УДК 004.492:004.032.26 DOI: 10.31673/2412-4338.2023.040109

Сосновий В.О., ***Лащевська Н.О.***
Державний університет інформаційно-комунікаційних технологій, Київ

РОЗРОБКА СТРУКТУРИ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ АНАЛІЗУ ВИЯВЛЕННЯ ВТОРГНЕНЬ

Анотація: *Належні рішення безпеки в інформаційно-комунікаційному світі мають вирішальне значення для забезпечення безпеки мережі, забезпечуючи захист мережі в режимі реального часу від уразливостей мережі та використання даних. Ефективна стратегія виявлення вторгнень здатна використовувати цілісний підхід для захисту критично важливих систем від несанкціонованого доступу чи атак. В роботі розглянуто останні наукові досягнення та дослідження стосовно аналізу виявлення вторгнень в мережу з використанням методів машинного навчання (МН). В статті описано комплексне рішення безпеки на основі машинного навчання (МН) для виявлення вторгнень в мережі з використанням комплексної контрольованої структури МН і методів вибору функцій ансамблю. Крім того, надано порівняльний аналіз кількох моделей МН і методів вибору функцій. В статті розроблено загальний механізм виявлення та досягнення вищої точності з мінімальною частотою помилкових позитивних результатів (ЧПР). В статті використовуються набори даних і результати показують, що модель виявлення може успішно ідентифікувати 99,3% вторгнень із найменшою кількістю помилок у 0,5%, що відображає кращі показники продуктивності порівняно з існуючими рішеннями. В статті об'єднано вибір функцій ансамблю та підходи машинного навчання ансамблю як механізм виявлення в СВВ для виявлення мережевих аномалій. Проведено експериментальне дослідження з наборами функцій, отриманими з дев'яти методів вибору функцій, а потім об'єднано ці набори функцій, щоб отримати мінімальну кількість функцій за допомогою голосування більшості. Проведено порівняльний аналіз наборів функцій. Використано контрольовані методи, які ефективніші з збалансованим набором даних. Щоб зробити навчальний набір даних збалансованим, спочатку було вибрано тип даних (доброякісні або атакуючі) з мінімальною кількістю екземплярів даних у цьому навчальному наборі даних. Реалізовано алгоритм вибору функцій ансамблю та класифікації ансамблю, щоб покращити загальну продуктивність запропонованої моделі машинного навчання. Запропоновані перспективи розвитку подальших досліджень.*

Ключові слова: *безпека мережі, виявлення шкідливих програм, система виявлення вторгнень, машинне навчання, вибір ознак ансамблю, порівняльний аналіз.*

Sosnovyy V.O., ***Lashchevska N.O.***
State University of Information and Communication Technologies, Kyiv

DEVELOPMENT OF A NEURAL NETWORK STRUCTURE FOR INTRUSION DETECTION ANALYSIS

Abstract: *The right security solutions in the information and communication world are critical to network security by providing real-time network protection against network vulnerabilities and data usage. An effective intrusion detection strategy is able to use a holistic approach to protect critical systems from unauthorized access or attacks. The paper examines the latest scientific achievements and research related to the analysis of detection of network intrusions using machine learning (ML) methods. The article describes a complex security solution based on machine learning (ML) for network intrusion detection using a complex controlled ML structure and ensemble feature selection methods. In addition, a comparative analysis of several MH models and function selection methods is provided. The article develops a general mechanism for detecting and achieving higher accuracy with a minimum frequency of false positive results (FPR). The paper uses datasets and the results show that the detection model can successfully identify 99.3% of intrusions with the lowest error rate of 0.5%, which shows better performance compared to existing solutions. The article*

combines the selection of ensemble functions and ensemble machine learning approaches as a detection mechanism in SBB to detect network anomalies. An experimental study was conducted with feature sets obtained from nine feature selection methods, and then these feature sets were combined to obtain the minimum number of features using majority voting. A comparative analysis of sets of functions was carried out. Controlled methods are used, which are more efficient with a balanced data set. To make the training dataset balanced, the data type (benign or attacking) with the minimum number of data instances in that training dataset was first selected. An ensemble feature selection and ensemble classification algorithm is implemented to improve the overall performance of the proposed machine learning model. Prospects for the development of further research are proposed.

Keywords: *network security, malware detection, intrusion detection system, machine learning, ensemble feature selection, comparative analysis.*

1. Вступ.

Зростання частоти кібератак є актуальною проблемою сучасності. Ці атаки завдають шкоди як окремим особам, так і підприємствам і можуть вплинути на конфіденційність, цілісність і доступність критичної інформації, яка передається через мережу. Підприємствам важливо захищати свої мережі від зловмисників, хакерів і мережевих загроз. Тому мережева система повинна включати кілька інструментів безпеки для захисту важливих даних і послуг від потенційних загроз. Система виявлення вторгнень (СВВ) — це програмна або апаратна реалізація для перевірки мстивих рухів або порушень стратегії в мережі. Система відстежує зловмисну активність або порушення безпеки в мережі та сповіщає адміністратора про потенційні загрози. Атаки на мережі з кожним днем стають фатальними. Щоб йти в ногу з кіберзагрозами, що постійно змінюються, сучасні мережеві бізнес-середовища вимагають високого рівня безпеки для безпечного та надійного обміну інформацією між організаціями. Традиційні СВВ (на основі сигнатур і на основі аномалій) не розроблені належним чином для адаптації до моделей мережевих вторгнень, що постійно змінюються. Впровадження ¬штучного інтелекту в СВВ має потенціал, щоб впоратися з новими шаблонами атак і забезпечити безпеку мережі.

В останні кілька десятиліть машинне навчання (МН), підмножина штучного інтелекту, широко використовувалося для покращення виявлення вторгнень шляхом автоматизації та прогнозування на просунутому рівні. Методи МН дозволяють працювати з наборами даних, які можна змінювати, відтворювати та розширювати. Ці технології ¬допомагають СВВ стати надійними, навчаючись і протидіючи видимим і невидимим атакам. Крім того, одна модель МН не завжди може передбачати точно, тоді як комбінація моделей МН (ансамблева МН) виявляє точніше. Ансамблеві класифікатори МН перевершують індивідуальні класифікатори у вирішенні різноманітних проблем класифікації. Крім того, вибір функцій (ВФ) — це процес вибору підмножини відповідних і релевантних функцій із великого простору функцій. Це відіграє життєво важливу роль у досягненні вищої точності, а також мінімізації часу навчання моделі. Проблеми оптимізації функцій і проектування СВВ з найкращими методами МН спонукали нас дослідити продуктивність різних методів ВФ. Крім того, ми були мотивовані вивчити методи ансамблю для вибору функцій і класифікації моделей, оскільки вони враховують рішення кількох кандидатів для оптимізації, а не індивідуальний вибір.

2. Аналіз останніх досліджень і публікацій. Алгоритми машинного навчання (МН) зосереджені на розробці комп’ютерних програм зі здатністю системи автоматично отримувати та покращувати функціональність без втручання людини. проаналізовано більше 50 статей про виявлення вторгнень за допомогою МН і проведено порівняльний аналіз використовуваних алгоритмів і наборів даних. Виявлення вторгнень з використанням сукупності парадигм програмного обчислення виконано порівняльний аналіз штучної нейронної мережі, опорний вектор машини, а також ансамбль цих двох моделей. Було зазначено, що ансамблевий класифікатор перевершує ці окремі класифікатори у виявленні вторгнень. Підхід до глибокого навчання для інтелектуальної системи виявлення вторгнень [1] запропоновано високомасштабовану та гібридну структуру глибинної нейронної мережі (ГНН), яка здатна

виявляти кібератаки в реальному часі шляхом моніторингу мережевого трафіку та подій на рівні хоста. Також було проведено дослідження й оцінку ефективності різних неглибоких і глибоких мереж ¬для систем виявлення вторгнень у мережу [2]. Обидва дослідження були оцінені за допомогою набору даних KDDCup' 99. Крім того, в роботі [3] запропоновано систему виявлення ботнету на основі глибинного навчання для виявлення та класифікації доменних імен, і цю систему можна розгорнути на рівні провайдера для моніторингу пристроїв IoT

Також аналіз останніх досліджень показав, що було розглянуто та запропоновано цілий перелік методів для класифікації потоків даних, визначено інструменти, основні підходи та техніки ансамблевого навчання, а також запропоновані перспективи в подальшому нейромережевому навчанні [4-6].

Ансамблевий адаптивний алгоритм для підвищення точності виявлення за допомогою набору даних NSL-KDD порівнюється з розробленим алгоритмом MultiTree. Дане порівняння дозволило запропонувати покращений СВВ з використанням обмеженої ВФ (25 підмножин і 35 підмножин характеристик) і ансамблевих моделей, заснованих на методах пакетування та підвищення, які відображали високу точність для набору даних NSL-KDD [7]. Запропонована в [8] структура ансамблю для контрольованих класифікаторів забезпечує величезне підвищення точності виявлення атак DDoS на основі чотирьох класифікаторів. МН Напівконтрольовані методи ВФ [9] використовують як позначені, так і немарковані дані, і проілюстрували ¬дві таксономії, засновані на ієрархічній структурі. Адамс і Белінг [10] провели огляд методів ВФ для Гауса. Отже дослідження літератури вказує на те, що методику ансамблю можна широко застосовувати у виборі функцій і класифікації моделей для отримання кращих результатів у виявленні вторгнень.

3. Мета і задачі дослідження. полягає в тому, щоб забезпечити загальну структуру виявлення вторгнень з вищою точністю за допомогою різних аномальних наборів даних, таких як NSL-KDD, UNSW-NB15 і CICIDS2017. Оскільки критичні характеристики є важливими в класифікації моделі, потрібно збільшити кількість методів ВФ порівняно з попередніми дослідженнями. Також об’єднати методи на основі фільтрів, оболонки та вбудовані методи ВФ у структуру ВФАН. Включити вибір функцій ансамблю (ВФАН) із класифікацією ансамблю (КАНМН), щоб отримати кращу точність із мінімальною кількістю помилок.

4. Результати дослідження.

В даній статті використано метод моделювання загроз STRIDE та розглянуто більшість із дев’яти високорівневих кроків STRIDE. Припустімо, що зловмисник планує виконати будь-яку неправомірну дію, наприклад:

1) порушити роботу системи, скомпрометувавши її або надсилаючи трафік із кількох різних скомпрометованих хостів чи ботнетів;
2) заливає жертву через підробку IP-адрес;
3) розширює доступ користувача, щоб отримати контроль над системою;
4) реплікує зловмисний код всередині цілі з зворотним доступом або без нього;
5) використовує вразливість системи за допомогою неправильного програмного забезпечення;
6) проникає через будь-який рівень структури OSI.

Припускається, що наступні атаки можуть статися, оскільки надані набори даних складаються з обмеженої кількості відомих зразків атак. Це атака користувача на корінь, атака віддаленого до локального, атака зондування, фаззери, аналіз, бекдори, експлойти, загальний, Reconnais sance, шеллкод, хробаки, атака грубою силою, атака Heartbleed, ботнет, розподілена атака на відмову в обслуговуванні (DDoS), веб-атака, інфільтраційна атака. Запропонований метод може лише захистити систему від виникнення атаки.

Система перевірки є одним із фундаментальних етапів будь-якої моделі загроз для оцінки ефективності. Процес перевірки моделі загроз можна класифікувати на дві групи:

теоретична та емпірична перевірка. Теоретична перевірка ґрунтується на існуючих теоріях, гіпотезах і моделюваннях, тоді як емпіричні докази базуються на перевірці за допомогою експериментів, досвіду та спостережень. У запропонованій моделі загрози застосовано підходи емпіричної перевірки, які більш конкретно базуються на експерименті. Ефективність загрози було виміряно шляхом збору та аналізу даних за допомогою комплексних експериментів. Підтверджено кінцевий результат, порівнявши різні експериментальні моделі на різних наборах даних.

Методологія розділена на п’ять основних етапів, а саме: збір даних, попередня обробка даних, вибір ознак ансамблю, класифікація моделі та виявлення аномалій.

Збір даних: продуктивність моделі машинного навчання значною мірою залежить від якості даних, які використовуються для навчання моделі. Дані містять відповідну інформацію про проблемну область і потребують очищення для подальшого аналізу. Тут використовується три добре відомі набори даних від різних дослідницьких організацій для навчання, тестування та перевірки запропонованої структури, і це NSL-KDD, CICIDS2017 і UNSW-NB15. Окрім поділу набору даних на тренувальні та тестові дані, є третя частина набору даних, яку називають «даними перевірки». Ці перевірочні дані імітують дані реального часу та використовуються для перевірки запропонованої моделі в СВВ.

Попередня обробка даних: необроблені набори даних потрібно очистити, продезінфікувати, трансформувати, нормалізувати та зменшити функції перед подачею в класифікатори МН. Щоб очистити та дезінфікувати набір даних, потрібно видалити рядки, що містять значення «,» або «inf». Для набору даних CICIDS2017 один пробіл перед кожною назвою об’єкта було видалено. Також видалено стовпець «id» у наборі даних UNSW-NB15. Оскільки моделі МН можуть обробляти числове введення, було перетворено нечислові дані в числові за допомогою техніки кодування даних для всіх трьох наборів даних. Щоб нормалізувати дані, ми використовуємо техніку масштабування MinMax, яка відповідає широкому діапазону значень у межах шкали від 0 до 1.

Таблиця1.

Дані для наборів даних NSL-KDD, UNSW-NB15 і CICIDS2017, які використовуються в експерименті

Набір даних	Навчання			Тестування			Перевірка		
	Всього	Доброякісних	Атака	Всього	Доброякісних	Атака	Всього	Доброякісна	Атака
NSL-KDD	125974	67373	58630	22544	12833	9711	1000	200	800
UNSW-NB15	112000	56000	56000	70000	35000	35000	1000	200	800
CICIDS2017	75000	38000	38000	25000	13000	12000	1000	200	800

Використано контрольовані методи, які ефективніші з збалансованим набором даних. Щоб зробити навчальний набір даних збалансованим, спочатку було вибрано тип даних (доброякісні або атакуючі) з мінімальною кількістю екземплярів даних у цьому навчальному наборі даних. Використовуючи цю мінімальну кількість, було взято майже однакову кількість нормальних (доброякісних) і аномальних (атака) екземплярів даних, щоб підготувати збалансований навчальний набір для наборів даних NSL-KDD і UNSW-NB15. Для підготовки тестового набору для цих наборів даних було враховано початкову кількість екземплярів тестових даних. Однак набір даних CICIDS2017 не має окремих наборів для навчання та

тестування. Щоб підготувати збалансований набір даних, потрібно дотримуватися тієї ж процедури, що й з двома наборами даних. Потім його розділити у співвідношенні 75 до 25, де 75% — дані навчання, а 25% — дані тестування. Кількість доброякісних даних і даних про атаку в трьох наборах даних показано в Таблиці 1.

Вибір функції ансамблю (АНВФ): вибір ансамблю ознак – це процес визначення найкращої підмножини ознак на основі методу голосування більшості. Процес ілюструється на рисунку 1, де використовується дев'ять окремих методів ВФ. Після цього ранжуємо функції відповідно до голосування більшістю, де більше половини методів ВФ вибирають функції. Основні внески цього дослідження полягають у використанні кількох алгоритмів вибору ознак, виконанні порівняльного аналізу між ними та групуванні їх за допомогою голосування більшості .

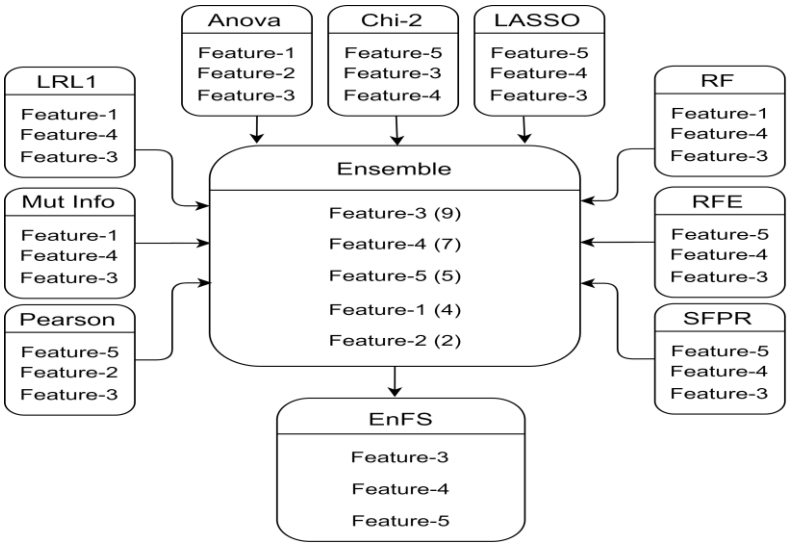

Рис. 1. Підхід до вибору ансамблевих функцій (АНВФ) з використанням голосування більшості.

На основі продуктивності було вибрано дев'ять найкращих методів ВФ зі списку п'ятнадцяти добре відомих методів. Це Anova, хі-квадрат, LASSO, логістична регресія зі штрафом L1, випадковий ліс, рекурсивне усунення ознак, кореляція Пірсона, взаємна інформація та SFPR. По-перше, було використано ці методи вибору функцій, щоб знайти оптимальні набори функцій, і керована сукупність моделей навчила п’ять окремих і шість моделей ансамблю для аналізу продуктивності за допомогою цих наборів функцій. Налаштувано методи вибору функцій, використовуючи кілька лічильників функцій, наприклад 5, 10, 15, 20, 25, 30 і повний набір. У більшості випадків для всіх трьох наборів даних підрахунок ознак 20 створив оптимальні набори функцій, які дали найкращі результати серед них. Потім було об’єднано всі дев’ять наборів функцій за допомогою методу голосування більшості та створили мінімальну кількість функцій. Усі функції, тобто; дев'ять наборів функцій з дев'яти методів ВФ, набір функцій з підходу АНВФ і повний набір функцій використовуються для навчання моделей, перелічених у керованій структурі ансамблю (SupAHMH) для цілей аналізу.

Класифікація моделі: у запропонованій груповій керованій структурі машинного навчання виконано двоетапну класифікацію, використовуючи послідовно індивідуальні та ансамблеві класифікатори, щоб розрізняти аномалії в наборах даних. Тренуємо п’ять контрольованих індивідуальних моделей, таких як логістична регресія (ЛР), дерево рішень (ДР), Naive Bayes (NB), нейронна мережа (НМ) і опорна векторна машина (ОВМ), використовуючи навчальні дані. Ці моделі тестуються за допомогою тестування даних та відповідні результати тесту (наприклад, 1 для аномального, 0 для доброякісного) генерують матрицю прогнозування. Ця матриця прогнозування, додана міткою з даних тестування,

формує дані для класифікаторів ансамблю. Цей набір даних поділяється на навчальні та тестові дані зі співвідношенням 70 до 30. Потім навчається та тестується шість класифікаторів ансамблю, а саме голосування за групою більшості , логістичну регресію ансамблю, групу Naive Bayes, нейронна мережа ансамблю, дерево рішень ансамблю і векторна машина підтримки ансамблю. Налаштувано гіперпараметри цих класифікаторів за допомогою алгоритму пошуку по сітці, щоб отримати гіперпараметри з найкращим значенням. Виконано 10-кратну перехресну перевірку під час навчання моделі для випадково розділеного набору даних, щоб уникнути переобладнання моделі. Після порівняльного аналізу ефективності цих одинадцяти моделей з використанням метрик оцінки було отримано найкращу модель для виявлення аномалії. Механізм об'єднання окремих класифікаторів за допомогою ансамблевої методики проілюстровано на рис. 2.

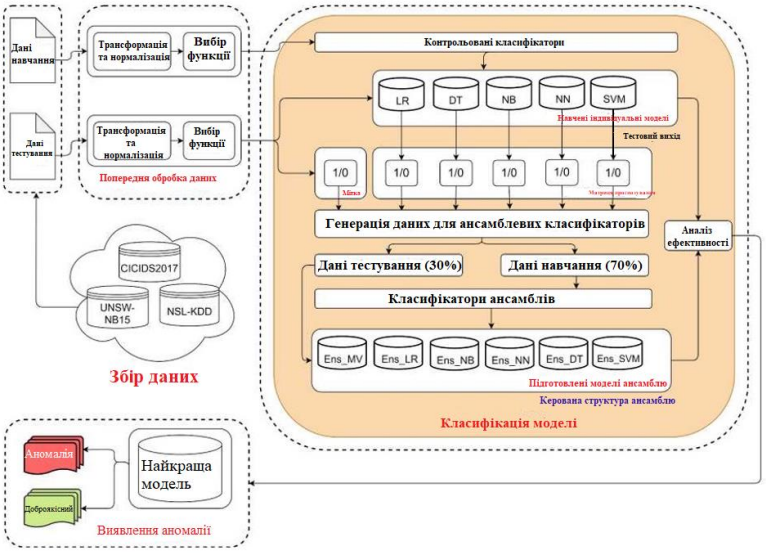

Рис. 2. Процес запропонованого методу.

Рис. 3. Структура ансамблевого машинного навчання як метод виявлення СВВ

Виявлення аномалії: в реальному світі СВВ, як показано на рис. 3, можна розмістити на шлюзі в захищеній мережі. СВВ складається з датчиків і препроцесора даних аудиту для перетворення вхідного трафіку в дані про діяльність. Використано «Дані перевірки», щоб імітувати потік даних у реальному часі. Як виявлення використовується модель із найефективнішою моделлю, отримана з контрольованої структури ансамблю модель СВВ.

Модель виявлення в системі виявлення аналізує дані перевірки та ідентифікує дані як аномальні або доброякісні. Після цього на основі правил із таблиці рішень механізм прийняття рішень виконує необхідні дії та повідомляє адміністратору мережі про потенційну загрозу.

Було реалізовано алгоритм вибору функцій ансамблю та класифікації ансамблю, щоб покращити загальну продуктивність запропонованої моделі машинного навчання. Як визначено в Алгоритмі , він приймає вхідні дані як функції та цілі для навчання та тестування моделей класифікації. Ініціалізовано список методів вибору функцій, моделі машинного навчання для індивідуальної класифікації, моделі машинного навчання для класифікації ансамблю та кількість найкращих функції. Крім того, було ініціалізовано структуру даних для зберігання результатів передбачення індивідуальної моделі тут. Підхід ансамблевого вибору та класифікації ознак виконується для всіх розглянутих найкращих наборів функцій, керованих зовнішнім циклом. K-найкращі ознаки, виділені індивідуальними алгоритмами вибору ознак, зберігаються для їх подальшого ансамблювання. Використовуємо навчальні функції та ціль, щоб отримати найкращий набір функцій із набору даних. Набір функцій ансамблю було розраховано на основі техніки голосування більшості, тобто ознака є важливою, якщо її вибрано принаймні половиною алгоритму вибору ознак. Наступний блок виконується для ідентифікації набору функцій ансамблю, і це кінець частини алгоритму, пов'язаної з вибором функцій ансамблю. Наступна частина алгоритму з класифікації ансамблю тренує моделі машинного навчання та зберігає результат класифікації в структурі даних, ініціалізованій на початку алгоритму. Накопичений результат було використано як новий набір даних, де ознаки є індивідуальними результатами класифікації, а ціль — це основна правда з набору даних. Новий набір даних було знову навчено та класифіковано всіма моделями машинного навчання, які розглядалися для експерименту. В наступному кроці аналізуємо результат, щоб оцінити продуктивність окремого класифікатора для визначення найкраща модель.

Аналіз складності алгоритму: перш ніж аналізувати складність алгоритму, потрібно визначити кілька термінів, а саме: n_{fm} як кількість методів вибору ознак, n_{mi} як кількість моделей МН для індивідуальної класифікації, n_{me} як кількість моделей МН для ансамблю класифікація, n_{fs} як кількість найкращих функцій, n_{fl} як кількість функцій у списку функцій, O(fs)як найвища часова складність серед усіх методів ВФ, O(mi) як найвища часова складність серед усіх окремі моделі МН і O(me) як найвищу часову складність серед усіх моделей МН ансамблю.

Складність алгоритму можна визначити як час, так і простір. Тому часова складність алгоритму така

$$\mathrm{T(n)} = n_{fs}\left[n_{fm}\left\{O(fs) + O\left(n_{fl}\right) + O\left(n_{fl}\right) + n_{mi}O(mi) + n_{me}O(me)\right\}\right]$$

Оскільки використано постійну кількість моделей МН і методів ВФ, спрощена версія часової складності дорівнює часовій складності, коли використовується один класифікатор і один метод ВФ.

$$\mathrm{T}(n) \approx O(fs) + MAX(O(mi), O(me))$$

В алгоритмі було використано 5 вхідних змінних, 5 статичних змінних, 6 ітераторів циклу, 1 карту та 8 інших змінних. У більшості випадків використовується змінну типу «список» і один тип хеш-карти.

У цьому дослідженні потрібні правильно встановлені оціночні метрики, щоб знайти найкращу ефективну модель, яку можна включити в СВВ. Чутливість, специфічність, точність, прецизійність, запам'ятовування та f-міра є добре відомими показниками оцінки ефективності. Вони походять від чотирьох основних термінів, таких як істинно позитивні,

хибнопозитивні, істинно негативні та хибнонегативні показники. Показники оцінки визначаються таким чином:

1) Точність [(TP + TN)/Total]: відсоток справжнього виявлення DDoS-атак від загальної кількості екземплярів даних.
2) Точність [TP/(FP + TP)]: це вимірювання того, як часто модель правильно визначає атаку DDoS.
3) Частота хибних спрацьовувань (FPR) [FP/(FP + TN)]: вказує, як часто модель викликає помилкову тривогу, визначаючи доброякісну атаку DDoS.
4) Recall [TP/(FN+TP)]: це вимірювання кількості DDoS-атак, які модель правильно визначає.
5) Відкликання також відоме як істинно-позитивний коефіцієнт, чутливість або рівень виявлення DDoS,
6) F1-Score [2 TP/(2 TN + FP + FN)]: середнє гармонічне значення точності та запам'ятовування.

Отримані результати експериментів з трьома наборами даних, NSL-KDD, UNSW-NB15 і CICIDS2017, у два етапи: вибір функцій і моделювання даних.

Показано функції та методи вибору функцій для NSL-KDD, UNSW-NB15 та CICIDS2017 відповідно, де 1 означає функцію (у рядку), вибрану відповідним методом ВФ (у стовпці), а 0 означає, що не вибрано. Останній стовпець підраховує загальний вибір методами ВФ для однієї функції. Функції перераховані в порядку зменшення загальної кількості. Оскільки використовується дев'ять методів ВФ, більшість перемагає, коли кількість перевищує 4,5 (тобто 5). Значення підрахунку для функцій нижче 5 вилучаються зі списку функцій і не відображаються. Ця техніка голосування за допомогою більшості використовується для вибору 20 із 43, 19 із 42 та 22 із 80 оптимізованих функцій із наборів даних NSL-KDD, UNSW-NB15 та CICCBB2017 відповідно. Запропонована методика АНВФ значно зменшує набір функцій для моделі класифікації на 53,5%, 54,8% і 72,5% відповідно для набору даних NSL-KDD, UNSW-NB15 і CICIDS2017.

З кожного з цих трьох наборів даних витягується одинадцять наборів функцій: дев'ять із них з дев'яти окремих методів ВФ, набір функцій із структури вибору функцій ансамблю та повний набір функцій (тобто вибір функцій не застосовувався). Використовуючи ці одинадцять наборів функцій, було навчено та перевірено класифікатори, які перераховані в груповій керованій структурі. Для кожного набору функцій керована структура ансамблю використовується для навчання п'яти окремих моделей і шести моделей ансамблю. Таким чином, загалом генерується 121 модель для одинадцяти наборів функцій. Експериментальні результати демонструють перевищення продуктивності запропонованої моделі класифікації ансамблю порівняно з індивідуальним класифікатором. У випадку набору даних NSL-KDD і CICIDS2017 запропонована структура SupАНВФ показує на 100% кращу точність класифікації для всіх індивідуальних і групових наборів функцій, тоді як 7 з 11 (63,3%) найефективніших класифікаторів були запропонованою моделлю машинного навчання ансамблю для UNSW-NB15 набір даних. Загалом запропонована модель ансамблевої класифікації перевершує індивідуальну модель класифікації для всіх наборів даних в експериментах.

5. Висновки.

У даній статті об'єднано вибір функцій ансамблю та підходи машинного навчання ансамблю як механізм виявлення в СВВ для виявлення мережевих аномалій. Для структури вибору функцій ансамблю спочатку було проведено експеримент з наборами функцій, отриманими з дев'яти методів вибору функцій, а потім об'єднано ці набори функцій, щоб отримати мінімальну кількість функцій за допомогою голосування більшості. Результати експериментів демонструють що набір функцій, отриманий за допомогою підходу АНВФ, має кращі результати порівняно з будь-яким окремим методом ВФ. Крім того, використано набір функцій в структурі МН під наглядом ансамблю, щоб знайти найкращу модель, яку можна включити в будь-яку СВВ. В статті було використано одинадцять наборів функцій (дев'ять із

дев'яти різних методів ВФ, набір функцій із АНВФ та повний набір функцій). Також проведено порівняльний аналіз цих наборів функцій, де набір АНВФ у більшості випадків перевершує окремі набори функцій. Крім того, для кожного набору функцій використано структуру МН під контролем ансамблю, щоб навчити одинадцять моделей (п'ять окремих і шість моделей ансамблю). Майже 80% комплектних моделей перевершують одиночні моделі. Для цього експерименту використано набори даних NSL-KDD, UNSW-NB15 і CICIDS2017.

Використання повного набору даних для всіх трьох наборів даних може призвести до довшого часу виконання експерименту. Тому використано зменшену кількість даних для цілей навчання, щоб збільшити час моделювання, і, таким чином, знижуючи ймовірність майже ідеальної продуктивності. У майбутньому планується включити повний обсяг даних, а також нові набори даних (включно з іншими доменами), щоб перевірити ефективність методу. Планується розглянути неконтрольоване навчання з груповою класифікацією, яка зміцнить даний метод. Крім того, змагальне машинне навчання (ЗМН) може бути додано як розширення досліджень у майбутньому, щоб уможлививши безпечне впровадження методів МН у змагальних налаштуваннях і, таким чином, зберегти безпеку всієї системи.

Список використаної літератури:

1. Lyu X., Ying F., Onpium P. Scene style conversion algorithm of AI digital host: a deep learning approach. 2023 2nd international conference on edge computing and applications (ICECAA),м.Namakkal,India,1921лип.2023р.
2. R. Vinayakumar, K. Soman, and P. Poornachandran, "Evaluating effec-tiveness of shallow and deep networks to intrusion detection system," in 2017 International Conference on Advances in Computing, Commu¬nications and Informatics (ICACCI), pp. 1282–1289, IEEE, 2017.
3. R. Vinayakumar, M. Alazab, S. Srinivasan, Q.-V. Pham, S. K. Padan- nayil, and K. Simran, "A visualized botnet detection system based deep learning for the internet of things networks of smart cities," IEEE Transactions on Industry Applications, vol. 56, no. 4, pp. 4436–4456, 2020.
4. H. M. Gomes, J. P. Barddal, F. Enembreck, and A. Bifet, "A survey on ensemble learning for data stream classification," ACM Computing Surveys (CSUR), vol. 50, no. 2, pp. 1–36, 2017.
5. O. Sagi and L. Rokach, "Ensemble learning: A survey," Wiley Interdisci-plinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1249, 2018.
6. X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," IEEE Access, vol. 7, pp. 82512–82521, 2019.
7. N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, and H. F. M. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," in Proceedings of the Australasian Computer Science Week Multiconference, pp. 1–6, 2018.
8. S. Das, A. M. Mahfouz, D. Venugopal, and S. Shiva, "Ddos intrusion detection through machine learning ensemble," in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 471–477, IEEE, 2019.
9. R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," Pattern Recogni-tion, vol. 64, pp. 141–158, 2017.
10. S. Adams and P. A. Beling, "A survey of feature selection methods for gaussian mixture models and hidden markov models," Artificial Intelligence Review, vol. 52, no. 3, pp. 1739–1779, 2019.
11. W. Xiong and R. Lagerstro¨m, "Threat modeling–a systematic literature review," Computers & security, vol. 84, pp. 53–69, 2019.
12. Achuthan K. Threat modeling and threat intelligence system for cloud using splunk. 2022 10th international symposium on digital forensics and security (ISDFS), Istanbul, Turkey, 6–7 June 2022. 2022. URL: https://doi.org/10.1109/isdfs55398.2022.9800787