові дані постачальників або експлуатацію вразливостей сторонніх сервісів чи продуктів.

Машинне навчання (ML) відіграє вирішальну роль у класифікації та виявленні кібератак, аналізуючи величезні обсяги даних та виявляючи закономірності, що вказують на шкідливу активність. Однією з суттєвих переваг ML є його здатність виявляти аномалії, які можуть вказувати на потенційні загрози, такі як атаки нульового дня та передові постійні загрози (APT). Крім того, ML автоматизує процес ідентифікації загроз, скорочуючи час, необхідний для виявлення та реагування на загрози, що допомагає зменшити збитки та мінімізувати час простою. Можливості ML розпізнавання образів особливо корисні для виявлення певних типів кібератак, тим самим підвищуючи точність виявлення загроз. Крім того, системи ML масштабовані та можуть обробляти великомасштабні дані, що робить їх придатними для підприємств з розгалуженими мережами. Крім того, моделі ML можуть постійно навчатися на нових даних, покращуючи свої можливості виявлення з часом. Це адаптивне навчання є важливим для того, щоб випереджати нові загрози. Ще однією перевагою ML у кібербезпеці є його здатність зменшувати кількість хибнопозитивних результатів у виявленні загроз, що дозволяє командам безпеки зосередитися на справжніх загрозах та уникнути непотрібних сповіщень. Автоматизуючи багато аспектів виявлення та реагування на загрози, ML також зменшує експлуатаційні витрати, пов'язані з ручним моніторингом безпеки та

реагуванням на інциденти.

Моделі машинного навчання (ML) навчаються з використанням наборів даних для класифікації різних типів кібератак. Однак ці набори даних часто незбалансовані, а це означає, що деякі типи атак значно недостатньо представлені порівняно з іншими. Цей дисбаланс створює критичну проблему, оскільки моделі, навчені на таких наборах даних, як правило, зміщені в бік поширеніших типів атак, що призводить до низьких показників виявлення рідкісніших, але потенційно небезпечніших атак. Наприклад, якщо набір даних містить переважну більшість даних про фішингові атаки, але дуже мало випадків експлойтів нульового дня, модель, ймовірно, чудово виявить фішинг, але не зможе розпізнати експлойти нульового дня. Для вирішення проблем дисбалансу дослідники та фахівці з кібербезпеки використовують такі методи, як доповнення даних, генерація синтетичних даних та вдосконалені алгоритми машинного навчання, розроблені для обробки незбалансованих даних. Ці методи допомагають забезпечити надійність моделей та їх здатність точно ідентифікувати як поширені, так і рідкісні кіберзагрози, тим самим забезпечуючи більш надійний захист від широкого спектру атак.

Однак, самі по собі ці зусилля не є особливо переконливими, якщо моделі машинного навчання також не використовуються для призначення ваг ознакам набору даних. Цей процес зважування є критично важливим, оскільки він допомагає визначити, які ознаки мають найбільший вплив на реакції моделі. Розуміючи важливість ознак, дослідники та практики можуть отримати глибше розуміння факторів, що зумовлюють кіберзагрози, що призводить до більш ефективних стратегій виявлення та запобігання. Без цього моделі залишаються чорними скриньками, пропонуючи обмежену цінність, що виходить за рамки простої класифікації.

Основною метою цього дослідження було вивчення та оцінка різних методів доповнення даних для підвищення ефективності моделей контрольованого навчання у виявленні кібератак. Шляхом створення

реалістичного набору даних, який імітує різні типи кібератак, дослідження мало на меті відобразити типові дисбаланси, що спостерігаються у справжніх загрозах кібербезпеці. Згодом застосування кількох методів доповнення даних, включаючи SMOTE та ADASYN, мало на меті виправити асиметричний розподіл класів, забезпечуючи більш збалансований набір даних для навчання моделей контрольованого машинного навчання.

Крім того, у дослідженні висунулася гіпотеза, що впровадження цих методів доповнення даних призведе до значного покращення можливостей виявлення різних моделей, таких як випадкові ліси та нейронні мережі. Ще однією ключовою гіпотезою було те, що аналіз важливості ознак запропонує критичне розуміння того, які атрибути найбільше впливають на прогностичні результати моделей, тим самим підвищуючи їхню інтерпретованість та допомагаючи вдосконалювати процеси інженерії ознак та їх відбору.

1.2 Внесок цієї роботи

Внесок цього дослідження полягає в наступному. Новий, реалістичний набір даних, що симулює різні типи кібератак, відображаючи складний та незбалансований характер реальних сценаріїв кібербезпеки;

Крім того, комплексне застосування кількох методів доповнення даних, включаючи SMOTE (техніку синтетичної меншості надмірної вибірки) та ADASYN (адаптивну синтетичну вибірку), спеціально адаптовано для покращення набору даних у контексті кібербезпеки;

Крім того, дослідження включає систематичну оцінку того, як різні моделі машинного навчання працюють після доповнення даних, визначаючи ті з них, які найефективніші у розпізнаванні різноманітних кіберзагроз;

Зрештою, це дослідження покращує розуміння прогностичних функцій та значною мірою сприяє прозорості та поясненню моделей, що є важливим для операційної довіри та ефективного розгортання в середовищах кібербезпеки.

1.3 Огляд літературних джерел

Кероване навчання широко використовується в системах виявлення вторгнень (IDS) завдяки своїй ефективності у використанні маркованих наборів даних для навчання прогностичних моделей.

В [1] надають комплексний огляд розгортання та інтеграції машинного навчання (ML) у кібербезпеці. Ключові внески включають висвітлення переваг ML над традиційними методами виявлення, керованими людиною, та визначення додаткових завдань кібербезпеки, які можна покращити за допомогою навчання з учителем. У дослідженні обговорюються такі внутрішні проблеми, як зсув концепцій, змагальні умови та конфіденційність даних, які впливають на реальне розгортання ML у кібербезпеці. Обмеження цього підходу включають повільні темпи інтеграції ML у виробничі середовища та необхідність постійних оновлень для обробки загроз, що розвиваються. У статті також представлені тематичні дослідження, що демонструють промислове застосування ML у кібербезпеці, підкреслюючи необхідність спільних зусиль між зацікавленими сторонами для просування ролі ML у цій галузі.

В [2] пояснюють, як методи навчання з учителем та без учителя, такі як логістична регресія та кластеризація, використовуються для виявлення вторгнень та аномалій, тоді як методи навчання з учителем, такі як згорткові нейронні мережі (CNN) та рекурентні нейронні мережі (RNN), ефективно виявляють шкідливе програмне забезпечення та кіберзагрози з високою точністю. Незважаючи на свій потенціал, машинне навчання та навчання з учителем стикаються з проблемами, включаючи потребу у великих наборах даних, високий рівень хибнопозитивних результатів та постійний розвиток кіберзагроз, що вимагає регулярних оновлень та людського нагляду. У роботі закликається до постійних досліджень, покращення наборів даних та інтегрованих методів штучного інтелекту, щоб випереджати кіберзлочинців. Нарешті, вона підкреслює важливість подальших досліджень застосування

штучного інтелекту в кібербезпеці, заохочення співпраці та розробки передових методів захисту цифрових середовищ.

У значущому дослідженні [3] було проаналізовано журнали мережевих підключень за допомогою різних моделей навчання з учителем, включаючи логістичну регресію, дерева рішень (DT), випадкові ліси (RF), SVM, наївний баєсівський алгоритм та дерева з градієнтним підсиленням. Вони представили новий набір даних UWF-ZeekData22, який є загальнодоступним на сайті Університету Західної Флориди. Цей набір даних маркується відповідно до фреймворку MITRE ATT&CK, хоча процес маркування чітко не задокументований. Він включає 18 атрибутів, причому шість основних характеристик, визначених на основі отримання інформації, - це історія підключень, протокол транспортного рівня, протокол прикладного рівня, кількість байтів корисного навантаження, надісланих ініціатором, IP-адреса призначення та кількість пакетів, надісланих ініціатором. Набір даних UWF-ZeekData22 в основному охоплює дві тактики: розвідку та виявлення, зі значним дисбалансом між кількістю спостережень для кожної тактики — 2087 для виявлення та 504 576 для розвідки. Основна увага дослідження зосереджена радше на класифікації тактик противника, ніж на конкретних методах атаки. Сильними сторонами дослідження [3]. є детальний огляд нового підходу до попередньої обробки даних журналу. Методологія передбачає групування числових значень (за винятком портів відправника та призначення) з використанням методу ковзного середнього. Номінальні ознаки перетворюються на числові мітки, IP-адреси класифікуються на основі стандартної мережевої класифікації першого октету, а порти групуються за діапазонами. Зазначеним обмеженням дослідження є його залежність від двійкової класифікації для оцінки продуктивності моделі. Результати дослідження підкреслюють, що методи на основі дерев, зокрема дерева рішень, дерева з градієнтним підсиленням та випадкові ліси, показали відмінні результати, досягнувши точності понад 99% разом з високою точністю, повнотою, F-оцінкою та показниками AUROC.

В [4] дослідили модель машинного навчання з наглядом, навчену та протестовану на двох різних наборах даних. Перший набір даних складається з даних реальних приватних мереж, зібраних авторами, а другий – це набір даних UNSW-NB15, розроблений Австралійським центром кібербезпеки. Дані з організаційного середовища вимагали значної підготовки, включаючи перетворення приватних IP-адрес на публічні в межах захоплених пакетів. Вартий уваги метод вилучення ознак включав аналіз пакетів з різними TCP-заголовками, такими як ICMP, SYN, SYN-ACK, NULL, FIN, XMAS (PSH-URG-FIN) та FIN-ACK, з вимірюваннями, що проводилися кожні дві секунди для кожної вихідної IP-адреси. Етапи попередньої обробки включали видалення надлишкових стовпців, стовпців з порожніми значеннями та стовпців з невеликою варіацією. Для категоріальних ознак було використано гаряче кодування. Такий підхід до обробки відсутніх значень та деталізації процесу маркування, використовуючи інструменти з відкритим кодом, такі як Snort та Suricata, для створення сповіщень з даних пакетів, є комплексним методом, де ці сповіщення служать мітками для навчання. В [4] використовували як методи фільтрації, так і методи обгортки для вибору ознак, прагнучи ефективного уточнення набору ознак. На противагу цьому, дослідження [3] виконало зменшення атрибутів за допомогою інформаційного приросту для оцінки важливості та релевантності ознак. Вони порівняли дві контрольовані моделі: ансамблеву модель, що складається з баєсівських класифікаторів, KNN, логістичної регресії та SVM, з CNN. Важливим висновком їхнього дослідження було те, що CNN перевершили ансамблеву модель в обох наборах даних, що підкреслює ефективність підходів глибокого навчання в обробці складних структур даних та шаблонів.

У своєму дослідженні [5] запропонували підхід глибокого навчання на основі ансамблю для підвищення продуктивності систем виявлення вторгнень (IDS). Ефективність запропонованої моделі оцінювалася з використанням кількох наборів даних, включаючи SDN-IoT, KDD-Cup-1999,

UNSW-NB15, WSN-DS та CICIDS-2017, хоча розглядалися лише зразки з останніх чотирьох. Ці набори даних були попередньо оброблені для забезпечення нормалізації. Моделями, реалізованими для класифікації та вилучення ознак, були RNN, LSTM та Gated Recurrent Units (GRU). Ці нейронні мережі продемонстрували високу ефективність у виявленні атак на всіх наборах даних. Наступний етап запропонованої методології включав зменшення розмірності за допомогою аналізу головних компонент ядра (KPCA) для покращення керованості та ефективності даних. Під час процесу об'єднання ознак ознаки, вилучені з різних моделей, були об'єднані для формування комплексного набору ознак. Цей об'єднаний набір ознак потім оброблявся класифікаторами базового рівня, зокрема SVM та Random Forests. Вихідні дані цих класифікаторів згодом були проаналізовані за допомогою метарівневого класифікатора, логістичної регресії, для завершення процесу виявлення. Показники ефективності цього дослідження підкреслили значний успіх у класифікації атак, досягнувши точності понад 98% на всіх наборах даних, окрім набору даних KDD-Cup-1999, де точність становила 89%.

Подальше дослідження моделей машинного навчання без вчителя провели [6], які використовували набори даних, розроблені Канадським інститутом кібербезпеки. На цьому етапі вони видалили надлишкові ознаки та виключили вибірки з відсутніми або нескінченними значеннями, а також дубліковані рядки. Примітно, що на відміну від дослідження [3], яке цитувалося раніше, такі ознаки, як IP-адреси, були відкинуті, щоб уникнути перенавчання. Вони застосували різноманітні методи масштабування ознак, включаючи StandardScaler, RobustScaler, QuantileTransformer та MinMaxScaler з бібліотеки scikit-learn. Ці кроки попередньої обробки були важливими для подальшого застосування таких моделей, як аналіз головних компонентів (PCA), ізоляційний ліс, автокодер та однокласова SVM, зосереджуючись на виявленні аномалій. Однак вони спостерігали значне падіння продуктивності моделі на новіших даних, зібраних у 2018 році, при цьому AUROC зменшився в середньому на 30,45%, а мінімальне падіння для однокласової

SVM становило 17,85%. Автори пояснили це зниження змінами в розподілі міток атак між наборами даних та неадекватністю гіперпараметрів моделі під час перевірки.

Ще один інноваційний підхід до обробки даних був запропонований в [7], спрямований на вирішення проблеми високого навантаження на пам'ять, яке зазвичай пов'язане зі зберіганням даних трафіку. Їхній метод зосереджений на аналізі лише початкових байтів перших кількох пакетів у потоці, що значно знижує вимоги до зберігання даних. Цей підхід демонструє практичне рішення однієї з фундаментальних проблем аналізу мережевого трафіку, підкреслюючи потенціал методів самостійного навчання в системах виявлення вторгнень (СВВ). Кожне з цих досліджень робить внесок у розвиток ландшафту застосувань машинного навчання в кібербезпеці, демонструючи різні стратегії підвищення ефективності та результативності СВВ.

У своєму інноваційному дослідженні [8] запропонували напівавторизований підхід, заснований на методах кластеризації. Основною метою було використання різних методів кластеризації для початкового маркування набору даних, що потім сприяло б навчанню контрольованих алгоритмів для класифікації.

Істотним обмеженням цього дослідження була його залежність від синтетичних наборів даних, згенерованих за допомогою моделювання. На відміну від інших досліджень, набір ознак, що використовувався тут, був відносно невеликим і включав такі змінні, як швидкість трафіку, затримка обробки та використання процесора сервера, які вважалися достатніми для виявлення розподілених атак типу "відмова в обслуговуванні" (DDoS). Розмір набору даних також був обмежений і містив лише 1000 спостережень, що може неадекватно відображати складніші сценарії атак. Тим не менш, результати цього набору даних порівнювалися з підмножиною набору даних CICIDS2017, яка спеціально зосереджена на DDoS-атаках.

Після збору та нормалізації набору даних було застосовано два методи

кластеризації – агломеративну кластеризацію та K-середні. Зокрема, K-середні були покращені за допомогою головних компонентів, отриманих за допомогою PCA. Наступний крок включав механізм голосування для узгодження результатів кластеризації: якщо спостереження було послідовно позначено в обох результатах, йому було призначено остаточний клас (доброякісний або DDoS); в іншому випадку воно позначалося як «Підозріле». Цей процес був вирішальним для створення надійно позначеного набору даних для подальшого навчання з учителем.

Використані алгоритми з учителем включали метод K-найближчих сусідів (KNN), метод опорних векторів (SVM) та випадковий ліс з гіперпараметрами, тонко налаштованими під час фази навчання. Точність класифікації на синтетичному наборі даних була вражаючою, особливо для випадкового лісу, який досяг коефіцієнта точності 96,66%. Ця методологія також була валідована на вищезгаданій підмножині CICIDS2017, досягнувши точності понад 86%, що підкреслює ефективність цього напівавторизованого підходу.

Щодо доповнення даних, в [9] детально розглянули методи доповнення, зосередившись на їх застосуванні в машинному навчанні для даних зображень. Були досліджені такі методи, як перевертання, обрізання, обертання та коригування колірного простору, детально описано, як вони допомагають у створенні різноманітних наборів даних з обмежених джерел даних. Ці методи є критично важливими для зменшення перенавчання моделі та підвищення надійності прогнозів.

В [10] обговорюють інтеграцію ШІ з традиційними стратегіями кібербезпеки, зазначаючи, як ШІ може значно покращити боротьбу з кіберзагрозами за допомогою таких технологій, як великі дані, блокчейн та поведінкова аналітика. У цій роботі детально наведено поглиблений аналіз як «розподілених» методів ШІ (таких як багатоагентні системи, штучні нейронні мережі, штучні імунні системи та генетичні алгоритми), так і «компактних» методів ШІ (включаючи системи машинного навчання,

експертні системи та нечітку логіку). Ці класифікації допомагають диференціювати методи ШІ на основі їхньої сфери застосування та складності.

У роботі [11] досліджували застосування генеративно-змагальних мереж (GAN) для створення синтетичних даних для кібербезпеки. Ключовий внесок включав всебічне вивчення можливостей GAN у генеруванні реалістичних даних про кібератаки та їх використання для покращення систем вторгнення вторгнення (IDS). Дослідження визначило такі проблеми, як ефективність даних, згенерованих GAN, у точному представленні реальних атак та необхідність подальшого дослідження стійкості моделей глибокого навчання, навчених на синтетичних даних. Обмеження включали постійні занепокоєння щодо якості та реалістичності синтетичних даних, що створюються GAN. У статті підкреслюється важливість синтетичних даних для подолання проблем конфіденційності та безпеки, пов'язаних з обміном даними в реальному світі.

В [12] представили метод покращення продуктивності IDS шляхом поєднання архітектур глибокого навчання з методами доповнення даних. Ключовий внесок включав використання чотирьох відомих наборів даних (UNSW-NB15, 5G-NIDD, FLNET2023 та CIC-IDS-2017) для демонстрації того, що прості моделі на основі CNN можуть досягти високої точності у виявленні вторгнень. Серед виділених обмежень – постійна проблема дисбалансу класів та незначне покращення продуктивності, що спостерігається при використанні складніших архітектур глибокого навчання порівняно з простішими моделями. У дослідженні наголошується на важливості якості даних та доповнення для покращення можливостей виявлення.

2 ОСНОВНІ ПОЛОЖЕННЯ

У сфері кібербезпеки, що швидко розвивається, машинне навчання (ML) відіграє вирішальну роль у покращенні виявлення та пом'якшення загроз. Ефективні методи доповнення даних, такі як SMOTE та ADASYN, є важливими для усунення дисбалансу класів у наборах даних кібербезпеки. Однак простого балансування наборів даних та навчання моделей ML недостатньо, якщо ці моделі також не можуть призначати ваги ознакам набору даних. Розуміння важливості ознак є ключовим для визначення того, які фактори найбільше впливають на реакції моделі, тим самим покращуючи інтерпретованість та ефективність моделі. У цьому розділі досліджуються різні методи доповнення даних, моделі навчання з учителем та значення зважування ознак у контексті кібербезпеки, пропонуючи розуміння оптимізації продуктивності моделі та прийняття рішень.

2.1 Методи доповнення даних

Для збагачення набору даних без втрати цінної інформації також можна використовувати передові методи, такі як генерація синтетичних даних за допомогою таких методів, як SMOTE (техніка синтетичної меншості надмірної вибірки [13]). Враховуючи вибірку x_i з класу меншин, SMOTE визначає своїх k-найближчих сусідів у просторі ознак. Нехай x_{nn} позначаємо одного з цих k-найближчих сусідів. Синтетичний зразок x_{new} генерується шляхом інтерполяції між x_i і x_{nn} використовуючи рівняння:

$$x_{\text{new}} = x_i + \lambda\,(x_{nn} - x_i)$$

де λ – випадкове число між 0 та 1. Цей крок інтерполяції створює нову

вибірку, яка є лінійною комбінацією вихідної вибірки та її сусіда, таким чином зберігаючи загальний розподіл даних, розширюючи при цьому клас меншості.

ADASYN (Адаптивна синтетична вибірка, [14]) розширює SMOTE, зосереджуючись більше на генерації синтетичних зразків поруч зі зразками класу меншин, які помилково класифіковані класифікатором. Математично ADASYN обчислює кількість синтетичних зразків для генерації для кожного зразка класу меншин x_i використовуючи розподіл щільності:

$$r_i = \frac{\gamma_i}{\sum_{i=1}^{N} \gamma_i}$$

де γ_i – кількість вибірок мажорного класу серед k-найближчих сусідів x_i Кількість синтетичних зразків G_i генерувати для кожного x_i пропорційна r_i.

У випадках, коли класи суттєво перетинаються, SMOTE та ADASYN можуть вводити синтетичні зразки в області, де класи недостатньо розділені. Це може призвести до збільшення кількості неправильних класифікацій, оскільки синтетичні зразки в областях, що перекриваються, можуть бути неправильно класифіковані, що знижує загальну продуктивність моделі. Крім того, введення синтетичних зразків в області, що перекриваються, може розмити межі прийняття рішень, що ускладнить для класифікатора розрізнення між класами.

Коли синтетичні зразки вносять шум, як SMOTE, так і ADASYN можуть страждати від зниження продуктивності. Зокрема, синтетичні зразки, які не є репрезентативними для фактичного розподілу даних, можуть вносити шум, що призводить до поганого узагальнення моделі. Крім того, наявність шумних зразків може призвести до надмірного налаштування моделі на синтетичні дані, знижуючи її здатність добре працювати з невидимими даними. Крім того, шум може негативно вплинути на точність і повноту, оскільки модель може давати більше хибнопозитивних і хибнонегативних

результатів.

Однак, для зменшення проблем перекриття класів та шуму можна застосувати кілька стратегій. Наприклад, дані можна попередньо обробити для видалення шуму перед застосуванням SMOTE або ADASYN. По-друге, можна використовувати більш складні методи (такі як Borderline-SMOTE або SVM-SMOTE), які зосереджені на генерації вибірок поблизу межі прийняття рішення. Нарешті, можна постійно оцінювати продуктивність та налаштовувати параметри методів надмірної вибірки, щоб мінімізувати внесення шуму.

Метод Borderline-SMOTE [15] спеціально орієнтований на вибірки класів меншин, які знаходяться близько до межі з класом більшості. Він використовує ту саму стратегію інтерполяції, що й SMOTE, але обмежує її тими вибірками меншин, найближчі сусіди яких включають вибірки класів більшості. Формула генерації синтетичних вибірок залишається такою ж, як і в методі SMOTE.

Доповнення даних за допомогою зв'язків Томека [16] – це метод, який використовується переважно для підвищення продуктивності класифікаторів на незбалансованих наборах даних. Він включає ідентифікацію пар екземплярів, які є найближчими сусідами, але належать до різних класів, та їх видалення для збільшення роздільності класів.

Між парою екземплярів існує зв'язок Томека x_i і x_j з різних класів, якщо немає екземпляра x_k такий, що $d(x_i,x_k)<d(x_i,x_j)$, де d представляє використовувану метрику відстані, часто евклідову відстань. Математично її можна охарактеризувати так.

$$(y_i \neq y_j) \wedge (\nexists x_k \in S : (d(x_i, x_k) < d(x_i, x_j) \vee d(x_j, x_k) < d(x_j, x_i)))$$

Цей метод особливо ефективний для задач бінарної класифікації та часто використовується як метод очищення даних, а не як метод передискретизації.

SMOTEENN [17] поєднує два підходи для вирішення проблеми дисбалансу класів у наборах даних машинного навчання: SMOTE (техніка синтетичної меншості надмірної вибірки) для надмірної вибірки класу меншості та ENN (відредагований найближчий сусід) для очищення даних шляхом недостатньої вибірки обох класів.

ENN видаляє будь-який зразок, у якого більшість k-найближчих сусідів належать до іншого класу. Для заданого зразка x_i, його видаляють, якщо

$$\frac{1}{k}\sum_{j=1}^{k} I\left(y_j \neq y_i\right) > 0.5$$

де I – індикаторна функція, y_j – мітка класу j-го найближчого сусіда, а y_i є міткою класу x_i.

SMOTEENN застосовує SMOTE для генерації синтетичних зразків, а потім використовує ENN для видалення будь-яких згенерованих або оригінальних зразків, які неправильно класифіковані їхніми найближчими сусідами. Ця комбінація допомагає уточнити межі класів далі, ніж використання лише SMOTE.

Нарешті, SMOTE-Tomek [18] – це гібридний метод, який поєднує підхід SMOTE для надмірної вибірки класу меншості з зв'язками Томека для очищення перекриваючихся вибірок між класами. Цей метод особливо ефективний для покращення класифікації незбалансованих наборів даних, як шляхом розширення класу меншості, так і шляхом підвищення роздільності класів. SMOTE-Tomek спочатку застосовує SMOTE для створення додаткових синтетичних вибірок для балансування розподілу класів. Згодом він застосовує метод зв'язків Томека для видалення будь-яких зв'язків Томека, виявлених між синтетичними та вихідними вибірками. Цей процес видалення допомагає зменшити шум і зробити класи більш чіткими, що корисно для подальшого процесу навчання.

Цікаво обговорити обчислювальні витрати методів доповнення даних

та їхній вплив на модель. Часова складність SMOTE становить $O(T\times k\times d)$, де T – кількість синтетичних зразків, k – кількість найближчих сусідів, а d – розмірність даних. Це призводить до помірного збільшення часу навчання через необхідність пошуку найближчих сусідів та генерації синтетичних зразків. ADASYN має подібну часову складність до SMOTE, але включає додаткові обчислення для визначення складності екземплярів, що призводить до дещо вищих обчислювальних витрат та додаткового часу обробки порівняно зі SMOTE. Borderline SMOTE має таку ж часову складність, як і SMOTE, але з додатковими кроками для ідентифікації граничних зразків. Це призводить до вищих обчислювальних витрат, оскільки ідентифікація граничних зразків вимагає додаткових обчислень. Часова складність для Tomek Links становить $O(n^2\times d)$, де n– кількість зразків, а d – розмірність. Це призводить до значного часу попередньої обробки через необхідність обчислення попарних відстаней між зразками, що впливає на загальний час навчання моделі. SMOTEENN поєднує витрати SMOTE та методу відредагованих найближчих сусідів (ENN), зазвичай $O(T\times k\times d)+O(n\times k)\times d)$. Високі обчислювальні витрати є результатом поєднання генерації синтетичних зразків та очищення за методом найближчого сусіда, що призводить до значного збільшення використання ресурсів. SMOTE Tomek поєднує витрати SMOTE та Tomek Links, зазвичай $O(T\times k\times d)+O(n^2\times d$). Таке поєднання призводить до дуже високих обчислювальних витрат через значні попарні обчислення відстаней та генерацію синтетичних вибірок, що суттєво впливає на час навчання та споживання ресурсів.

Обчислювальні витрати цих методів безпосередньо впливають на час навчання моделі та використання ресурсів. Методи з вищою часовою складністю, такі як SMOTE Tomek та SMOTEENN, значно збільшують час попередньої обробки та подовжують загальний час навчання. Високі обчислювальні витрати призводять до збільшення використання процесора та пам'яті, що може бути обмежуючим фактором для великих наборів даних або складних моделей. Методи з квадратичною часовою складністю (наприклад,

Tomek Links) можуть погано масштабуватися з великими наборами даних.

Доповнення даних має вирішальне значення в кібербезпеці для створення більш повних наборів даних, які можуть допомогти в кращому навчанні моделей машинного та глибокого навчання. У науковій літературі пропонуються різні дослідження, зосереджені навколо моделей машинного навчання, що застосовуються в кібербезпеці.

2.2 Моделі навчання з учителем

У цій роботі розглядаються деякі моделі навчання з учителем, такі як наївний байєсівський метод, KNN, XGBoost (XGB), градієнтний бустінг-машина (GBM), логістична регресія та випадковий ліс, а також моделі глибокого навчання, такі як RNN та LSTM, на наборах даних кібератаки.

Наївний Байєсівський класифікатор – це ймовірнісний класифікатор, заснований на теоремі Байєса з припущенням незалежності між ознаками. Його сильні сторони включають швидкість та ефективність, особливо з великими наборами даних, та хорошу роботу з невеликими обсягами навчальних даних. Однак він припускає незалежність між ознаками, що рідко буває вірним у реальних даних, і не підходить для наборів даних з високо корельованими ознаками. KNN – це простий непараметричний алгоритм, який класифікує вибірку на основі класу більшості серед її k-найближчих сусідів. KNN простий у реалізації та розумінні, і він ефективний для невеликих наборів даних з чітко визначеними класами. Його обмеження полягають у тому, що він є обчислювально ресурсоємним з великими наборами даних, а його продуктивність може погіршитися з високовимірними даними. XGB – це оптимізована розподілена бібліотека градієнтного підвищення, розроблена для високої ефективності та гнучкості. XGB пропонує високу продуктивність і точність, і добре обробляє відсутні значення та великі набори даних. Однак він вимагає ретельного налаштування гіперпараметрів і може бути схильним до перенавчання, якщо

його не регуляризувати належним чином. GBM будує адитивну модель поетапно вперед, оптимізуючи диференційовані функції втрат. Метод випадкового лісу (GBM) відомий своєю високою точністю та стійкістю, і він ефективний як для задач регресії, так і для задач класифікації. Основними обмеженнями є його обчислювальна ресурсоємність та повільне навчання, а також схильність до перенавчання без належного налаштування. Логістична регресія – це лінійна модель, що використовується для бінарної класифікації, яка прогнозує ймовірність категоріальної залежної змінної. Вона проста та інтерпретується, ефективна для бінарної та поліноміальної класифікації. Однак вона припускає лінійний зв'язок між ознаками та логарифмічними шансами результату, і не підходить для складних наборів даних з нелінійними зв'язками. Класифікатор випадкового лісу – це метод ансамблю навчання, який будує кілька дерев рішень та об'єднує їх для отримання точнішого та стабільнішого прогнозу. Випадковий ліс добре обробляє великі набори даних з вищою розмірністю та стійкий до перенавчання завдяки своїй ансамблевій природі. Однак він є обчислювально ресурсоємним, особливо з великою кількістю дерев, і менш інтерпретується, ніж одинарні дерева рішень.

RNN – це клас нейронних мереж, де з'єднання між вузлами утворюють орієнтований граф вздовж часової послідовності, що дозволяє їм демонструвати часову динамічну поведінку. RNN ефективні для задач прогнозування послідовностей і можуть добре обробляти дані часових рядів і послідовні дані. Однак вони схильні до проблеми зниклого градієнта, що ускладнює навчання, і вимагають значних обчислювальних ресурсів. Нарешті, LSTM (мережі з обмеженою точністю передачі даних) – це особливий вид RNN, здатний навчатися довгостроковим залежностям і пом'якшувати проблему зниклого градієнта. LSTM здатні навчатися довгостроковим залежностям і ефективні для часових рядів і послідовних даних. Однак вони обчислювально дорогі та повільніші в навчанні, а також вимагають значного налаштування гіперпараметрів.

2.3 Метрики

Точність – це широко використовувана метрика, яка відображає загальну правильність прогнозів моделі. Вона розраховується як частка правильно передбачених випадків відносно загальної кількості випадків у наборі даних наступним чином:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Хоча точність і дає загальне уявлення про продуктивність моделі, вона може вводити в оману в сценаріях з незбалансованими наборами даних.

Точність, з іншого боку, зосереджується на частці істинно позитивних прогнозів серед усіх передбачуваних позитивних результатів. Вона, по суті, вимірює здатність моделі точно ідентифікувати позитивні випадки:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Високе значення точності вказує на низький рівень хибнопозитивних результатів, що свідчить про здатність моделі розрізняти позитивні та негативні випадки.

Відповідність, яку також називають чутливістю або істинним коефіцієнтом позитивних результатів, вимірює здатність моделі виявляти всі відповідні позитивні випадки. Вона розраховується як відношення правильно передбачених позитивних спостережень до загальної кількості фактичних позитивних спостережень у наборі даних:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Високе значення повноти означає низький рівень хибнонегативних результатів, що свідчить про ефективність моделі у виявленні всіх відповідних позитивних випадків.

F1-оцінка усуває потенційні недоліки покладання виключно на точність або повноту, надаючи гармонійне середнє обох показників. Цей консолідований показник пропонує збалансовану оцінку, особливо цінну в ситуаціях із незбалансованим розподілом класів. F1-оцінка розраховується наступним чином:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Значення F1 коливається від 0 до 1, де значення 1 означає ідеальний сценарій, коли як точність, так і повнота є ідеальними.

Зрештою, підтримка стосується загальної кількості фактичних випадків для кожного класу в наборі даних. Хоча вона не є прямим показником продуктивності моделі, підтримка забезпечує вирішальний контекст для інтерпретації інших показників. Розуміючи розподіл класів та кількість точок даних на клас (підтримка), ми можемо ефективніше оцінити значущість розрахованих значень точності, повноти та F1-оцінки.

2.4 Ваги ознак

CatBoost [22] – це алгоритм машинного навчання, розроблений компанією Yandex, який є частиною сімейства алгоритмів градієнтного бустингу. Термін «CatBoost» відображає його здатність ефективно обробляти категоріальні ознаки та його природу як алгоритму бустингу. Він спеціально

розроблений для забезпечення високої продуктивності з акцентом на швидкість та точність, що особливо вигідно при роботі з категоріальними даними. Ця модель оптимізує процес градієнтного бустингу за допомогою використання симетричних дерев та дерев-неявників, підвищуючи як швидкість, так і точність, одночасно зменшуючи перенавчання. Це робить модель особливо надійною та придатною для великих наборів даних, з реалізацією, яка підтримує прискорення на графічному процесорі та багатоядерну обробку.

Ключовою особливістю CatBoost є його здатність надавати уявлення про важливість функцій моделі. У сфері кібербезпеки процес присвоєння ваг змінним під час аналізу даних має першорядне значення з кількох переконливих причин. По-перше, кібербезпека вимагає вивчення величезних наборів даних для виявлення аномалій, потенційних загроз та існуючих вразливостей. Присвоєння ваг змінним дозволяє розрізнити найбільш впливові фактори, що сприяють потенційним порушенням безпеки. Ці набуті знання дозволяють фахівцям з кібербезпеки пріоритезувати свої зусилля на найважливіших аспектах, що зрештою підвищує ефективність стратегій виявлення та пом'якшення загроз.

По-друге, постійно мінливий характер кіберзагроз створює постійні труднощі. Зловмисники постійно розробляють нові методи використання вразливостей системи. Завдяки оцінці змінних ваг, моделі безпеки можуть динамічно адаптуватися та оновлюватися, щоб відображати мінливий ландшафт загроз. Такий динамічний підхід забезпечує постійну актуальність та надійність заходів безпеки перед обличчям нових загроз.

Зрештою, оцінка змінних ваг також сприяє пояснимості та інтерпретації моделей машинного навчання. У контексті кібербезпеки розуміння обґрунтування конкретної генерації сповіщень є критично важливим. Така прозорість не лише сприяє зміцненню довіри із зацікавленими сторонами, але й допомагає в судово-медичних розслідуваннях для відстеження походження та методологій, що використовуються в кібератаках.

3 ТЕСТУВАННЯ МЕТОДУ

3.1 Методологія

Методологія, що використовувалася в цьому дослідженні, була спрямована на вирішення проблем, що виникають через незбалансовані набори даних у виявленні кібератак. Спочатку було створено реалістичний набір даних для моделювання різних типів кібератак, що відображає дисбаланси, які зазвичай спостерігаються в реальних кіберзагрозах. Цей набір даних слугує основою для оцінки ефективності різних методів доповнення даних.

Щоб виправити асиметрію розподілу класів, у дослідженні було застосовано кілька методів доповнення даних, зокрема SMOTE (техніка синтетичної меншості надмірної вибірки) та ADASYN (адаптивна синтетична вибірка). Ці методи генерують синтетичні вибірки для балансування набору даних, тим самим забезпечуючи більш справедливий розподіл класів для навчання моделей машинного навчання з учителем.

Після доповнення, різні моделі машинного навчання з учителем були навчені та оцінені на збалансованому наборі даних. Продуктивність цих моделей оцінювалася за допомогою стандартних показників, таких як точність, прецизійність, повнота та F1-оцінка, щоб визначити покращення, що з'явилися завдяки методам доповнення даних.

Крім того, у дослідженні проаналізували важливість ознак, щоб визначити, які атрибути найбільше впливають на прогностичні результати моделей. Цей аналіз не лише покращує інтерпретованість моделей, але й надає цінну інформацію для вдосконалення процесів розробки та вибору ознак, тим самим оптимізуючи продуктивність моделі.

3.2 Середовище для тестування

У лабораторному середовищі для створення та керування віртуальними мережами використовувалося VMware vSphere версії 8.0.2.00300. Ця конфігурація є критично важливою для ефективної розробки та тестування рішень мережевої безпеки. Сервер Security Onion версії 2.4.60 був налаштований з двома мережевими інтерфейсами. Один інтерфейс був підключений до внутрішньої мережі, а інший використовувався для зв'язку з контейнерами Docker. Сервер був оснащений 12 процесорами, 24ГБ оперативної пам'яті та жорстким диском на 250ГБ, налаштованим як Thick Provision Lazy Zeroed. Була налаштована стандартна оціночна версія Windows Server 2022 (21H2) з 2 процесорами, 12ГБ оперативної пам'яті та жорстким диском на 90ГБ.

Клієнт Windows 10 Pro (версія 22H2) працював з 2 процесорами, 4ГБ оперативної пам'яті та жорстким диском на 48 ГБ. Його мережевий інтерфейс був підключений до тієї ж внутрішньої мережі, що сприяло проведенню різних експериментів з мережевої безпеки. Клієнт Ubuntu 22.04.4 LTS був частиною мережі. Кожна система була налаштована з 2 процесорами, 3ГБ оперативної пам'яті та жорстким диском на 25ГБ. Ці клієнти взаємодіяли з іншими мережевими компонентами для імітації реального трафіку та сценаріїв атак. Головний веб-сервер (APACHE01) та зворотний проксі-сервер працювали на Ubuntu 22.04.4 LTS з аналогічними апаратними конфігураціями, як і клієнти Ubuntu. Вони відіграють вирішальну роль у розміщенні та забезпеченні безпеки веб-додатків.

Для імітації реалістичного мережевого трафіку було реалізовано кілька генераторів. Серед них генератор для створення трафіку з внутрішньої мережі до Інтернету на основі «шумного» проєкту. Цей проєкт передбачає рекурсивний збір та відвідування посилань із зазначених кореневих URL-адрес, доки не зникнуть доступні посилання або не буде досягнуто часу очікування.

APACHE01 захищено зворотним проксі-сервером (APACHE-REVERSE-PROXY), який підключений до Інтернету через брандмауер, що підвищує безпеку розміщених програм. Крім того, трафік та активність контролюються за допомогою таких інструментів, як Security Onion, щоб отримати аналітику мережевого трафіку та потенційних загроз безпеці.

3.3 Набір даних

Характеристики набору даних та їх значення зображено в таблиці 3.1

Таблиця 3.1 – Опис змінних у фреймі даних кібератаки

Назва змінної	Опис
resp_pkts	Кількість пакетів, надісланих відповідачем під час з'єднання.
service	Тип служби, до якої здійснюється доступ (наприклад, HTTP, FTP).
local_resp	Вказує, чи є відповідач локальним у мережі.
protocol	Мережевий протокол, що використовується в з'єднанні (наприклад, TCP, UDP).
duration	Тривалість з'єднання в секундах.
conn_state	Стан з’єднання (наприклад, встановлено, закрито).
orig_pkts	Кількість пакетів, надісланих ініціатором під час з'єднання.
dest_port	Номер порту призначення з’єднання.
orig_bytes	Кількість байтів, надісланих ініціатором під час з'єднання.
local_orig	Вказує, чи є ініціатор локальним для мережі.
resp_bytes	Кількість байтів, надісланих відповідачем під час з'єднання.
src_port	Номер вихідного порту з'єднання.
techniques_mitre	Методи MITRE ATTACK, пов'язані з кібератакою.

Спочатку набір даних складався з 436404 рядків, які після перевірки було скорочено до 307658 (багато спостережень було продубльовано, а ознаки {duration, orig_bytes, resp_bytes} включали 98844 NaN значення) та нормалізація. Ця попередня обробка довела загальну кількість рядків у наборі даних до 208735.

Зокрема, змінна techniques_mitre приймає такі значення:

- network_service_discovery;
- benign;
- reconnaissance_vulnerability_scanning;
- reconnaissance_wordlist_scanning;
- remote_system_discovery;
- domain_trust_discovery;
- account_discovery_domain;
- reconnaissance_scan_ip_blocks.

Виявлення мережевих сервісів – це процес ідентифікації та характеристики сервісів, що працюють на мережевих пристроях. Зловмисники використовують такі методи, як сканування портів та перерахування сервісів, для ідентифікації відкритих портів, сервісів прослуховування та відповідних їм версій програмного забезпечення. Ця інформація допомагає точно визначити потенційні вразливості в мережі та побудувати комплексну карту топології мережі.

Важливо розрізняти доброякісну діяльність та зловмисну мережеву розвідку. Доброякісна діяльність охоплює дії, властиві нормальній роботі системи, включаючи легітимні оновлення програмного забезпечення, процедури планового технічного обслуговування та стандартну поведінку користувачів. На противагу цьому, зловмисна мережева розвідка, як детально описано в наступних розділах, передбачає навмисні спроби використання вразливостей та компрометації безпеки системи.

Розвідувальне сканування вразливостей передбачає систематичне дослідження цільових систем для виявлення вразливостей, що можуть бути

використані. Зловмисники використовують цей метод для виявлення застарілого програмного забезпечення, неправильних конфігурацій та інших прогалин у безпеці. Основна мета – зібрати інформацію, яку згодом можна використовувати для отримання несанкціонованого доступу або виконання зловмисних дій.

Цей метод передбачає використання заздалегідь визначених списків слів або фраз (списків слів) для систематичного дослідження потенційних точок інтересу в цільовому середовищі. Зловмисники використовують списки слів для проведення атак методом грубої сили або спроб вгадати критичну інформацію, таку як імена користувачів, паролі, URL-адреси та інші конфіденційні дані. Розвідувальне сканування списків слів часто доповнює інші розвідувальні дії для підвищення ефективності та точності атаки.

Віддалене виявлення систем – це процес збору інформації про віддалені системи в мережі. Це може включати ідентифікацію активних хостів, мережевих спільних ресурсів та доступних ресурсів. Методи, що використовуються для віддаленого виявлення систем, включають ping-сканування, сканування портів та запити мережевих служб. Кінцева мета – відображення структури мережі та визначення потенційних цілей для подальших спроб зловмисників.

У доменних середовищах зловмисники використовують виявлення довірчих відносин домену, щоб зрозуміти довірчі відносини, встановлені між різними доменами. Розуміння цих довірчих відносин може надати зловмисникам шляхи для горизонтального переміщення та ескалації привілеїв. Це може включати ідентифікацію довірених доменів, контролерів доменів та будь-яких міждоменних політик, що регулюють контроль доступу.

Виявлення облікових записів (доменів) – це метод, який використовують зловмисники для переліку облікових записів користувачів у доменному середовищі. Це включає виявлення імен користувачів, пов’язаних

груп користувачів та відповідних дозволів. Отриману інформацію можна використовувати для планування атак, що включають крадіжку облікових даних, ескалацію привілеїв та горизонтальне переміщення в межах скомпрометованого домену. Звичайні методи виявлення облікових записів включають запити до Active Directory та використання вбудованих команд домену.

Розвідувальне сканування IP-блоків передбачає систематичне сканування великих діапазонів IP-адрес для виявлення активних пристроїв і служб. Зловмисники використовують цей метод для картографування цільової мережевої інфраструктури та визначення потенційних цілей для подальшого використання. Цей тип сканування може виявити критично важливу інформацію, таку як кількість активних хостів, використовувані операційні системи та мережеві пристрої, присутні в цільовому середовищі.

Зрештою, виявлення групових політик стосується процесу ідентифікації та аналізу об'єктів групових політик (GPO) у середовищі домену Windows. Зловмисники досліджують GPO, щоб отримати уявлення про конфігурації безпеки, адміністративні шаблони та політики користувачів. Цю інформацію можна використовувати для виявлення неправильних конфігурацій, розуміння розгорнутих засобів контролю безпеки та визначення потенційних слабких місць, які можна використати для досягнення їхніх зловмисних цілей.

Згідно з таблицею 3.2, набір даних видається незбалансованим через значну диспропорцію у зустрічальності різних ознак, зокрема в межах «розподілу techniques_mitre». Незбалансовані набори даних часто зустрічаються в машинному навчанні та статистичному аналізі та можуть призвести до упереджених моделей, які неадекватно представляють класи меншин. Ознака «network_service_discovery» демонструє переважне домінування зі 144 279 зустрічаннями, що майже в 2,4 рази більше, ніж у наступної за частотою категорії, «benig», яка має 60 997 зустрічань. Ця домінантна ознака може призвести до того, що прогностичні моделі

демонструватимуть сильну упередженість у бік прогнозування цієї категорії, потенційно за рахунок точності в інших менш частих категоріях.

Таблиця 3.2 – Розподіл techniques_mitre

Розподіл techniques_mitre	Випадки
network_service_discovery	144,279
benign	60,997
reconnaissance_vulnerability_scanning	1581
reconnaissance_wordlist_scanning	715
remote_system_discovery	554
domain_trust_discovery	411
account_discovery_domain	84
reconnaissance_scan_ip_blocks	80
group_policy_discovery	34

Більше того, такі категорії, як «group_policy_discovery», «reconnaissance_scan_ip_blocks» та «account_discovery_domain», надзвичайно недостатньо представлені, лише 34, 80 та 84 рази відповідно. Таке розріджене представлення ускладнює процес навчання статистичних моделей, оскільки даних недостатньо для досягнення гарної продуктивності узагальнення нових або невидимих даних, що потрапляють до цих категорій. Широкий діапазон розподілу ознак, від найбільш частих до найменш частих (144279 випадків проти 34 випадків), підкреслює разючий дисбаланс, що вказує не лише на перекіс у бік певних ознак, але й на значну недостатню представленість інших.

Алгоритми машинного навчання зазвичай працюють краще, коли кількість екземплярів для кожного класу приблизно однакова. Незбалансований набір даних може призвести до моделей, які зміщені в бік класів з більшою кількістю екземплярів, збільшуючи ймовірність неправильної класифікації екземплярів меншинних класів. Це може серйозно

вплинути на точність моделі, зокрема на її здатність виявляти менш часті, але потенційно важливі категорії.

3.4. Доповнення даних

Усунення цього дисбалансу може включати використання таких методів, як надмірна вибірка класів меншин, недостатня вибірка класів більшості або використання таких підходів, як SMOTE, ADASYN, Borderline-SMOTE, Tomel-Links, SMOTEENN та SMOTE Temek.

В таблиці 3.3 пропонується вичерпне порівняння різних методів доповнення даних, застосованих до незбалансованого набору даних, класифікованого як «techniques_mitre». Ці методи включають SMOTE, ADASYN, Borderline-SMOTE, Tomek Links, SMOTEENN та комбінацію SMOTE та Tomek Links, кожен з яких адаптований для зміни розподілу класів меншин та більшості шляхом генерації синтетичних даних або очищення даних.

Застосування цих методів доповнення мало на меті нормалізувати показники зустрічальності в різних категоріях до цільового числа, приблизно 115474, для більшості методів, що вказує на рівень, встановлений для досягнення балансу класів.

Таблиця 3.3 – Порівняння методів доповнення даних.

Techniques_ MITRE	Original	SMOTE	ADASYN	Borderline - SMOTE	Tomek-Links	SMOTEENN	SMOTE Tomek
1	2	3	4	5	6	7	8
Benign	144,279	115,474	115,870	115,474	48,353	107,187	114,001
Account Discovery Domain	60,997	115,474	115,480	115,474	40	114,351	115,334
Domain Trust Discovery	1581	115,474	115,533	115,474	238	113,080	115,030

Продовження таблиці 3.3

1	2	3	4	5	6	7	8
Group Policy Discovery	715	115,474	115,473	115,474	25	114,503	115,258
Network Service Discovery	554	115,474	115,474	115,474	115,467	115,405	115,470
Reconnaissance Scan IP Blocks	411	115,474	115,478	115,474	62	115,357	115,473
Reconnaissance Vulnerability Scanning	84	115,474	115,751	115,474	1004	111,585	114,734
Reconnaissance Wordlist Scanning	80	115,474	115,475	115,474	577	115,474	115,474
Remote System Discovery	34	115,474	115,475	115,474	431	113,998	115,324

Застосування цих методів доповнення мало на меті нормалізувати показники зустрічальності в різних категоріях до цільового числа, приблизно 115474, для більшості методів, що вказує на рівень, встановлений для досягнення балансу класів.

3.5 Моделі машинного навчання

Моделі машинного навчання, такі як наївний баєсівський метод, метод K-найближчих сусідів (KNN), XGBoost (XGB), метод градієнтного підсилення (GBM), логістична регресія та класифікатор випадкового лісу, часто є кращими в прогнозній аналітиці завдяки своїм різноманітним сильним сторонам та застосовності в широкому спектрі задач. Кожна модель має унікальний набір можливостей, що робить її придатною для різних типів даних та прогностичних завдань. Перевага цих моделей у різних аналітичних сценаріях випливає з їхньої здатності балансувати точність та обчислювальну ефективність, водночас надаючи рішення, які легко інтерпретувати та

впроваджувати в реальних додатках.

У таблиці 3.4 показано важливі тенденції щодо поведінки цих моделей в умовах тестування.

Таблиця 3.4 – Значення точності для різних класифікаторів та методів доповнення даних.

Classifier	SMOTE	ADASYN	Borderline SMOTE	Tomek Links	SMOTEENN	SMOTE Tomek
Naïve Bayes	0.497	0.453	0.602	0.718	0.668	0.659
KNN	0.824	0.978	0.981	0.992	0.993	0.990
XGB	0.838	0.925	0.942	0.993	0.981	0.977
GBM	0.842	0.940	0.953	0.989	0.989	0.984
RF	0.833	0.985	0.985	0.994	0.998	0.996
Logistic	0.738	0.741	0.847	0.807	0.860	0.851
RNN	0.759	0.823	0.888	0.979	0.647	0.797
LSTM	0.819	0.875	0.916	0.982	0.945	0.944

Наївний байєсівський метод, який традиційно цінується за свою простоту та ефективність обробки великих наборів даних, демонструє помірну точність. Це очікувано, враховуючи його припущення про незалежність ознак, що не завжди може бути вірним у реальних наборах даних. Продуктивність KNN загалом краща, що відображає її здатність адаптувати свою стратегію класифікації на основі локальної структури даних. Однак, залежність KNN від масштабування ознак та прокляття розмірності іноді можуть негативно впливати на її продуктивність.

XGB та GBM, обидві моделі підвищення точності, демонструють високу точність, що підкреслює їхню силу в роботі зі складними наборами даних, які включають нелінійні зв'язки між ознаками. Ці моделі базуються на помилках попередніх дерев і, отже, можуть адаптивно покращувати свої

прогнози. Висока продуктивність свідчить про їхню стійкість, але вона також висвітлює необхідність ретельного налаштування параметрів, щоб уникнути надмірного підлаштування до навчальних даних.

Логістична регресія забезпечує прийнятну точність, що корисно в сценаріях, що потребують оцінки ймовірності для бінарних результатів. Її продуктивність, як правило, менш конкурентоспроможна порівняно з ансамблевими методами, але пропонує цінну інформацію завдяки своїй інтерпретованості.

Випадковий ліс зазвичай демонструє чудову точність завдяки своїй здатності зменшувати перенавчання шляхом усереднення кількох дерев рішень. Ця модель ефективна для обробки різних типів даних, включаючи незбалансовані набори даних.

Хоча результати свідчать про те, що ансамблеві методи, такі як XGB, GBM та Random Forest, як правило, забезпечують вищу точність, це слід збалансувати з розумінням того, що висока точність іноді може бути результатом перенавчання. Хоча перенавчання не є основним предметом цього обговорення, воно неявно доречне при інтерпретації високої точності складних моделей.

Зрештою, до розширеного набору даних («tomek_links.csv»), отриманого за допомогою підходу Tomek links, було застосовано модель CatBoost. Результат показано на рисунку 3.1

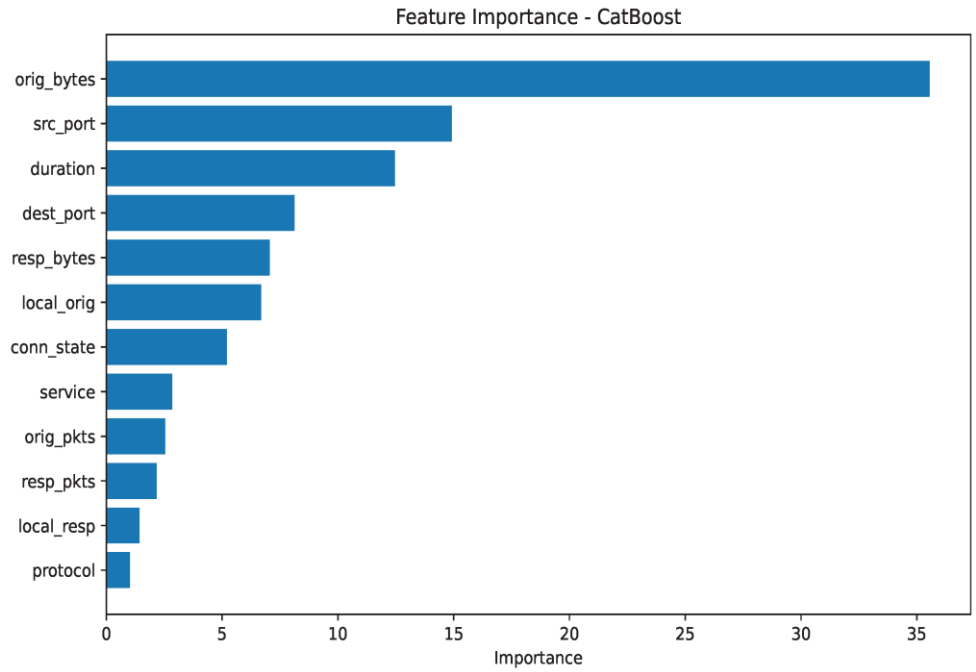

Рисунок 3.1 – Важливість ознак

3.6 Параметри моделі

У таблиці 3.5 наведено параметри, що використовуються кожною моделлю. Для GBM ключові гіперпараметри включають кількість оцінок, швидкість навчання та максимальну глибину. Їх налаштування включає наступне:

- n_estimators - більше число зазвичай збільшує складність моделі. Перехресна перевірка допомагає знайти оптимальний баланс, щоб уникнути перенавчання;

- learning_rate - контролює внесок кожного дерева. Нижчі значення зазвичай вимагають більшої кількості дерев;

- max_depth - обмежує глибину окремих дерев для контролю перенавчання.

Для KNN основним гіперпараметром є кількість сусідів:

- _neighbors - невелике число може призвести до шумних прогнозів, тоді як велике число може згладити прогноз, але ігнорувати локальні нюанси. Пошук по сітці з перехресною перевіркою зазвичай використовується для

визначення оптимального значення.

Таблиця 3.5 – Моделі та їх параметри

Model	Parameters
GBM	n_estimators = 100, learning_rate = 0.1, max_depth = 3
KNN	n_neighbors = 5
Logistic Regression	max_iter = 10,000, class_weight = 'balanced'
LSTM	optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy']
GaussianNB	none
Random Forest	n_estimators = 100
RNN	optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy']
XGB	n_estimators = 100, learning_rate = 0.1, max_depth = 3,
	eval_metric = 'mlogloss'

Ключові гіперпараметри включають максимальну кількість ітерацій та ваги класів:

- max_iter: забезпечує збіжність. Вищі значення дозволяють розв'язувачу виконати більше ітерацій для збіжності, що особливо корисно для складних наборів даних;

- class_weight: балансує набір даних, коригуючи ваги обернено пропорційно частотам класів. Це особливо важливо для незбалансованих наборів даних.

Критичні гіперпараметри для LSTM включають оптимізатор, функцію втрат та метрики:

- оптимізатор: «Adam» зазвичай використовується завдяки своїм можливостям адаптивного навчання;

- втрата: «categorical_crossentropy» використовується для задач

класифікації з кількома класами;

- показники: «accuracy» – це стандартний показник для оцінки ефективності класифікації.

GaussianNB зазвичай не потребує налаштування гіперпараметрів, оскільки це проста ймовірнісна модель. Натомість, важливі гіперпараметри включають кількість оцінок:

- n_estimators - кількість дерев у лісі. Більша кількість дерев зазвичай покращує продуктивність, але збільшує час обчислення.

Подібно до LSTM, важливі гіперпараметри включають оптимізатор, функцію втрат та метрики:

- оптимізатор - «Adam» є кращим за свою ефективність та продуктивність;
- втрата - «categorical_crossentropy» для багатокласової класифікації;
- метрики - «accuracy» для оцінки ефективності.

Нарешті, для XGBoost ключові гіперпараметри включають кількість оцінювачів, швидкість навчання, максимальну глибину та метрику оцінювання:

- n_estimators: визначає кількість раундів підвищення;
- learning_rate - нижчі показники вимагають більше раундів підвищення;
- max_depth - контролює глибину кожного дерева, щоб запобігти перенавчанню;
- eval_metric - «mlogloss» використовується для класифікації за кількома класами.

ВИСНОВКИ

Впровадження передових методів доповнення даних у моделях навчання з учителем значно підвищує стійкість та ефективність систем виявлення кібератак. Це дослідження демонструє, що завдяки інтеграції посилань SMOTE, ADASYN та Tomek можна не лише покращити точність прогнозування, але й суттєво покращити узагальнюваність моделей для різноманітних та мінливих ландшафтів кіберзагроз.

Крім того, наші висновки підкреслюють важливість використання гібридного підходу до доповнення даних, який ретельно вирішує проблеми незбалансованих наборів даних, поширених у програмах кібербезпеки. Використовуючи ці методи, ми успішно мінімізували потенціал перенавчання та покращили рівень виявлення кібератак.

Оскільки кіберзагрози продовжують розвиватися у складності та витонченості, здатність систем виявлення адаптуватися та реагувати з тонким розумінням стає дедалі важливішою. Методи, розроблені та протестовані в цьому дослідженні, підтримують більш складні, адаптивні реакції на кіберзагрози, надаючи фахівцям з безпеки інструменти, які є як реактивними, так і превентивно адаптивними.

Оновлення моделей у відповідь на кіберзагрози, що розвиваються, вимагає динамічного та безперервного підходу, щоб забезпечити ефективність систем виявлення проти нових та складних схем атак. Однією з потенційних стратегій є впровадження системи безперервного навчання. У цій системі модель періодично перенавчається з використанням останніх даних, що допомагає враховувати найновіші схеми загроз та аномалії, що спостерігаються в мережевому трафіку. Інша стратегія передбачає використання методів ансамбльового навчання. Поєднуючи кілька моделей, кожна з яких навчена на різних аспектах або часових рамках даних, система може досягти більшої стійкості до різних стратегій атак. Ансамблеві методи також можуть включати

нові моделі, навчені на останніх даних, що дозволяє системі інтегрувати свіжі знання, не відкидаючи повністю знання зі старих моделей. Крім того, впровадження циклу зворотного зв'язку від центру операцій безпеки (SOC) може бути дуже корисним. Коли виявляється потенційна загроза, SOC може надати зворотний зв'язок про те, чи це був істинно позитивний результат, чи хибнопозитивний. Цей зворотний зв'язок можна використовувати для точного налаштування моделі, покращуючи її точність з часом. Методи доповнення даних відіграють вирішальну роль в оновленні моделей. Постійно генеруючи синтетичні дані, що відображають найновіші схеми та сценарії атак, навчальний набір даних можна розширити та урізноманітнити. Такий підхід допомагає підтримувати ефективність моделі проти широкого спектру загроз, включаючи ті, які можуть бути не поширені в історичних даних. Виявлення аномалій можна покращити, інтегруючи методи навчання без учителя разом із методами навчання з учителем. У той час як моделі з учителем навчаються на маркованих даних, моделі без учителя можуть виявляти нові та незвичайні закономірності без попередніх знань. Поєднуючи ці підходи, система виявлення може адаптуватися до нових методів атаки, які відхиляються від відомих закономірностей.

У майбутніх роботах буде досліджено вплив обчислювальних витрат та ефективності методів доповнення, що забезпечить глибше розуміння їхнього практичного застосування. Буде проведено подальший аналіз адаптованих стратегій навчання, чутливих до витрат, де вартість неправильної класифікації класів меншин встановлюється вищою, ніж вартість класифікації класів більшості, щоб змусити модель приділяти більше уваги недостатньо представленим класам. Ці заходи мають вирішальне значення для побудови надійних моделей, які добре працюють у всіх категоріях і не є упередженими до більшості. Нарешті, будуть включені складніші механізми пояснення, що використовують консолідовані методи, такі як LIME (локальні інтерпретовані моделі - агностичні пояснення) та SHAP (адитивні пояснення Шеплі), для глибшого розуміння кібератак.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Apruzzese, G.; Laskov, P.; Montes de Oca, E.; Mallouli, W.; Brdalo Rapa, L.; Grammatopoulos, A.V.; Di Franco, F. The role of machine learning in cybersecurity. Digit. Threat. Res. Pract. 2023, 4, 1–38.

2. Mijwil, M.; Salem, I.E.; Ismaeel, M.M. The significance of machine learning and deep learning techniques in cybersecurity: A comprehensive review. Iraqi J. Comput. Sci. Math. 2023, 4, 87–101.

3. Bagui, S.; Mink, D.; Bagui, S.; Ghosh, T.; McElroy, T.; Paredes, E.; Khasnavis, N.; Plenkers, R. Detecting reconnaissance and discovery tactics from the MITRE ATT&CK framework in Zeek conn logs using spark's machine learning in the big data framework. Sensors 2022, 22, 7999.

4. О.С. Ляшенко, І.А. Великодний, В.Г. Знайдюк, О.Д. Журило. Модель та методи виявлення широкомасштабної атаки в середовищі IoT / Системи управління, навігації та зв'язку. Том 1 № 75 (2024), С.127-132

5. Ravi, V.; Chaganti, R.; Alazab, M. Recurrent deep learning-based feature fusion ensemble meta-classifier approach for intelligent network intrusion detection system. Comput. Electr. Eng. 2022, 102, 108156.

6. Verkerken, M.; D'hooge, L.; Wauters, T.; Volckaert, B.; De Turck, F. Towards model generalization for intrusion detection: Unsupervised machine learning techniques. J. Netw. Syst. Manag. 2022, 30, 1–25.

7. Hwang, R.H.; Peng, M.C.; Huang, C.W.; Lin, P.C.; Nguyen, V.L. An unsupervised deep learning model for early network traffic anomaly detection. IEEE Access 2020, 8, 30387–30399.

8. Aamir, M.; Zaidi, S.M.A. Clustering based semi-supervised machine learning for DDoS attack classification. J. King Saud-Univ.-Comput. Inf. Sci. 2021, 33, 436–446.

9. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. Glob. Trans. Proc. 2022, 3, 91–99.

10. Naik, B.; Mehta, A.; Yagnik, H.; Shah, M. The impacts of artificial intelligence techniques in augmentation of cybersecurity: A comprehensive review. Complex Intell. Syst. 2022, 8, 1763–1780.

11. Agrawal, G.; Kaur, A.; Myneni, S. A review of generative models in generating synthetic attack data for cybersecurity. Electronics 2024, 13, 322.

12. Mohammad, R.; Saeed, F.; Almazroi, A.A.; Alsubaei, F.S.; Almazroi, A.A. Enhancing Intrusion Detection Systems Using a Deep Learning and Data Augmentation Approach. Systems 2024, 12, 79.

13. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. 2018, 61, 863–905.

14. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; IEEE: Piscataway Township, NJ, USA, 2008; pp. 1322–1328.

15. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887

16. Swana, E.F.; Doorsamy, W.; Bokoro, P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors 2022, 22, 3246.

17. He, H.; Garcia, E.A. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 2009, 21, 1263–1284.

18. Yang, H.; Li, M. Software Defect Prediction Based on SMOTE-Tomek and XGBoost. In Bio-Inspired Computing: Theories and Applications; Pan, L., Cui, Z., Cai, J., Li, L., Eds.; Springer: Singapore, 2022; pp. 12–31.

19. Handa, A.; Sharma, A.; Shukla, S.K. Machine learning in cybersecurity: A review. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019, 9, e1306.

20. Dasgupta, D.; Akhtar, Z.; Sen, S. Machine learning in cybersecurity: A comprehensive survey. J. Def. Model. Simul. 2022, 19, 57–106.

21. Баєв І.С. Методи збільшення даних для покращення контрольованого навчання при виявлення кібератак // / Теоретичне та практичне застосування результатів сучасної науки: матеріали IX Міжнародноїстудентської наукової конференції, м.Умань, 13червня, 2025рік / ГО «Молодіжна наукова ліга». - Вінниця: ТОВ «УКРЛОГОСГруп», 2025.

22. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 2018, 31, 6639–6649