КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue
DOI 10.28925/2663-4023.2025.30.963
УДК 004.056.53:004.8]:334.716-021.412.1

Гамза Дмитро Євгенійович
аспірант кафедри систем та технологій кібербезпеки
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0005-0947-2420
d.gamza@stud.duikt.edu.ua

ВПЛИВ ОПТИМІЗАЦІЇ ДАТАСЕТУ CSE–CIC–IDS2018 НА ЕФЕКТИВНІСТЬ ГІБРИДНОЇ СТЕКІНГОВОЇ МОДЕЛІ ВИЯВЛЕННЯ МЕРЕЖЕВИХ ВТОРГНЕНЬ

Анотація. У цій статті представлено розширений порівняльний аналіз ефективності гібридної стекінгової моделі, призначеної для виявлення мережевих вторгнень, де особливий акцент зроблено на трансформації показників продуктивності до та після впровадження комплексного методу попередньої обробки сучасного датасету CSE–CIC–IDS2018. Запропонований підхід до підготовки даних базується на синергії трьох стратегічних компонентів: алгоритму SMOTE для інтелектуального балансування класів шляхом генерації синтетичних зразків міноритарних атак, методу Min–Max нормалізації для масштабування ознакового простору до діапазону [0, 1], що забезпечує рівномірний внесок кожного параметра у процес навчання, та методу головних компонент (PCA) для агресивного зниження розмірності даних без втрати ключової дисперсії. Для досягнення максимальної об'єктивності та верифікації результатів було проведено масштабний експериментальний цикл, що охоплював навчання й тестування ключових фундаментальних алгоритмів машинного навчання, а також десяти унікальних конфігурацій гібридного метакласифікатора на основі стекінгового ансамблю. Експериментально доведено, що така глибока оптимізація вхідних даних дозволяє гібридній моделі подолати проблему «перенавчання» на мажоритарних класах та значно підвищити аналітичну потужність, що відобразилося у зростанні показника Accuracy на 3,87% та F1–міри на 5,11%. Найбільш критичним для практичного застосування результатом стало радикальне скорочення часу прогнозування на 76,0%, що фактично знімає обчислювальні бар’єри для інтеграції складних ансамблевих методів у високонавантажені системи виявлення вторгнень, які працюють у режимі реального часу. Таким чином, інтеграція SMOTE, Min–Max нормалізації та PCA визначена як фундаментальна архітектурна передумова для створення стійких до кіберзагроз систем нового покоління, здатних ефективно виявляти аномалії в умовах високої інтенсивності мережевого трафіку.

Ключові слова: кібербезпека, загрози, виявлення мережевих вторгнень, CSE–CIC–IDS2018, SMOTE, Min–Max нормалізація, аналіз головних компонент (PCA), стекінг, гібридна модель, машинне навчання.

ВСТУП

Зростання кількості та складності кіберзагроз у сучасних корпоративних мережах обумовлює нагальну потребу в розвитку ефективних систем виявлення вторгнень (IDS). Дослідження свідчать, що традиційні сигнатурні підходи втрачають ефективність проти нових класів атак, тоді як методи машинного навчання демонструють здатність до узагальнення на раніше невідомі загрози [12]. Проте практична застосовність ML–моделей у задачах кібербезпеки суттєво залежить від якості навчальних даних: реальний мережевий трафік має властивості, які ускладнюють безпосереднє застосування стандартних алгоритмів [11].

Одним із найбільш репрезентативних відкритих датасетів для задач виявлення вторгнень є датасет CSE–CIC–IDS2018 [1], суттєво перевершує застарілі набори даних, зокрема KDD Cup 99 за реалістичністю трафіку та різноманітністю векторів атак [15]. Утім, навіть цей датасет у первинному вигляді характеризується трьома системними проблемами, характерні для більшості реальних наборів даних мережевої безпеки: критичний дисбаланс класів (частка легітимного трафіку перевищує 85%), надмірна розмірність ознакового простору (понад 80 атрибутів із сильною мультиколінеарністю) та гетерогенність масштабів числових значень. Ці проблеми відомі в літературі як одні з ключових перешкод для побудови надійних IDS–класифікаторів [8, 14].

Метою даної статті є кількісна оцінка впливу комплексного методу попередньої обробки – що поєднує SMOTE [2], Min–Max нормалізацію та PCA [3] – на ефективність гібридної стекінгової моделі [4] класифікації мережевого трафіку на датасеті CSE–CIC–IDS2018 [1]. Для об'єктивної оцінки проведено повний цикл порівняльного експерименту: навчання та тестування ідентичних моделей окремо на необроблених та оптимізованих даних. Наукова новизна дослідження полягає у кількісному підтвердженні того, що переваги стекінгової архітектури реалізуються лише за умови належної якості вхідних даних.

ДАТАСЕТ ТА МЕТОДИ ПОПЕРЕДНЬОЇ ОБРОБКИ

Характеристика датасету CSE–CIC–IDS2018

Датасет CSE–CIC–IDS2018 сформовано у контрольованому лабораторному середовищі, що імітує корпоративну мережу з 50 робочими станціями та виділеними атакувальними хостами [1]. Генерація трафіку здійснювалася на основі поведінкових профілів користувачів (B-Profiles) реальної поведінки користувачів, що відрізняє цей набір від синтетичних датасетів попередніх поколінь, таких як KDD Cup 99 та DARPA 1999, ізольованість яких від реального трафіку була предметом критики [15]. Набір охоплює сім категорій атак: Brute Force, Heartbleed, Botnet, DoS, DDoS, Web Attacks та Infiltration. Кожен мережевий потік описується 80 ознаками, що охоплюють статистику довжини пакетів, часові характеристики (IAT (Inter-Arrival Time)), прапорці TCP та показники швидкості передачі.

Загальний обсяг набору становить понад 16 мільйонів записів, що теоретично забезпечує достатню статистичну потужність для навчання складних ансамблевих моделей. Водночас дослідники зазначають, що масштаб набору сам по собі не компенсує проблему дисбалансу класів: модель, навчена на датасеті з 90% прикладів одного класу, схильна до систематичного ігнорування міноритарних класів навіть за наявності мільйонів записів [14].

Критичною проблемою первинного датасету є аномальний дисбаланс класів: частка легітимного трафіку (Benign) перевищує 85%, тоді як окремі класи атак (Infiltration, певні типи Botnet) представлені лише кількома сотнями записів. Дана проблема широко досліджена в контексті IDS: класифікатор, навчений на незбалансованих даних, оптимізується під мінімізацію загальної помилки, жертвуючи чутливістю до міноритарних класів [8]. Для задач виявлення вторгнень це означає, що модель фактично «ігнорує» рідкісні, але критичні типи атак – саме ті, виявлення яких є метою системи. Відомо також, що цей ефект підсилюється в стекінгових ансамблях, оскільки зміщені прогнози базових моделей накопичуються на рівні мета–класифікатора [4].

Метод комплексної оптимізації датасету

Для усунення виявлених недоліків первинних даних застосовано триетапний метод попередньої обробки, що є узгодженим із рекомендаціями щодо підготовки даних для IDS–систем на основі машинного навчання [12].

Перший етап – подолання нерівномірності розподілу класів на основі алгоритму SMOTE (Synthetic Minority Over-sampling Technique). На відміну від наївного дублювання записів міноритарного класу (random oversampling), що призводить до перенавчання [14], SMOTE генерує нові синтетичні зразки шляхом інтерполяції між існуючими об'єктами та їх k–найближчими сусідами у просторі ознак [2]. Новий зразок обчислюється за формулою:

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

де $\lambda \in [0, 1]$ – рівномірно розподілена випадкова величина. Така стратегія розширює межі рішень для міноритарних класів, а не просто копіює існуючі точки, що підвищує узагальнюючу здатність класифікатора [10].

Другий етап – масштабування ознакового простору методом Min–Max нормалізації. Кожна ознака приводиться до діапазону [0, 1] за формулою:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Необхідність цього кроку обумовлена гетерогенністю природи мережевих ознак: Flow Duration вимірюється мільйонами мікросекунд, тоді як прапорці TCP є бінарними значеннями. Без нормалізації алгоритми, що використовують евклідову метрику – зокрема SVM [13] та KNN – надають домінуючий вплив ознакам з великими абсолютними значеннями, фактично ігноруючи бінарні прапорці, які часто є найбільш інформативними при виявленні сканування портів та DoS–атак. Аналогічна проблема характерна для градієнтних методів у нейронних мережах, де великі значення дестабілізують процес оптимізації.

Третій етап – зниження розмірності методом PCA (Principal Component Analysis). Метод головних компонент знаходить лінійні комбінації вихідних ознак – головні компоненти – впорядковані за обсягом поясненої дисперсії [3]. Застосування PCA до попередньо нормалізованого простору ознак CSE–CIC–IDS2018 дозволило редукувати 80–вимірний простір до 18 ортогональних компонент, що зберігають 95% загальної дисперсії. Ортогональність компонент за означенням усуває мультиколінеарність, яка була ідентифікована як одна з головних перешкод при роботі з цим датасетом. Практичним наслідком редукції стало зниження обчислювального навантаження на 65%, що є критичним для досягнення прийнятного часу інференсу в системах реального часу [11].

АРХІТЕКТУРА ГІБРИДНОЇ СТЕКІНГОВОЇ МОДЕЛІ

Гібридна модель побудована на архітектурі стекінгу (Stacking Ensemble). Ключова ідея стекінгу, запропонованого Волпертом [4], полягає в тому, що замість об'єднання прогнозів базових моделей за фіксованим правилом (голосування, усереднення), навчається окремий мета–класифікатор, який адаптивно зважує їх внески. Це дозволяє автоматично компенсувати систематичні помилки окремих алгоритмів. Перший рівень

включає десять різнорідних алгоритмів: SVM як лінійний роздільник у просторі ознак [13], Random Forest як ансамбль декореляційних дерев [9], KNN як непараметричний класифікатор відстані, алгоритми градієнтного бустингу XGBoost [5], LightGBM [6] та CatBoost [7] як представники сімейства адаптивних ансамблів, а також Extra Trees, MLP, Logistic Regression та Naive Bayes. Різнорідність базових моделей є принциповою умовою ефективності стекінгу: алгоритми мають помилятися на різних підмножинах об'єктів.

Розподіл вибірки здійснювався у співвідношенні 70/15/15 для навчання, валідації та тестування відповідно. Для оцінки ефективності використовувались дві ключові метрики: Accuracy як загальна частка правильних класифікацій та F1–score (macro) як їх гармонічне середнє між точністю та повнотою по всіх класах. F1–score є більш релевантною метрикою для незбалансованих задач виявлення вторгнень, оскільки однаково штрафує за хибно–позитивні спрацьовування та пропущені атаки [8]. Додатково фіксувався середній час прогнозування на один запит (мс) як показник придатності до розгортання в реальному часі [11]. Усі експерименти виконувались на ідентичному апаратному забезпеченні: Intel Core i7–13700HX 3.7 ГГц, 64 ГБ DDR4, NVIDIA GeForce RTX 4070, ОС Ubuntu 24.04 LTS.

Для аналізу впливу попередньої обробки проведено два ідентичних цикли навчання: перший – на первинному (необробленому) датасеті, другий – на оптимізованому за допомогою описаного вище методу.

РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНОГО ДОСЛІДЖЕННЯ

Показники базових алгоритмів без оптимізації датасету

Таблиця 1 відображає результати класифікації базових алгоритмів, навчених на первинному датасеті CSE–CIC–IDS2018 без будь–якої попередньої обробки.

Таблиця 1

Результати класифікації базових моделей на необробленому датасеті (до оптимізації)

Модель	Accuracy	F1–score	Час прогнозу (мс)
SVM	0.863	0.831	18.4
Random Forest	0.921	0.904	13.7
KNN	0.847	0.812	26.3
XGBoost	0.934	0.921	16.8
LightGBM	0.926	0.912	11.9
CatBoost	0.931	0.918	13.1
Extra Trees	0.914	0.899	14.2
MLP	0.878	0.853	22.6
Logistic Regression	0.841	0.806	9.3
Naive Bayes	0.793	0.754	3.8

Примітка: навчання виконано на первинному датасеті без застосування SMOTE, Min–Max нормалізації та PCA.

Аналіз результатів підтверджує теоретично очікуваний ефект незбалансованих даних [8, 14]: загальна точність (Accuracy) варіює від 0.793 до 0.934, однак цей показник є оманливим індикатором якості IDS. Більш показовим є стабільний і значний розрив між Accuracy та F1-score: для Naive Bayes він становить 0.039, для Logistic Regression – 0.035, для KNN – 0.035. Згідно з теорією навчання на незбалансованих даних [14], такий розрив свідчить про те, що моделі мінімізують загальну помилку коштом чутливості до міноритарних класів – тобто саме тих атак, виявлення яких є метою IDS. Окремо слід відзначити критично високі показники часу прогнозу для алгоритмів, чутливих до розмірності простору: KNN – 26.3 мс, MLP – 22.6 мс, SVM – 18.4 мс. Для SVM це є наслідком обчислення ядерних функцій у 80–вимірному просторі [13], для KNN – квадратичної залежності від кількості ознак при пошуку найближчих сусідів.

Показники гібридної стекінгової моделі без оптимізації датасету

У Таблиці 2 представлено результати гібридних стекінгових моделей, сформованих на основі базових класифікаторів, навчених на необробленому датасеті.

Таблиця 2

Порівняння гібридних стекінгових моделей на необробленому датасеті (до оптимізації)

Базові моделі	Мета–класифікатор	Accuracy	F1–score	Час прогнозу (мс)
XGBoost + CatBoost + LightGBM	XGBoost	0.9441	0.9187	28.6
XGBoost + CatBoost + Random Forest	XGBoost	0.9433	0.9172	30.1
XGBoost + CatBoost + LightGBM	Gradient Boosting	0.9418	0.9151	29.3
XGBoost + CatBoost + Random Forest	Gradient Boosting	0.9409	0.9138	31.4
XGBoost + CatBoost + LightGBM	Random Forest	0.9396	0.9124	29.8
XGBoost + CatBoost + Random Forest	Random Forest	0.9381	0.9107	31.7
XGBoost + CatBoost + LightGBM	Logistic Regression	0.9352	0.9068	27.9
XGBoost + CatBoost + Random Forest	Logistic Regression	0.9339	0.9047	29.4
CatBoost + LightGBM + Extra Trees	XGBoost	0.9314	0.9013	27.1
CatBoost + LightGBM + Extra Trees	Gradient Boosting	0.9291	0.8986	28.4

Примітка: навчання виконано на первинному датасеті без застосування SMOTE, Min–Max нормалізації та PCA.

Стекінгова архітектура на необробленому датасеті підтверджує загальновідому перевагу ансамблевих методів [4]: найкраща конфігурація (XGBoost + CatBoost + LightGBM, мета–класифікатор XGBoost) перевищує найкращу базову модель за Accuracy на 1.07% (0.9441 проти 0.934). Однак два критичних обмеження унеможливлюють практичне застосування. По–перше, час прогнозування усіх конфігурацій лежить у діапазоні 27.1–31.7 мс – у 4–5 разів більше за прийнятний поріг для IDS, що означатиме повну затримку реакції системи на атаку [11]. По–друге, F1–score найкращої конфігурації (0.9187) залишається відносно низьким попри стекінг, що є прямим наслідком зміщених прогнозів базових моделей: коли кожна базова модель систематично недооцінює міноритарні класи через дисбаланс [8], мета–класифікатор отримує скорочений та спотворений сигнал і відтворює ці зміщення на виході.

Показники базових алгоритмів після оптимізації датасету

Таблиця 3 відображає аналогічні показники базових алгоритмів після застосування методу комплексної оптимізації датасету (SMOTE + Min–Max + PCA).

Таблиця 3

Результати класифікації базових моделей на оптимізованому датасеті (після застосування SMOTE, Min–Max, PCA)

Модель	Accuracy	F1–score	Час прогнозу (мс)
SVM	0.91	0.89	11.3
Random Forest	0.94	0.93	8.2
KNN	0.90	0.88	14.7
XGBoost	0.95	0.94	10.1
LightGBM	0.94	0.93	6.8
CatBoost	0.95	0.94	7.5
Extra Trees	0.93	0.92	7.9
MLP	0.92	0.91	13.2
Logistic Regression	0.89	0.87	5.4
Naive Bayes	0.85	0.82	2.1

Примітка: датасет попередньо оброблено методом SMOTE (балансування), Min–Max нормалізацією та PCA (18 компонент).

Порівняння таблиць 1 і 3 наочно демонструє диференційований ефект оптимізації залежно від природи алгоритму. Найбільший приріст Accuracy спостерігається у моделей, які теоретично найбільш вразливі до гетерогенності масштабів та дисбалансу: KNN – з 0.847 до 0.90 (+6.3%), SVM – з 0.863 до 0.91 (+5.4%), Logistic Regression – з 0.841 до 0.89 (+5.8%), MLP – з 0.878 до 0.92 (+4.8%). Це узгоджується з теоретичними обґрунтуваннями: SVM оптимізує гіперплощину розділення у метричному просторі і є чутливим до масштабу ознак [13]; KNN безпосередньо обчислює евклідові відстані, що робить його повністю залежним від нормалізації; градієнтний спуск у MLP нестабільний при великих та різнорідних значеннях вхідних ознак. SMOTE [2] забезпечив суттєве покращення F1–score для всіх моделей, оскільки усунув систематичне зміщення в бік

мажоритарного класу [14]. Алгоритми градієнтного бустингу (XGBoost: +1.7%, LightGBM: +1.5%, CatBoost: +2.0%) показали помірніший приріст точності – що очікувано, адже вони мають вбудовані механізми зважування класів і менш чутливі до масштабу – однак отримали значний виграш у часі інференсу завдяки PCA [3]: XGBoost з 16.8 до 10.1 мс, LightGBM – з 11.9 до 6.8 мс, CatBoost – з 13.1 до 7.5 мс.

Показники гібридної стекінгової моделі після оптимізації датасету

Ключові результати дослідження представлені у Таблиці 4, що демонструє значення метрик гібридних стекінгових моделей, навчених на оптимізованому датасеті.

Таблиця 4.

Порівняння гібридних стекінгових моделей на оптимізованому датасеті (після застосування SMOTE, Min–Max, PCA)

Базові моделі	Мета–класифікатор	Accuracy	F1–score	Час прогнозу (мс)
XGBoost + CatBoost + LightGBM	XGBoost	0.9807	0.9657	7.16
XGBoost + CatBoost + Random Forest	XGBoost	0.9801	0.9648	7.57
XGBoost + CatBoost + LightGBM	Gradient Boosting	0.9796	0.9639	7.40
XGBoost + CatBoost + Random Forest	Gradient Boosting	0.9791	0.9631	7.83
XGBoost + CatBoost + LightGBM	Random Forest	0.9784	0.9643	7.48
XGBoost + CatBoost + Random Forest	Random Forest	0.9779	0.9628	7.91
XGBoost + CatBoost + LightGBM	Logistic Regression	0.9762	0.9611	6.91
XGBoost + CatBoost + Random Forest	Logistic Regression	0.9755	0.9598	7.31
CatBoost + LightGBM + Extra Trees	XGBoost	0.9741	0.9587	6.51
CatBoost + LightGBM + Extra Trees	Gradient Boosting	0.9733	0.9572	6.73

Примітка: датасет попередньо оброблено методом SMOTE, Min–Max нормалізацією та PCA (18 компонент).

Після оптимізації датасету найкраща конфігурація гібридної моделі (XGBoost + CatBoost + LightGBM, мета–класифікатор XGBoost) досягає Accuracy 0.9807 та F1–score 0.9657 – приріст на 3.87% та 5.11% відповідно порівняно з моделлю, навченою на необробленому датасеті. Суттєво, що приріст F1–score перевищує приріст Accuracy, що є характерною ознакою успішного балансування класів [2, 10]: покращення відбулося насамперед за рахунок підвищення чутливості до рідкісних класів атак. Особливо

показовим є скорочення мінімального часу прогнозування з 27.1 до 6.51 мс (−76.0%), що є прямим ефектом PCA [3]: зменшення розмірності з 80 до 18 ознак скоротило обсяг обчислень кожного базового алгоритму, транслювавшись у пропорційне зниження часу всього стекінгового пайплайну. Значення 6.51 мс відповідає вимогам до IDS/IPS у режимі реального часу [11], тоді як 27.1 мс до оптимізації унеможливлювало такий режим.

ПОРІВНЯЛЬНИЙ АНАЛІЗ РЕЗУЛЬТАТІВ

Зведений порівняльний аналіз ефекту оптимізації представлено у Таблиці 5.

Таблиця 5.

Зведене порівняння ключових показників до та після оптимізації датасету

Параметр	До оптимізації	Після оптимізації	Абс. приріст	Відн. приріст, %
Найкраща Accuracy (гібр.)	0.9441	0.9807	+0.0366	+3.87%
Найкращий F1–score (гібр.)	0.9187	0.9657	+0.0470	+5.11%
Середня Accuracy гібридних моделей	0.9377	0.9775	+0.0398	+4.24%
Середній F1–score гібридних моделей	0.9099	0.9621	+0.0522	+5.74%
Мін. час прогнозу (гібр., мс)	27.1	6.51	−20.6	−76.0%
Середній Accuracy базових моделей	0.885	0.918	+0.033	+3.73%
Середній F1–score базових моделей	0.861	0.903	+0.042	+4.88%
Приріст Accuracy (стекінг vs. базові)	+0.053	+0.060	+0.007	+12.4%

Аналіз зведеної таблиці дозволяє виокремити кілька принципово важливих спостережень.

По–перше, вплив оптимізації на базові алгоритми є диференційованим і теоретично передбачуваним. Алгоритми, засновані на евклідовій метриці або градієнтних методах (KNN, SVM [13], MLP, Logistic Regression), чутливі як до масштабу ознак, так і до дисбалансу класів, тому демонструють найбільший приріст Accuracy (4.8–6.3%). Деревоподібні та бустингові алгоритми (Random Forest [9], XGBoost [5], LightGBM [6], CatBoost [7]) мають вбудовані механізми інваріантності до масштабу і показують помірніший приріст точності (1.5–2.0%), однак отримують ключовий виграш у часі інференсу – в середньому −43% – завдяки зниженню розмірності через PCA [3].

По–друге, стекінгова архітектура демонструє значно більшу чутливість до якості вхідних даних, ніж окремі базові класифікатори. Ця властивість теоретично обґрунтована природою стекінгу [4]: мета–класифікатор навчається на прогнозах базових моделей, тому систематичне зміщення останніх (обумовлене дисбалансом класів [8]) транслюється у спотворений навчальний сигнал для мета–рівня і не може бути скоригованим лише вибором складнішого мета–алгоритму. Після застосування SMOTE

[2] базові моделі отримують збалансоване навчальне середовище, що усуває систематичне зміщення і надає мета–класифікатору репрезентативний сигнал для всіх класів атак.

По–третє, більший приріст F1–score (+5.11%) порівняно з Accuracy (+3.87%) є статистично значущим індикатором: він свідчить про те, що оптимізація насамперед покращила виявлення рідкісних класів атак, а не лише загальну точність на мажоритарному класі. Це є закономірним і цільовим ефектом SMOTE [2, 10]. У контексті систем виявлення вторгнень F1–score є пріоритетною метрикою, оскільки пропущена атака (хибно–негативний результат) тягне за собою компрометацію системи, тоді як хибнопозитивне спрацьовування є лише операційною незручністю [12]. Отже, підвищення F1–score є більш цінним результатом, ніж еквівалентне підвищення Accuracy.

По–четверте, скорочення мінімального часу прогнозування на 76.0% (з 27.1 до 6.51 мс) є принциповою зміною класу придатності системи. Дослідники практичного розгортання IDS зазначають, що затримка понад 20–25 мс робить систему непридатною для роботи в режимі реального часу в сучасних мережах [11]. Значення 6.51 мс, досягнуте завдяки PCA–редукції розмірності [3], дозволяє класифікувати до 153 тисяч мережевих потоків за секунду на одному ядрі. Суттєво, що цей ефект масштабується: архітектура стекінгу [4] природно підтримує паралелізацію базових класифікаторів, що при використанні багатоядерних процесорів пропорційно збільшує пропускну здатність системи.

ВИСНОВКИ

У статті проведено комплексне порівняльне дослідження ефективності гібридної стекінгової моделі виявлення мережевих вторгнень до та після застосування методу оптимізації датасету CSE–CIC–IDS2018, що включає SMOTE, Min–Max нормалізацію та PCA. Отримані результати дозволяють сформулювати такі висновки:

Комплексна оптимізація датасету CSE–CIC–IDS2018 забезпечує статистично значущий приріст ефективності гібридної стекінг–моделі: Accuracy зростає з 0.9441 до 0.9807 (+3.87%), F1–score – з 0.9187 до 0.9657 (+5.11%), час прогнозування скорочується з 27.1 до 6.51 мс (−76.0%).

Вплив оптимізації на базові алгоритми є диференційованим: алгоритми, чутливі до масштабу ознак та дисбалансу класів (KNN, SVM, MLP, Logistic Regression), демонструють суттєвий приріст Accuracy від 4.8% до 6.3%, тоді як алгоритми градієнтного бустингу (XGBoost, LightGBM, CatBoost) – помірніший приріст точності (1.5–2.0%), компенсований значним зниженням часу інференсу (до − 43%) завдяки PCA.

Гібридна стекінгова архітектура є принципово більш чутливою до якості вхідних даних порівняно з ізольованими класифікаторами, оскільки систематичні похибки базових моделей накопичуються на рівні мета–класифікатора.

Досягнутий мінімальний час прогнозування 6.51 мс після оптимізації відповідає вимогам до IDS/IPS–систем реального часу, тоді як значення 27.1 мс (до оптимізації) унеможливлює практичне розгортання в умовах корпоративних мереж з інтенсивним трафіком.

Результати підтверджують, що попередня обробка даних є невід'ємним елементом повного циклу розробки IDS–систем на основі машинного навчання і не може бути замінена або компенсована вибором більш складної архітектури класифікатора.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)*, 108–116.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.
3. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*(2065), 1–16.
4. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.
7. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*, 6638–6648.
8. Lim, M., & Al-Hussain, A. (2019). Class imbalance problem in intrusion detection systems: A survey. *IEEE Access*, *7*, 90561–90578.
9. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
10. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer, Cham.
11. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *Proceedings of the IEEE Symposium on Security and Privacy*, 305–316.
12. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176.
13. Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.
14. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.
15. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1–6.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue

Dmytro Hamza
Postgraduate student, Department of Cybersecurity Systems and Technologies
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0009-0005-0947-2420
d.gamza@stud.duikt.edu.ua

IMPACT OF OPTIMIZATION OF THE CSE–CIC–IDS2018 DATASET ON THE EFFICIENCY OF THE HYBRID STICKING MODEL FOR NETWORK INTRUSION DETECTION

Abstract. This paper presents an extensive comparative analysis of the performance of a hybrid stacking model designed for network intrusion detection, with a special emphasis on the transformation of performance indicators before and after the implementation of a comprehensive preprocessing method for the modern CSE–CIC–IDS2018 dataset. The proposed data preparation approach is based on the synergy of three strategic components: the SMOTE algorithm for intelligent class balancing by generating synthetic minority attack samples, the Min–Max normalization method for scaling the feature space to the range [0, 1], which ensures a uniform contribution of each parameter to the training process, and the Principal Component Analysis (PCA) method for aggressively reducing the dimensionality of the data without losing key variance. To achieve maximum objectivity and verify the results, a large-scale experimental cycle was conducted, covering the training and testing of key fundamental machine learning algorithms, as well as ten unique configurations of the hybrid stack ensemble-based metaclassifier. It has been experimentally proven that such deep optimization of input data allows the hybrid model to overcome the problem of "overtraining" on majority classes and significantly increase the analytical power, which was reflected in an increase in accuracy by 3.87% and F1-measure by 5.11%. The most important result for practical application was a radical reduction in prediction time by 76.0%, which effectively removes computational barriers for integrating complex ensemble methods into high-load intrusion detection systems operating in real-time. Thus, the integration of SMOTE, Min-Max normalization and PCA is defined as a fundamental architectural prerequisite for creating new generation systems resistant to cyber threats, capable of effectively detecting anomalies in conditions of high network traffic intensity

Keywords: cybersecurity, threats, network intrusion detection, CSE–CIC–IDS2018, SMOTE, Min-Max normalization, principal component analysis (PCA), stacking, hybrid model, machine learning.

REFERENCES

1. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)*, 108–116.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.
3. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*(2065), 1–16.
4. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.
7. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems, 31*, 6638–6648.
8. Lim, M., & Al-Hussain, A. (2019). Class imbalance problem in intrusion detection systems: A survey. *IEEE Access, 7*, 90561–90578.
9. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

10. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer, Cham.
11. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *Proceedings of the IEEE Symposium on Security and Privacy*, 305–316.
12. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176.
13. Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.
14. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.
15. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1–6.