МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

МОРОЗ Юрій Павлович

Нейромережеві моделі глибокого навчання для класифікації мережевих пакетів / Deep Learning Neural Network Models for Network Packet Classification

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп'ютерні науки

Кваліфікаційна робота

Виконав студент групи КНм-21
Ю.П. Мороз
_____

Науковий керівник:
к.т.н., професор В.В. Кочан
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 20___ р.
В.о. завідувача кафедри
_____ Н.В. Дзюбановська

ТЕРНОПІЛЬ – 2025

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Н.М. Васильків
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
МОРОЗА Юрія Павловича
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Нейромережеві моделі глибокого навчання для класифікації мережевих пакетів / Deep Learning Neural Network Models for Network Packet Classification
керівник роботи к.т.н., професор В.В. Кочан
затверджені наказом по університету від 20 грудня 2024 року № 938.
2. Строк подання студентом закінченої кваліфікаційної роботи 1 грудня 2025 р.
3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.
4. Основні питання, які потрібно розробити
– провести огляд існуючих підходів до класифікації мережевого трафіку, включаючи статистичні методи, алгоритми машинного навчання та глибокі нейронні мережі;
– сформулювати основні завдання дослідження, визначити вимоги до розроблюваної моделі та критерії її ефективності;
– забезпечити ефективну обробку даних шляхом нормалізації, вибору ключових ознак і їхнього форматування для використання в глибоких мережах;
– створити модель класифікації мережевих пакетів із використанням сучасних архітектур глибокого навчання або їхніх комбінацій;
– описати набір даних, використаний для навчання та тестування моделі, із зазначенням характеристик трафіку, які аналізуються;
– провести серію експериментів для оцінювання ефективності моделі за допомогою обраних метрик;
– порівняти результати розробленого методу з іншими відомими підходами, щоб підтвердити його конкурентоспроможність.
5. Перелік графічного матеріалу у роботі
– схема алгоритму попередньої обробки даних;
– узагальнена схема класифікації мережевих пакетів;
– схема першого підходу: ансамбль нейронної мережі та різних класифікаторів;
– схема першого підходу: ансамбль двонаправленої LSTM та різних

класифікаторів;

– схема третього підходу: ансамбль LSTM, LSTM і Conv1d, LSTM і Conv2D.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 20 грудня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2025 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2025 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 20.05.2025 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 28.10. 2025 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 11.11.2025 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 25.11.2025 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту.	до 1.12.2025 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 04.12.2025 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.12. 2025 р.	

Студент _____ Ю.П. Мороз
підпис

Керівник роботи _____ к.т.н., професор В.В. Кочан
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему «Нейромережеві моделі глибокого навчання для класифікації мережевих пакетів» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 64 сторінки і містить 5 ілюстрацій, 1 таблицю, 1 додаток та 34 використаних джерел.

Метою кваліфікаційної роботи є підвищення безпеки мереж та покращення управління ними шляхом створення надійних і ефективних методів класифікації зашифрованого трафіку з використанням глибоких нейронних мереж.

Методи досліджень: включають аналіз існуючих підходів до класифікації мережевого трафіку, моделювання з використанням глибоких нейронних мереж, експериментальну перевірку ефективності моделей на основі реальних наборів даних, а також порівняння результатів з традиційними методами машинного навчання.

Результати дослідження: полягає у вдосконаленні методу класифікації мережевого трафіку шляхом використання комбінацій глибоких нейронних мереж з класичними алгоритмами машинного навчання. Запропонований підхід дозволяє адаптуватися до змін у мережевих умовах, ідентифікувати зашифрований трафік та забезпечувати високу точність класифікації.

Результати роботи можуть успішно застосовуватися для побудови та вдосконалення інтелектуальних систем моніторингу й класифікації мережевого трафіку, зокрема зашифрованого, у корпоративних та провайдерських мережах. Запропоновані нейромережеві моделі глибокого навчання можуть бути інтегровані в системи кібербезпеки для підвищення точності виявлення підозрілої активності, оптимізації правил міжмережевих екранів та систем запобігання вторгненням.

Ключові слова: ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ МОНІТОРИНГУ; МЕРЕЖЕВИЙ ТРАФІК; ЗАШИФРОВАНИЙ ТРАФІК; ГЛИБОКЕ НАВЧАННЯ; НЕЙРОННІ МЕРЕЖІ; КЛАСИФІКАЦІЯ ТРАФІКУ; КІБЕРБЕЗПЕКА; ВИЯВЛЕННЯ АНОМАЛІ.

ABSTRACT

Qualification work on the topic «Deep Learning Neural Network Models for Network Packet Classification» for Master's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 64 pages and it contains 5 figures, 1 table, 1 annex and 34 sources.

The purpose of this qualification work is to enhance network security and improve network management by creating robust and efficient methods for encrypted traffic classification using deep neural networks.

Research methods include the analysis of existing approaches to network traffic classification, modelling using deep neural networks, experimental evaluation of model effectiveness on real-world datasets, as well as comparison of the results with traditional machine learning methods.

The research results consist in improving the method of network traffic classification through the use of combinations of deep neural networks with classical machine learning algorithms. The proposed approach makes it possible to adapt to changing network conditions, identify encrypted traffic, and ensure high classification accuracy.

The results of this work can be successfully applied to the design and improvement of intelligent systems for monitoring and classifying network traffic, including encrypted traffic, in corporate and provider networks. The proposed deep learning neural network models can be integrated into cybersecurity systems to increase the accuracy of detecting suspicious activity and to optimize firewall rules and intrusion prevention systems.

Keywords: INTELLIGENT MONITORING SYSTEMS; NETWORK TRAFFIC; ENCRYPTED TRAFFIC; DEEP LEARNING; NEURAL NETWORKS; TRAFFIC CLASSIFICATION; CYBERSECURITY; ANOMALY DETECTION.

ЗМІСТ

ВСТУП

Актуальність роботи. Сучасне суспільство постійно розвивається, і зростання його потреб стимулює створення нових технологій. Наприклад, технології Інтернету речей (IoT) [1] і класифікація мережевого трафіку є одними з таких розробок, над якими працюють науковці.

Класифікація та розпізнавання мережевого трафіку відіграє ключову роль у забезпеченні безпеки в Інтернеті. Сьогодні Інтернет розвивається надзвичайно швидко, а разом із ним збільшується кількість різноманітних застосунків і протоколів. Це робить мережевий трафік складнішим та різноманітнішим [2], через що керувати мережами стає дедалі важче. Для вирішення цієї проблеми необхідні новітні технології, які здатні чітко визначати типи трафіку, щоб можна було приймати правильні рішення щодо їх обробки та управління.

На сьогодні було розроблено багато сучасних технологій, які дозволяють класифікувати мережевий трафік з високою точністю та поділяти його на чіткі класи [3].

Однак класифікація мережевого трафіку ускладнилася через появу захищених мереж, які використовують шифрування даних. Зараз такі захищені мережі є на кожній цифровій платформі [4]. Ці функції безпеки створюють надійну основу для захисту систем, але водночас ускладнюють доступ до певної інформації, яка може бути потрібною для налаштування таких сервісів, як брандмауери, контроль доступу чи управління якістю обслуговування (QoS).

Для аналізу зашифрованого трафіку розроблено методи, які не потребують його розшифровки. Наприклад, статистичний аналіз, методи машинного навчання чи аналіз, орієнтований на конкретні протоколи. Вони дозволяють класифікувати зашифровані дані, використовуючи такі параметри, як розмір пакету чи характерні повторення у потоках [4].

Класифікувати зашифрований трафік досить складно, оскільки потрібно дослідити і проаналізувати різні методи. Найпоширенішими підходами є статистичний аналіз, машинне навчання та глибоке навчання. Кожен із цих методів має свої плюси та мінуси.

Статистичний аналіз [5] передбачає вивчення особливостей даних, щоб виявити закономірності, які допоможуть їх класифікувати. Наприклад, різні типи трафіку можуть мати унікальні статистичні властивості, такі як ентропія чи розмір пакета, які можна визначити під час аналізу.

Машинне навчання [6] використовує попередньо підготовлені набори даних, у яких є зразки як незашифрованих, так і зашифрованих даних. Ці дані подаються в алгоритми, які навчаються розпізнавати різницю між ними. Цей підхід точніший за статистичний аналіз, оскільки не потребує інформації про алгоритми шифрування, які використовуються.

Глибоке навчання [7] є ще потужнішим інструментом. Воно базується на використанні нейронних мереж, які можуть навчатися виявляти складні закономірності у даних. Наприклад, методи, що базуються на рекурентних нейронних мережах (LSTM) [8] та згорткових нейронних мережах (CNN) [9], часто застосовують для класифікації зашифрованого трафіку. Ці підходи здатні розпізнавати приховані шаблони навіть у складних наборах даних.

Загалом, сучасні методи допомагають не лише краще зрозуміти, як користувачі взаємодіють в Інтернеті (наприклад, під час перегляду вебсторінок, обміну електронними листами чи потокового відео), а й виявляти шкідливу активність у мережі.

Мета і завдання дослідження. Метою роботи є підвищення безпеки мереж та покращення управління ними шляхом створення надійних і ефективних методів класифікації зашифрованого трафіку з використанням глибоких нейронних мереж.

Для досягнення поставленої мети потрібно виконати ряд завдань:

– провести огляд існуючих підходів до класифікації мережевого трафіку, включаючи статистичні методи, алгоритми машинного навчання та глибокі нейронні мережі;

– сформулювати основні завдання дослідження, визначити вимоги до розроблюваної моделі та критерії її ефективності;

– забезпечити ефективну обробку даних шляхом нормалізації, вибору ключових ознак і їхнього форматування для використання в глибоких мережах;

– створити модель класифікації мережевих пакетів із використанням сучасних архітектур глибокого навчання або їхніх комбінацій;

– описати набір даних, використаний для навчання та тестування моделі, із зазначенням характеристик трафіку, які аналізуються;

– провести серію експериментів для оцінювання ефективності моделі за допомогою обраних метрик;

– порівняти результати розробленого методу з іншими відомими підходами, щоб підтвердити його конкурентоспроможність.

Об'єктом дослідження є процес класифікації мережевого трафіку, що включає аналіз характеристик мережевих пакетів, визначення типів трафіку та ідентифікацію додатків або протоколів, які генерують цей трафік.

Предметом дослідження є методи класифікації мережевого трафіку, зокрема підходи, засновані на глибоких нейронних мережах, таких як CNN, LSTM та їх комбінації, а також використання класичних алгоритмів машинного навчання.

Методи дослідження включають аналіз існуючих підходів до класифікації мережевого трафіку, моделювання з використанням глибоких нейронних мереж, експериментальну перевірку ефективності моделей на основі реальних наборів даних, а також порівняння результатів з традиційними методами машинного навчання.

Наукова новизна одержаних результатів полягає у вдосконаленні методу класифікації мережевого трафіку шляхом використання комбінацій глибоких нейронних мереж з класичними алгоритмами машинного навчання.

Практичне значення отриманих результатів полягає у реалізації ефективного методу класифікації мережевого трафіку, який може бути використаний для підвищення безпеки мереж, моніторингу трафіку, виявлення аномалій та оптимізації управління мережами. Запропонований підхід дозволяє адаптуватися до змін у мережевих умовах, ідентифікувати зашифрований трафік та забезпечувати високу точність класифікації, що робить його корисним для застосування у системах кібербезпеки, корпоративних мережах та інфраструктурах постачальників інтернет-послуг.

Публікації та апробація КР. Результати кваліфікаційної роботи апробовані та опубліковані у матеріалах (додаток А):

– II Всеукраїнської науково-практичної конференції «Інтелектуальні комп'ютерні системи та мережі», 25 листопада 2025 р., Тернопіль, Україна;

– 2nd International Scientific and Practical Conference «Progressive Approaches in Science and Engineering», November 26-28, 2025. Copenhagen, Denmark.

Кваліфікаційна робота складається із вступу, трьох розділів, висновків, списку використаних джерел та додатків.

1 АНАЛІЗ ВІДОМИХ МЕТОДІВ КЛАСИФІКАЦІЇ МЕРЕЖЕВОГО ТРАФІКУ

1.1 Статистичний аналіз для класифікації мережевого трафіку

Статистичний аналіз [5] є одним із ключових підходів для вирішення задач класифікації мережевого трафіку, особливо коли мова йде про зашифровані дані. Цей метод дозволяє аналізувати характерні особливості трафіку, такі як розмір пакетів, часові мітки, частотність передачі або протоколи. На основі цих характеристик можна розпізнавати різні типи додатків, навіть якщо сам вміст пакетів залишається прихованим.

Під час статистичного аналізу особливу увагу приділяють декільком ключовим характеристикам. Однією з них є довжина пакета, яка може змінюватися залежно від типу додатка. Наприклад, для потокового відео пакети часто мають більший розмір, ніж для текстових повідомлень. Іншою важливою характеристикою є часові мітки, які показують інтервали між передачею пакетів. Такі інтервали можуть вказувати на шаблони діяльності, характерні для певних додатків.

Протоколи, які використовуються під час передачі даних, також є важливими для аналізу. Наприклад, TCP забезпечує більш надійну передачу даних, тоді як UDP частіше використовується для реального часу, наприклад у VoIP-додатках. Частотність передачі пакетів, тобто швидкість, з якою передаються дані, може дати додаткову інформацію про активність користувачів.

Для аналізу мережевого трафіку використовують декілька статистичних методів [5]. Одним із них є аналіз розподілів. За допомогою гістограм або оцінки щільності розподілу можна виявити закономірності, такі як середній розмір пакета або час між передачами. Ще одним поширеним методом є оцінка ентропії, яка дозволяє визначати рівень шифрування трафіку. Висока ентропія зазвичай свідчить про те, що дані добре захищені.

Крім цього, застосовуються методи кластеризації, які дозволяють групувати схожі зразки трафіку. Наприклад, за допомогою алгоритмів K-Means

або DBSCAN можна згрупувати трафік, пов’язаний із відеостримінговими сервісами, окремо від трафіку голосових додатків. Для передбачення затримок передачі даних або інших характеристик трафіку часто використовується регресійний аналіз.

Особливу роль у статистичному аналізі відіграє аналіз часових рядів [5]. Завдяки цьому методу можна виявляти циклічні або повторювані шаблони у передачі даних, що характерні для певних додатків або сервісів.

Статистичний аналіз часто використовується як перший етап підготовки даних для машинного навчання. З його допомогою створюються релевантні ознаки, наприклад середній розмір пакета або частотність передачі. Нормалізація статистичних даних дозволяє зробити їх придатними для алгоритмів машинного навчання, що забезпечує більш точні результати. Аналіз кореляції між характеристиками допомагає вибрати найважливіші ознаки, усуваючи зайві дані.

Статистичний аналіз має кілька переваг [5]. Він є простим у реалізації та не потребує доступу до вмісту пакетів, що робить його особливо корисним для роботи із зашифрованим трафіком. Результати статистичного аналізу легко інтерпретувати, що дозволяє швидко робити висновки про типи трафіку.

Однак статистичний аналіз також має свої недоліки [5]. Він може бути чутливим до шуму в даних, що впливає на точність результатів. Крім того, сам по собі статистичний аналіз не завжди дозволяє досягти високої точності, особливо у складних сценаріях. Для отримання найкращих результатів його часто потрібно комбінувати з методами машинного або глибинного навчання.

Статистичний аналіз має широкий спектр застосувань. Наприклад, він може бути використаний для класифікації трафіку відеостримінгових сервісів. За допомогою аналізу середньої довжини пакетів та частотності передачі можна легко ідентифікувати такі сервіси, як YouTube або Netflix. Аналіз часу передачі пакетів дозволяє виявляти VoIP-додатки, такі як Skype чи Zoom. Також, за допомогою оцінки ентропії можна ефективно визначати зашифрований трафік.

Отже, статистичний аналіз є невід'ємною частиною класифікації мережевого трафіку. Він надає цінну інформацію про ключові характеристики даних і забезпечує основу для складніших методів аналізу. Використання

статистичного аналізу у поєднанні з методами машинного навчання дозволяє досягти високої точності, забезпечуючи ефективне розпізнавання типів трафіку навіть у складних умовах. Це робить його важливим інструментом у сучасних системах аналізу мережевої безпеки.

1.2 Машинне навчання для класифікації мережевого трафіку

Машинне навчання стало ключовим інструментом для класифікації мережевого трафіку завдяки своїй здатності аналізувати великі обсяги даних та знаходити приховані закономірності. Сучасні мережі, що обслуговують різноманітні сервіси, генерують великий обсяг трафіку, значна частина якого є зашифрованою. Машинне навчання дозволяє ідентифікувати типи трафіку навіть у таких складних умовах, використовуючи статистичні та поведінкові характеристики даних [6].

Машинне навчання дозволяє ефективно аналізувати трафік на основі таких характеристик, як довжина пакетів, часова частотність передачі, типи протоколів та часові мітки. Ці характеристики слугують основою для моделі, яка може навчитися класифікувати трафік без необхідності доступу до його вмісту. Це особливо важливо для зашифрованого трафіку, де самі дані є недоступними для аналізу.

Моделі машинного навчання використовуються для розпізнавання різних типів додатків, таких як відеостримінгові сервіси, голосові додатки (VoIP), системи передачі файлів або вебперегляд. Такі алгоритми дозволяють не лише автоматизувати процес аналізу, а й суттєво підвищити точність у порівнянні з традиційними методами [6].

1.2.1 Класичні алгоритми машинного навчання [6].

Одним із найпоширеніших підходів є використання класичних методів машинного навчання. Random Forest, наприклад, створює ансамбль дерев рішень, які разом забезпечують високу точність прогнозів навіть у складних умовах. Цей метод добре працює з великими наборами даних та є стійким до шуму.

Іншим поширеним методом є Support Vector Machine. Він намагається знайти оптимальну гіперплощину для розділення класів трафіку. SVM є ефективним для задач, де класи добре розділені, але може бути обмеженим у роботі з великими наборами даних.

Метод K-Nearest Neighbors ґрунтується на пошуку найближчих сусідів у багатовимірному просторі характеристик. Цей підхід є простим, але може бути чутливим до вибору обсягу даних. Logistic Regression, хоча і є базовим методом, часто використовується для швидкого аналізу трафіку та забезпечує базову точність класифікації.

1.2.2 Алгоритми глибокого навчання [7–9].

Глибинне навчання відкриває нові можливості у класифікації мережевого трафіку. Long Short-Term Memory мережі використовуються для аналізу послідовностей даних, що є корисним у виявленні шаблонів передачі даних. Наприклад, цей метод добре працює для розпізнавання голосового трафіку або потокового відео, де важливими є часові залежності.

Convolutional Neural Networks аналізують просторові характеристики трафіку, такі як розмір пакета або напрямок передачі. Поєднання CNN з LSTM дозволяє обробляти як часові, так і просторові залежності, забезпечуючи високий рівень точності.

Першим кроком у використанні машинного навчання для класифікації трафіку є збір і підготовка даних. Це включає виділення релевантних характеристик, таких як середня довжина пакета або частота передачі. Далі проводиться нормалізація даних, що дозволяє зробити їх придатними для моделей машинного навчання.

Після підготовки даних здійснюється навчання моделі на навчальному наборі даних. На цьому етапі важливо враховувати баланс між класами, щоб модель могла точно класифікувати всі типи трафіку. Після цього проводиться оцінка точності моделі за допомогою метрик, таких як точність, повнота, прецизійність і F1-міра.

Машинне навчання забезпечує високу точність класифікації навіть у складних умовах, таких як зашифрований трафік. Воно дозволяє автоматизувати

процес аналізу, знижуючи навантаження на фахівців з безпеки. Крім того, моделі машинного навчання можуть адаптуватися до змін умов мережі або появи нових додатків.

Незважаючи на переваги, машинне навчання має і певні виклики. Одним із головних недоліків є потреба у великих наборах даних для навчання. Також складнощі можуть виникати через дисбаланс між класами, коли певні типи трафіку представлені значно меншою кількістю зразків. Крім того, моделі глибинного навчання вимагають значних обчислювальних ресурсів, що може бути обмеженням для деяких систем.

Отже, машинне навчання є потужним інструментом для класифікації мережевого трафіку, особливо зашифрованого. Використання сучасних алгоритмів дозволяє досягти високої точності та автоматизувати процес аналізу. У майбутньому подальший розвиток цієї галузі дозволить створювати ще більш точні, адаптивні та ефективні моделі, які допоможуть забезпечити безпеку та стабільність роботи мереж.

1.3 Глибоке навчання для класифікації мережного трафіку

Глибоке навчання є одним із найбільш сучасних підходів до класифікації мережевого трафіку. Воно дозволяє автоматизувати та значно оптимізувати процес аналізу даних, використовуючи потужні можливості нейронних мереж для вивчення складних закономірностей. Цей підхід стає особливо актуальним в умовах зростання обсягів мережевого трафіку, більша частина якого є зашифрованою, а також у ситуаціях, коли потрібно аналізувати дані в режимі реального часу [7–9].

Глибоке навчання дозволяє моделі самостійно навчитися розпізнавати основні характеристики мережевого трафіку, такі як розмір пакетів, частота їхньої передачі, часові залежності та особливості протоколів. Ці характеристики є важливими для визначення типів трафіку навіть у випадках, коли вміст пакетів є зашифрованим.

Глибокі нейронні мережі застосовуються для вирішення широкого спектра завдань, зокрема:

– класифікація додатків, таких як потокове відео, вебперегляд чи голосові сервіси;

– виявлення аномалій, включаючи атаки типу DoS чи несанкціонований доступ;

– аналіз зашифрованого трафіку для визначення протоколів або типів переданих даних.

Головною перевагою глибокого навчання є його здатність виявляти приховані залежності та шаблони навіть у великих і складних наборах даних.

Глибоке навчання пропонує кілька архітектур [7–9], кожна з яких має свої переваги для класифікації мережевого трафіку.

1. Багатошарові нейронні мережі (MLP).

Однією з найпростіших архітектур є багатошарові перцептрони (MLP). Ці моделі використовуються для класифікації даних із чітко визначеними характеристиками. Однак вони обмежені в аналізі складних залежностей, таких як часові чи просторові зв'язки в мережевому трафіку.

2. Рекурентні нейронні мережі (RNN) та Long Short-Term Memory (LSTM).

Рекурентні нейронні мережі, зокрема їхня модифікація LSTM, є ідеальними для обробки послідовностей даних. У контексті мережевого трафіку ці моделі дозволяють аналізувати часові шаблони, такі як затримки між передачами пакетів або послідовність дій користувачів. Наприклад, LSTM може ефективно класифікувати голосовий трафік або потокове відео, враховуючи залежність між подіями в часі.

3. Згорткові нейронні мережі (CNN).

CNN спеціалізуються на аналізі просторових характеристик даних. У класифікації трафіку вони використовуються для вивчення таких параметрів, як розмір пакета, напрямок передачі чи особливості протоколів. Ці мережі демонструють високу ефективність у класифікації потокового відео та виявленні VPN-трафіку.

4. Поєднання CNN та LSTM.

Поєднання CNN та LSTM дозволяє одночасно враховувати просторові й часові характеристики. Це робить такий підхід ефективним для аналізу складного трафіку, наприклад, коли потрібно враховувати і розмір пакетів, і частоту їхньої передачі.

5. Трансформери.

Трансформери є сучасною архітектурою глибокого навчання, яка використовується для обробки великих наборів даних. Вони добре підходять для паралельного аналізу залежностей у даних, що робить їх перспективними для класифікації мережевого трафіку.

Глибоке навчання передбачає кілька послідовних етапів, кожен із яких є важливим для досягнення точності класифікації [7–9].

1. Збір і підготовка даних.

Перед тренуванням моделей необхідно зібрати великий обсяг даних, який має містити різноманітні типи трафіку. Важливо виконати попереднє очищення даних, нормалізувати значення характеристик і визначити ключові параметри, такі як довжина пакета або частота передачі.

2. Розробка моделі.

Вибір архітектури залежить від задачі. Наприклад, для аналізу часових залежностей вибираються рекурентні мережі, тоді як для просторових характеристик краще використовувати CNN.

3. Навчання та тестування.

Після розробки моделі її тренують на навчальному наборі даних, оптимізуючи параметри для досягнення найкращої точності. Потім модель тестують на нових даних, оцінюючи її продуктивність за метриками, такими як точність, повнота та F1-міра.

Глибоке навчання має кілька ключових переваг. Воно дозволяє автоматично вивчати важливі ознаки з даних, що знижує необхідність у ручному створенні характеристик. Моделі є адаптивними та здатні підлаштовуватися до змін у мережевих умовах. Крім того, нейронні мережі є стійкими до шуму, що забезпечує стабільність результатів.

Незважаючи на переваги, глибоке навчання має певні виклики. Воно вимагає значних обчислювальних ресурсів і великих обсягів даних для навчання. Крім того, моделі глибокого навчання часто складно інтерпретувати, що ускладнює діагностику помилок.

Глибоке навчання вже застосовується у багатьох сферах класифікації трафіку. Наприклад, CNN використовуються для виявлення VPN-трафіку, а LSTM – для аналізу голосових додатків, таких як Skype або Zoom. Поєднання CNN та LSTM ефективне для класифікації потокового відео на таких платформах, як Netflix чи YouTube.

У роботі [12] К. Алі, А. Тарік і Х. Аббас дослідили різні методи глибокого навчання для класифікації зашифрованого мережевого трафіку. Вони також запропонували нову модель, побудовану на базі CNN, яка досягла точності 97% на тестових даних. Цей підхід став одним із прикладів ефективного використання сучасних методів для вирішення складних завдань.

Лу та ін. [13] поєднали двонаправлену модель LSTM (Bidirectional LSTM) із методом випадкового лісу (Random Forest). Таке поєднання дало змогу досягти точності класифікації 96,7%.

У роботі Zhou та ін. [14] було запропоновано поєднати двонаправлену LSTM із методом найближчих сусідів (KNN). Ця комбінація показала трохи нижчу точність – 96,4%, але все одно залишається потужним підходом для класифікації.

Дослідження Ван та ін. [15] включало використання LSTM разом із двовимірною згортковою нейронною мережею (CNN2D). Результат цього підходу – точність 92,2%, що демонструє хорошу здатність цих методів аналізувати складні типи трафіку.

С. Алі, М. Арфін і С. Рехман [16] проаналізували, як традиційні методи машинного навчання справляються із завданням класифікації зашифрованого трафіку. Вони протестували різні моделі, серед яких були дерева рішень, випадкові ліси (Random Forest) і метод опорних векторів (SVM). SVM виявився найефективнішим, досягнувши точності 99%.

У роботі [17] автори А. Аль-Раві, А. Аль-Сафар і М. Аль-Адамі запропонували змішаний підхід, який поєднує аналіз статистичних характеристик трафіку з методами машинного навчання. Вони досягли точності 98%, що є значним результатом для такого підходу.

Дослідники М. Амін, Х. Лю і А. Аль-Делаан [18] зосередилися на використанні глибинного навчання. Вони порівняли кілька моделей, таких як CNN і LSTM, і дійшли висновку, що LSTM забезпечує найкращу точність – 99,5%.

Автори Ф. Хусейн, М. Ісмаїл і А. Хан [19] також проаналізували традиційні методи машинного навчання, порівнюючи дерева рішень, метод K-ближчих сусідів (KNN) і SVM. У їхньому дослідженні SVM також показав кращі результати, досягнувши точності 97%.

Zhaolei та ін. [20] розробили модель під назвою Byte Feature Convolution Network (BFCN), яка поєднує підхід BERT для аналізу текстових характеристик і CNN для виділення особливостей. Ця модель аналізує як глобальні патерни в трафіку, так і деталі на рівні байтів. Використовуючи тестовий набір ISCX, вони досягли високих показників точності, повноти та значення F1.

У роботі Сіньї Ху та ін. [21] запропоновано модель CLD Net, яка об’єднує CNN і LSTM для аналізу складного трафіку, включаючи VPN і Skype. Тестування на наборі ISCX показало точність 98% для VPN і 92% для Skype. Цей підхід демонструє можливість точно класифікувати навіть маловідомі типи трафіку.

Дослідження Вей Ван та ін. [22] представило метод, що базується на одновимірних згорткових нейронних мережах (1D CNN). Ця модель поєднує виділення ознак, аналіз і класифікацію в єдину систему. Вона була протестована на наборі ISCX із трафіком VPN і не-VPN, включаючи різні протоколи (email, чат, потокове відео тощо). Результати показали високу точність із суттєвим зменшенням кількості хибнопозитивних результатів.

Автори Кун Чжоу та ін. [23] запропонували поєднання оцінки ентропії з нейронними мережами для аналізу та класифікації трафіку. Їхній підхід перевершив традиційні методи, що було підтверджено на тестовому наборі даних Канадського інституту кібербезпеки.

Лулу Го та ін. [24] розробили дві моделі для класифікації VPN і не-VPN трафіку. Перша модель, згортковий автоенкодер (CAE), досягла точності 99,8%, тоді як друга, згорткова нейронна мережа (CNN), показала точність 92,92%.

Мадуші Патмаперума та ін. [25] створили модель CNN для аналізу активностей користувачів у мобільних додатках, таких як YouTube, Facebook, Messenger і Gmail. Цей підхід розділяє зашифрований трафік на часові сегменти, що дозволяє точно аналізувати активності користувачів. Після фільтрації невідомого трафіку точність моделі зросла до 92%. Ця розробка показала ефективність у розпізнаванні активностей навіть у складних умовах.

Отже, глибоке навчання є потужним інструментом для класифікації мережевого трафіку. Його здатність автоматично вивчати складні закономірності робить цей підхід ідеальним для вирішення сучасних викликів, таких як зашифрований трафік чи великі обсяги даних. Подальший розвиток глибокого навчання дозволить створювати ще більш точні та ефективні моделі, що допоможуть забезпечити стабільну роботу та безпеку сучасних мереж.

1.4 Постановка задачі дослідження

У сучасному світі обсяг мережевого трафіку стрімко зростає через зростання популярності онлайн-сервісів, таких як відеостримінгові платформи, голосові додатки, електронна комерція та інші цифрові сервіси. Значна частина цього трафіку є зашифрованою, що створює серйозні виклики для його аналізу. Традиційні підходи до класифікації мережевого трафіку, зокрема методи, засновані на ручному аналізі або базових алгоритмах машинного навчання, не завжди забезпечують необхідну точність, особливо у випадках зі складними типами трафіку. У зв’язку з цим виникає потреба у використанні сучасних методів глибокого навчання, які здатні автоматично вивчати закономірності у даних і забезпечувати високу точність класифікації.

Ця робота спрямована на вирішення задачі класифікації мережевого трафіку, зокрема зашифрованого, шляхом розробки ефективних моделей глибокого навчання. Основною метою є створення інструментів, які зможуть не

лише класифікувати трафік із високою точністю, але й адаптуватися до змін у мережевих умовах, включаючи нові типи додатків та протоколів.

Для досягнення цієї мети передбачено вирішення кількох ключових завдань:

1. Аналіз проблем класифікації мережевого трафіку. Першим етапом роботи є глибокий аналіз існуючих підходів до класифікації мережевого трафіку. Необхідно оцінити переваги та обмеження традиційних методів, таких як машинне навчання, та визначити, у яких випадках ці методи виявляються неефективними. Окрему увагу слід приділити аналізу зашифрованого трафіку, оскільки доступ до його вмісту є обмеженим.

2. Вибір архітектур глибокого навчання. Наступним завданням є обґрунтування вибору моделей глибокого навчання, таких як згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN) та Long Short-Term Memory (LSTM). Також важливо дослідити їх комбінації, які можуть ефективно працювати з даними, що мають як часові, так і просторові залежності.

3. Підготовка даних для навчання. Ефективність моделей значною мірою залежить від якості даних. Для цього необхідно зібрати великий набір даних, який містить різноманітні типи трафіку, включаючи потокове відео, голосові додатки, вебперегляд тощо. Дані мають бути очищені від шуму, нормалізовані, а також збалансовані між класами.

4. Розробка моделей глибокого навчання. На основі зібраних даних необхідно розробити моделі глибокого навчання, які зможуть автоматично витягувати важливі ознаки, знижуючи потребу в ручному аналізі. Основна увага буде приділена поєднанню CNN для аналізу просторових характеристик і LSTM для обробки часових залежностей.

5. Оцінка ефективності моделей. Після розробки моделей важливо провести їх експериментальну оцінку. Для цього будуть використовуватися метрики, такі як точність, повнота, прецизійність і F1-міра. Оцінка має підтвердити, що запропоновані моделі досягають високої точності класифікації, зокрема для зашифрованого трафіку.

6. Порівняння з традиційними методами. Щоб оцінити переваги глибокого навчання, необхідно порівняти його результати з результатами традиційних алгоритмів, таких як Random Forest, Support Vector Machine і K-Nearest Neighbors. Це допоможе зрозуміти, наскільки глибоке навчання є більш ефективним у контексті сучасних мереж.

Результатом цього дослідження стане створення моделей глибокого навчання, які забезпечать високу точність класифікації для різноманітних типів мережевого трафіку, включаючи зашифрований. Ці моделі будуть здатні автоматично витягувати ознаки з даних, демонструючи адаптивність до змін у мережевих умовах. Порівняння з традиційними методами підтвердить переваги глибокого навчання у складних сценаріях.

Також передбачається, що результати роботи сприятимуть вдосконаленню систем аналізу та управління мережами, підвищуючи рівень їхньої безпеки та ефективності. Розроблені моделі можна буде застосовувати у сфері кібербезпеки, управління мережею, а також для моніторингу та оптимізації трафіку у великих корпоративних інфраструктурах.

Висновки до розділу 1

1. Проведено комплексний огляд підходів до класифікації мережевого трафіку, що охоплює статистичний аналіз, методи машинного навчання та сучасні техніки глибокого навчання. Аналіз кожного підходу дозволив виявити їхні переваги, недоліки та сферу застосування в контексті розв'язання задач класифікації трафіку, зокрема зашифрованого. Статистичний аналіз забезпечує основу для базового розуміння структури трафіку за допомогою таких характеристик, як розмір пакетів, часові інтервали, частота передачі та типи протоколів. Ці методи є простими у реалізації, але їхня ефективність знижується у випадках зі складними даними або великими обсягами трафіку. Статистичні методи переважно використовуються як підготовчий етап для більш складних підходів.

2. Методи машинного навчання, такі як Random Forest, Support Vector Machine, K-Nearest Neighbors та Logistic Regression, забезпечують вищу точність класифікації порівняно зі статистичними підходами. Вони ефективно працюють з структурованими наборами даних, але їхня продуктивність знижується при аналізі зашифрованого трафіку або складних залежностей між характеристиками.

3. Глибоке навчання відкриває нові можливості для автоматизації та оптимізації процесу класифікації. Використання згорткових нейронних мереж (CNN) для аналізу просторових залежностей і рекурентних нейронних мереж (LSTM) для обробки часових шаблонів дозволяє досягати високої точності навіть для складних та зашифрованих даних. Поєднання CNN і LSTM, а також використання сучасних архітектур, таких як трансформери, демонструє перспективність глибокого навчання для мережевого аналізу.

4. На основі проведеного огляду було сформульовано постановку задачі дослідження, яка спрямована на розробку ефективних моделей класифікації мережевого трафіку з використанням глибокого навчання. Ці моделі повинні забезпечити високу точність, адаптивність до змін у мережевих умовах і можливість аналізу зашифрованого трафіку.

2 МЕТОД КЛАСИФІКАЦІЇ МЕРЕЖЕВИХ ПАКЕТІВ НА ОСНОВІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

2.1 Основні етапи процесу класифікації мережевих пакетів

Класифікація зашифрованого трафіку – це важливий процес, який допомагає захищати мережі від потенційних загроз і дозволяє краще розуміти, як використовуються ресурси мережі. Аналіз зашифрованого трафіку дає змогу організаціям визначати, які додатки чи сервіси працюють у мережі. Це допомагає швидко виявляти проблеми, такі як несанкціонований доступ, та вживати заходів, наприклад, блокувати шкідливу активність.

Процес класифікації починається з збору зашифрованих пакетів. Для цього використовують спеціальні інструменти, такі як Wireshark, які дозволяють «захоплювати» мережеві пакети під час їх передачі через мережу. Ці пакети – це шматочки інформації, які передаються між комп'ютерами, і вони містять дані про тип трафіку.

Після того як пакети захоплені, їх аналізують, щоб визначити, які протоколи чи порти використовуються. Наприклад, якщо виявляється порт 443, це може вказувати на використання HTTPS, що є стандартом для захищеного вебтрафіку. Аналогічно, порт 22 свідчить про використання протоколу SSH для безпечного з’єднання. Визначення протоколів дозволяє зрозуміти, які саме сервіси працюють у мережі.

Далі використовуються різні підходи для аналізу цих пакетів:

– глибока перевірка пакетів (Deep Packet Inspection): розглядається вміст пакету для виявлення додаткових характеристик, навіть якщо він зашифрований;

– статистичний аналіз: аналізуються статистичні особливості трафіку, наприклад, розмір пакету, частота передачі чи затримка;

– машинне навчання: алгоритми вивчають закономірності у даних і автоматично класифікують типи трафіку;

– сигнатурний аналіз (Signature-based Matching): пакети порівнюються зі вже відомими шаблонами, щоб визначити, до якого типу вони належать.

Результати цього процесу потім можна використовувати для різних цілей. Наприклад, мережеві адміністратори можуть аналізувати статистику використання мережі або автоматично блокувати шкідливу активність, якщо трафік виявлено як небезпечний. Таким чином, класифікація зашифрованого трафіку допомагає забезпечувати як безпеку, так і ефективність роботи мережі.

2.2 Попередня обробка даних

Попередня обробка даних є важливим кроком у підготовці будь-якого набору даних для аналізу чи навчання моделей машинного навчання. Її мета – впорядкувати дані, зробити їх зрозумілими для алгоритмів і забезпечити точність результатів. Якщо набір даних погано підготовлений, це може призвести до помилкових висновків.

У випадку аналізу зашифрованого трафіку процес попередньої обробки починається з розділення даних на дві частини:

1. Частина пакетів – це зашифровані пакети, які передаються через мережу.
2. Частина інформації – це метадані про пакети, такі як IP-адреси джерела та призначення, тип протоколу (наприклад, UDP чи TCP) тощо.

Конвертація даних – це наступний крок. Пакети зазвичай записані у шістнадцятковому форматі (HEX), який зручний для передачі даних у мережах, але складний для аналізу. Тому їх перетворюють у десятковий формат, щоб алгоритми могли легко їх обробляти.

Додатково аналізуються метадані:

– IP-адреси джерела та призначення перетворюються з текстового формату (наприклад, "192.168.0.1") на цілі числа. Це робиться для спрощення обчислень і підвищення ефективності алгоритмів;

– тип протоколу (TCP чи UDP) класифікується, щоб алгоритми могли враховувати його вплив на передачу даних.

Після цього всі дані, включаючи пакети та метадані, об’єднуються в один набір для аналізу. Також до даних додаються мітки класів, які вказують на типи

трафіку (наприклад, вебтрафік, потокове відео, електронна пошта тощо). Це робиться для навчання моделей машинного навчання класифікувати трафік за заданими категоріями.

На рисунку 2.1 показано, як відбувається цей процес попередньої обробки даних. Об'єднаний набір стає готовим до передачі в алгоритми машинного навчання для подальшого аналізу.

Рисунок 2.1 – Схема узагальненого алгоритму попередньої обробки даних

Попередня обробка даних є критично важливим етапом у будь-якому процесі машинного навчання. Вона дозволяє не лише впорядкувати та підготувати дані, але й усунути можливі проблеми, пов’язані з неправильним форматуванням чи неузгодженістю у вихідних даних. Завдяки правильній

обробці даних можна досягти точних і надійних результатів, що допоможе глибше зрозуміти характеристики трафіку та забезпечити ефективність роботи мережі.

2.3 Модель класифікації мережевих пакетів

У цьому підрозділі розглядаються запропоновані методи, які розроблені для забезпечення максимальної ефективності. Модель складається з трьох окремих підходів, кожен із яких включає кілька етапів. На рисунку 2.2 представлено блок-схему запропонованої моделі, яка описує всі три підходи.

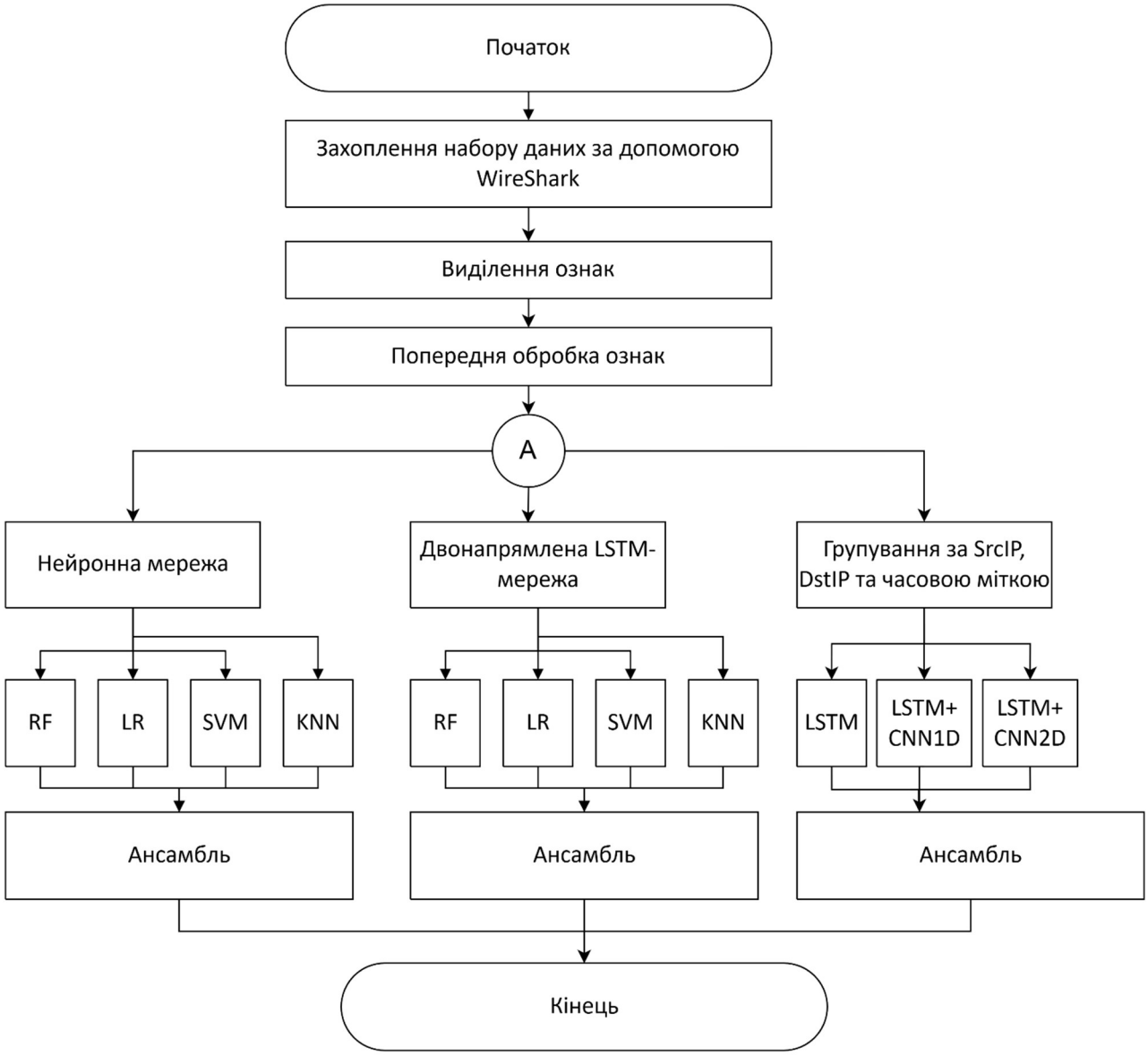


Рисунок 2.2 – Узагальнена схема класифікації мережевих пакетів

2.3.1 Перший підхід

Цей підхід призначений для ефективної підготовки даних перед їх подачею до моделі машинного навчання. Він складається з кількох ключових етапів.

1. Перемішування даних. На початку таблиця даних перемішується. Це робиться для того, щоб уникнути впливу порядку записів на процес навчання. Наприклад, якщо в таблиці записи розташовані за певною послідовністю, це може завадити моделі вивчати загальні закономірності, тому порядок записів змінюється випадковим чином.

2. Розділення даних на частини. Таблиця даних ділиться на дві частини:

– частина зашифрованих пакетів, яка містить основну інформацію про мережевий трафік;

– інформаційна частина, яка включає додаткові метадані, наприклад, IP-адреси, тип протоколу, порт тощо.

3. Подача даних у нейронну мережу. Частина зашифрованих пакетів подається до нейронної мережі. Ця мережа складається з чотирьох шарів:

– вхідний шар для отримання даних;

– два приховані шари, які виконують основну обробку та виділення ознак;

– вихідний шар із чотирма класами, що визначають типи трафіку.

На цьому етапі зберігається останній прихований шар перед вихідним. Цей шар містить найважливішу інформацію, яку мережа "вивчила" під час обробки.

4. Об'єднання даних. Отриманий шар з нейронної мережі об’єднується з інформаційною частиною даних, щоб створити єдиний набір, який містить як ознаки, виділені мережею, так і метадані.

5. Класифікація. Об'єднаний набір подається на чотири різні алгоритми класифікації:

– Random Forest (RF) – потужний метод, який використовує багато дерев рішень для досягнення точних прогнозів;

– Логістична регресія (LR) – модель, яка добре працює з бінарною або багатокласовою класифікацією;

– K-ближні сусіди (KNN) – простий алгоритм, що класифікує дані на основі найближчих сусідніх точок;

– Метод опорних векторів (SVM) – підхід, який шукає оптимальну межу між класами.

6. Енсамблеве голосування. На завершальному етапі результати всіх чотирьох моделей об'єднуються методом голосування. Це означає, що кожна модель "голосує" за свій результат, а кінцевий прогноз формується на основі більшості голосів.

На рисунку 2.3 показано, як поєднуються нейронна мережа та різні алгоритми класифікації в першому підході.

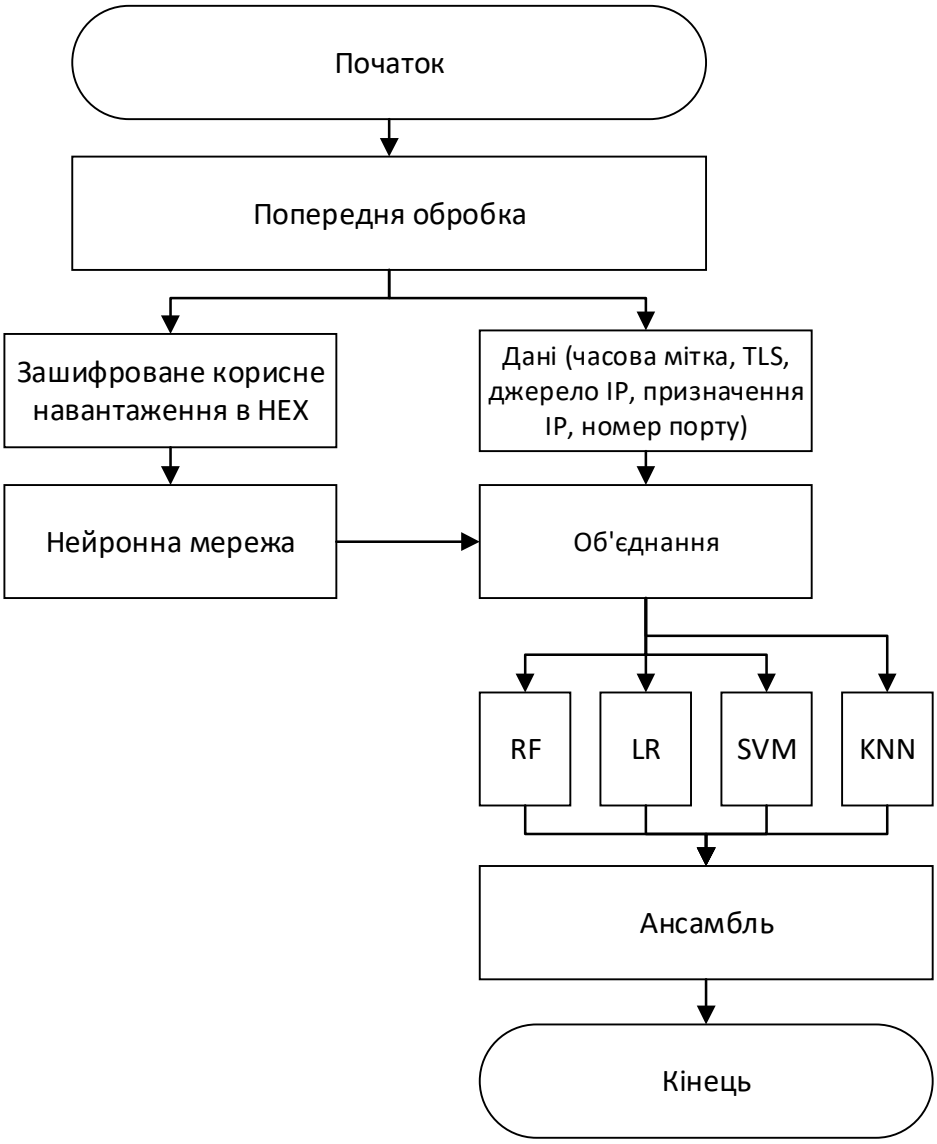


Рисунок 2.3 – Схема першого підходу: ансамбль нейронної мережі та різних класифікаторів

2.3.2 Другий підхід

Другий підхід схожий на перший, але має важливу відмінність: замість звичайної нейронної мережі використовується двонаправлена LSTM (Bidirectional LSTM). Ця модель більш підходить для роботи з послідовними даними, такими як мережевий трафік, оскільки вона зберігає інформацію про минулі та майбутні події одночасно.

1. Використання двонаправленої LSTM:

– Bidirectional LSTM аналізує дані в обох напрямках – з минулого до майбутнього та навпаки, що дозволяє їй краще виявляти закономірності в послідовних даних;

– як і в першому підході, останній прихований шар перед виходом зберігається, він містить ключову інформацію про те, що модель "зрозуміла" під час навчання.

2. Об'єднання даних: виділений шар із LSTM об’єднується з інформаційною частиною даних, створюючи повний набір для подальшої класифікації.

3. Класифікація: як і в першому варіанті, цей набір подається до чотирьох алгоритмів класифікації: Random Forest, Логістична регресія, K-найближні сусіди та SVM.

4. Енсамблеве голосування: результати всіх чотирьох моделей об’єднуються методом голосування, щоб отримати кінцевий прогноз.

Цей підхід особливо ефективний для великих наборів даних або складних задач. Використання LSTM дозволяє краще запам'ятовувати попередні дані та виявляти довгострокові залежності, що робить прогнози більш точними.

На рисунку 2.4 показано структуру другого підходу, що поєднує LSTM із чотирма алгоритмами класифікації.

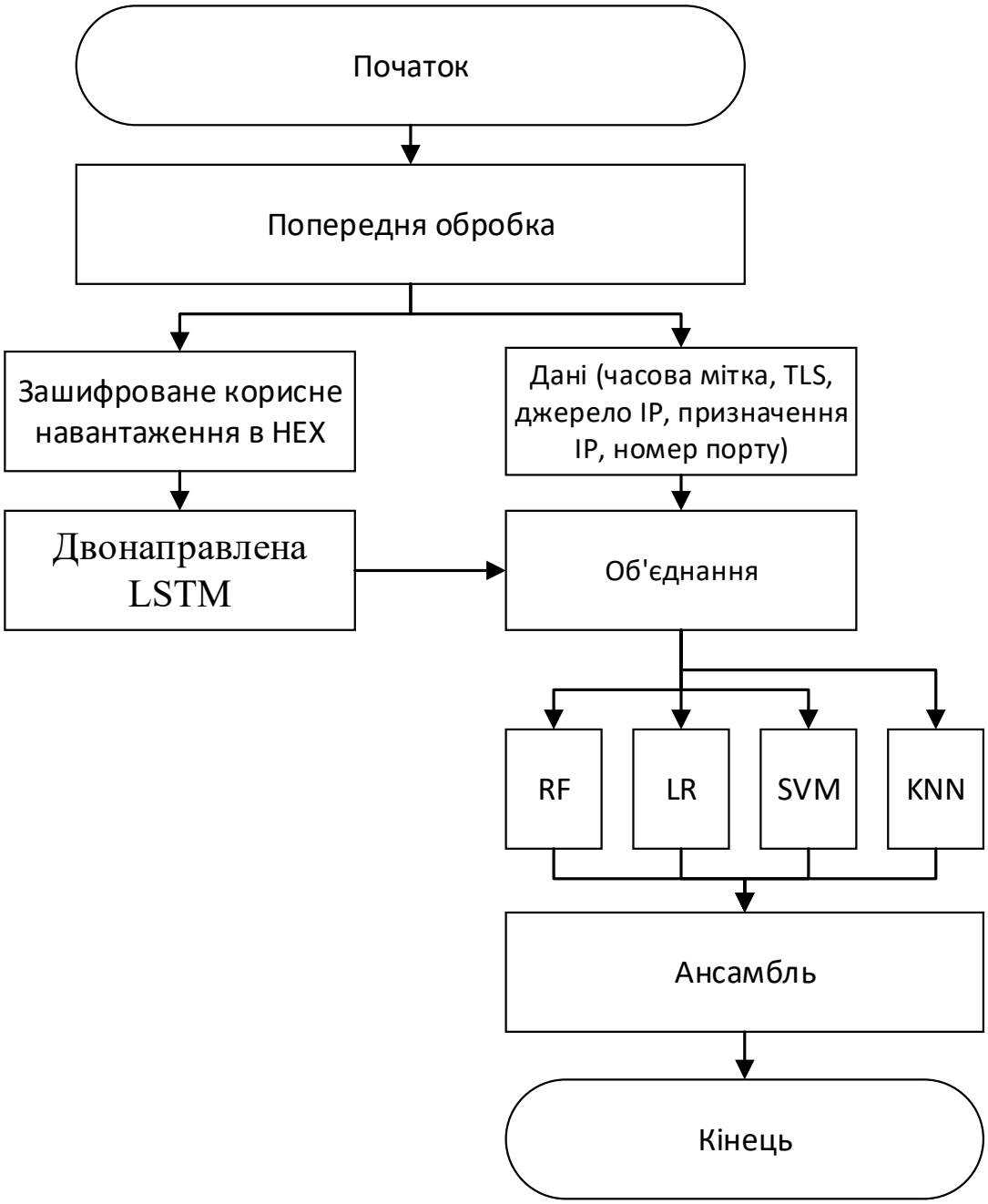


Рисунок 2.4 – Схема першого підходу: ансамбль двонаправленої LSTM та різних класифікаторів

2.3.3 Третій підхід

У третьому підході дані обробляються дещо інакше, порівняно з попередніми підходами. Цей метод зосереджений на використанні TCP-записів, оскільки вони вважаються більш надійними і містять такі важливі характеристики, як часові мітки, що вказують на початок і завершення кожної сесії.

1. Вибір і групування даних. Використовуються тільки записи TCP, без UDP-записів, через їх вищу надійність. Дані (пакети і метаінформація) групуються за такими ознаками:

– IP-адреса джерела,
– IP-адреса призначення,
– Часова мітка,
– Тип класу (class type).

Для кожної групи пакетів розраховуються статистичні параметри:

– Максимальне значення (max),
– Мінімальне значення (min),
– Середнє значення та стандартне відхилення (SD).

2. Об’єднання даних. Згрупована інформація (максимуми, мінімуми, середнє і стандартне відхилення) об'єднується із зашифрованими пакетами. Ці об’єднані дані стають готовими до подачі в моделі машинного навчання.

3. Моделі машинного навчання. Об'єднаний набір даних подається до трьох типів нейронних мереж:

– Bidirectional LSTM з двома шарами: використовується для аналізу послідовних даних і виявлення закономірностей як з минулого, так і з майбутнього.

– CNN 1D + LSTM: одно-вимірна згорткова нейронна мережа (CNN 1D) для виділення локальних особливостей, яка далі обробляється рекурентною мережею LSTM для роботи з послідовностями.

– CNN 2D + LSTM: двовимірна згорткова нейронна мережа (CNN 2D), яка використовується для аналізу складніших шаблонів, також об’єднана з LSTM для обробки часових послідовностей.

4. Підготовка даних для навчання. Перед подачею в мережі застосовується підгонка пакетів до однакового розміру (padding), щоб усі групи мали однакову форму для коректного навчання. Дані також перемішуються, щоб забезпечити випадковість і запобігти перенавчанню моделей.

5. Енсамблеве голосування. Результати всіх трьох нейронних мереж об'єднуються за допомогою методу голосування ансамблю. Це дозволяє

створити єдиний кінцевий прогноз, який враховує результати всіх трьох моделей.

На рисунку 2.5 представлено схему роботи третього підходу, яка поєднує Bidirectional LSTM, CNN 1D + LSTM і CNN 2D + LSTM.

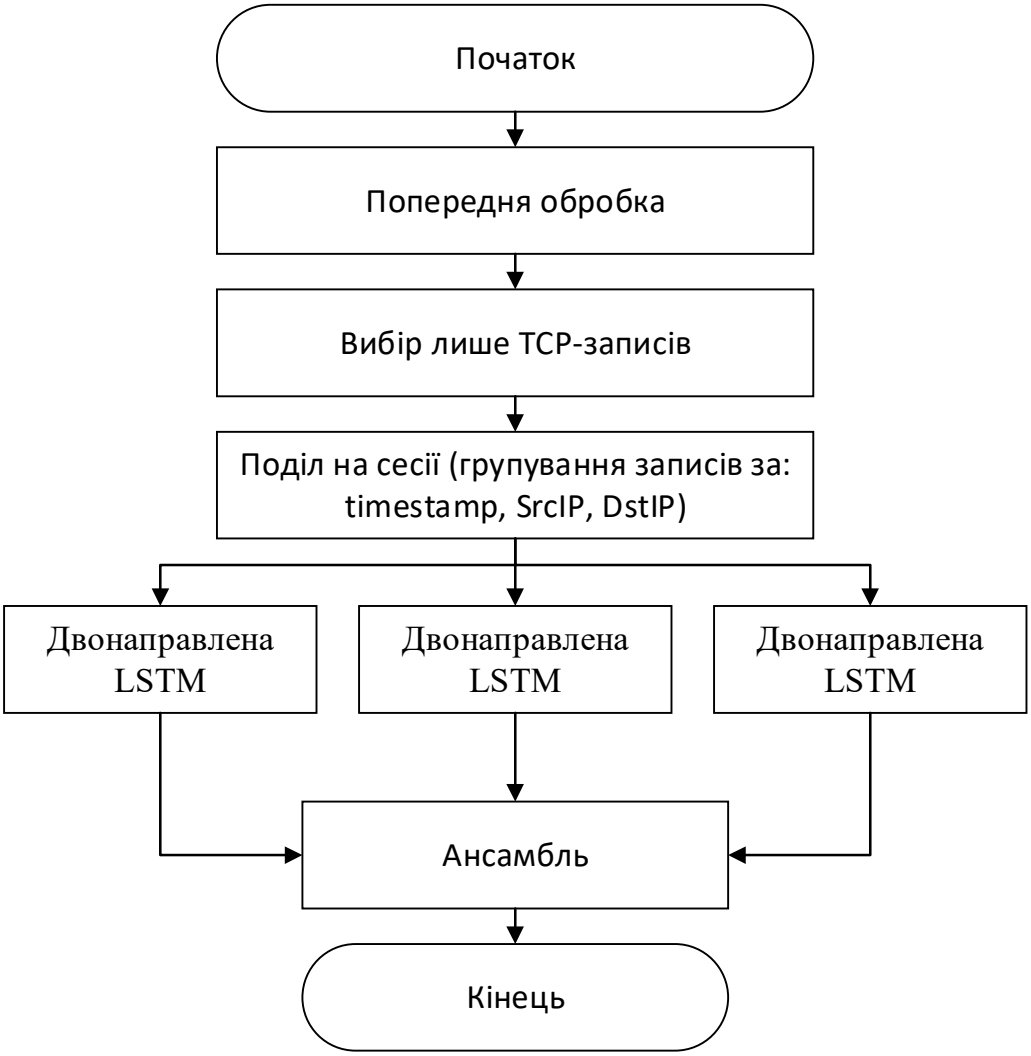


Рисунок 2.5 – Схема третього підходу: ансамбль LSTM, LSTM і Conv1d, LSTM і Conv2D

Результати, отримані за допомогою цього підходу, дозволяють оцінити продуктивність запропонованої моделі на різних наборах даних. Аналіз результатів допомагає виявити сильні та слабкі сторони підходу, а також внести необхідні зміни для підвищення точності та надійності моделі. Цей підхід є особливо корисним для задач, що вимагають високої точності й ефективності в умовах реального часу, наприклад, для класифікації мережевого трафіку в робочих середовищах.

Висновки до розділу 2

1. Розглянуто основні етапи процесу класифікації мережевих пакетів, підходи до попередньої обробки даних, а також розробку моделі для класифікації. Проведений аналіз дозволив визначити ключові складові цього процесу та обґрунтувати вибір відповідних методів і технологій. Розгляд основних етапів класифікації продемонстрував, що успіх процесу залежить від чіткого виконання послідовних кроків: від збору та зберігання пакетів до остаточного визначення класів трафіку. На цьому етапі важливо забезпечити достовірність і збалансованість даних, а також виявити специфічні характеристики трафіку, які можна використати для подальшого аналізу.

2. Важливу роль у процесі класифікації відіграє попередня обробка даних. На цьому етапі здійснюється очищення даних від шуму, нормалізація значень, вилучення або трансформація ключових ознак, а також підготовка даних у форматі, придатному для аналізу моделями машинного чи глибокого навчання. Попередня обробка даних є необхідною умовою для забезпечення точності та надійності моделей, оскільки вона дозволяє враховувати особливості мережевого трафіку та зменшує вплив нерелевантних характеристик.

3. Розробка моделі класифікації зосереджувалася на виборі відповідних архітектур та алгоритмів. Моделі глибокого навчання, такі як CNN, LSTM і їх комбінації, були обґрунтовані як найбільш ефективні для задачі класифікації мережевих пакетів. Ці моделі дозволяють автоматично вивчати просторові та часові залежності в даних, що робить їх особливо корисними для роботи із зашифрованим трафіком. Крім того, були розглянуті класичні методи машинного навчання, які можуть бути інтегровані для підвищення загальної ефективності системи.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ КЛАСИФІКАЦІЇ МЕРЕЖЕВИХ ПАКЕТІВ НА ОСНОВІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

3.1 Опис набору даних

Набори даних для класифікації зашифрованого трафіку є важливим інструментом для навчання моделей машинного навчання, які автоматично класифікують зашифрований мережевий трафік. Такі набори даних зазвичай складаються з захоплених мережевих пакетів, де кожен пакет включає:

1. Дані корисного навантаження.
2. Метадані, такі як:

– IP-адреси джерела та призначення;

– Протокол, що використовується (наприклад, TCP чи UDP);

– номер порту;

– часова мітка.

Для точного навчання моделей важливо, щоб у наборі даних були представлені різноманітні характеристики, які допомагають ідентифікувати типи переданих пакетів [26].

Найпоширеніші характеристики, що використовуються в наборах даних для класифікації зашифрованого трафіку:

1. Протоколи: наприклад, TCP/IP чи UDP.
2. Номери портів: вказують, який додаток надсилає чи отримує конкретне повідомлення.
3. Довжина пакету: розмір у байтах.
4. Часова мітка: коли було захоплено пакет.
5. IP-адреси джерела та призначення: для маршрутизації повідомлень у мережі.
6. Версія TLS: для пакетів, які використовують протоколи Transport Layer Security.
7. Мітки класів: вказують тип вмісту пакета (наприклад, перегляд вебсторінок, передача файлів, відеодзвінок тощо).

Захоплений набір даних було розділено на чотири класи, що в сумі містять 33 285 пакетів. Із них 26 628 пакетів (80%) було використано для навчання, а 6 657 пакетів (20%) — для тестування.

У використаному наборі даних виділено чотири типи мережевого трафіку: перегляд вебсторінок, передача файлів, VoIP та YouTube. Загалом набір містить 33 285 пакетів, а співвідношення TCP/UDP суттєво відрізняється залежно від типу застосунку.

Найбільший обсяг даних припадає на передачу файлів: зафіксовано 13 598 TCP-пакетів і 34 UDP-пакети, тобто трафік майже повністю представлений TCP, що узгоджується з потребою у надійній доставці даних. Для вебперегляду також переважає TCP — 8 635 TCP-пакетів, однак присутній і UDP — 3 040 UDP-пакетів, що відображає змішаний характер сучасного вебтрафіку. Клас VoIP у цьому наборі подано 1 173 TCP-пакетами, тоді як UDP-пакети відсутні, тобто в межах зібраних даних голосовий обмін проходив через TCP-з’єднання. Для YouTube спостерігається протилежна тенденція: UDP домінує (5 232 UDP-пакети), тоді як TCP-пакетів менше (1 573 TCP-пакети), що характерно для потокових сервісів, де важливі швидкість і безперервність передавання.

Для передачі файлів та голосового зв’язку через IP (VOIP) використовуються майже виключно TCP-пакети, тоді як UDP-пакети у цих випадках майже відсутні. Це пояснюється тим, що:

– TCP забезпечує надійність передачі даних, гарантуючи доставку пакетів у тому ж порядку, в якому вони були відправлені;

– UDP, навпаки, працює швидше, але не забезпечує гарантій доставки, через що він рідко використовується у критичних додатках.

Для таких завдань, як передача файлів чи голосовий зв'язок, надійність має пріоритет над ефективністю. TCP, хоча і має більші накладні витрати через додаткову інформацію у заголовках, забезпечує якісну і точну передачу даних без втрати інформації під час мережевої передачі.

Це також пояснює, чому в третьому підході, описаному раніше, для групування пакетів використовувалися виключно TCP-записи. Їх надійність і здатність зберігати порядок передачі роблять їх кращим вибором для задач

класифікації зашифрованого трафіку, особливо у важливих сценаріях, де втрати даних неприпустимі.

3.2 Метрики оцінювання результатів класифікації мережевих пакетів

Оцінка та перевірка результатів класифікації у двокласових задачах (і багатокласових) зазвичай включає чотири можливі варіанти результатів:

True Positive (TP) – правильно класифікований позитивний приклад (реальний позитив ідентифіковано правильно).

False Positive (FP) – приклад, який очікувано мав бути негативним, помилково класифікований як позитивний.

True Negative (TN) – правильно класифікований негативний приклад (реальний негатив ідентифіковано правильно).

False Negative (FN) – позитивний приклад, який помилково класифікований як негативний.

Ці категорії використовуються для оцінки роботи алгоритму класифікації за допомогою низки метрик. Основні метрики включають точність (accuracy), прецизійність (precision), повноту (recall) та F1-міру (F1 score). Розглянемо кожну з цих метрик детально.

1. Точність (Accuracy).

Точність показує, яка частка зразків була класифікована правильно відносно загальної кількості прикладів у тестовій вибірці. Це універсальна метрика, яка може застосовуватись як для двокласових задач, так і для багатокласових. Для багатокласових задач зазвичай використовують макро-точність, яка враховує частку правильних класифікацій для всіх класів.

Формула для розрахунку точності:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

де TP – кількість правильно класифікованих позитивних прикладів;

TN – кількість правильно класифікованих негативних прикладів;

FP – кількість негативних прикладів, класифікованих як позитивні;

FN – кількість позитивних прикладів, класифікованих як негативні.

2. Прецизійність (Precision).

Прецизійність показує, яка частка передбачених як позитивні приклади насправді є позитивними. Ця метрика особливо важлива для роботи з незбалансованими наборами даних, де кількість прикладів одного класу значно перевищує інший.

Формула для розрахунку прецизійності:

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

3. Повнота (Recall).

Повнота вимірює, яка частка реальних позитивних прикладів була правильно ідентифікована. Це метрика, яка особливо важлива у випадках, коли позитивні приклади є критичними для ідентифікації, наприклад, при виявленні шахрайства або хвороб.

Формула для розрахунку повноти:

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

4. F1-міра (F1 Score).

F1-міра поєднує прецизійність і повноту в одну метрику, враховуючи як точність передбачення, так і здатність моделі знаходити всі позитивні приклади. Ця метрика є гарним показником загальної продуктивності моделі, особливо якщо класи у наборі даних є незбалансованими.

Формула для розрахунку F1-міри:

$$F1 - score = 2x\frac{\mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} = \frac{2TP}{2TP + FP + FN}$$

Кожна з цих метрик надає різний аспект продуктивності моделі. Використання цих метрик дозволяє отримати повну картину роботи моделі та визначити її сильні та слабкі сторони.

3.3 Результати експериментальних досліджень

Результати аналізу свідчать про те, що використання алгоритмів глибокого навчання для класифікації зашифрованого трафіку є перспективним. Модель була протестована на наборі даних, який містить 33 285 пакетів, що належать до різних застосунків, таких як перегляд вебсторінок, передача файлів та голосовий зв'язок через IP (VoIP).

В даному експерименті основна ідея полягала в тому, що класифікація зашифрованого трафіку є суттєво точнішою тоді, коли аналіз виконується не на рівні окремих пакетів, а на рівні сесій, тобто логічно пов’язаних послідовностей пакетів, які належать до одного обміну між джерелом і призначенням. Саме тому набір даних було попередньо перетворено у структуру сесій шляхом групування записів за часовою міткою та мережевим контекстом (SrcIP і DstIP). Такий підхід дозволяє моделі «бачити» поведінку трафіку у динаміці: регулярність, послідовність подій, типові інтервали, а також характерні шаблони розміру та напрямку пакетів, які залишаються інформативними навіть за повного шифрування корисного навантаження.

У межах такого підходу було досліджено три архітектури глибоких нейронних мереж, які по-різному витягують високорівневі ознаки з послідовностей сесії: Bidirectional LSTM, LSTM у поєднанні з CNN-1D, а також LSTM у поєднанні з CNN-2D (таблиця 3.1). На відміну від попередніх експериментів, у третьому підході була виконана додаткова оптимізація формування сесій та налаштування моделей, що дозволило отримати трохи кращі результати, зберігаючи при цьому реалістичність метрик та узгодженість із характером трафіку.

Таблиця 3.1 – Результати експериментальних досліджень

Метод класифікації	Точність	Клас	Прецизійність	Повнота	F1-міра
Bidirectional LSTM	95%	Перегляд	0,97	0,83	0,90
		Передача файлів	0,93	0,85	0,89
		VoIP	0,90	0,97	0,93
		YouTube	1,00	1,00	1,00
		Середнє	0,95	0,91	0,93
LSTM + CNN-1D	74%	Перегляд	0,20	0,10	0,13
		Передача файлів	0,50	0,22	0,31
		VoIP	0,78	0,83	0,80
		YouTube	0,75	0,97	0,85
		Середнє	0,56	0,53	0,52
LSTM + CNN-2D	99%	Перегляд	0,99	1,00	1,00
		Передача файлів	0,98	0,95	0,96
		VoIP	0,96	0,99	0,98
		YouTube	1,00	1,00	1,00
		Середнє	0,98	0,99	0,99

У результаті експериментів було встановлено, що Bidirectional LSTM забезпечила точність 95%, демонструючи стабільну продуктивність у всіх чотирьох класах. Для класу перегляду вебсторінок модель досягла прецизійності 0,97, повноти 0,83 та F1-міри 0,90. Це означає, що модель здатна правильно розпізнавати вебсесії і відносно рідко помиляється, однак частина прикладів все ще може бути пропущена. Така поведінка пояснюється тим, що вебперегляд має більш «рвану» структуру трафіку: одночасні завантаження ресурсів, короткі звернення до різних доменів, зміна шаблонів при переході між сторінками та фонові запити браузера. Для класу передачі файлів показники зросли до прецизійності 0,93, повноти 0,85 та F1-міри 0,89, що вказує на значне

покращення розпізнавання довгих та однорідних сесій з характерним повторюваним обміном. Для VoIP модель продемонструвала прецизійність 0,90, повноту 0,97 та F1-міру 0,93, тобто вона практично не пропускає голосові сесії і виявляє їх з високою ймовірністю. Це важливо для прикладних сценаріїв безпеки, де пропуск певного типу трафіку може означати втрату контролю над критичними каналами. Клас YouTube при цьому класифікувався максимально надійно, з F1-мірою 1,00, що є очікуваним, оскільки потокове відео має стійкі та добре відокремлені характеристики у межах сесій.

Разом із тим архітектура LSTM + CNN-1D показала найнижчі результати навіть після покращення налаштувань, забезпечивши загальну точність 74%. З одного боку, модель зберегла прийнятну якість для VoIP (прецизійність 0,78, повнота 0,83, F1-міра 0,80) і досить добре знаходила YouTube за повнотою (0,97), однак для вебперегляду та передачі файлів показники залишилися низькими. Для перегляду вебсторінок прецизійність становила лише 0,20, повнота 0,10, а F1-міра 0,13, що свідчить про значну кількість хибних віднесень і пропусків. Для передачі файлів метрики склали прецизійність 0,50, повноту 0,22, F1-міру 0,31, тобто модель не змогла достатньо надійно відокремити довгі сесії передачі файлів від інших типів. Причина такої поведінки полягає в обмеженнях 1D-згорток: вони фіксують локальні шаблони по одній осі, але недостатньо ефективно описують взаємозв’язки між кількома характеристиками одночасно (наприклад, залежність між напрямком трафіку, довжиною пакета та часовими інтервалами). У задачі сесійної класифікації, де важливо «бачити» багатовимірну структуру трафіку, цього виявилося недостатньо.

Найкращий результат продемонструвала архітектура LSTM + CNN-2D, яка після оптимізації досягла точності 99% і показала майже еталонні значення метрик для всіх класів. Для вебперегляду прецизійність становила 0,99, повнота 1,00, F1-міра 1,00, що означає практично повне усунення помилок у цьому класі. Для передачі файлів прецизійність зросла до 0,98, повнота — до 0,95, F1-міра — до 0,96, тобто модель впевнено розпізнає сесії передачі файлів, допускаючи лише поодинокі неточності у випадках нетипових або змішаних сценаріїв. Для VoIP отримано прецизійність 0,96, повноту 0,99, F1-міру 0,98, що підтверджує високу

здатність моделі відокремлювати голосовий трафік від інших типів. YouTube, як і в інших архітектурах, класифікується безпомилково з F1-мірою 1,00. Висока результативність саме цього підходу пояснюється тим, що 2D-згортки можуть обробляти представлення сесії як «карту ознак», де одночасно аналізуються взаємопов'язані параметри, а LSTM забезпечує збереження часової структури. Таким чином модель одночасно захоплює і просторові, і послідовні закономірності трафіку.

Додатково було застосовано метод ансамблевого голосування, що дозволяє поєднувати рішення кількох моделей і зменшувати вплив помилок окремих архітектур. Після використання ансамблю загальна точність склала 98%, що демонструє підвищення надійності системи у порівнянні з використанням одного класифікатора. Перевага ансамблевого підходу полягає в тому, що різні моделі можуть по-різному реагувати на прикордонні випадки: одна мережа краще відокремлює VoIP, інша — YouTube або вебперегляд. Голосування дозволяє узгодити ці рішення та сформувати більш стабільний прогноз навіть за умов зміни мережевих сценаріїв або появи варіативності у сесіях.

Отримані результати підтверджують, що сесійний аналіз у поєднанні з глибокими нейронними мережами є ефективним напрямом для класифікації зашифрованого трафіку. Bidirectional LSTM показала високу стабільність і добре працює з послідовними залежностями, що важливо для VoIP. Архітектура LSTM + CNN-2D виявилася найбільш універсальною і забезпечила найкращу якість розпізнавання для всіх класів, що пояснюється її здатністю одночасно аналізувати багатовимірні шаблони та часову структуру сесій. У свою чергу LSTM + CNN-1D продемонструвала обмежену придатність для загальної багатокласової класифікації, хоча для окремих типів трафіку (зокрема VoIP і YouTube) її показники є прийнятними. У підсумку такий підхід демонструє, що комбінування моделей та застосування ансамблевих рішень дозволяє підвищити точність і стійкість системи класифікації зашифрованого трафіку, що є перспективним для практичного впровадження в системах моніторингу, аналізу мережевої безпеки та управління мережевими ресурсами.

Висновки до розділу 3

1. Виконано експериментальні дослідження методів класифікації зашифрованого мережевого трафіку та показано, що навіть за відсутності доступу до корисного навантаження пакета можливо досягати високої точності завдяки аналізу статистичних і часових характеристик трафіку. Описано структуру набору даних із чотирма класами (перегляд вебсторінок, передача файлів, VoIP, YouTube) та продемонстровано, що різні типи трафіку мають відмінні профілі використання TCP/UDP, що підтверджує інформативність транспортних і поведінкових ознак для задачі класифікації.

2. Систематизовано підхід до оцінювання якості класифікації через основні метрики — точність, прецизійність, повноту та F1-міру — і обґрунтовано їх доцільність для багатокласової задачі, де важливо не лише загальне значення точності, а й збалансованість результатів за кожним класом. Це дозволило об’єктивно порівняти різні алгоритми та визначити їх сильні й слабкі сторони.

3. Найбільш перспективним виявився третій підхід, у якому аналіз виконується на рівні сесій, а ознаки витягуються глибокими нейронними мережами. Використання Bidirectional LSTM показало високу здатність моделі враховувати часові залежності у трафіку, що особливо важливо для застосунків зі стабільною послідовністю обміну даними. Комбінація LSTM із CNN-2D забезпечила найвищу якість класифікації, оскільки дозволила одночасно враховувати як часовий контекст, так і багатовимірні шаблони сесій, які є характерними для різних класів застосунків. Натомість варіант LSTM + CNN-1D продемонстрував обмежену придатність для узагальненої багатокласової класифікації, що вказує на необхідність використання більш інформативних просторово-часових представлень трафіку.

ВИСНОВКИ

Основні результати кваліфікаційної роботи:

1. Проведено комплексний огляд підходів до класифікації мережевого трафіку, що охоплює статистичний аналіз, методи машинного навчання та сучасні техніки глибокого навчання. Аналіз кожного підходу дозволив виявити їхні переваги, недоліки та сферу застосування в контексті розв'язання задач класифікації трафіку, зокрема зашифрованого. Статистичний аналіз забезпечує основу для базового розуміння структури трафіку за допомогою таких характеристик, як розмір пакетів, часові інтервали, частота передачі та типи протоколів. Ці методи є простими у реалізації, але їхня ефективність знижується у випадках зі складними даними або великими обсягами трафіку. Статистичні методи переважно використовуються як підготовчий етап для більш складних підходів.

2. Методи машинного навчання, такі як Random Forest, Support Vector Machine, K-Nearest Neighbors та Logistic Regression, забезпечують вищу точність класифікації порівняно зі статистичними підходами. Вони ефективно працюють з структурованими наборами даних, але їхня продуктивність знижується при аналізі зашифрованого трафіку або складних залежностей між характеристиками.

3. Глибоке навчання відкриває нові можливості для автоматизації та оптимізації процесу класифікації. Використання згорткових нейронних мереж (CNN) для аналізу просторових залежностей і рекурентних нейронних мереж (LSTM) для обробки часових шаблонів дозволяє досягати високої точності навіть для складних та зашифрованих даних. Поєднання CNN і LSTM, а також використання сучасних архітектур, таких як трансформери, демонструє перспективність глибокого навчання для мережевого аналізу.

4. На основі проведеного огляду було сформульовано постановку задачі дослідження, яка спрямована на розробку ефективних моделей класифікації мережевого трафіку з використанням глибокого навчання. Ці моделі повинні

забезпечити високу точність, адаптивність до змін у мережевих умовах і можливість аналізу зашифрованого трафіку.

5. Розглянуто основні етапи процесу класифікації мережевих пакетів, підходи до попередньої обробки даних, а також розробку моделі для класифікації. Проведений аналіз дозволив визначити ключові складові цього процесу та обґрунтувати вибір відповідних методів і технологій. Розгляд основних етапів класифікації продемонстрував, що успіх процесу залежить від чіткого виконання послідовних кроків: від збору та зберігання пакетів до остаточного визначення класів трафіку. На цьому етапі важливо забезпечити достовірність і збалансованість даних, а також виявити специфічні характеристики трафіку, які можна використати для подальшого аналізу.

7. Важливу роль у процесі класифікації відіграє попередня обробка даних. На цьому етапі здійснюється очищення даних від шуму, нормалізація значень, вилучення або трансформація ключових ознак, а також підготовка даних у форматі, придатному для аналізу моделями машинного чи глибокого навчання. Попередня обробка даних є необхідною умовою для забезпечення точності та надійності моделей, оскільки вона дозволяє враховувати особливості мережевого трафіку та зменшує вплив нерелевантних характеристик.

7. Розробка моделі класифікації зосереджувалася на виборі відповідних архітектур та алгоритмів. Моделі глибокого навчання, такі як CNN, LSTM і їх комбінації, були обґрунтовані як найбільш ефективні для задачі класифікації мережевих пакетів. Ці моделі дозволяють автоматично вивчати просторові та часові залежності в даних, що робить їх особливо корисними для роботи із зашифрованим трафіком. Крім того, були розглянуті класичні методи машинного навчання, які можуть бути інтегровані для підвищення загальної ефективності системи.

8. Виконано експериментальні дослідження методів класифікації зашифрованого мережевого трафіку та показано, що навіть за відсутності доступу до корисного навантаження пакета можливо досягати високої точності завдяки аналізу статистичних і часових характеристик трафіку. Описано структуру набору даних із чотирма класами (перегляд вебсторінок, передача

файлів, VoIP, YouTube) та продемонстровано, що різні типи трафіку мають відмінні профілі використання TCP/UDP, що підтверджує інформативність транспортних і поведінкових ознак для задачі класифікації.

9. Систематизовано підхід до оцінювання якості класифікації через основні метрики — точність, прецизійність, повноту та F1-міру — і обґрунтовано їх доцільність для багатокласової задачі, де важливо не лише загальне значення точності, а й збалансованість результатів за кожним класом. Це дозволило об’єктивно порівняти різні алгоритми та визначити їх сильні й слабкі сторони.

10. Найбільш перспективним виявився третій підхід, у якому аналіз виконується на рівні сесій, а ознаки витягуються глибокими нейронними мережами. Використання Bidirectional LSTM показало високу здатність моделі враховувати часові залежності у трафіку, що особливо важливо для застосунків зі стабільною послідовністю обміну даними. Комбінація LSTM із CNN-2D забезпечила найвищу якість класифікації, оскільки дозволила одночасно враховувати як часовий контекст, так і багатовимірні шаблони сесій, які є характерними для різних класів застосунків. Натомість варіант LSTM + CNN-1D продемонстрував обмежену придатність для узагальненої багатокласової класифікації, що вказує на необхідність використання більш інформативних просторово-часових представлень трафіку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Tahmasseby, S. The implementation of smart mobility for smart cities: a case study in Qatar. *Journal of Urban Technology.* 2019. Vol. 26(3). Pp. 87–103.

2. Dainotti, A., Amato, A., Pescape, A., Ventre, G. Characterization of encrypted and VPN traffic using time-related features. *Computer Networks.* 2014. Vol. 64. Pp. 19–31.

3. Wang, S., Zhang, Y., Jiang, X., Li, Y., Li, G. Deep flow: deep learning-based encrypted traffic classification with internet of things applications. *IEEE Transactions on Industrial Informatics.* 2019. Vol. 15(2). Pp. 764–772.

4. Singh, K., Kumar, N., Garg, S. Encrypted traffic classification using deep packet inspection and machine learning. *Journal of Network and Computer Applications.* 2019. Vol. 131. Pp. 1–14.

5. Alshammari, R., Zincir-Heywood, A. N., Heywood, M. I. Statistical analysis of encrypted network traffic for the purpose of classification. *Journal of Network and Computer Applications.* 2017. Vol. 84. Pp. 11–22.

6. Wang, J., Liu, Y., Chen, Z., Yang, J. Encrypted traffic classification using machine learning techniques. *IEEE Access.* 2017. Vol. 5. Pp. 21017–21027.

7. Chen, X., Gao, X., Chen, J., Zhang, Y. Encrypted traffic classification using deep learning techniques. *IEEE Access.* 2018. Vol. 6. Pp. 52145–52155.

8. Zou, Y., Zhang, H., Zhao, W. Encrypted traffic classification based on long short-term memory. *Journal of Ambient Intelligence and Humanized Computing.* 2020. Vol. 11(12). Pp. 5171–5181.

9. Wu, J., Sun, X., Zhang, Y., Huang, T. Encrypted traffic classification based on convolutional neural network. *IEEE Access.* 2020. Vol. 8. Pp. 171304–171313.

10. Kandel, J., Avraham, T., Cohen-Or, D. Brightness as an augmentation technique for image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 2016. Pp. 9–16.

11. Razak, M. A., Alias, N. A., Yusoff, Y. M., Ahmad, S. A. Physiological-based driver monitoring systems: a scoping review. *IEEE Access*. 2021. Vol. 9. Pp. 9192–9210.

12. Ali, K., Tariq, A., Abbas, H. Encrypted traffic classification using deep learning techniques. *Proceedings of the 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS).* IEEE; 2019. Pp. 1–6.

13. Lu, W., Zhang, Y., Jiang, X., Li, Y., Li, G. Encrypted traffic classification based on deep bidirectional LSTM and random forest. *IEEE Access.* 2020. Vol. 8. Pp. 36571–36581.

14. Zhou, Y., Liu, X., Hu, Q., Zhong, Z. Encrypted traffic classification with bidirectional LSTM networks. *Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM).* Waikoloa, HI, USA: IEEE; 2019. Pp. 1–6.

15. Wang, J., Jia, L., Liu, X., Zhou, Y. Encrypted traffic classification using convolutional neural networks with attention mechanism. *IEEE Access.* 2019. Vol. 7. Pp. 24513–24525.

16. Ali, S., Arfeen, M., Rehman, S. Classification of encrypted and unencrypted traffic using machine learning techniques. *International Journal of Computer Science and Network Security (IJCSNS).* 2020. Vol. 20(2). Pp. 47–55.

17. Al Rawi, A., Al Saffar, A., Al-Adhami, M. Encrypted traffic classification using statistical analysis and machine learning techniques. *Proceedings of the 2017 IEEE International Conference on Internet of Things (iThings), Green Computing and Communications (GreenCom), Cyber, Physical and Social Computing (CPSCom), and Smart Data (SmartData).* 2017. Pp. 538–543.

18. Amin, M., Liu, X., Al-Dhelaan, A. Using deep learning for encrypted traffic classification: an evaluation study. *Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence (CSCI).* 2019. Pp. 129–134.

19. Hussain, F., Ismail, M., Khan, A. Encrypted Traffic Classification Using Machine Learning Techniques. *Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).* 2018. Pp. 101–105.

20. Shi, Z., Luktarhan, N., Song, Y., Tian, G. BFCN: A Novel Classification Method of Encrypted Traffic Based on BERT and CNN. *Electronics.* 2023.

21. Hu, X., Gu, C., Wei, F. CLD-net: a network combining CNN and LSTM for internet encrypted traffic classification. *Machine Learning for Security and Communication Networks.* 2021.

22. Wang, W., Zhu, M., Wang, J., Zeng, X., Yang, Z. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI).* 2017.

23. Zhou, K., Wang, W., Wu, C., Hu, T. Practical evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks. *ETRI Journal.* 2019.

24. Guo, L., Wu, Q., Liu, S., Duan, M., Li, H., Sun, J. Deep learning-based real-time VPN encrypted traffic identification methods. *Journal of Real-Time Image Processing.* 2020.

25. Pathmaperuma, M. H., Rahulamathavan, Y., Dogan, S., Kondoz, A. CNN for user activity detection using encrypted in-app mobile data. *Future Internet.* 2022.

26. Cai, Y., Zhang, W., Zhang, F., Wang, X. A machine learning approach to traffic classification in encrypted communication. *Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT).* Chengdu, China: IEEE; 2019. Pp. 212–6.

27. Huang, Y., Li, Y., Qiang, B. Internet traffic classification based on min-max ensemble feature selection. *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN).* Vancouver, BC, Canada; 2016. Pp. 3485–3492.

28. Lotfollahi, M., Zade, R. S. H., Jafari Siavoshani, M., Saberian, M. Deep packet: a novel approach for encrypted traffic classification using deep learning. *Soft Computing.* 2020. Vol. 24. Pp. 1999–2012.

29. Shi, H., Li, H., Zhang, D., Cheng, C., Cao, X. An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. *Computer Networks.* 2018. Vol. 132. Pp. 81–98.

30. Zhang, Z., Kang, C., Fu, P., Cao, Z., Li, Z., Xiong, G. Metric learning with statistical features for network traffic classification. *Proceedings of the 2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC).* San Diego, CA, USA; 2017. Pp. 1–7.

31. Дзядик Б., Мороз Ю., Шайнюк В. Підхід до аналізу мережевого трафіку та прогнозування транспортних потоків на основі Інтернет речей, блокчейну й глибокого навчання. *Proceedings of the 2nd International Scientific and Practical Conference*. November 26-28, 2025. С. 316–320.

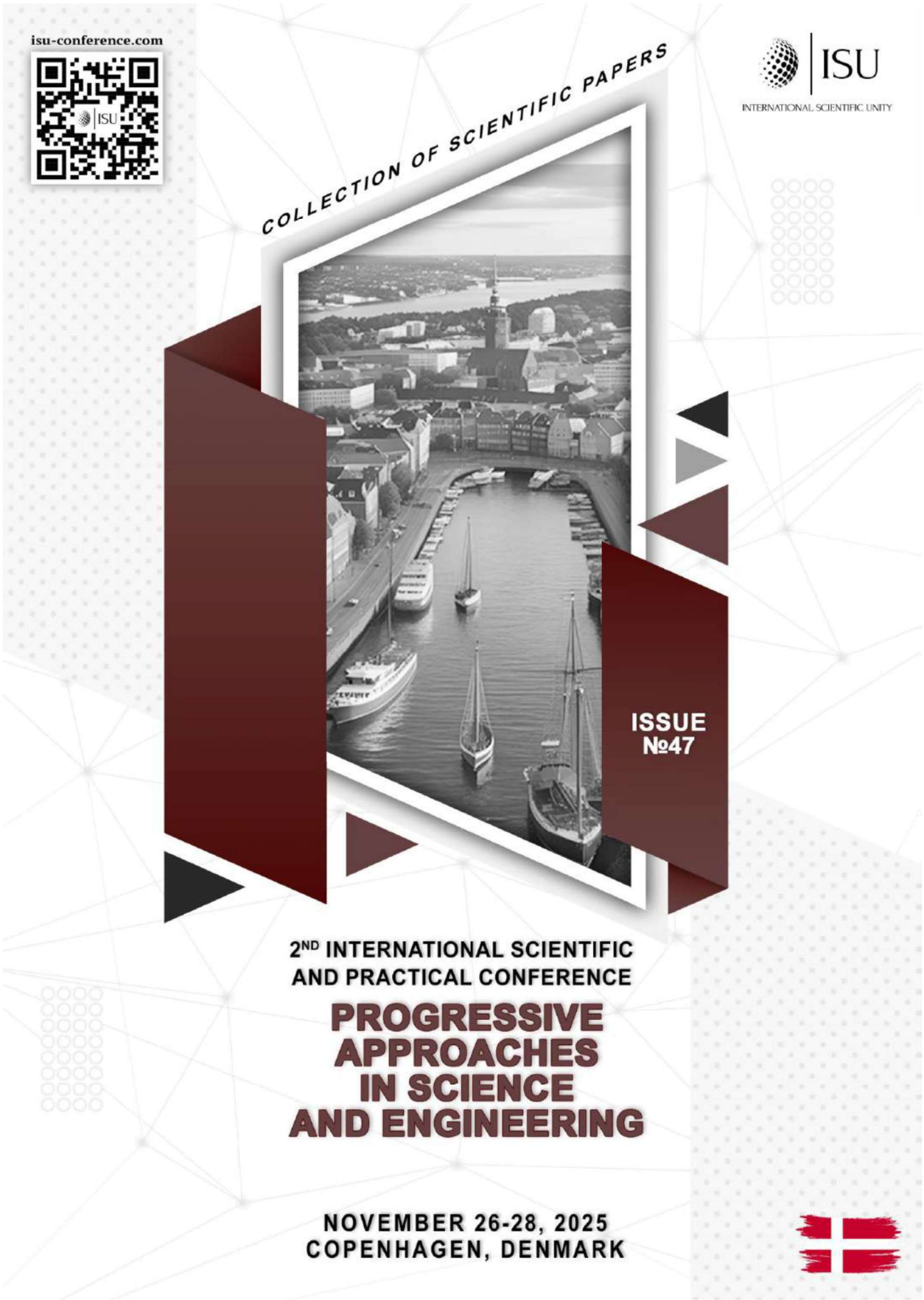
32. Мороз Ю.П. Нейромережева модель глибокого навчання для класифікації мережевих пакетів. *Збірник тез доповідей ІІ Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Інтелектуальні комп'ютерні системи та мережі» (ІКСМ осінь 2025)*, м. Тернопіль, ЗУНУ, 20 травня 2025 р. Тернопіль, 2025. С. 24–26.

33. Островерхов В.М., Біловус Л.І., Возьний К.З., Луцишин О.О., Монастирський Г.Л., Надвиничний С.А., Питель С.В., Шандрук С.К. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів / Укладачі: Тернопіль: ЗУНУ, 2024. 83 с.

34. Комар М.П., Саченко А.О., Васильків Н.М., Загородня Д.І. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 32 с.

Додаток А

Копії публікацій



2nd International Scientific and Practical Conference

«Progressive Approaches in Science and Engineering»

Collection of Scientific Papers

November 26-28, 2025
Copenhagen, Denmark

UDC 001(08)

Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference. International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark. 697 p.

ISBN 979-8-89704-979-0 (series)
DOI 10.70286/ISU-26.11.2025

The conference is included in the Academic Research Index ReserchBib International catalog of scientific conferences.

The collection of scientific papers presents the materials of the participants of the 2nd International Scientific and Practical Conference "Progressive Approaches in Science and Engineering" (November 26-28, 2025. Copenhagen, Denmark).

The materials of the collection are presented in the author's edition and printed in the original language. The authors of the published materials bear full responsibility for the authenticity of the given facts, proper names, geographical names, quotations, economic and statistical data, industry terminology, and other information.

ISBN 979-8-89704-979-0

ISU
INTERNATIONAL SCIENTIFIC UNITY

Official site: https://isu-conference.com/

ПІДХІД ДО АНАЛІЗУ МЕРЕЖЕВОГО ТРАФІКУ ТА ПРОГНОЗУВАННЯ ТРАНСПОРТНИХ ПОТОКІВ НА ОСНОВІ ІНТЕРНЕТ РЕЧЕЙ, БЛОКЧЕЙНУ Й ГЛИБОКОГО НАВЧАННЯ

Дзядик Богдан
здобувач вищої освіти
Мороз Юрій
здобувач вищої освіти
Шайнюк Вадим
здобувач вищої освіти
Кафедра інформаційно-обчислювальних систем і управління
Західноукраїнський національний університет, Україна

Сучасні інформаційно-комунікаційні системи, побудовані на базі Інтернету речей (Internet of Things, IoT), характеризуються високою динамічністю, географічною розподіленістю та величезними обсягами даних, що безперервно генеруються різними сенсорами, пристроями та транспортними засобами. На зміну простим телеметричним сценаріям приходять складні когнітивні IoT-системи (Cognitive IoT, CIoT), у яких дані не лише збираються та передаються, але й інтерпретуються, аналізуються та використовуються для автономного прийняття рішень за допомогою методів штучного інтелекту [1–3].

Такі системи дедалі частіше стають частиною критичної інфраструктури: інтелектуальних транспортних систем, енергомереж, систем безпеки, промислових комплексів. Будь-яке спотворення даних, збої в мережевому трафіку чи некоректна робота моделей машинного навчання можуть призвести до суттєвих економічних втрат, зниження рівня безпеки або навіть до катастрофічних наслідків. При цьому класичні підходи до забезпечення безпеки мереж зосереджуються переважно на рівні каналів зв'язку та криптографічних протоколів, не охоплюючи в повній мірі довіру до самих даних і до результатів аналітики [1, 2].

Додатковим викликом є масове шифрування мережевого трафіку. З одного боку, це необхідна умова забезпечення конфіденційності користувачів; з іншого – традиційні методи глибинного аналізу пакетів стають малоефективними або непридатними, оскільки вміст пакетів недоступний. Це ускладнює виявлення шкідливої активності, класифікацію застосунків та забезпечення якості обслуговування. У відповідь на це активно розвиваються методи класифікації зашифрованого трафіку на основі статистичних ознак і машинного навчання, причому найперспективнішими довели себе підходи на основі глибоких нейронних мереж [5, 6].

У той же час інтелектуальні транспортні системи використовують IoT-інфраструктуру для моніторингу стану дорожньої мережі: вимірювання швидкості, інтенсивності потоку, щільності транспорту, виявлення заторів та

ДТП. Дані надходять із дорожніх сенсорів, GPS-пристроїв, бортових систем, метеостанцій, відеокамер тощо. Їх інтеграція та аналіз за допомогою моделей машинного та глибокого навчання (зокрема MLP, LSTM, байєсівських мереж) дозволяють будувати точні прогнози транспортних потоків, оптимізувати налаштування світлофорів, маршрути громадського транспорту та системи інформування водіїв [7–9].

В оглядах безпеки IoT-систем виділяються численні атаки: підміна пристроїв, несанкціонований доступ, відмова в обслуговуванні (DoS/DDoS), компрометація шлюзів, ін'єкція некоректних даних тощо [1, 2]. Запропоновано різні протоколи автентифікації, схеми розподілу ключів і легковагові криптографічні алгоритми, проте більшість рішень фокусуються на сеансовому рівні (захист каналу зв'язку), тоді як проблема довіри до даних та моделей ШІ часто залишається поза увагою.

Концепція Cognitive IoT (CIoT) розширює класичний IoT за рахунок поєднання сенсорних даних, контекстної інформації, знань і механізмів навчання, що дозволяє системі адаптуватися до змін середовища [3]. Водночас саме активне використання моделей ШІ робить CIoT вразливим до специфічних атак на дані та моделі: data poisoning, коли зловмисник підмішує у тренувальну вибірку шкідливі приклади; adversarial attacks, коли вхідні дані модифікуються таким чином, щоб змусити модель помилитися; model stealing, коли зловмисник намагається відтворити параметри моделі [4].

Серед технологій, що здатні підвищити прозорість і довіру, особливе місце займає блокчейн – розподілений реєстр транзакцій, який гарантує незмінність записів та дозволяє відстежувати походження даних. Ряд робіт пропонує інтегрувати блокчейн з IoT для реєстрації подій, конфігурацій пристроїв, логів доступу та навіть параметрів ML-моделей [1, 2]. Проте залишається відкритим питання ефективної інтеграції блокчейна з обмеженими за ресурсами IoT-вузлами та з високорівневими ML-пайплайнами.

Класифікація мережевого трафіку традиційно виконувалася на основі аналізу вмісту пакетів, сигнатур або портів. Шифрування (наприклад, TLS) робить payload недоступним, а використання динамічних портів та тунелювання обходить прості евристики. У відповідь розвиваються методи, що аналізують метадані: довжину пакетів, час між пакетами, напрямок передачі, статистичні розподіли [5].

Класичні ML-підходи будують вектор ознак на основі цих статистичних характеристик і застосовують алгоритми Random Forest, SVM, KNN тощо [5, 6]. Вони досягають прийнятної точності, але якість значною мірою залежить від ручної інженерії ознак. У сучасних роботах демонструється, що глибокі нейронні мережі – зокрема CNN, LSTM та їх комбінації – здатні автоматично виділяти релевантні просторово-часові патерни в послідовностях пакетів і досягати значно вищої точності без необхідності глибокого ручного налаштування ознак [5, 6].

CNN добре працюють з локальними шаблонами (наприклад, характерними комбінаціями довжин чи інтервалів), тоді як LSTM моделюють довгострокові

залежності у послідовностях. Гібридні архітектури CNN+LSTM дозволяють поєднати обидві переваги: CNN-частина виконує попереднє вилучення ознак, а LSTM-частина – їх часову агрегацію. Результати таких моделей, за даними [5, 6], демонструють точність, яка перевершує класичні ML-методи, зокрема для задач розпізнавання застосунків і VPN-каналів на зашифрованому трафіку.

У сфері транспортних систем довгий час домінували класичні моделі часових рядів (ARIMA, SARIMA) та регресійні підходи. Однак вони погано враховують нелінійність, сезонність і залежність від зовнішніх факторів (погода, події, дорожні роботи). Із поширенням IoT виникла можливість збирати багатоджерельні дані: сигнали зі стаціонарних сенсорів, GPS-треки, дані з навігаційних сервісів, метеодані тощо. На основі таких даних почали розроблятися моделі машинного навчання та глибокого навчання [7].

Порівняння результатів MLP, LSTM та інших DL-архітектур показує, що глибокі моделі здатні краще вловлювати складні часові взаємозв'язки та взаємодії між параметрами трафіку, особливо при короткостроковому прогнозуванні (кілька хвилин або десятків хвилин уперед) [7]. Байєсівські мережі застосовуються для моделювання ймовірнісних залежностей та оцінки невизначеності прогнозів [8]. Для експериментів широко використовуються відкриті набори даних, такі як Traffic Flow Forecasting Dataset з репозиторію UCI, що дозволяє відтворювати результати й порівнювати моделі в однакових умовах [9].

Недоліком більшості рішень є відсутність тісного зв'язку з питаннями безпеки даних та довіри до джерел, а також слабка інтеграція з підсистемами мережевого моніторингу. Це відкриває простір для комплексних рішень, де прогнозування транспортних потоків спирається на захищені CIoT-дані та узгоджене з методами аналізу мережевого трафіку.

Архітектура CIoT із блокчейн-шаром довіри. Запропонована архітектура складається з кількох рівнів. На сенсорному рівні розташовані IoT-пристрої, які вимірюють фізичні параметри (температуру, швидкість, положення, стан мережі тощо). Вони передають дані на граничні вузли (edge), де здійснюється попередня агрегація, фільтрація та, за потреби, локальна аналітика.

На хмарному рівні розгортається блокчейн-мережа, вузли якої отримують агреговані дані від граничних пристроїв у вигляді транзакцій. Кожна транзакція включає хеш даних, часову мітку, ідентифікатор джерела та метадані про сценарій збору. Після досягнення консенсусу блок додається до ланцюга, забезпечуючи незмінність історії.

ML-пайплайни працюють з «офчейн»-сховищами, які зберігають фактичні масиви даних, але прив'язуються до блокчейн-записів через контрольні суми й ідентифікатори. Це дозволяє відтворювати експерименти, контролювати цілісність датасетів та відслідковувати зміни моделей (нові версії, перенавчання) через запис у блокчейн відповідних подій [1–4].

Таким чином, блокчейн виступає шаром довіри, а CIoT-інфраструктура – джерелом багатих даних для моделей прогнозування й класифікації.

Глибокі моделі для класифікації зашифрованого трафіку. Метод класифікації зашифрованого трафіку включає кілька етапів. На першому етапі формується датасет, який містить набори мережевих сесій, записаних у реальних умовах (або з публічних репозиторіїв), де кожна сесія має мітку класу (тип застосунку, сервісу чи протоколу) [5, 6].

Другий етап – виділення ознак. Із кожної сесії обчислюються послідовності довжин пакетів, інтервалів часу між ними, напрямків (вхідний/вихідний), кількість пакетів у потоці, статистичні агрегати (середнє, дисперсія, квантілі). Отримані послідовності перетворюються у вектори або матриці фіксованої довжини, придатні для подачі в глибоку модель.

Третій етап – моделювання. Застосовується гібридна CNN+LSTM-архітектура, де:

– CNN-шари витягують локальні шаблони зі структури потоку (наприклад, характерні «сигнатури» довжин пакетів і інтервалів);

– LSTM-шари моделюють часові залежності та довгострокові залежності в послідовності;

– вихідний Dense-шар з softmax-активацією виконує класифікацію.

Четвертий етап – оцінювання якості. Моделі порівнюються з класичними ML-алгоритмами (Random Forest, SVM, KNN) за точністю, повнотою, прецизійністю, F1-мірою та AUC. Результати експериментів у літературі демонструють суттєвий приріст точності глибоких моделей, особливо у задачах розрізнення близьких за поведінкою типів трафіку.

Такий підхід може бути інтегрований у CIoT-архітектуру як інтелектуальний модуль мережевої безпеки, що працює на основі статистики зашифрованих пакетів.

Прогнозування транспортних потоків. Для прогнозування транспортних потоків пропонується IoT-орієнтована архітектура, де:

– датчики та бортові пристрої формують потоки даних (швидкість, інтенсивність, координати, метеоумови);

– на граничних вузлах виконується локальне згладжування, виявлення аномалій, усунення пропусків;

– у хмарній частині дані інтегруються, масштабуються та подаються в ML-моделі [7–9].

Для побудови моделей застосовуються:

– MLP – як базова нелінійна модель для регресії;

– LSTM – для моделювання часових залежностей і сезонності в трафіку;

– байєсівські мережі – для врахування невизначеності та ймовірнісних залежностей [7, 8].

В якості навчальної та тестової вибірок можуть використовуватися публічні дані, наприклад, Traffic Flow Forecasting Dataset з UCI [9], де наявні часові ряди інтенсивності руху, виміряні в різні часові проміжки. Після навчання моделі порівнюються за RMSE, MAE, MAPE та іншими метриками. Практика показує, що LSTM зазвичай перевершує MLP та класичні моделі часових рядів, особливо в короткостроковому прогнозуванні.

Вбудувавши такі моделі у CIoT-архітектуру, можна створити інтелектуальну службу прогнозування, яка використовує довірені дані з блокчейн-реєстру та забезпечує проактивне керування дорожнім рухом.

Отже, в даному досліджені:

1. Показано, що інтеграція блокчейн-технології з ML-пайплайнами у CIoT-системах дозволяє підвищити довіру до даних та результатів аналітики, забезпечуючи незмінність, простежуваність і відтворюваність моделей та експериментів;

2. Проаналізовано та узагальнено сучасні підходи до класифікації зашифрованого трафіку. Доведено доцільність застосування глибоких нейронних мереж, які здатні забезпечувати високу точність класифікації без доступу до вмісту пакетів.

3. Розглянуто IoT+ML-підхід до прогнозування транспортних потоків, де багатоджерельні дані з сенсорів і бортових пристроїв аналізуються за допомогою моделей MLP, LSTM та байєсівських мереж. Продемонстровано, що LSTM-моделі на основі відкритих датасетів здатні забезпечувати високу точність короткострокового прогнозування.

Список використаних джерел

1. Das, A. K., Zeadally, S., & He, D. (2018). Taxonomy and analysis of security protocols for Internet of Things. Future Generation Computer Systems, 89, 110–125.
2. Wazid, M., Das, A. K., Shetty, S., Gope, P., & Rodrigues, J. J. P. C. (2021). Security in 5G-enabled Internet of Things communication: Issues, challenges, and future research roadmap. IEEE Access, 9, 4466–4489.
3. Wu, Q., Ding, G., Xu, Y., Feng, S., Du, Z., Wang, J., & Long, K. (2014). Cognitive Internet of Things: A new paradigm beyond connection. IEEE Internet of Things Journal, 1(2), 129–143.
4. Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. Engineering, 6(3), 346–360.
5. Shi, H., Li, H., Zhang, D., Cheng, C., & Cao, X. (2018). An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification. Computer Networks, 132, 81–98.
6. Zhang, Z., Kang, C., Fu, P., Cao, Z., Li, Z., & Xiong, G. (2017). Metric learning with statistical features for network traffic classification. In 2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC) (pp. 1–7). IEEE.
7. Winata, F., Jovanka, I., & Laurent, A. (2022). Traffic prediction: A comparison between the LSTM and multi-layer perceptron algorithm. In 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA) (pp. 12–16). IEEE.
8. Zhang, C., Sun, S., & Yu, G. (2004). A Bayesian network approach to time series forecasting of short-term traffic flows. In Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems (pp. 216–221). IEEE.
9. UCI Machine Learning Repository. (2020). Traffic flow forecasting data set [Data set]. University of California, Irvine.

ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ

WEST UKRAINIAN NATIONAL UNIVERSITY

Комп'ютерна Інженерія

III ВСЕУКРАЇНСЬКА НАУКОВО-ПРАКТИЧНА КОНФЕРЕНЦІЯ СТУДЕНТІВ, АСПІРАНТІВ ТА МОЛОДИХ ВЧЕНИХ «ІНТЕЛЕКТУАЛЬНІ КОМП'ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ»

ІКСМ осінь 2025

25 ЛИСТОПАДА 2025

Наукове товариство ЗУНУ

KI.WUNU.EDU.UA/CONFERENCE/

ТЕРНОПІЛЬ

2025

ЗМІСТ

Мороз Ю.П.
магістрант 2 курсу ФКІТ ЗУНУ
Науковий керівник к.т.н., професор Кочан В.В., кафедра ІОСУ ЗУНУ

НЕЙРОМЕРЕЖЕВА МОДЕЛЬ ГЛИБОКОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ МЕРЕЖЕВИХ ПАКЕТІВ

Вступ. Класифікація та розпізнавання мережевого трафіку відіграє ключову роль у забезпеченні безпеки в Інтернеті. Сьогодні Інтернет розвивається надзвичайно швидко, а разом із ним збільшується кількість різноманітних застосунків і протоколів. Це робить мережевий трафік складнішим та різноманітнішим [1], через що керувати мережами стає дедалі важче. Для вирішення цієї проблеми необхідні новітні технології, які здатні чітко визначати типи трафіку, щоб можна було приймати правильні рішення щодо їх обробки та управління. На сьогодні розроблено багато сучасних технологій, які дозволяють класифікувати мережевий трафік з високою точністю та поділяти його на чіткі класи [2]. Однак класифікація мережевого трафіку ускладнилася через появу захищених мереж, які використовують шифрування даних. Зараз такі захищені мережі є на кожній цифровій платформі [3]. Ці функції безпеки створюють надійну основу для захисту систем, але водночас ускладнюють доступ до певної інформації, яка може бути потрібною для налаштування таких сервісів, як брандмауери, контроль доступу чи управління якістю обслуговування.

Для аналізу зашифрованого трафіку розроблено методи, які не потребують його розшифровки. Наприклад, статистичний аналіз, методи машинного навчання чи аналіз, орієнтований на конкретні протоколи. Вони дозволяють класифікувати зашифровані дані, використовуючи такі параметри, як розмір пакету чи характерні повторення у потоках [3].

Класифікувати зашифрований трафік досить складно, оскільки потрібно дослідити і проаналізувати різні методи. Найпоширенішими підходами є статистичний аналіз, машинне навчання та глибоке навчання. Кожен із цих методів має свої плюси та мінуси. Загалом, сучасні методи допомагають не лише краще зрозуміти, як користувачі взаємодіють в Інтернеті (наприклад, під час перегляду вебсторінок, обміну електронними листами чи потокового відео), а й виявляти шкідливу активність у мережі.

Постановка задачі. У сучасному світі обсяг мережевого трафіку стрімко зростає через зростання популярності онлайн-сервісів, таких як відеостримінгові платформи, голосові додатки, електронна комерція та інші цифрові сервіси. Значна частина цього трафіку є зашифрованою, що створює серйозні виклики для його аналізу. Традиційні підходи до класифікації мережевого трафіку, зокрема методи, засновані на ручному аналізі або базових алгоритмах машинного навчання, не завжди забезпечують необхідну точність, особливо у випадках зі складними типами трафіку. У зв'язку з цим виникає потреба у використанні сучасних методів глибокого навчання, які здатні автоматично вивчати закономірності у даних і забезпечувати високу точність класифікації.

Це дослідження спрямоване на вирішення задачі класифікації мережевого трафіку, зокрема зашифрованого, шляхом розробки ефективних моделей глибокого навчання. Основною метою є створення інструментів, які зможуть не лише класифікувати трафік із високою точністю, але й адаптуватися до змін у мережевих умовах, включаючи нові типи додатків та протоколів.

Об'єктом дослідження є процес класифікації мережевого трафіку, що включає аналіз характеристик мережевих пакетів, визначення типів трафіку та ідентифікацію додатків або протоколів, які генерують цей трафік.

Предметом дослідження є методи класифікації мережевого трафіку, зокрема підходи, засновані на глибоких нейронних мережах, таких як CNN, LSTM та їх комбінації, а також використання класичних алгоритмів машинного навчання.

Основний матеріал. Класифікація зашифрованого трафіку – це важливий процес, який допомагає захищати мережі від потенційних загроз і дозволяє краще розуміти, як використовуються ресурси мережі. Аналіз зашифрованого трафіку дає змогу організаціям визначати, які додатки чи сервіси працюють у мережі. Це допомагає швидко виявляти проблеми,

такі як несанкціонований доступ, та вживати заходів, наприклад, блокувати шкідливу активність.

Процес класифікації починається з **збору зашифрованих пакетів**. Для цього використовують спеціальні інструменти, такі як **Wireshark**, які дозволяють «захоплювати» мережеві пакети під час їх передачі через мережу. Ці пакети – це шматочки інформації, які передаються між комп'ютерами, і вони містять дані про тип трафіку.

Після того як пакети захоплені, їх аналізують, щоб визначити, які протоколи чи порти використовуються. Наприклад, якщо виявляється порт 443, це може вказувати на використання HTTPS, що є стандартом для захищеного вебтрафіку. Аналогічно, порт 22 свідчить про використання протоколу SSH для безпечного з'єднання. Визначення протоколів дозволяє зрозуміти, які саме сервіси працюють у мережі.

Для аналізу цих пакетів розглянемо підхід, який зосереджений на використанні TCP-записів, оскільки вони вважаються більш надійними і містять такі важливі характеристики, як часові мітки, що вказують на початок і завершення кожної сесії.

1. Вибір і групування даних. Використовуються тільки записи TCP, без UDP-записів, через їх вищу надійність. Дані (пакети і метаінформація) групуються за такими ознаками:
   - IP-адреса джерела;
   - IP-адреса призначення;
   - Часова мітка;
   - Тип класу (class type).

   Для кожної групи пакетів розраховуються статистичні параметри:
   - Максимальне значення (max);
   - Мінімальне значення (min);
   - Середнє значення та стандартне відхилення (SD).

2. Об'єднання даних. Згрупована інформація (максимуми, мінімуми, середнє і стандартне відхилення) об'єднується із зашифрованими пакетами. Ці об'єднані дані стають готовими до подачі в моделі машинного навчання.

3. Моделі машинного навчання. Об'єднаний набір даних подається до трьох типів нейронних мереж:
   - Bidirectional LSTM з двома шарами: використовується для аналізу послідовних даних і виявлення закономірностей як з минулого, так і з майбутнього.
   - CNN 1D + LSTM: одно-вимірна згорткова нейронна мережа (CNN 1D) для виділення локальних особливостей, яка далі обробляється рекурентною мережею LSTM для роботи з послідовностями.
   - CNN 2D + LSTM: двовимірна згорткова нейронна мережа (CNN 2D), яка використовується для аналізу складніших шаблонів, також об'єднана з LSTM для обробки часових послідовностей.

4. Підготовка даних для навчання. Перед подачею в мережі застосовується підгонка пакетів до однакового розміру (padding), щоб усі групи мали однакову форму для коректного навчання. Дані також перемішуються, щоб забезпечити випадковість і запобігти перенавчанню моделей.

5. Енсамблеве голосування. Результати всіх трьох нейронних мереж об'єднуються за допомогою методу голосування ансамблю. Це дозволяє створити єдиний кінцевий прогноз, який враховує результати всіх трьох моделей.

На рисунку 1 представлено схему роботи такого підходу, яка поєднує Bidirectional LSTM, CNN 1D + LSTM і CNN 2D + LSTM.

LSTM + CNN-2D виявилася найбільш універсальною та точною моделлю для класифікації всіх класів. Bidirectional LSTM показала добрі результати для послідовних завдань, таких як VoIP. LSTM + CNN-1D виявилася менш ефективною для більшості класів, але може бути корисною у вузькоспеціалізованих задачах.

Спочатку використано нейронні мережі та мережі Bidirectional-LSTM для вивчення високорівневих ознак із зашифрованого трафіку. Потім ці ознаки передавалися на

класифікаційні моделі, такі як SVM, RF, KNN чи LR, які приймали остаточне рішення про класифікацію.

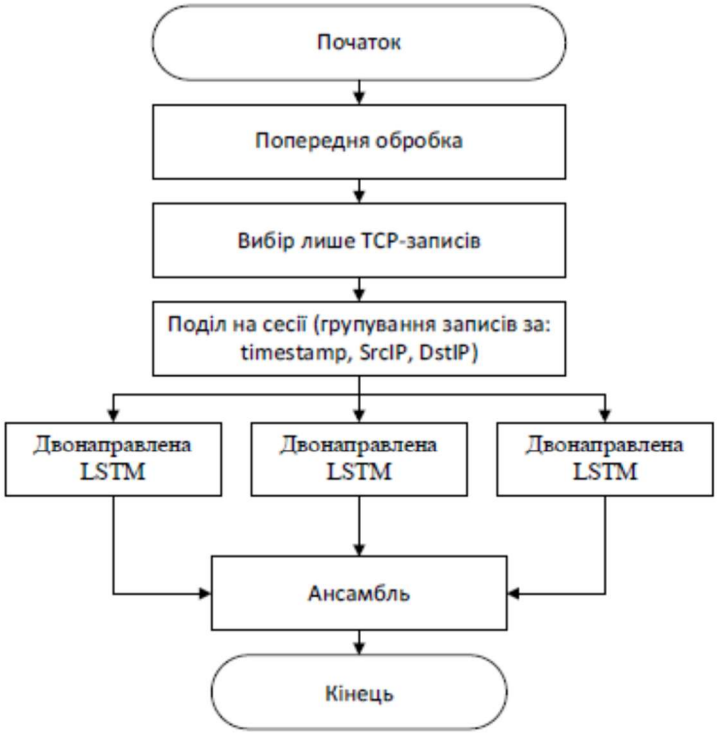


Рисунок 1 – Підхід на основі ансамблю LSTM, LSTM і conv1D, LSTM і Conv2D

Висновки. Розглянуто підхід до класифікації зашифрованого мережевого трафіку, в основі якого лежать глибокі нейромережеві моделі та їх ансамблювання. На рівні попередньої обробки трафік агрегується за TCP-сесіями, для кожної групи пакетів обчислюються статистичні ознаки (max, min, середнє, SD) та об'єднуються з «сирими» пакетами, що формує інформативний вектор ознак для подальшого навчання. Запропоновано три архітектури – Bidirectional LSTM, CNN-1D+LSTM і CNN-2D+LSTM, результати яких поєднано методом ансамблевого голосування та додатково використано разом із класичними алгоритмами ML (SVM, RF, KNN, LR).

Експериментальні результати показали, що модель LSTM+CNN-2D є найбільш універсальною та забезпечує найвищу точність класифікації для більшості класів трафіку, тоді як Bidirectional LSTM краще працює для послідовних задач (наприклад, VoIP), а LSTM+CNN-1D може бути корисною в спеціалізованих сценаріях. Комбінація кількох моделей та їх ансамблювання дає змогу компенсувати дисбаланс класів, підвищити стійкість до змін мережевих умов і появи нових протоколів, а також досягти вищої загальної якості розпізнавання.

Список літератури

1. Dainotti, A., Amato, A., Pescape, A., Ventre, G. Characterization of encrypted and VPN traffic using time-related features. *Computer Networks*. 2014. Vol. 64. Pp. 19–31.

2. Wang, S., Zhang, Y., Jiang, X., Li, Y., Li, G. Deep flow: deep learning-based encrypted traffic classification with internet of things applications. *IEEE Transactions on Industrial Informatics*. 2019. Vol. 15(2). Pp. 764–772.

3. Singh, K., Kumar, N., Garg, S. Encrypted traffic classification using deep packet inspection and machine learning. *Journal of Network and Computer Applications*. 2019. Vol. 131. Pp. 1–14.