УДК 004.7

DOI: 10.18372/2073-4751.80.19779

Кудренко С.О., к.т.н.,
orcid.org/ 0000-0002-0759-3908,
e-mail: stanislava@i.ua,
Столяр А.Л..,
orcid.org/0000-0002-7669-1202,
e-mail: stoliarannanau@gmail.com,
Макєєв І.Г.,
orcid.org/0009-0009-8679-5652,
e-mail: 8390988@stud.kai.edu.ua

ФУНКЦІОНАЛЬНИЙ ВІДБІР ReliefF ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ ТРАФІКУ

Державний університет «Київський авіаційний інститут»

Вступ

Кількість інцидентів, пов'язаних із несанкціонованим доступом, розподіленими атаками відмови в обслуговуванні (*DDoS*) та витоком даних, постійно зростає. Традиційні методи виявлення аномалій на основі правил або статистичного аналізу часто не здатні адаптуватися до швидко змінюваних шаблонів трафіку та виявляти нові загрози. У зв'язку з цим виникає потреба у застосуванні інтелектуальних методів аналізу даних, здатних ефективно відрізняти нормальну мережеву активність від потенційно шкідливої.

Один із перспективних підходів до розв'язання цієї проблеми базується на застосуванні генетичних алгоритмів у поєднанні з методами вибору ознак, такими як *ReliefF*. Генетичні алгоритми, натхненні принципами природного добору, демонструють високу ефективність у розв'язанні оптимізаційних задач, дозволяючи виділяти найбільш релевантні ознаки для класифікації мережевого трафіку. Вибір ознак є критично важливим для підвищення точності та швидкості систем виявлення аномалій, оскільки зайві або нерелевантні ознаки можуть знижувати продуктивність моделі та спричиняти велику кількість хибних спрацьовувань.

Останні дослідження [1-12] підтверджують ефективність використання гібридних методів для виявлення мережевих атак. Наприклад, у роботі [1] представили гібридний підхід, що поєднує генетичні алгоритми з оптимізаційними методами та нейронними мережами, що дозволило значно підвищити точність виявлення аномалій. У дослідженні [2] продемонстровано, що використання *ReliefF* у поєднанні з глибокими нейронними мережами суттєво зменшує розмірність вхідних даних без втрати інформаційної цінності.

Перспективність такого підходу полягає в його здатності адаптивно оновлюватися відповідно до змін у трафіку, що дозволяє ефективно виявляти як відомі, так і нові аномалії. Крім того, використання алгоритму *ReliefF* для вибору найбільш значущих ознак дозволяє скоротити витрати на обчислення та підвищити продуктивність системи.

Мета

Метою даної роботи є дослідження ефективності застосування генетичних стратегій у поєднанні з *ReliefF* для виявлення аномалій трафіку у динамічних мережах. У ході дослідження буде проведено аналіз релевантності вибраних ознак, їхнього впливу на точність виявлення атак.

Принцип роботи ReliefF

ReliefF дозволяє визначити, які характеристики найбільше впливають на класифікацію трафіку, оцінюючи значущість кожної ознаки залежно від її здатності розрізняти нормальні та аномальні зразки. Основний механізм роботи алгоритму базується на знаходженні найближчих

сусідів для випадково вибраного зразка та оцінці того, наскільки певна ознака відрізняється між об'єктами одного класу (нормальний трафік або аномальний) та різних класів. Якщо значення певної характеристики значно змінюється між нормальним і аномальним трафіком, *ReliefF* призначає їй вищу вагу, а якщо характеристика однаково зустрічається в обох класах, її вплив знижується.

У контексті аналізу мережевого трафіку застосування *ReliefF* починається з попередньої обробки даних, де відбувається нормалізація числових параметрів, перетворення категорійних ознак у числові значення та усунення зайвих атрибутів. Після цього алгоритм оцінює вплив кожної ознаки на здатність класифікаційної моделі виявляти аномалії. Наприклад, такі характеристики, як кількість з'єднань із певним хостом за короткий проміжок часу, частота змін портів або кількість помилкових запитів, можуть отримати високі вагові значення, оскільки вони часто вказують на підозрілу активність, наприклад *DDoS*-атаку або сканування мережі.

Після виконання *ReliefF* отримані ваги використовуються для відбору найкращих характеристик, які будуть використані в подальшому навчанні моделі виявлення аномалій. Це дозволяє зменшити розмірність вхідних даних без втрати інформативності, що сприяє покращенню продуктивності класифікатора та прискоренню обчислень. Оскільки системи моніторингу мережевого трафіку часто працюють у реальному часі, швидкість аналізу відіграє важливу роль у забезпеченні оперативного реагування на загрози.

Основна ідея *ReliefF* (рис. 1) полягає у вимірюванні значущості кожної ознаки, виходячи з її здатності розрізняти схожі та різні зразки. Для цього використовуються принципи пошуку найближчих сусідів, як у методах *kNN*.

Ознаки ранжуються за важливістю, і вибираються найбільш значущі для класифікації.

Чим більше значення ваги для ознаки – тим вищий її вплив на розрізнення аномального та нормального трафіку.

Для ілюстрації використання *ReliefF* застосуємо його до мережевого трафіку з *NSL-KDD*, щоб вибрати найкращі ознаки для подальшого аналізу. *NSL-KDD* – це вдосконалена версія класичного набору даних *KDD Cup* 1999, призначена для тестування та навчання моделей виявлення мережевих вторгнень (*Intrusion Detection Systems, IDS*). *NSL-KDD* використовується в дослідженнях машинного навчання та кібербезпеки для розробки методів виявлення аномалій у мережевому трафіку.

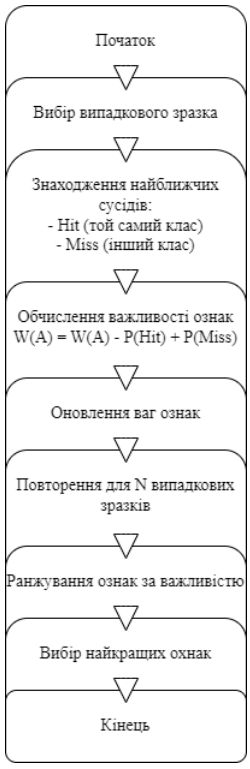

Рис. 1. Принцип роботи ReliefF

Аномальний трафік у *NSL-KDD* поділяється на чотири основні категорії атак:

- *DoS* (*Denial of Service*) – атаки, спрямовані на перевантаження ресурсів, наприклад *Smurf, Neptune, Teardrop*.
- *Probe* (*Scanning attacks*) – спроби збору інформації про систему, наприклад *Satan, Nmap, Ipsweep*.
- *R2L* (*Remote-to-Local*) – віддалені атаки для отримання доступу до системи, наприклад *Guess_Password, Warezclient*.

- *U2R* (*User-to-Root*) – отримання адміністративного доступу, наприклад *Rootkit, Buffer Overflow*.

Групи характеристик у *NSL-KDD*:

1. Базові характеристики трафіку:

- *duration* – тривалість з’єднання
- *src_bytes* – обсяг вихідного трафіку
- *dst_bytes* – обсяг вхідного трафіку

2. Характеристики, пов’язані з *TCP/IP*:

- *protocol_type* – тип протоколу (*TCP, UDP, ICMP*)
- *service* – тип сервісу (*HTTP, FTP, SMTP*)
- *flag* – статус *TCP*-з'єднання

3. Кількісні характеристики зв’язків:

- *count* – кількість з’єднань за останній час
- *same_srv_rate* – відсоток з'єднань із тим самим сервісом
- *dst_host_count* – кількість з'єднань із певним хостом

4. Спеціальні характеристики, що свідчать про атаку:

- *is_guest_login* – ознака гостьового входу
- *serror_rate* – відсоток з’єднань із помилкою сервісу
- *srv_serror_rate* – помилки сервісу у поточній сесії

У нашій роботі *NSL-KDD* використовується для:

- Попередньої обробки мережевого трафіку.
- Застосування *ReliefF* для вибору найважливіших ознак.
- Навчання моделі виявлення аномалій на основі генетичних алгоритмів та класифікаторів.
- Оцінки точності моделі та її здатності виявляти різні типи атак.

Цей датасет є надійним джерелом для тестування інтелектуальних алгоритмів безпеки, що робить його ідеальним для імітаційного моделювання.

Моделювання виявлення аномалій у трафіку на основі ReliefF та оптимізації ознак

Моделювання виявлення аномалій у мережевому трафіку складалося з кількох основних етапів (рис. 2), що включали підготовку даних, вибір ключових ознак і навчання класифікаційної моделі.

Процес розпочався із завантаження та попередньої обробки набору даних *NSL-KDD*, який містить інформацію про нормальні та аномальні запити. Оскільки вихідні дані містили категорійні змінні, такі як тип протоколу, сервіс та статус *TCP*-з'єднання, було проведено їх кодування за допомогою *LabelEncoder* для перетворення в числовий формат. Числові ознаки, що мали різні діапазони значень, були нормалізовані за допомогою *MinMaxScaler*, що забезпечило рівномірне представлення всіх параметрів у моделі. Після цього набір даних був розділений на навчальну (80%) та тестову (20%) вибірки, що дозволило оцінити якість класифікації на незалежних зразках.

Після підготовки даних було проведено навчання базової моделі *Random Forest* без вибору ознак. Цей крок забезпечив еталонні значення продуктивності, з якими в подальшому порівнювали результати моделей з вибором ознак. Далі застосовувався метод *SelectKBest* для відбору найбільш значущих характеристик трафіку.

У процесі моделювання було використано *SelectKBest*, який є одним із методів відбору ознак у машинному навчанні. Його основне призначення – оцінка впливу кожної ознаки на цільову змінну та відбір найбільш релевантних характеристик, що покращують якість класифікації.

Відбір ознак було виконано на основі статистичного критерію *ANOVA F-value* (*f_classif*), який визначає, наскільки значущою є кожна характеристика для розділення класів (нормальний трафік та аномальний). Чим вище значення цього критерію, тим більший внесок ознаки у точність класифікації.

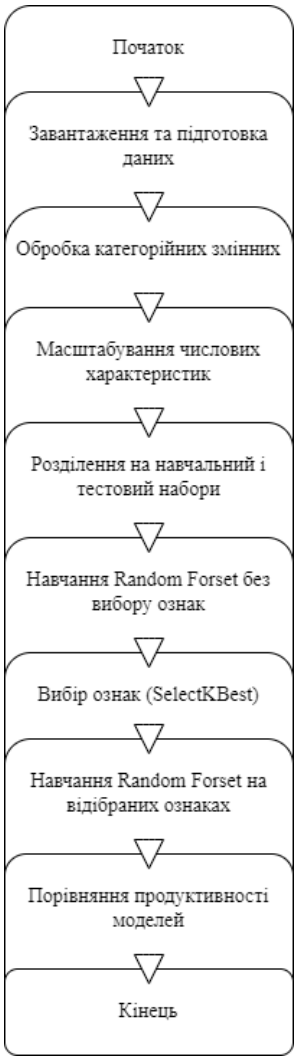

Рис. 2. Основні кроки процесу моделювання виявлення аномалій трафіку на основі ReliefF

На першому етапі було обрано 10 найкращих ознак, що дозволило зменшити розмірність даних і прискорити навчання моделі. Після аналізу було прийнято рішення збільшити кількість відібраних ознак до 20, що дозволило покращити точність без значного збільшення обчислювальних витрат.

Було проведено навчання класифікаційної моделі на відібраних наборах ознак та виконано порівняння продуктивності з повним набором характеристик. Отримані результати продемонстрували, що використання *SelectKBest* дозволяє суттєво скоротити час навчання, при цьому зберігаючи високий рівень точності виявлення аномального трафіку.

Спочатку було відібрано 10 найкращих ознак, що дозволило суттєво зменшити розмірність вхідних даних і прискорити навчання моделі. Проте подальший аналіз показав, що збільшення кількості відібраних ознак до 20 дозволяє досягти кращого балансу між точністю та швидкістю.

Після вибору ключових ознак було повторно проведено навчання моделі *Random Forest*, але вже з використанням обмеженого набору характеристик. Для кожного підходу було виміряно точність класифікації та час навчання. Отримані результати порівнювалися між собою: було виявлено, що застосування *ReliefF* дозволяє суттєво скоротити час навчання, при цьому збереження 20 найважливіших ознак забезпечує продуктивність, близьку до повного набору даних.

На завершальному етапі було виконано порівняння всіх варіантів, і результати були представлені у вигляді графіків. Виявлено, що *ReliefF* у поєднанні з *SelectKBest* дає змогу значно зменшити розмірність моделі, прискорюючи процес аналізу без критичного падіння точності, що є важливим для систем виявлення аномалій у мережевому трафіку.

Діаграма продуктивності моделі (рис. 3) ілюструє вплив вибору ознак на точність класифікації аномального трафіку та час навчання алгоритму. Вона містить три варіанти конфігурацій: без *ReliefF* (повний набір ознак), з *ReliefF* (10 ознак) та з *ReliefF* (20 ознак). Додатково, на діаграмі представлено лінію часу навчання, що дозволяє оцінити ефективність зменшення розмірності ознак.

Перший стовпчик відображає результати моделі без *ReliefF*, яка використовує повний набір характеристик мережевого трафіку для виявлення аномалій. У цьому випадку точність класифікації є найвищою, оскільки модель має доступ до всіх параметрів, включаючи потокові характеристики, рівень використання мережевих ресурсів та взаємодії між вузлами. Однак цей підхід має значний недолік – високе навантаження на обчислювальні ресурси, що може бути неприйнятним у

системах моніторингу мереж у реальному часі.

Другий стовпчик відображає продуктивність моделі з *ReliefF* та 10 відібраними ознаками. Відбір ключових параметрів значно скорочує час навчання, роблячи модель придатною для швидкого аналізу трафіку в динамічних мережах. Проте, зменшення кількості ознак призводить до часткової втрати точності, що може впливати на ефективність виявлення складних атак, таких як *R2L* (віддалений доступ) або *U2R* (підвищення привілеїв).

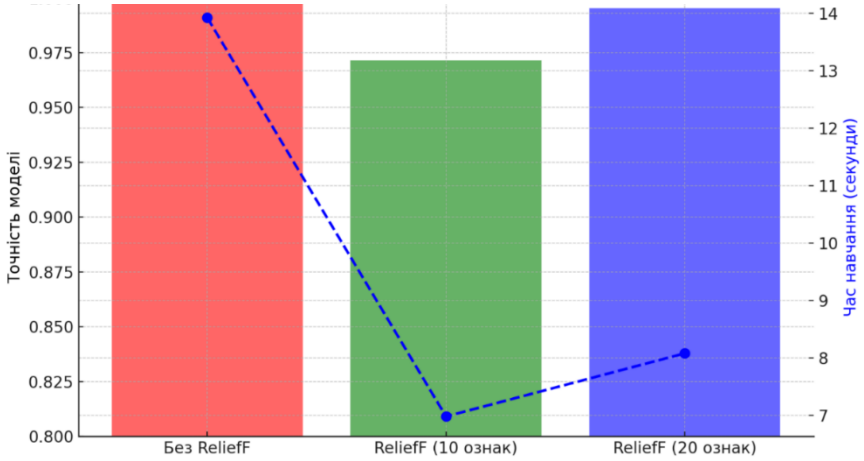

Рис. 3. Діаграма продуктивності моделі з ReliefF і без ReliefF

Третій стовпчик на діаграмі відображає результати моделі з *ReliefF* та 20 відібраними ознаками, що забезпечує оптимальний баланс між продуктивністю та точністю виявлення аномалій. Час навчання залишається значно меншим порівняно з моделлю без *ReliefF*, а точність наближається до вихідного рівня. Завдяки збільшенню кількості відібраних ознак модель краще розпізнає аномалії в мережевому трафіку, включаючи *DoS*-атаки, сканування портів та аномальну поведінку користувачів, що робить її ефективнішою для використання у реальних сценаріях кібербезпеки.

Аналіз діаграми продуктивності підтверджує, що *ReliefF* дозволяє суттєво скоротити час навчання моделі без критичного зниження точності, що є важливим фактором для застосування в реальних умовах. Це робить його ефективним методом оптимізації для систем виявлення аномалій, які працюють у режимі реального часу.

Висновки

У ході дослідження було розглянуто підхід до виявлення аномалій у мережевому трафіку на основі поєднання методу *ReliefF* для вибору ознак та алгоритму *Random Forest* для класифікації. Було проведено моделювання з використанням набору даних *NSL-KDD*, що дозволило оцінити вплив вибору ключових характеристик на продуктивність та ефективність класифікації.

Аналіз отриманих результатів показав, що застосування *ReliefF* дозволяє суттєво зменшити розмірність даних, що призводить до прискорення навчання моделі. При цьому точність класифікації залежить від кількості відібраних ознак. Модель без *ReliefF*, що використовує повний набір характеристик, продемонструвала найвищу точність, проте її навчання потребувало значних обчислювальних ресурсів. Водночас модель з *ReliefF* та 10 відібраними ознаками показала значне скорочення часу навчання, проте зниження точності свідчило про втрату частини корисної інформації. Оптимальним варіантом виявилася модель з *ReliefF* та 20 відібраними ознаками, яка забезпечила компроміс між продуктивністю та точністю виявлення аномалій.

Отримані результати підтвердили ефективність *ReliefF* для аналізу мережевого трафіку, особливо в контексті систем виявлення загроз, що працюють у режимі

реального часу. Скорочення кількості характеристик без значного погіршення якості класифікації дозволяє оптимізувати роботу мережевих *IDS*-систем, забезпечуючи швидке та ефективне виявлення атак, таких як *DDoS*, сканування портів, аномальна поведінка користувачів та вторгнення через експлойти.

Таким чином, *ReliefF* у поєднанні з методами машинного навчання може бути використаний для підвищення ефективності аналізу мережевого трафіку, зменшення обчислювальних витрат та покращення адаптації системи до нових загроз. У подальших дослідженнях доцільним є тестування даного підходу на інших наборах даних, а також використання гібридних методів відбору ознак, що можуть додатково покращити продуктивність системи виявлення аномалій.

Література

1. Singh I. et al. An improved Network Intrusion Detection System using a Novel Hybrid Genetic Coati-Pelican Optimization Algorithms with an Ensemble Fuzzy c-means and CNN-RF. *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* : proceedings, MYSORE, India, 26–27 July 2024 / IEEE. 2024. P. 1–8. DOI: 10.1109/APCIT62007.2024.10673641.

2. Alam M. S., Patel Y., Gurram Y. GMC-168 Hybrid Approach of Data Mining and Deep Learning for Network Intrusion Classification in Big Data. URL: https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1510&context=cday.

3. Jeyakarthic M., Selvakumar T. Hybrid Dynamic Kernel Neural Learning for Efficient Anomaly Detection. *Wireless Sensor Networks*. 2025. Vol. 12(1). P. 106–120. DOI: 10.22247/ijcna/2025/08.

4. Benmalek M., Seddiki A. Particle swarm optimization-enhanced machine learning and deep learning techniques for Internet of Things intrusion detection. *Data Science and Management*. 2025. URL: https://doi.org/10.1016/j.dsm.2025.02.005.

5. Ramezani R. et al. Bench to Bedside: AI and Remote Patient Monitoring. *Frontiers in Digital Health*. 2025. Vol. 7. URL: https://doi.org/10.3389/fdgth.2025.1584443.

6. Ghantous M., ElHarras S., ElAasser M. Use of AI and Intelligent Algorithms. *Lecture Notes in Networks and Systems. Vol. 1268. Intelligent Systems, Blockchain, and Communication Technologies. Selected Papers From the International Conference on Intelligent Systems, Blockchain, and Communication Technologies (ISBCom24) - Volume 1. Conference proceedings* / ed. by A. Abdelgawad, A. Jamil, A. A. Hameed. 2025. P. 454–467.

7. Logeswari G. et al. An improved synergistic dual-layer feature selection algorithm with two type classifier for efficient intrusion detection in IoT environment. *Scientific Reports*. 2025. Vol. 15. 8050. URL: https://doi.org/10.1038/s41598-025-91663-z.

8. Pradeesh S., Jeyakarthic M., Thirumalairaj A. Enhanced Hybrid Approach for Multi-Class DDoS Attack Detection and Classification in Software-Defined Networks Using Remote Sensing and Data Analytics. *Remote Sensing in Earth Systems Sciences*. 2025. DOI: 10.1007/s41976-025-00204-9.

9. Ghimire A. et al. Enhancing Cybersecurity in Critical Infrastructure with LLM-Assisted Explainable IoT Systems. arXiv: 2503.03180. 2025. 5 p.

10. Ansar N. et al. A Robust Hybrid Rf-Bilstm Approach Reinforcing Iot Networks Against Cyber Threats. URL: https://ssrn.com/abstract=5165735.

11. Chaudhary P., Singh A. K., Gupta B. B. Dynamic multiphase DDoS attack identification and mitigation framework to secure SDN-based fog-empowered consumer IoT Networks. *Computers and Electrical Engineering*. 2025. Vol. 123, part C. 110226. DOI: 10.1016/j.compeleceng.2025.110226.

12. Karrothu A., Sriramakrishnan G. V., Jegdic K. Gazelle-Dingo Optimization and Ensemble Classification: A Hybrid Approach for Intrusion Detection in Fog Computing. *Transactions on Emerging Telecommunications Technologies*. 2025. Vol. 36, iss. 3. 70084. DOI: 10.1002/ett.70084.

Кудренко С.О., Столяр А.Л., Макєєв І.Г.

ФУНКЦІОНАЛЬНИЙ ВІДБІР ReliefF ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ ТРАФІКУ

Стаття присвячена проблемі виявлення аномалій у мережевому трафіку з використанням методів машинного навчання та оптимізації вибору ознак. Оскільки сучасні комп'ютерні мережі постійно ускладнюються, а обсяги переданих даних зростають, ефективне виявлення загроз потребує швидких і точних алгоритмів аналізу трафіку. Особливу складність становить відокремлення нормальної мережевої активності від потенційно шкідливих атак, що можуть маскуватися під звичайні запити.

У роботі розглянуто використання методу ReliefF для вибору ключових ознак мережевого трафіку та його вплив на продуктивність класифікаційної моделі Random Forest. Досліджується питання оптимізації вибору характеристик трафіку, що дозволяє зменшити обчислювальні витрати та прискорити процес навчання, не втрачаючи важливих індикаторів аномальної активності. Особлива увага приділена аналізу продуктивності системи за різних обсягів відібраних ознак, що дає змогу знайти оптимальний баланс між точністю та швидкістю аналізу.

Запропонований підхід орієнтований на автоматизовані системи виявлення вторгнень (IDS), які працюють у режимі реального часу та повинні швидко реагувати на потенційні атаки, такі як DDoS, сканування портів, несанкціонований доступ та експлойти. Дослідження спрямоване на підвищення ефективності існуючих методів аналізу мережевого трафіку та адаптацію алгоритмів вибору ознак до умов високонавантажених мереж.

***Ключові слова**: ReliefF; Random Forest; вибір ознак; аномалії трафіку; мережеві атаки; виявлення загроз; NSL-KDD.*

Kudrenko S.O., Stoliar A.L., Makieiev I.G.

ReliefF FUNCTIONAL SELECTION FOR TRAFFIC ANOMALIES DETECTION

This paper addresses the problem of network traffic anomaly detection using machine learning techniques and feature selection optimization. As modern computer networks become increasingly complex and data volumes grow, efficient threat detection requires fast and accurate traffic analysis algorithms. A major challenge lies in distinguishing normal network activity from potentially malicious attacks, which may be disguised as legitimate traffic.

The study explores the use of the ReliefF method for selecting key network traffic features and its impact on the performance of the Random Forest classification model. The focus is on optimizing feature selection to reduce computational costs and speed up training while preserving essential indicators of anomalous activity. Special attention is given to evaluating system performance under different feature selection scenarios, allowing for an optimal trade-off between accuracy and processing speed.

The proposed approach is designed for automated intrusion detection systems (IDS) operating in real-time, requiring rapid responses to potential threats such as DDoS attacks, port scanning, unauthorized access, and exploits. The research aims to enhance the efficiency of existing network traffic analysis methods and adapt feature selection algorithms for high-load network environments.

***Keywords**: ReliefF; Random Forest; feature selection; traffic anomalies; network attacks; threat detection; NSL-KDD.*