Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерної інженерії та інформаційних систем

КВАЛІФІКАЦІЙНА РОБОТА

Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики
Назва теми

Рівень вищої освіти *другий (магістерський)*

Галузь знань *12 «Інформаційні технології»*
Шифр, назва

Спеціальність *123 «Комп'ютерна інженерія»*
Шифр, назва

Освітня програма *«Комп'ютерна інженерія та програмування»*
Назва

Шифр КвРКІ 240113.24.01.12 ПЗ

Виконав здобувач ІІ курсу, група КІ2м-24-1 _____ Володимир ДУДНИК
Підпис Ініціали, прізвище

Керівник д. техн. наук, професор _____ Василь ЯЦКІВ
Науковий ступінь, учене звання Підпис Ініціали, прізвище

Нормоконтролер д. техн. наук, професор _____ Сергій ЛИСЕНКО
Науковий ступінь, учене звання Підпис Ініціали, прізвище

До захисту допускаю:
завідувач кафедри КІІС _____ Ольга ПАВЛОВА
«01» травня 2026 р. підпис Ініціали, прізвище
дата

Хмельницький 2026

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ

Рівень вищої освіти ДРУГИЙ (МАГІСТЕРСЬКИЙ)

Галузь знань 12 ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

Спеціальність 123 КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

Освітня програма «КОМП'ЮТЕРНА ІНЖЕНЕРІЯ ТА ПРОГРАМУВАННЯ»

ЗАТВЕРДЖУЮ
Завідувачка кафедри КІІС
Ольга ПАВЛОВА

" 12 " 01 2026 р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

Дуднику Володимиру Миколайовичу
Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

Керівник проекту (роботи) Яцків Василь Васильович, д-р техн. наук, професор
Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 12.01.2026 р. № 6

2. Термін подання здобувачем роботи на кафедру 01.05.2026 р.

3. Вихідні дані до роботи Завдання на кваліфікаційну роботу

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)

Розроблення методу аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики.

Алгоритмічна реалізація та дослідження ефективності розробленого методу аналізу трафіку.

Програмна реалізація та експериментальна перевірка розробленого методу аналізу трафіку комп'ютерних мереж.

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

Архітектура ПЗ проєкту

Архітектура ПЗ для кіберфізичної системи

Апаратне забезпечення проєкту

6. Консультанти розділів кваліфікаційної роботи

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

7. Дата видачі завдання «_12_»_____01_____2026 р.

КАЛЕНДАРНИЙ ПЛАН

№з/п	Назва етапів (розділів) дипломного проєкту (роботи)	Термін виконання етапів проєкту (роботи)	Примітка
1	Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи з керівником	12.01.2026	виконано
2	Ознайомлення з предметною областю; формулювання мети та задач дослідження; визначення об'єкта та предмета дослідження	15.01.2026	виконано
3	Робота над розділом 1 – теоретичні основи аналізу мережевого трафіку та виявлення аномалій	01.02.2026	виконано
4	Робота над розділом 2 – розроблення моделі аналізу мережевого трафіку на основі ентропійних характеристик	01.03.2026	виконано
5	Робота над науковою статтею	01.03.2026	виконано
6	Робота над розділом 3 – розроблення методу аналізу мережевого трафіку на основі ентропійних характеристик	29.03.2026	виконано
7	Робота над розділом 4 – експериментальна перевірка методу аналізу мережевого трафіку	01.04.2026	виконано
8	Оформлення пояснювальної записки згідно вимог	25.04.2026	виконано
9	Попередній захист ДРМ	29.04.2026	виконано
10	Захист ДРМ на засіданні ЕК	01.05.2026	

Здобувач _____ Володимир ДУДНИК
Підпис Ім'я, ПРІЗВИЩЕ

Керівник кваліфікаційної роботи _____ Василь ЯЦКІВ
Підпис Ім'я, ПРІЗВИЩЕ

РЕФЕРАТ

Тема кваліфікаційної роботи магістра: Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики.

Автор роботи: Дудник Володимир Миколайович

Керівник роботи: Яцків Василь Васильович, д-р техн. наук, професор

Пояснювальна записка: 81 с., 17 рис., 4 табл., 1 дод., 90 джерел.

МЕРЕЖЕВИЙ ТРАФІК, АНАЛІЗ ТРАФІКУ, ЕНТРОПІЙНІ ХАРАКТЕРИСТИКИ, БАГАТОВИМІРНА МАТЕМАТИЧНА СТАТИСТИКА, ВИЯВЛЕННЯ АНОМАЛІЙ, PCA, КРИТЕРІЙ ХОТЕЛЛІНГА, КІБЕРЗАГРОЗИ.

Об'єктом дослідження є процеси передавання та аналізу трафіку в комп'ютерних мережах.

Предметом дослідження є методи, моделі та програмно-технічні засоби аналізу мережевого трафіку на основі ентропійних характеристик і багатовимірної математичної статистики.

Метою кваліфікаційної роботи магістра є підвищення точності виявлення аномалій шляхом розроблення методу та програмно-технічних засобів аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик і методів багатовимірної математичної статистики.

Для розв'язання поставлених задач використовувалися методи системного аналізу, теорії інформації, ентропійного аналізу мережевого трафіку, багатовимірної математичної статистики, методу головних компонент, критерію Хотеллінга T^2, статистичного моделювання, алгоритмічного та програмного проєктування.

Наукова новизна отриманих результатів:

– набув подальшого розвитку метод аналізу мережевого трафіку для виявлення аномальних станів, який, на відміну від існуючих підходів, базується на узгодженому використанні ентропійних характеристик структури трафіку та

багатовимірного статистичного аналізу їх спільної динаміки, що дозволяє підвищити чутливість до структурних змін мережевого потоку;

– удосконалено модель опису станів мережного трафіку, яка подає кожне часове вікно у вигляді вектора інформативних ознак і визначає аномальні стани за статистичною мірою відхилення від профілю нормального режиму, що забезпечує формалізований перехід від спостережуваних даних до прийняття рішення;

– набуло подальшого розвитку поєднання віконної обробки трафіку з багатовимірним оцінюванням, яке дозволяє враховувати часову динаміку змін і взаємозв'язки між ознаками, що підвищує стійкість методу до випадкових коливань і зменшує кількість хибних спрацювань;

– удосконалено підхід до оцінювання ефективності методів аналізу мережевого трафіку, який, на відміну від стандартного використання окремих метрик, передбачає комплексний аналіз інтегральної статистики, ROC- і PR-кривих, а також розподілу затримки виявлення, що дозволяє більш повно оцінити якість детекції та швидкість реагування системи.

На основі проведених досліджень розроблена модульна архітектура та компоненти програмного забезпечення реалізації методу аналізу мережного трафіку, що включають модулі підготовки даних, формування часових вікон, обчислення ентропійних характеристик, багатовимірного статистичного аналізу, прийняття рішення та експериментальної перевірки.

Практична значимість отриманих результатів полягає у можливості використання розробленого методу та його програмної реалізації в системах моніторингу мережевої інфраструктури, виявлення кіберзагроз, аналізу навантаження та підтримки прийняття рішень щодо адміністрування комп'ютерних мереж.

У першому розділі проаналізовано мережевий трафік як об'єкт моніторингу, класи деградацій та аномалій, відомі методи й засоби виявлення відхилень, а також обґрунтовано доцільність використання ентропійних характеристик і багатовимірної статистики для аналізу трафіку.

У другому розділі виконано постановку задачі дослідження, формалізацію процесу аналізу мережевого трафіку, обґрунтовано вибір ентропійних і багатовимірних статистичних методів.

У третьому розділі розроблено алгоритмічну реалізацію методу, зокрема алгоритми формування часових вікон і підготовки вибірки, обчислення ентропійних характеристик.

У четвертому розділі обґрунтовано вибір засобів програмної реалізації, реалізовано модулі підготовки даних, формування часових вікон та обчислення ентропійних характеристик, модулі багатовимірного статистичного аналізу і прийняття рішення, а також описано вхідні дані, умови проведення експериментів та сценарії перевірки методу.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

БД – база даних

БПР – блок прийняття рішень

ПЗ – програмне забезпечення

СВВ – система виявлення вторгнень

DDoS – Distributed Denial of Service, розподілена атака відмови в обслуговуванні

IDS – Intrusion Detection System, система виявлення вторгнень

IPFIX – Internet Protocol Flow Information Export, протокол експорту інформації про потоки

NetFlow – технологія збору та аналізу мережевих потоків

CAP – Packet Capture, формат збереження мережевого трафіку

PCAPNG – Packet Capture Next Generation, розширений формат збереження мережевого трафіку

PCA – Principal Component Analysis, метод головних компонент

MSPC – Multivariate Statistical Process Control, багатовимірний статистичний контроль процесів

KL – Kullback–Leibler divergence, дивергенція Кульбака–Лейблера

JSD – Jensen–Shannon divergence, дивергенція Дженсена–Шеннона

EWMA – Exponentially Weighted Moving Average, експоненціально зважене ковзне середнє

CSV – Comma-Separated Values, текстовий формат табличних даних

Scapy – бібліотека Python для аналізу та обробки мережевих пакетів

TShark – консольний аналізатор мережевого трафіку Wireshark

ВСТУП

Сучасний етап розвитку інформаційних технологій характеризується стрімким зростанням обсягів передавання даних, ускладненням архітектур комп'ютерних мереж та підвищенням вимог до їхньої надійності, продуктивності й безпеки. Комп'ютерні мережі є основою функціонування підприємств, державних установ, фінансових сервісів, освітніх платформ і систем критичної інфраструктури. За таких умов особливого значення набуває своєчасний і точний аналіз мережевого трафіку, який дає змогу виявляти аномалії, прогнозувати перевантаження, ідентифікувати загрози та забезпечувати ефективне управління мережевими ресурсами.

Традиційні підходи до аналізу трафіку переважно базуються на статистичних показниках, сигнатурних методах або простому моніторингу інтенсивності пакетних потоків. Проте в умовах зростаючої різноманітності мережевих сервісів, динамічної зміни характеру навантаження та появи складних кіберзагроз таких підходів часто недостатньо для повного опису стану мережі. Це зумовлює необхідність застосування більш гнучких та інформативних методів, здатних враховувати як невизначеність і структурну складність трафіку, так і багатофакторний характер його параметрів.

Одним із перспективних напрямів дослідження є використання ентропійних характеристик, які дозволяють кількісно оцінювати ступінь упорядкованості, випадковості та інформаційної насиченості мережевих потоків. Ентропійний підхід є ефективним для виявлення прихованих змін у поведінці трафіку, що можуть бути пов'язані з аномаліями, атаками або змінами режимів функціонування мережі. Водночас багатовимірна математична статистика надає апарат для комплексного аналізу сукупності взаємопов'язаних параметрів трафіку, виявлення закономірностей, класифікації станів мережі та побудови моделей прийняття рішень.

Поєднання ентропійних характеристик і методів багатовимірної математичної статистики створює підґрунтя для розроблення більш точних і

адаптивних методів аналізу трафіку комп’ютерних мереж. Такий підхід дозволяє не лише фіксувати факт відхилення від норми, а й глибше досліджувати внутрішню структуру мережевих процесів, що є важливим для побудови сучасних програмно-технічних засобів моніторингу та аналітики.

Актуальність теми дипломної роботи зумовлена необхідністю підвищення ефективності аналізу мережевого трафіку в умовах збільшення обсягів даних, різнорідності інформаційних потоків та зростання вимог до кібербезпеки. Використання ентропійних показників і багатовимірних статистичних методів відкриває можливості для вдосконалення існуючих рішень у сфері моніторингу комп’ютерних мереж та створення нових програмно-технічних засобів для своєчасного виявлення аномальних станів.

Метою кваліфікаційної роботи магістра є підвищення точності виявлення аномалій шляхом розроблення методу та програмно-технічних засобів аналізу трафіку комп’ютерних мереж на основі ентропійних характеристик і методів багатовимірної математичної статистики.

Об’єктом дослідження є процеси передавання та аналізу трафіку в комп’ютерних мережах.

Предметом дослідження є методи, моделі та програмно-технічні засоби аналізу мережевого трафіку на основі ентропійних характеристик і багатовимірної математичної статистики.

Наукова новизна отриманих результатів:

– набув подальшого розвитку метод аналізу мережевого трафіку для виявлення аномальних станів, який, на відміну від існуючих підходів, базується на узгодженому використанні ентропійних характеристик структури трафіку та багатовимірного статистичного аналізу їх спільної динаміки, що дозволяє підвищити чутливість до структурних змін мережевого потоку;

– удосконалено систему аналізу трафіку комп’ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики, яка визначає аномальні стани за статистичною мірою відхилення від профілю

нормального режиму, що забезпечує формалізований перехід від спостережуваних даних до прийняття рішення.

У першому розділі виконано аналіз сучасного стану задачі аналізу мережевого трафіку. Розглянуто трафік комп'ютерних мереж як об'єкт моніторингу, досліджено основні класи аномальних станів і деградацій, проаналізовано відомі методи та засоби виявлення відхилень, а також обґрунтовано доцільність використання ентропійних характеристик і багатовимірної математичної статистики для аналізу стану мережі. Показано, що ентропійні підходи є ефективними для виявлення структурних змін у трафіку, тоді як багатовимірні статистичні методи дають змогу враховувати взаємозв'язки між ознаками, зменшувати кількість хибних спрацювань і формувати більш стійке рішення щодо стану мережевого середовища.

У другому розділі виконано постановку задачі дослідження та формалізацію процесу аналізу мережевого трафіку. Обґрунтовано вибір ентропійних характеристик, дивергентних мір і багатовимірних статистичних методів як основи побудови моделі. Сформовано систему інформативних ознак мережевого трафіку, розроблено математичну модель опису станів трафіку, у межах якої кожне часове вікно подається у вигляді вектора ознак, а аномальний стан визначається за статистичною мірою відхилення від профілю нормального режиму. Також побудовано структурну модель аналізу трафіку, яка відображає послідовність переходу від вхідних мережевих даних до формування рішення про стан мережі. Таким чином, у другому розділі створено теоретичну та модельну основу для подальшої розробки методу аналізу.

У третьому розділі розроблено метод аналізу мережевого трафіку на основі ентропійних характеристик і багатовимірної математичної статистики. Метод побудовано як послідовність взаємопов'язаних процедур, що охоплюють формування часових вікон, підготовку вибірки, обчислення ентропійних характеристик, формування вектора ознак, виконання багатовимірного статистичного аналізу та прийняття рішення щодо стану трафіку. Визначено логіку взаємодії окремих етапів методу, описано алгоритмічну схему переходу від сирих

мережевих даних до інтегральної статистики аномальності, а також обґрунтовано використання статистичного критерію для фіксації моменту переходу системи до аномального стану. Розроблений метод забезпечує комплексне врахування як структурних характеристик трафіку, так і спільної динаміки зміни ознак у часі.

У четвертому розділі обґрунтовано вибір засобів програмної реалізації розробленого методу та описано структуру програмного забезпечення, яке реалізує підготовку даних, віконну обробку, обчислення ентропійних характеристик, багатовимірне статистичне оцінювання й прийняття рішення. Окрему увагу приділено організації експериментальної перевірки методу, вибору вхідних даних, формуванню сценаріїв дослідження та побудові системи оцінювання результатів. Ефективність методу проаналізовано за допомогою інтегральної статистики аномальності, часових рядів ентропійних ознак, ROC- та PR-кривих, а також розподілу затримки виявлення. Такий підхід дозволив оцінити не лише якість відокремлення нормальних і аномальних станів, але й інтерпретованість результатів, стійкість до вибору параметрів і практичну придатність методу для задач мережевого моніторингу.

Практичне значення одержаних результатів полягає у можливості використання розробленого методу та програмно-технічних засобів у системах моніторингу мережевої інфраструктури, виявлення кіберзагроз, аналізу навантаження та підтримки прийняття рішень щодо адміністрування комп’ютерних мереж.

1 АНАЛІЗ СУЧАСНОГО СТАНУ ЗАДАЧІ АНАЛІЗУ ТРАФІКУ КОМП'ЮТЕРНИХ МЕРЕЖ

1.1 Трафік комп'ютерних мереж як об'єкт аналізу: характеристики, параметри та чинники впливу

Мережевий трафік – це обсяг даних, що передаються комп'ютерною мережею за певний час. Іншими словами, трафік відображає інтенсивність обміну даними між вузлами мережі у вигляді потоків пакетів, які маршрутизуються від відправника до одержувача та знову збираються у вихідні повідомлення на стороні отримувача [1]. Трафік поділяють за напрямком на вихідний (від внутрішніх вузлів назовні) та вхідний (зовнішні дані, що надходять до мережі), а також на внутрішній трафік (між вузлами всередині локальної мережі) і зовнішній (між локальною мережею та Інтернет) [2].

Характер трафіку може суттєво різнитися залежно від типів переданих даних: наприклад, реальному часі (аудіо- та відеострими, VoIP) притаманна чутливість до затримок і вимагається стабільна швидкість доставки, тоді як нереальночасовий трафік (електронна пошта, файлообмін) менш критичний до затримок. Параметри мережевого трафіку. Для опису мережевого трафіку використовують низку кількісних параметрів (метрик). Ключовою характеристикою є пропускна здатність або ширина каналу – максимальний обсяг даних, який може передаватися мережею за одиницю часу, вимірюється, як правило, у бітах за секунду (bps) [3].

Із фактично досягнутою пропускною здатністю пов'язаний параметр пропускна спроможність або продуктивність (throughput), що відбиває реальну швидкість передачі даних між кінцевими точками. Для оцінки якості трафіку важливі також затримки (latency) – час проходження пакета від джерела до отримувача, джиттер (jitter) – варіабельність затримки (коливання часу доставки пакетів), та втрата пакетів – частка пакетів, що не досягли адресата. Ці параметри впливають на продуктивність мережевих застосунків: наприклад, високі затримки чи втрати погіршують якість VoIP-зв'язку або потокового відео [4]. Додатково розглядають такі статистичні характеристики трафіку, як середній розмір пакетів,

кількість активних з’єднань або сесій, число переданих пакетів за інтервал часу, щільність трафіку в певні години тощо [5].

Сукупність цих показників формує профіль трафіку мережі в нормальному стані, з яким можна порівнювати поточні вимірювання при моніторингу. Динаміка та фактори впливу. Мережевий трафік є динамічним об’єктом моніторингу – його інтенсивність і структура змінюються з часом доби, днями тижня, залежно від активності користувачів та роботи сервісів. Для більшості мереж притаманні діурнальні цикли – регулярні добові коливання навантаження (наприклад, вдень трафік в офісній мережі значно вищий, ніж уночі) [6]. На обсяг і характер трафіку впливають такі фактори, як кількість одночасних користувачів і запитів, типи застосунків (наприклад, відеоконференції генерують значно більше даних, ніж текстові чати), конфігурація мережі та обладнання (пропускна здатність каналів, продуктивність маршрутизаторів), а також позаштатні ситуації та аномальні події.

До типових причин пікового навантаження або деградації трафіку належать різке збільшення кількості користувачів чи запитів, брак пропускної здатності каналу, обмеження апаратних ресурсів (наприклад, перевантаження маршрутизатора), неефективні налаштування мережевих протоколів і маршрутизації, а також раптові сплески трафіку (наприклад, вірусна активність чи флешмоб). Зокрема, якщо попит на мережеві ресурси перевищує доступну пропускну здатність, виникає конгестія (перевантаження мережі), що проявляється у збільшенні затримок, втратах пакетів та зниженні швидкості передачі даних.

Для підтримання стабільної роботи мережі організації впроваджують моніторинг трафіку та керування навантаженням – зокрема, планування ємності каналів, пріоритезацію критичного трафіку (QoS), шейпінг (штучне обмеження) потоків та інші заходи. Необхідність моніторингу трафіку. Моніторинг мережевого трафіку – це процес безперервного спостереження за параметрами руху даних у мережі з метою забезпечення її продуктивності, надійності та безпеки. Регулярний збір і аналіз статистики трафіку дозволяє виявляти проблеми (перевантаження, відмови вузлів, аномальні піки) на ранніх етапах і вживати заходів для їх усунення.

1.2 Аномальні стани мережного трафіку та їх прояви у статистичних і часових характеристиках

У контексті мережевого моніторингу під аномалією розуміють будь-яку нетипову зміну характеристик трафіку, що відхиляється від встановленого профілю нормальної роботи. Аномальний трафік може бути спричинений як зовнішніми атаками або зловмисною активністю, так і ненавмисними збоями, відмовами обладнання або незвичними поведінковими патернами користувачів. Важливою задачею є класифікація можливих аномалій, оскільки різні класи інцидентів мають різну природу та потребують різних підходів до виявлення і реагування. Основні класи мережевих аномалій [7].

Відмови та збої мережевого обладнання (outages) падіння сегмента мережі, вимкнення або перезавантаження маршрутизатора/комутатора, обрив каналу зв'язку тощо. Такі події призводять до різкого зниження або повної відсутності трафіку в певному напрямку. У часових рядах показників (наприклад, трафік інтерфейсу) відмова проявиться як раптове падіння до нуля або значного мінімуму, що може тривати до відновлення роботи елемента.

Флеш-крауд (flash crowd) різкий легітимний сплеск активності користувачів, коли велика кількість клієнтів одночасно звертається до ресурсу. Причини - наприклад, популярна трансляція, акція чи новинна подія. На відміну від атаки, флеш-крауд не є зловмисним, але викликає схожі симптоми: стрибок обсягу трафіку до значень, значно вищих за повсякденні. Такий сплеск може тривати від кількох хвилин до кількох днів. У часовому вимірі він проявляється як високочастотна компонента (різке зростання локальної дисперсії сигналу трафіку) на фоні більш повільних добових коливань. Статистично флеш-наплив відображається у збільшенні середнього значення та дисперсії потоків, але при цьому розподіл адрес/портів може залишатися подібним до нормального (усі користувачі діють легітимно, просто їх більше).

Кібератаки та зловмисна активність – це цілеспрямовані дії, що генерують аномальний трафік. Сюди відносять атаки типу DDoS (розподілена відмова в

обслуговуванні) - масове надсилання трафіку на ціль, щоб вичерпати її ресурси. DDoS проявляється як різкий пік інтенсивності трафіку, спрямованого на одну або кілька IP-адрес жертви, причому характерна особливість - концентрація трафіку: одна адреса отримувача домінує в загальному потоці, тоді як адрес відправників дуже багато (багато джерел атакують одну ціль) [8]. Статистичним індикатором є, наприклад, різке зменшення ентропії адрес призначення (усі пакети йдуть на одну ціль) при одночасному збільшенні кількості унікальних IP-джерел. Інший вид - сканування портів або адрес (network scans), коли зловмисник надсилає невеликий обсяг пакетів, але до багатьох різних адрес або портів з метою знайти вразливі вузли.

1.3 Аналіз відомих методів виявлення відхилень у трафіку комп'ютерних мереж

Проблема виявлення аномалій у мережевому трафіку досліджується вже кілька десятиліть, починаючи з класичної роботи Деннінг (Denning, 1987) по виявленню вторгнень через аномалії в системних журналах. За цей час накопичено значну кількість методів та підходів. Загалом, методи виявлення аномального трафіку можна розділити на дві великі категорії: методи, що базуються на знаннях про атаки (signature-based або misuse detection), та методи, що базуються на пошуку аномалій (anomaly detection). Перші покладаються на наперед відомі шаблони зловмисної активності (сигнатури) і порівнюють поточний трафік з базою відомих атак; вони ефективні проти відомих загроз, але безсилі проти нових, невідомих типів атак.

Другі ж методи створюють модель «легітимного» або нормального профілю трафіку, а потім визначають як аномалії всі випадки, що достатньо від цього профілю відхиляються [9]. Аномалійні методи здатні виявляти нові та невідомі загрози, але зазвичай стикаються з проблемою неправдивих спрацювань (коли значне відхилення не є атакою). В сучасних системах часто поєднуються обидва підходи: спочатку перевіряються відомі сигнатури (щоб швидко відфільтрувати

класичні атаки), а на решті трафіку застосовуються аномалійні детектори для пошуку нових загроз [10].

Найпростіший підхід - встановлення порогів на ключові метрики трафіку. Наприклад, правило «якщо інтенсивність трафіку перевищує N Мбіт/с - сигнал тривоги». Такі детектори за граничними значеннями легко реалізувати, вони дають миттєве спрацювання при перевищенні відомих лімітів. Однак статичні пороги погано адаптуються до змінних умов: нормальний піковий трафік у вечірній час може перевищувати денний поріг і давати хибні спрацювання, і навпаки - повільно зростаюча атака може не перевищити грубий поріг. Тому порогові методи часто доповнюють адаптивними моделями: будується статистичний профіль нормальною поведінки (середні, стандартні відхилення метрик за попередній період) і пороги встановлюються динамічно як, наприклад, [середнє + 3σ]. Використовуються контрольні карти Шухарта, критерії Граббса, IQR (інтерквартильний розмах) та інші статистичні тести для виявлення «викидів» у рядах даних [11-12]. Більш чутливими є кумулятивні суми (CUSUM) - алгоритми, що відстежують накопичене відхилення метрики від базової лінії та сигналізують при виході за поріг. Порогові та прості статистичні методи швидкі та інтерпретовані (зрозуміло, який показник перевищив норму), але часто виявляють лише об’ємні аномалії (волюметричні атаки) і пропускають більш тонкі.

Ці підходи розглядають історичні дані трафіку як часовий ряд і намагаються передбачити «очікуване» значення у наступний момент, а потім порівнюють з фактичним. Якщо фактичне значно відрізняється - виникає деградація характеристик мережевого трафіку. Використовуються моделі прогнозування на кшталт ARIMA, експоненціального згладжування (модель Холта-Вінтерса) для сезонних даних, а також спектральні методи. Наприклад, трафік мережі часто має добову періодичність, тож модель Холта-Вінтерса добре прогнозує його нормальний хід і дозволяє помітити екстремальні відхилення [13].

Водночас, складність цих моделей зростає при наявності багатьох впливових факторів; вони потребують якісних історичних даних для тренування і можуть давати похибку прогнозу, яку важко відрізнити від справжньої аномалії.

Це підклас статистичних методів, що базуються на оцінці ентропії та споріднених мір невпорядкованості розподілів даних. Ідея в тому, що нормальний трафік має відносно стабільний розподіл певних характеристик (адрес, портів, пакетів), а аномалії порушують цю стабільність, змінюючи ентропію. Як було згадано, можна відстежувати ентропію IP-адрес джерел, адрес призначення, портів, типів протоколів тощо. Значні відхилення ентропії від історичних значень свідчать про структурні зміни трафіку. Наприклад, при DDoS-атаці ентропія адрес призначення різко падає (усі пакети на одну ціль), а при скануванні - зростає (багато різних цілей) порівняно з нормою. Інформаційно-теоретичні методи включають також дивергенцію Кульбака-Лейблера (KLD) - міру різниці між поточним розподілом і еталонним (нормальним) розподілом [14].

Вони розглядають вектор показників трафіку одночасно і шукають аномалії у багатовимірному просторі. Класичним підходом тут є методи головних компонент (PCA), запропоновані в мережевому контексті Лакхіною та ін. PCA дозволяє знизити вимірність даних (виділити кілька головних компонент, що пояснюють більшість варіацій нормального трафіку) і відфільтрувати «шум» - все, що не описується цими компонентами, вважається потенційно аномальним. Під час моніторингу нові вектори вимірювань проектуються на нормальний підпростір, і вимірюється квадрат відстані Махаланобіса (критерій Хотеллінга T2) або норма відхилення в ортогональному (аномальному) підпросторі; якщо ця величина перевищує поріг - реєструється деградація характеристик мережевого трафіку. Багатовимірні методи здатні врахувати кореляції між різними показниками. Наприклад, одночасне невелике зростання трафіку на порту 80 і спад на порту 443 може в сумі бути аномальним з точки зору співвідношення, хоча окремо кожна зміна не перевищує порогу [15].

Такий підхід успішно використовується в системах MSPC (Multivariate Statistical Process Control), запозичених з промислових процесів, і отримав розвиток у мережевій сфері. Недоліком PCA є чутливість до вибору параметрів і до «маскування» аномалій при навчанні: якщо в навчальних даних присутня крупна аномалія, метод може помилково прийняти її за частину норми. Попри це, при

належній реалізації (врахування методології MSPC) PCA показав свою ефективність для мережевих задач, особливо в поєднанні з методами передобробки (нормалізація даних, виділення трендів). Окрім PCA, до багатовимірних статистичних відносять і пряме застосування критеріїв для багатовимірних розподілів - наприклад, критерій Хотеллінга T2 може застосовуватися без пониження розмірності: будується модель багатовимірного нормального розподілу для вектора ознак трафіку, і кожен новий вектор тестується на ймовірність належати до цього розподілу. В експериментах показано, що такий підхід може досягати високих показників виявлення при низькому рівні хибнопозитивних спрацьовувань [16].

1.4 Ентропійні методи аналізу мережного трафіку: особливості застосування, переваги та обмеження

Поняття ентропії у мережевому трафіку. Ентропія в інформаційній теорії - це міра невизначеності або хаотичності розподілу випадкової величини. Для дискретної випадкової величини X, що набуває значень x з імовірностями p(x), класична ентропія Шеннона визначається як :

$$H(X) = -\sum xp(x)log2p(x). \quad (1.1)$$

В контексті мережевого трафіку ентропія використовується для кількісної оцінки «розмаїття» певних характеристик трафіку [17]. Наприклад, якщо розглядати розподіл IP-адрес джерел у потоці пакетів, то ентропія покаже, наскільки рівномірно розподілені пакети між різними джерелами. Високе значення ентропії означає, що багато різних адрес рівноправно присутні (система дуже хаотична, невизначена), низьке значення - що домінують кілька адрес, і розподіл впорядкований (малий хаос). Формально, ентропія досягає максимуму Hmax = log2 N, коли всі N можливих значень рівноймовірні, і мінімуму 0, коли вся маса ймовірності зосереджена в одному значенні. Таким чином, ентропія дає одне число,

яке характеризує структуру трафіку з точки зору певного атрибуту. У застосуванні до мережевого моніторингу найчастіше обчислюють ентропію за такими атрибутами пакетів чи потоків:

- ентропія IP-адрес джерел (HsrcIP);
- ентропія IP-адрес призначення (HdstIP);
- ентропія номерів портів джерел (Hsport) та призначення (Hdport);
- ентропія протоколів (TCP, UDP, ICMP, ін.);
- інколи ентропія розмірів пакетів чи інших специфічних ознак.

Кожен з цих показників дає окремий погляд на «різноманітність» трафіку. В нормальному стані мережі ентропії знаходяться у деякому типовому діапазоні значень, що відповідає звичайній діяльності користувачів. Наприклад, для веб-сервера може бути нормально, що HdstIP низька (усі пакети йдуть до одного сервера), а HsrcIP висока (багато різних клієнтів). Система моніторингу може заздалегідь виміряти і зафіксувати цей профіль ентропій, тобто нормальні мінімальні та максимальні значення кожного показника. Виявлення аномалій за допомогою ентропії. При настанні аномалії розподіли атрибутів змінюються, що призводить до зміни відповідних ентропій. Метод ентропійного детектування полягає у відстеженні цих змін і виявленні, коли ентропія виходить за межі нормального діапазону [18].

Практична реалізація зазвичай така: мережевий трафік агрегується у певних інтервалах часу (наприклад, по 1 секунді або 1 хвилині), і для кожного інтервалу обчислюються значення ентропії за вибраними характеристиками. Далі ці значення порівнюються з еталонними (наприклад, ковзним середнім або профілем, збереженим за попередній період). Якщо ентропія вийшла за встановлений поріг аномальності - фіксується інцидент. Іноді замість жорстких порогів контролюються темпи зміни ентропії: наприклад, якщо ентропія за секунду впала більш ніж на 50% від середнього - це сигнал [19].

1.5 Аналіз існуючих підходів до виявлення аномальних станів трафіку в комп’ютерних мережах

Сучасні мережі захищаються та моніторяться за допомогою спеціалізованих систем, які реалізують методи, обговорені в попередньому розділі. Існує широкий спектр таких програмно-технічних засобів - від відкритих ПЗ для мережевого аналізу до потужних комерційних платформ і апаратних пристроїв, що здійснюють моніторинг трафіку в режимі реального часу. Розглянемо найбільш відомі рішення та їх характеристики. Сигнатурні системи виявлення атак (IDS/IPS). Найпоширенішими інструментами мережевої безпеки є IDS (Intrusion Detection System) та IPS (Intrusion Prevention System), що працюють за принципом зіставлення трафіку з базою сигнатур відомих атак. Приклад - відкрита система Snort та її аналоги (Suricata, Bro/Zeek) [20].

Snort - провідний open-source рушій, який використовує набір правил (сигнатур), що описують шаблони небезпечної активності (послідовності байтів, специфічні заголовки пакетів, черговість пакетів в сесії тощо). Snort проглядає кожен пакет мережі і при збігу з правилом генерує оповіщення або блокує пакет (в режимі IPS). Ці системи дуже ефективні проти відомих атак - їх перевага у точності (низький відсоток хибних спрацювань на шаблонних атаках) і зрозумілості (кожна спрацьована сигнатура відповідає конкретному описаному експлойту або вірусу). Проте, сигнатурні IDS не виявляють нових, невідомих атак або варіацій атак, що не мають записаної сигнатури [21].

Вони також можуть бути обходжені методами евразії (наприклад, фрагментація пакетів, шифрування корисного навантаження), якщо сигнатури не враховують таких трюків. З точки зору даної роботи, сигнатурні системи корисні для контрасту - вони являють собою інший полюс підходів, відмінний від аномалійного. Тому більшість комплексних засобів комбінують сигнатурний та поведінковий аналіз. Системи поведінкового аналізу та виявлення аномалій (NBA/NBAD, Network Behavior Anomaly Detection). Це засоби, котрі моніторять

мережевий трафік і виявляють відхилення від нормальної поведінки. Типовий представник - рішення класу Cisco Stealthwatch (Secure Network Analytics).

Stealthwatch аналізує телеметрію мережі (в основному NetFlow-записи) та проводить постійний моніторинг усього трафіку в реальному часі, будуючи базові лінії нормальної активності для кожного хоста. За допомогою контекстно-орієнтованого аналізу система автоматично виявляє аномальні поведінкові відхилення від цих базових профілів. Зокрема, Stealthwatch здатний визначати широкий спектр атак: від відомих (шкідливе ПЗ, DDoS) до нульових днів (невідомих атак) та внутрішніх загроз [22]. Це досягається за рахунок машинного навчання і евристик: система накопичує статистику (хто з ким зазвичай взаємодіє, які обсяги передає, які порти використовує) і при відхиленнях (нові незвичні зв'язки, збільшення трафіку до нетипових ресурсів, інші маркери) генерує алерти. Такі платформи, як Stealthwatch, стали популярними в корпоративному сегменті, оскільки дозволяють бачити атаки, які пройшли повз сигнатурний захист, і в цілому підвищують видимість мережевих процесів. Аналогічні функції пропонуються і іншими виробниками: напр. Darktrace - рішення на базі штучного інтелекту, яке позиціюється як «мережева імунна система». Darktrace використовує неконтрольоване машинне навчання (cluster analysis, deep learning) для побудови нормальної моделі мережевого середовища і здатне виявляти аномалії без попередніх знань про атаки [23].

В оглядах ринку Darktrace називають лідером в галузі поведенкового мережевого моніторингу, керованого штучним інтелектом. Його характерна риса - самонавчання: система протягом декількох тижнів «вивчає» мережу, а потім уже може сигналізувати про будь-які відхилення (напр., новий пристрій, який почав сканувати мережу, чи користувач, який раптом завантажує нетипово великий обсяг даних). Подібні NDR (Network Detection and Response) рішення є і в інших постачальників - Aruba IntroSpect, Vectra AI, Microsoft Defender for IoT тощо - всі вони засновані на ідеї аналізу поведінки. Інструменти потокового аналізу. Ближче до мережевого рівня діють системи, що аналізують NetFlow/sFlow дані для виявлення аномалій. Частково ми вже згадали Stealthwatch (NetFlow-нагрузка),

існують також open-source інструменти: наприклад, NfSen/NFDUMP з плагінами для виявлення аномалій, ntopng - моніторинг, який може показувати підозрілу активність [24].

В наукових колах створювались прототипи, як-от система від Tamura et al. (в рамках LADS - Lightweight Anomaly Detection System), де NetFlow-записи оброблялися та сортувалися за відхиленнями (heavy hitters, ентропія) і в режимі реального часу показували потенційні атаки. Ці інструменти часто використовують ентропійний аналіз. Як ми розглядали, програмні комплекси на основі NetFlow можуть підраховувати ентропію по потоку даних з маршрутизаторів та спрацьовувати при її аномальних коливаннях. Наприклад, у роботі описано прототип, де NetFlow-дані збираються з мережі (через TAP/SPAN порти), заносяться в БД, і потім аналізуються: фільтруються потоки за напрямками, протоколами, обчислюються значення ентропії; на етапі детектування ентропія порівнюється з профілем і при виході за межі [min, max] - фіксується аномалія. Порогові значення ентропії наводяться: нижче 0 або вище 1 (очевидно, після нормування) - ознака ненормальної концентрації чи розпорошеності трафіку. Такий підхід довів свою дієвість для DDoS: повідомляється, що значення ентропії під час атаки відрізняються на 90% від нормальних, що легко вловити. Апаратні засоби. Для операторів зв’язку та великих мереж часто використовуються апаратні рішення для захисту від DDoS та інших аномалій [25].

Приклад - системи Arbor Networks (нині Netscout) Pravail APS або Cisco Guard, Radware DefensePro тощо. Вони встановлюються на периметрі мережі і відстежують статистику на лініях зв’язку в реальному часі, використовуючи апаратне прискорення, щоб аналізувати пакети на високих швидкостях (десятки Гбіт/с). Такі пристрої застосовують як сигнатурні (відома атака - шаблон), так і аномалійні методи: наприклад, можуть автоматично виявити аномальні сплески трафіку певного типу і включити фільтрацію. Деякі використовують адаптивні фільтри: спочатку помічається, що, скажімо, трафік UDP на порт X виріс у 100 разів, тоді пристрій починає відкидати частину цього трафіку або перенаправляти його на «scrubbing center» - центр очищення, де трафік ретельніше аналізується.

Такі системи фокусуються переважно на волюметричних аномаліях (DDoS), менше - на «тонких» проникненнях. SIEM та засоби кореляції подій [26].

Хоча не є чисто мережевими продуктами, системи управління подіями безпеки (Splunk, IBM QRadar, ArcSight) також відіграють роль у виявленні аномального мережевого трафіку. Вони збирають різноманітні логи (в тому числі з мережевих пристроїв) і можуть генерувати оповіщення на основі кореляції подій і поведінкових правил. Наприклад, SIEM може помітити, що «хост А почав встановлювати з'єднання з нетиповою кількістю зовнішніх адрес» - це вже індикатор аномалії, який SIEM визначив на основі аналізу записів файрвола чи проксі. Зараз у багатьох SIEM з'являються модулі UEBA (User and Entity Behavior Analytics), що застосовують машинне навчання для виявлення аномалій у поведінці користувачів та пристроїв. Це фактично аналог NBAD, але на рівні логів і подій. Open-source інструменти для наукових досліджень. Окрім Snort/Suricata, треба згадати Zeek (раніше Bro) - потужний фреймворк мережевого моніторингу. Zeek не використовує сигнатури трафіку, натомість це скриптова платформа: вона аналізує мережеві протоколи і надає аналітику високорівневі події (сесія встановлена, файл передано, DNS-запит тощо). На основі цих подій можна писати скрипти, що детектують аномальні ситуації (наприклад, якщо один клієнт відкрив 1000 сесій за хвилину - потенційне сканування, чи якщо з хоста пішов трафік на непритаманні порти). Zeek надає гнучкість для реалізації різних методів, у тому числі аномалійних [27-29].

Хоча Zeek менш відомий поза колом спеціалістів, він широко використовується в дослідницьких проектах і навіть комерційних рішеннях як складова. Існують і вузькоспеціалізовані open-source бібліотеки, наприклад, PyOD (Python Outlier Detection) - набір алгоритмів для пошуку аномалій, який застосовується дослідниками при побудові своїх рішень (але сам по собі не є завершеним інструментом). Так само sklearn та TensorFlow/PyTorch використовуються, щоб розробляти ML-моделі, які потім інтегрують у системи моніторингу. Порівняння та висновки щодо існуючих засобів.

1.6 Багатовимірна математична статистика в аналізі мережного трафіку

Багатовимірний аналіз трафіку означає розгляд одразу кількох метричних характеристик або ознак трафіку спільно, з метою виявити аномальні взаємозв’язки чи відхилення у багатовимірному просторі. На відміну від одновимірних методів (де кожна метрика аналізується незалежно, наприклад окремо порівнюється трафік на порту 80 з порогом, або ентропія адрес з нормою), багатовимірні підходи дозволяють вловити кореляційні залежності між характеристиками та помітити аномалії, які проявляються лише у комбінації показників. Передумови використання багатовимірної статистики. У нормальному режимі роботи мережі багато показників трафіку мають узгоджену поведінку [30].

Наприклад, якщо збільшується кількість активних сесій, зазвичай зростає і обсяг переданих даних; якщо на веб-сервері основний трафік переключився з HTTP (порт 80) на HTTPS (порт 443), то спостерігається одночасне зниження трафіку на 80 і підвищення на 443 - поодинці ці зміни можуть не перевищувати пороги, але разом формують характерний шаблон. Аномалії часто руйнують типовий кореляційний патерн. Наприклад, при деяких атаках-спуфінгах (підробці IP-адрес) може різко зрости число унікальних IP-джерел, але не змінитися кількість MAC-адрес на канальному рівні - у нормі таких розбіжностей не буває, бо зазвичай кожна IP відповідає реальному пристрою з унікальним MAC. Неконсистентність - приклад багатовимірної аномалії. Щоб її зафіксувати, потрібно аналізувати зв’язок між двома різними параметрами (IP та MAC) одночасно. Метод головних компонент (PCA). Серед багатовимірних методів в аналізі мережевого трафіку найбільш відомий підхід на основі аналізу головних компонент, запропонований Лакхіною та співавт. і вперше застосований до мережних вимірювань у середині 2000-х. PCA - це техніка зменшення розмірності даних: маючи набір спостережень векторного параметра x =(x1 ,x2,...,xm), вона шукає лінійні комбінації компонент, які найкраще пояснюють варіацію даних. Перші кілька головних компонент (ГК) зазвичай пояснюють більшу частину змін у даних, і в припущенні, що ці зміни відповідають нормальній поведінці, їх називають нормальним підпростором [31].

Якщо відхилення велике - xnew є аномалією. У контексті мережі як вектор x можна брати, наприклад, набір показників трафіку за певний інтервал часу: x1 - обсяг трафіку на 1-му лінку, x2 - на 2-му, ..., xm - на m-му. Лакхіна та ін. застосували PCA до матриці трафіку між джерелами і призначеннями (Origin-Destination flows) у операторській мережі, де кожен вимір - це обсяг трафіку між певною парою вузлів. Їм вдалося показати, що мережевий трафік має дійсно низькорозмірну структуру: кілька головних компонент (наприклад, 5-10) пояснюють ~95% варіації, і ці компоненти відповідають регулярним патернам навантаження (добові цикли, постійні великі потоки) [32].

Подальші роботи з вдосконалення PCA для мереж (наприклад, Camacho et al., 2015) врахували принципи MSPC: з центруванням даних, масштабуванням, побудовою контрольних карт для статистик T2 (для внутрішнього підпростору) та Q (норма залишку поза підпростором), адаптивним оновленням моделі при зміні мережевих умов. Сьогодні PCA розглядають як один з елементів комплексних систем аналізу, котрий дає змогу зменшити шум і виділити потенційно підозрілі напрями варіації, але зазвичай у парі з іншими методами. Критерій Хотеллінга T2. Це класичний статистичний критерій для багатовимірних даних, який має пряму геометричну інтерпретацію: T2 вимірює відстань точки від центру розподілу з урахуванням коваріацій [33]. Якщо дані x мають багатовимірний нормальний розподіл з середнім μ і коваріаційною матрицею S, то статистика Хотеллінга обчислюється як $T2 = (x-\mu)TS-1(x-\mu)$.

Так у роботі Lee et al. (2008) застосували кластеризацію для виявлення DDoS-атак: нормальні профілі вузлів утворили кластери, а заражені машини з флуктуаціями трафіку виділялися як окремі точки. Ефективність кластеризації сильно залежить від правильного вибору ознак і попередньої обробки, і зазвичай вона не дає настільки чіткої межі між нормою й аномалією, як статистичні методи з чіткими порогами. Також можна згадати байєсівські моделі - побудову багатовимірного ймовірнісного розподілу нормального трафіку (можливо, непараметричного, з використанням ядерного оцінювання щільності) і обчислення апріорної ймовірності для нових точок. Цей напрям менш популярний, бо висока

розмірність ускладнює щільнісне оцінювання. Частково його замінили автоенкодери та інші моделі глибокого навчання, які фактично теж оцінюють щільність у прихованій формі. Комбінування багатовимірного підходу з ентропійним. Варто відзначити, що підходи не є взаємовиключними [34]

Переваги багатовимірних методів:

(1) Вони враховують комплексну картину і можуть значно знизити кількість хибних спрацювань, відсікаючи відхилення, які не підтверджуються іншими показниками. Як зазначається, багато ML-методів окремо можуть давати багато помилкових аномалій, бо не враховують взаємозв'язки між ознаками. Багатовимірний підхід ці взаємозв'язки явно моделює;

(2) Вони виявляють аномалії, невидимі в одному вимірі - це особливо стосується малопотужних, але високоспецифічних атак;

(3) При правильній статистичній інтерпретації (Hotelling T2 чи контрольні карти) можна встановити чіткі пороги з заданим рівнем значущості, тобто калібрувати систему під прийнятний рівень помилкових спрацювань;

(4) Можна ідентифікувати тип аномалії через аналіз вкладення ознак: наприклад, PCA дає власні вектори, за якими можна приблизно зрозуміти, яка комбінація показників «стріляє» - це допомагає в діагностиці (хоча й складніше, ніж з ентропією, але все ж).

Недоліки багатовимірних методів:

(1) Вони можуть бути чутливі до високовимірності та шуму. Якщо включити забагато нерелевантних ознак, модель «розмивається» і може не виявити нічого (або треба дуже великий обсяг навчальних даних). Тому завжди постає задача вибору ознак або розмірності;

(2) Обчислювальна складність - обчислення коваріацій, власних векторів, інверсії матриць - усе це може бути складно реалізувати в режимі реального часу на швидкостях багатогігабітних потоків;

(3) Адаптивність - мережа змінюється, кореляції можуть дрейфувати (наприклад, впровадження нового сервісу змінить структуру трафіку одразу по

кількох показниках). Модель треба періодично оновлювати, причому обережно, щоб не навчитися на аномалії;

(4) Складність впровадження на відміну від простих порогів, багатовимірні методи важче налаштувати і пояснити операторам. Це впливає на довіру до системи: адміністратор може воліти простіший метод, який він розуміє, навіть якщо той менш точний.

1.7 Висновки та постановка задачі

У першому розділі здійснено поглиблений аналіз мережевого трафіку як об’єкту моніторингу, розглянуто природу можливих аномалій та сучасні методи і засоби їх виявлення. Комп’ютерний мережевий трафік володіє статистичною структурою і типовими патернами (циклічністю, розподілом адрес та портів), які порушуються під час аномалій. Аномалії у трафіку поділяються на кілька класів - мережеві відмови, флеш-навантаження, кібератаки, помилки вимірювання - кожен з яких має свої прояви у часових (сплески, провали) та статистичних характеристиках (відхилення середніх, дисперсій, ентропії, кореляцій). Це підтверджує доцільність багатостороннього моніторингу трафіку за різними показниками, аби вчасно виявляти різні типи інцидентів. Існує широкий спектр методів: від простих порогових до складних алгоритмів штучного інтелекту. Порогові та одномірні статистичні методи легкі у реалізації, але неефективні проти хитромудрих або малопомітних аномалій. Сигнатурні IDS відмінно ловлять відомі атаки, але нездатні знайти нові загрози. Інформаційно-теоретичні (ентропійні) методи показали себе дієвими у виявленні структурних аномалій, даючи інтерпретовані ознаки атак типу DDoS, сканування тощо. Багатовимірні статистичні методи (PCA, Hotelling T2) дозволяють відстежувати комплексні відхилення й знижувати хибні тривоги за рахунок врахування кореляцій між показниками. Машинне навчання і особливо глибокі нейромережі здатні до високої адаптивності та детектування навіть слабовиражених аномалій, проте викликають проблеми з поясненістю результатів і потребують великих даних для навчання.

Жоден окремий метод не є панацеєю - нагальною є задача інтеграції кількох підходів, щоб використати їх переваги компенсуючи недоліки. Проаналізовано реальні системи моніторингу: від Snort/Suricata (сигнатури) до Cisco Stealthwatch і Darktrace (поведінковий аналіз на основі статистики та AI). Встановлено, що сучасні системи дедалі більше зміщуються у бік ентропійно-статистичних та ML методів, вводячи функції базового профілю трафіку, виявлення аномальної поведінки хостів і застосування алгоритмів навчання для автоматики. Втім, багато з цих рішень - комерційні закриті платформи; у відкритому доступі є окремі інструменти, проте відсутнє готове інтегроване рішення, яке б виконувало, наприклад, виявлення аномалій на базі ентропійних ознак у багатовимірному просторі в режимі реального часу. Це вказує на науково-практичну нішу, яку доцільно заповнити.

Виявлені тенденції та прогалини дозволяють сформулювати основну мету і завдання даної дипломної роботи. Загальна мета полягає у розробці методу та програмно-технічних засобів аналізу мережевого трафіку на основі ентропійних характеристик та багатовимірної математичної статистики. Для досягнення цієї мети необхідно вирішити такі конкретні завдання:

Потрібно визначити, які ентропійні показники (розподіли IP, портів, ін.) є найбільш інформативними для різних типів аномалій, та як інтегрувати їх у багатовимірну модель. Планується застосувати методи MSPC (Multivariate Statistical Process Control) - зокрема, критерій Хотеллінга або PCA - до вектора ентропій і інших статистичних ознак трафіку, щоб отримати узагальнений критерій аномальності трафіку.

Алгоритм має працювати в режимі онлайн (або наближеному до реального часу) на вхідному потоці даних, обчислювати необхідні статистики і ентропії за рухомими вікнами часу, порівнювати з нормальним профілем та приймати рішення про наявність аномалії. Слід передбачити механізми адаптації профілю під поступові зміни трафіку, щоб зменшити залежність від початкового навчання.

2 РОЗРОБЛЕННЯ МОДЕЛІ АНАЛІЗУ ТРАФІКУ КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ТА БАГАТОВИМІРНОЇ МАТЕМАТИЧНОЇ СТАТИСТИКИ

2.1 Постановка задачі дослідження та формалізація процесу аналізу мережного трафіку

Метою другого розділу є побудова формальної моделі аналізу мережного трафіку, яка пов'язує спостережувані мережеві записи з інтегральною оцінкою стану мережі. На відміну від оглядового першого розділу, де розглянуто природу аномалій і сучасні підходи до їх виявлення, у цьому розділі предметна область переводиться у формалізований опис. Такий перехід є необхідним, оскільки подальше розроблення методу в третьому розділі можливе лише тоді, коли чітко визначено, що саме вважається об'єктом спостереження, у якій формі подається стан трафіку та за якими ознаками приймається рішення про нормальний або аномальний режим функціонування мережі.

Об'єктом моделювання у даній роботі є мережевий трафік комп'ютерної мережі як послідовність подій передавання даних, що спостерігаються упродовж часу. Залежно від режиму збору інформації окремою подією може бути пакет, запис потоку або агрегований запис сеансу. Для цілей моделювання суттєвим є не формат подання, а наявність часової мітки та набору атрибутів, за якими можливо побудувати статистичний опис трафіку. До таких атрибутів належать адреси джерела і призначення, порти, протоколи, обсяги переданих байтів, кількість пакетів, тривалість потоку та інші доступні характеристики, які не вимагають аналізу корисного навантаження.

Оскільки миттєвий стан трафіку не може бути надійно оцінений за одиничною подією, базовою одиницею спостереження доцільно вважати часове вікно. У межах кожного вікна збирається множина мережевих записів, для якої обчислюються статистичні та ентропійні характеристики. Такий підхід дозволяє згладити випадкові флуктуації окремих пакетів, перейти від рівня подій до рівня станів і зробити подальший аналіз стійкішим до випадкових коливань. Крім того,

віконна організація є природною для задачі онлайн-моніторингу, оскільки дає змогу періодично оновлювати оцінку стану мережі без повторної обробки всієї історії спостережень.

Нехай W(t) позначає множину записів трафіку, що потрапили у часове вікно з центром або початком у момент t. Тоді кожному вікну ставиться у відповідність вектор ознак X(t) = {x1(t), x2(t), ..., xm(t)}, де m - кількість інформативних характеристик, обраних для опису стану трафіку. Компоненти цього вектора можуть мати різну природу: частина з них відображає обсяг та інтенсивність трафіку, частина - різноманітність структурних характеристик, а частина - відхилення поточних розподілів від еталонного профілю. Вектор X(t) розглядається далі як формальний образ стану мережі у даному вікні спостереження.

$$X(t) = \{x1(t), x2(t), \ldots, xm(t)\} \tag{2.1}$$

Задача аналізу трафіку в такій постановці зводиться до побудови відображення F, яке за вектором ознак X(t) формує оцінку стану S(t). У найпростішому випадку S(t) належить множині {N, A}, де N відповідає нормальному режиму функціонування, а A - аномальному. У розширеній постановці оцінка може бути багаторівневою і відображати не тільки сам факт відхилення, а й ступінь його вираженості, тобто рівень ризику або інтенсивність аномального процесу. Інтерпретація є корисною для практичного моніторингу, оскільки дозволяє відокремлювати слабкі відхилення від критичних інцидентів.

Під аномальним станом у роботі розумітиметься статистично значуще відхилення вектора X(t) від профілю нормального трафіку, сформованого за даними штатного режиму роботи мережі. Важливо підкреслити, що мова йде не про довільну незвичну подію, а саме про таку зміну структури або інтенсивності трафіку, яка порушує характерні взаємозв’язки між ознаками. Завдяки цьому модель орієнтується не лише на великі сплески навантаження, а й на приховані

зміни розподілів, які можуть супроводжувати сканування, низькоінтенсивні атаки, деградацію сервісів або збої мережевої інфраструктури.

Формалізація процесу аналізу трафіку повинна також враховувати практичні обмеження. По-перше, модель має спиратися на дані, які реально можна збирати на маршрутизаторі, комутаторі, мережевому сенсорі або системі потокового експорту, не вдаючись до дорогої повної інспекції вмісту пакетів. По-друге, вона має бути придатною для роботи в умовах нестаціонарності трафіку, коли профіль нормального стану змінюється внаслідок добових циклів, зміни кількості користувачів або введення нових сервісів. По-третє, модель повинна забезпечувати таке подання стану, яке може бути використане як у режимі офлайн-дослідження, так і в режимі, наближеному до реального часу.

З урахуванням зазначеного процес аналізу трафіку в загальному вигляді доцільно подати як послідовність взаємопов’язаних перетворень.

На першому етапі мережеві записи нормалізуються та впорядковуються за часом. На другому етапі вони агрегуються у часові вікна.

На третьому етапі для кожного вікна будуються емпіричні розподіли за обраними атрибутами, після чого обчислюються ентропійні, дивергентні та агреговані статистичні ознаки. На четвертому етапі ознаки поєднуються у багатовимірний вектор стану, що порівнюється з моделлю нормального режиму (рисунок 2.1).

Таким чином, у межах другого розділу предметна задача виявлення аномалій переводиться у формальну задачу побудови моделі станів мережного трафіку. Модель повинна поєднати часову агрегацію, статистичний опис структури трафіку, формування вектора ознак і правило визначення відхилення від норми. Подальші підрозділи конкретизують ці елементи, починаючи з обґрунтування вибору ентропійних характеристик і закінчуючи побудовою математичної та структурної моделі аналізу.

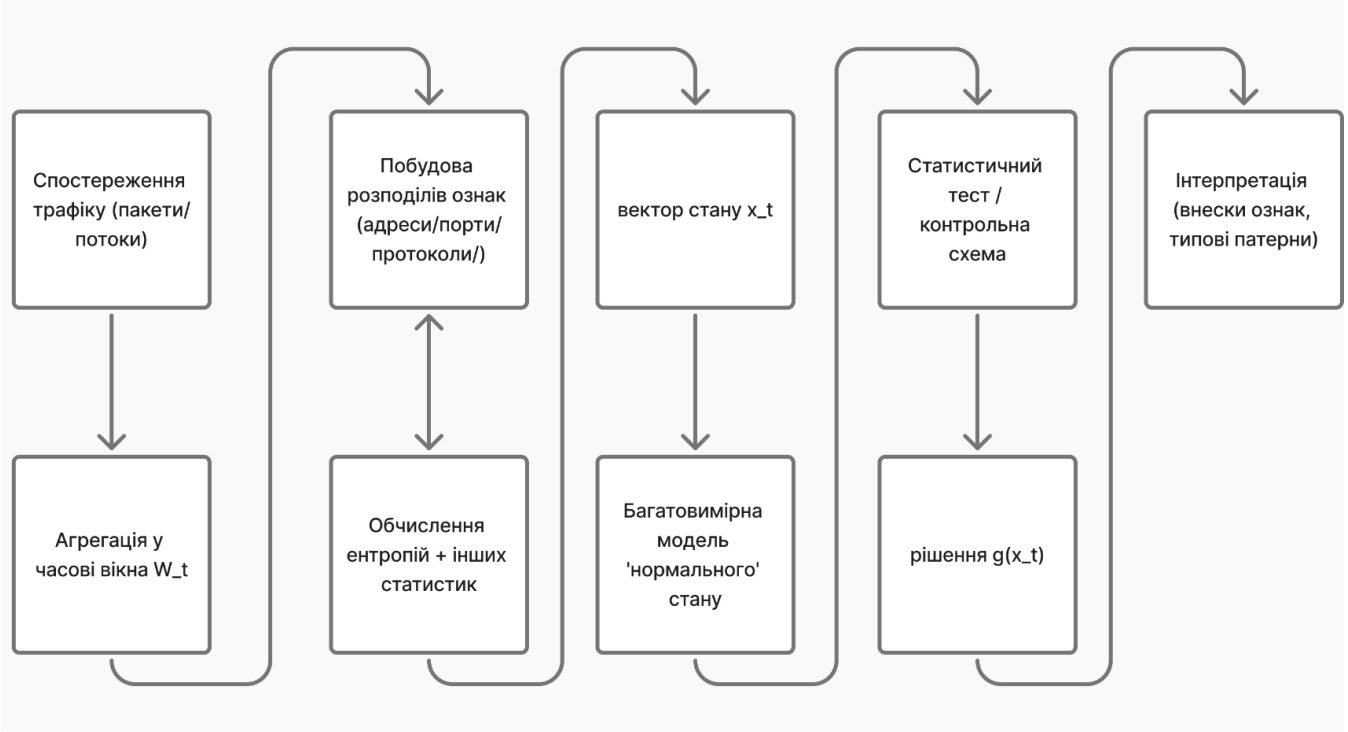


Рисунок 2.1 – Узагальнена схема процесу аналізу мережного трафіку

2.2 Обґрунтування вибору ентропійних характеристик і багатовимірних статистичних методів для аналізу трафіку

Ключова ідея пропонованої моделі полягає в тому, що стан мережі не можна адекватно описати лише однією об’ємною характеристикою, наприклад кількістю пакетів або байтів за інтервал часу. Для реального трафіку важливою є не тільки інтенсивність, а й структура взаємодій між вузлами, сервісами та транспортними портами. У багатьох практичних сценаріях саме зміна структури трафіку є першою ознакою небажаного процесу: при DDoS-атаці різко зростає концентрація трафіку на певній цілі, при скануванні розширюється множина адрес або портів призначення, а при збої сервісу змінюється баланс між окремими типами з’єднань. Для побудови моделі стану потрібні характеристики, здатні відображати внутрішню організацію потоків даних.

Ентропійні характеристики відповідають цій вимозі, оскільки дають змогу стисло описати ступінь невизначеності та різноманітності розподілу ознак трафіку. Якщо у межах вікна спостереження розподіл значень певного атрибута є

рівномірним, ентропія набуває вищих значень. Якщо ж основна маса спостережень концентрується на небагатьох значеннях, ентропія зменшується. Інтерпретація добре відповідає природі мережевих аномалій, де багато сценаріїв супроводжуються або надмірною концентрацією, або навпаки нетиповим розсіюванням комунікацій. Тому ентропія є зручною базовою мірою для фіксації структурних змін трафіку.

Для дискретної ознаки a, що у вікні W(t) набуває значень з емпіричними ймовірностями $p1(t), p2(t), \ldots, pk(t)$, базова ентропійна характеристика може бути визначена у вигляді $H_a(t) = -\Sigma\, p_i(t)\, log2\, p_i(t)$. У цьому виразі кожна ймовірність відображає частку спостережень, що припадає на відповідне значення ознаки в межах поточного вікна. Чим більш рівномірно розподілені значення, тим більшою є ентропія. Якщо ж домінує невелика кількість значень, ентропія зменшується. Для практичного аналізу ця формула зручна тим, що не потребує складних припущень щодо закону розподілу і безпосередньо обчислюється з емпіричних частот.

$$H_a(t) = -\Sigma\, p_i(t)\, log2\, p_i(t) \qquad (2.2)$$

У контексті аналізу мережевого трафіку найбільшу інформативність мають ентропії адрес джерела та призначення, транспортних портів джерела і призначення, а також ентропія протоколів. Ентропія адрес джерела дозволяє оцінити ступінь розосередженості активності відправників. Ентропія адрес призначення є індикатором концентрації або розподілу навантаження за вузлами мережі. Ентропія портів відображає зміну сервісної структури трафіку, а ентропія протоколів дає змогу відстежувати нетипове зміщення між TCP, UDP, ICMP та іншими типами передавання. У сукупності ці показники формують стислий, але змістовний опис поведінки трафіку.

Разом з тим абсолютного значення ентропії не завжди достатньо для виявлення аномалій. У практиці моніторингу важливим є не тільки поточний рівень невизначеності, а й зміна форми розподілу відносно базового режиму. З цієї

причини доцільно розглядати також дивергентні характеристики, що порівнюють поточний емпіричний розподіл з еталонним або адаптивно оновлюваним профілем. Дивергенція Кульбака-Лейблера та дивергенція Дженсена-Шеннона дозволяють виміряти, наскільки поточний розподіл ознаки відійшов від нормального, навіть у тих випадках, коли саме значення ентропії змінилося незначно. Таким чином, ентропійні та дивергентні ознаки доцільно використовувати не ізольовано, а у взаємодоповнювальній ролі.

Ще однією причиною вибору ентропійного підходу є його придатність до роботи з даними, що агрегуються у часових вікнах. Побудова гістограм значень та обчислення ентропії не потребують збереження всіх сирих подій на етапі аналізу. Достатньо підтримувати лічильники частот, що особливо важливо для систем, орієнтованих на роботу в реальному або близькому до реального часі. Крім того, ентропія є інтерпретованим показником: її зростання або спад можна пов’язати зі зміною концентрації чи різноманітності трафіку, що полегшує подальше пояснення рішень системи моніторингу.

Однак мережевий трафік є багатофакторним об’єктом, і жодна окрема ентропійна характеристика не може повністю охопити всі прояви аномального стану. Наприклад, зменшення ентропії адрес призначення може сигналізувати про концентрацію трафіку на одній цілі, але без урахування інтенсивності це ще не дає повної картини. Аналогічно зростання ентропії портів може бути пов’язане як із легітимною активністю кількох сервісів, так і зі скануванням. Тому модель повинна розглядати набір ознак одночасно та оцінювати не лише їхні окремі значення, а й взаємозв’язки між ними.

Таку можливість надають методи багатовимірної математичної статистики. На відміну від одновимірного порогового аналізу, багатовимірний підхід працює в просторі ознак і враховує коваріаційну структуру даних. Це означає, що система реагує не тільки на значне відхилення окремої змінної, а й на нетипову комбінацію кількох ознак, кожна з яких окремо може залишатися в межах допустимого діапазону. Для мережевого моніторингу така властивість є принципово важливою,

оскільки багато атак і деградацій проявляються саме як порушення звичних співвідношень між параметрами трафіку.

У межах пропонованої моделі доцільно використовувати багатовимірні методи у двох ролях. По-перше, вони забезпечують стандартизацію та інтеграцію різнорідних ознак у єдиний простір аналізу. По-друге, вони формують статистику відхилення від моделі нормального стану. Для цього можуть застосовуватися коваріаційно-орієнтовані міри, зокрема відстань Махаланобіса або статистика Хотеллінга T^2. За великої кількості ознак або наявності сильної корельованості між ними доцільним є також використання аналізу головних компонент як засобу зниження розмірності й виділення інформативного підпростору станів.

Обґрунтування використання багатовимірних методів пов’язане ще й з тим, що вони дозволяють сформувати модель норми не як набір незалежних інтервалів для кожної ознаки, а як область у просторі ознак. Область відповідає типовим поєднанням значень у нормальному режимі. Якщо поточний вектор виходить за межі цієї області, фіксується відхилення. Геометрично це відповідає переходу від набору окремих порогів до багатовимірної поверхні прийняття рішення. Перевага такого підходу полягає в суттєво вищій чутливості до складних і комбінованих аномалій (рисунок 2.2).

Ентропійні характеристики доцільно використовувати як базовий інструмент виявлення змін структури трафіку, а багатовимірну статистику - як засіб інтегрального оцінювання стану мережі з урахуванням взаємозв’язків між ознаками. Поєднання цих двох складових створює підґрунтя для побудови моделі, що є водночас інтерпретованою, обчислювально придатною та чутливою до широкого спектра аномальних режимів. Комбінація надалі дозволяє перейти від набору окремих характеристик до цілісного опису стану трафіку, придатного для прийняття рішення.

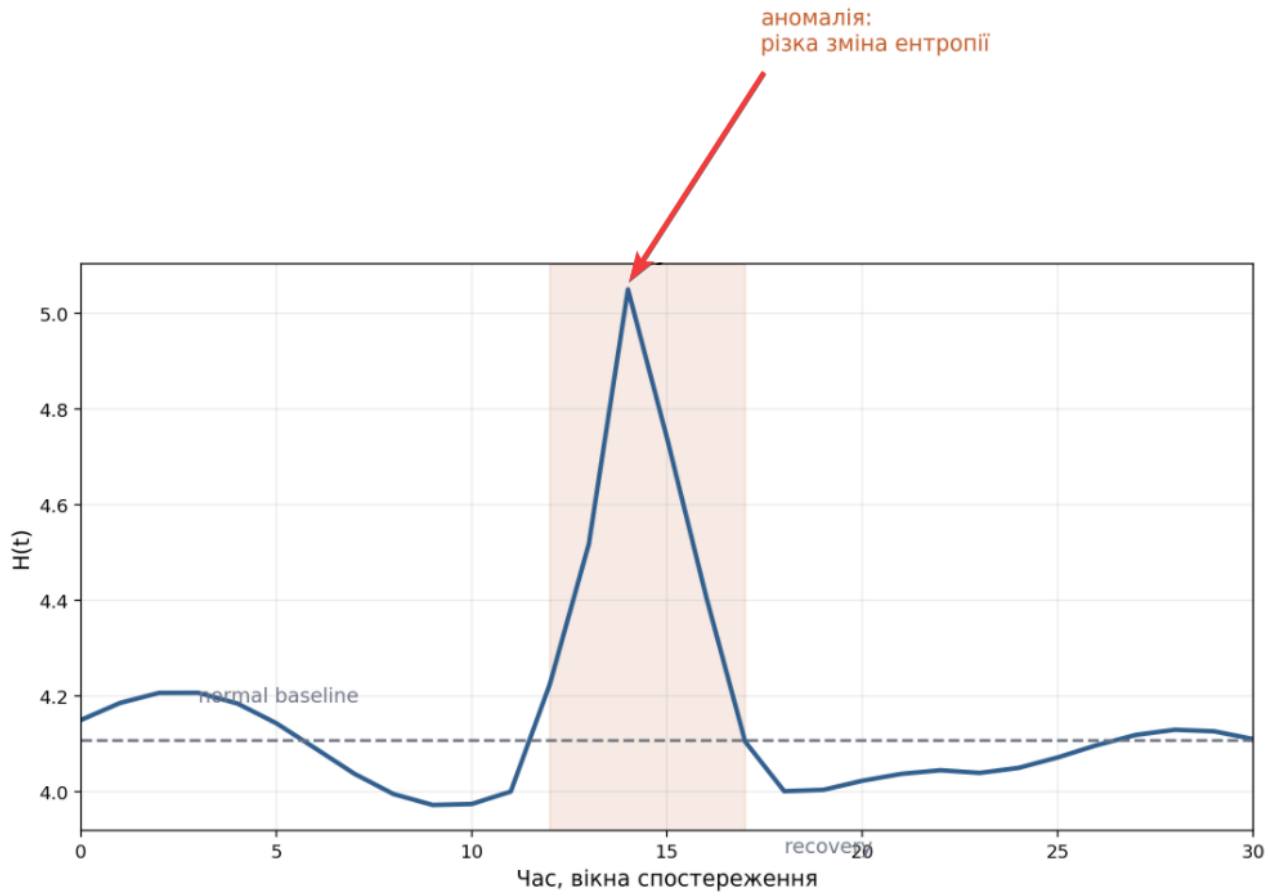


Рисунок 2.2 – Схематичний приклад профілю ентропії у ковзному часовому вікні

2.3 Формування системи інформативних ознак мережного трафіку

Після обґрунтування вибору ентропійного та багатовимірного апарату наступним кроком є формування системи інформативних ознак, тобто такого набору характеристик, який достатньо повно описує стан трафіку і водночас може бути обчислений на практиці. Система ознак не повинна бути надмірною, оскільки зайві або слабоінформативні параметри ускладнюють побудову моделі та знижують її стійкість. Водночас вона не може бути і надто вузькою, бо тоді частина аномальних проявів залишиться поза увагою. Ключовим завданням цього підрозділу є добір мінімально достатнього, але змістовного вектора характеристик стану мережевого трафіку.

Базовими критеріями відбору ознак є інформативність, обчислюваність, інтерпретованість і стійкість. Інформативність означає здатність ознаки реагувати на зміни режиму функціонування мережі. Обчислюваність вимагає, щоб значення

ознаки можна було отримати з доступних заголовкових полів пакетів або потокових записів без звернення до корисного навантаження. Інтерпретованість потрібна для того, щоб результати аналізу могли бути пояснені оператору мережі. Стійкість означає відносну нечутливість до малих випадкових флуктуацій, що не відображають реальних порушень у роботі мережі.

Першу групу інформативних ознак доцільно сформувати з об'ємних і інтенсивнісних характеристик. До неї належать кількість записів або пакетів у вікні, загальний обсяг переданих байтів, кількість потоків, середній розмір пакета, частота надходження нових з'єднань, тривалість активних потоків та інші агреговані показники. Ці ознаки характеризують рівень навантаження на мережу і дозволяють виявляти об'ємні аномалії, пов'язані з раптовим зростанням або спадом трафіку. Разом із тим вони не дають достатнього уявлення про внутрішню організацію комунікацій, тому повинні розглядатися лише як одна з частин загального вектора стану.

Другу групу становлять ентропійні структурні ознаки, які відображають різноманітність і розподіл категоріальних атрибутів. Для задачі аналізу мережевого трафіку доцільно виділити ентропію адрес джерела, ентропію адрес призначення, ентропію транспортних портів джерела і призначення, а також ентропію протоколів. За потреби ці характеристики можуть будуватися не лише для повних адрес, а й для їхніх агрегованих подань, наприклад підмереж або класів сервісів. Такий підхід дозволяє зробити аналіз стійкішим у мережах із великою кардинальністю адресного простору. Структурні ентропійні характеристики є особливо корисними для фіксації концентрації трафіку, нетипового розпорошення активності або різкої зміни сервісної структури взаємодій.

Третю групу доцільно сформувати з порівняльних характеристик, що описують відмінність поточного розподілу від базового профілю. Якщо у вікні спостереження для певної ознаки побудовано емпіричний розподіл, його можна зіставити з еталонним розподілом, сформованим на навчальному фрагменті нормального трафіку або отриманим шляхом адаптивного згладжування. Результатом такого зіставлення стає числова міра відхилення, зокрема дивергенція

Кульбака-Лейблера чи Дженсена-Шеннона. На відміну від ентропії, яка відображає внутрішню невизначеність поточного розподілу, дивергенція фіксує саме зсув відносно норми. Включення цих характеристик до вектора стану підсилює чутливість моделі до повільних, але систематичних змін профілю трафіку.

Четверту групу ознак утворюють додаткові показники різноманітності й співвідношень, які добре доповнюють ентропійний опис. До них належать кількість унікальних адрес джерела і призначення, кількість унікальних портів, співвідношення вхідного та вихідного трафіку, частка коротких або довгих потоків, співвідношення TCP і UDP, а також інші агрегати, чутливі до специфічних сценаріїв роботи мережі. Наприклад, distinct count адрес і портів дає пряме уявлення про число унікальних значень у вікні, тоді як ентропія доповнює його інформацією про те, наскільки рівномірно розподілені спостереження між цими значеннями. Завдяки поєднанню цих ознак модель уникає ситуації, коли подібні значення ентропії приховують різну реальну структуру трафіку.

Формування системи ознак вимагає також врахування масштабу мережі та характеру спостережуваних даних. У великих мережах із високою кардинальністю адрес і портів повна деталізація всіх значень може призводити до надмірних обчислювальних витрат. Тому на етапі побудови моделі доцільно передбачити узагальнення окремих полів або дискретизацію числових характеристик. Узагальнення сформованої системи інформативних ознак наведено в таблиці 2.1.

Після вибору окремих характеристик необхідно визначити принцип формування підсумкового вектора стану. Для кожного вікна спостереження всі обчислені ознаки об’єднуються в один багатовимірний опис X(t). При цьому вектор повинен містити як показники інтенсивності, так і структурні ентропійні ознаки та показники відхилення від базового профілю. Комбінована побудова є принциповою: якщо залишити лише об’ємні характеристики, модель втрачатиме чутливість до прихованих структурних змін; якщо залишити лише ентропії, вона може недооцінювати великі об’ємні аномалії. Поєднання різних типів ознак забезпечує повніший опис стану мережі.

Таблиця 2.1 – Групи інформативних ознак мережного трафіку та їх аналітичне призначення

Група ознак	Приклади показників	Аналітичне призначення
Об’ємні та інтенсивнісні	Кількість пакетів, байтів, потоків, середній розмір пакета	Характеризують загальне навантаження та інтенсивність обміну даними
Ентропійні структурні	Ентропія srcIP, dstIP, srcPort, dstPort, protocol	Відображають рівень концентрації або розсіювання трафіку за категоріальними атрибутами
Порівняльні	KL-divergence, Jensen–Shannon divergence, відхилення від базового рівня	Показують міру зміни поточного розподілу відносно еталонного профілю
Додаткові поведінкові	Кількість унікальних адрес і портів, співвідношення вхідного й вихідного трафіку	Уточнюють характер аномалії та доповнюють ентропійний опис трафіку

Система інформативних ознак у межах пропонованої моделі повинна включати кілька взаємодоповнювальних груп характеристик і будуватися за принципом мінімальної достатності. Вона має забезпечувати можливість практичного обчислення, інтерпретації та подальшого багатовимірного аналізу. Сформований у такий спосіб вектор ознак є основою математичної моделі станів мережного трафіку, побудова якої розглядається в наступному підрозділі.

2.4 Розроблення математичної моделі опису станів мережного трафіку

Після визначення складу інформативних ознак постає задача побудови математичної моделі, яка формально описує нормальний і аномальний стани

мережного трафіку. На цьому етапі необхідно перейти від словесного опису поведінки мережі до компактного математичного представлення, придатного для подальшого статистичного аналізу. Центральною ідеєю є розгляд кожного часового вікна як точки у багатовимірному просторі ознак. Якщо на основі вибірки штатних режимів функціонування мережі побудувати модель нормального профілю, тоді відхилення нових точок від цього профілю можуть використовуватися як формальна ознака аномального стану.

Нехай X(t) є m-вимірним вектором ознак, сформованим для вікна W(t). Сукупність векторів, отриманих на фрагментах трафіку, що інтерпретуються як нормальний режим роботи мережі, утворює навчальну вибірку нормального стану. На її основі оцінюються середній вектор μ та коваріаційна матриця S, які описують характерне положення і форму області нормального трафіку в просторі ознак. Таким чином, нормальний стан подається не окремим числом чи набором незалежних порогів, а багатовимірним статистичним профілем, що враховує взаємозв’язки між ознаками.

У найпростішому варіанті модель нормального режиму може бути записана як $X(t) = \mu + \varepsilon(t)$, де $\varepsilon(t)$ описує природні коливання ознак відносно базового профілю. Якщо припустити, що такі коливання мають скінченну коваріацію і переважно залишаються в межах характерної області нормального стану, тоді задача виявлення аномалій зводиться до оцінювання міри віддаленості поточного вектора X(t) від профілю (μ, S). Чим більшою є ця віддаленість, тим менш імовірним є припущення, що поточне вікно відповідає штатному режиму роботи мережі.

Для багатовимірного вимірювання відхилення доцільно використовувати коваріаційно-орієнтовану статистику. Найбільш природною у цій постановці є квадратична форма $D^2(t) = (X(t) - \mu)^T S^{-1} (X(t) - \mu)$, яка є узагальненням евклідової відстані на випадок корельованих ознак. На відміну від покомпонентного порівняння з незалежними порогами, така статистика враховує напрямки природної мінливості та силу зв’язку між окремими характеристиками. Якщо дві ознаки за нормального режиму змінюються узгоджено, то модель не буде

трактувати їхню спільну зміну як аномалію, доки вона залишається в межах характерної коваріаційної структури.

$$D^2(t) = (X(t) - \mu)^T S^{-1}(X(t) - \mu), \quad (2.3)$$

Порогове правило класифікації в цій постановці можна записати як S(t) = N, якщо $D^2(t) \leq h$, і S(t) = A, якщо $D^2(t) > h$, де h є граничним значенням статистики відхилення. Вибір порога може здійснюватися або на основі теоретичних наближень, або емпірично за квантилями статистики, отриманої на навчальній вибірці нормальних вікон. Другий підхід часто є більш практичним, оскільки реальний мережевий трафік не завжди задовольняє строгим припущенням багатовимірної нормальності. Водночас сам формальний принцип залишається незмінним: аномалія визначається як вихід точки за межі області, характерної для нормального профілю мережі.

Особливістю мережевого трафіку є наявність корельованих і частково надлишкових ознак. Наприклад, кількість пакетів, обсяг байтів, кількість унікальних портів та ентропія портів часто змінюються не незалежно, а узгоджено. У таких умовах математична модель може бути підсилена використанням аналізу головних компонент. PCA дозволяє перейти від початкового простору ознак до нового ортогонального базису, в якому основна варіативність нормального режиму описується невеликою кількістю компонент. У межах моделі це дає змогу або зменшити розмірність вектора стану, або розділити простір на нормальний та залишковий підпростори для точнішого виявлення аномалій.

Застосування PCA доцільне насамперед тоді, коли розмірність вектора стану є значною або коли між ознаками існує сильна лінійна залежність. Після стандартизації початкових ознак вектор X(t) проектується на простір головних компонент, де формується нове подання Z(t). Частина компонент, що пояснює основну частку дисперсії навчальної вибірки, інтерпретується як нормальний підпростір. Відхилення у цьому просторі можуть оцінюватися за статистикою

Хотеллінга T^2, тоді як залишок у відкинутому підпросторі - за статистикою SPE або Q. Побудова робить модель чутливішою до аномалій, які порушують звичну структуру залежностей між ознаками.

Окремо слід врахувати нестаціонарність мережевого трафіку. Навіть за відсутності атак чи збоїв нормальний профіль мережі змінюється залежно від часу доби, дня тижня, сезонних чинників та розвитку самої інформаційної системи. Тому математична модель не повинна трактувати будь-яке повільне зміщення профілю як аномалію.

Для цього базовий вектор μ та, за потреби, інші параметри моделі можуть оновлюватися адаптивно, наприклад за допомогою експоненціального згладжування. Такий підхід дозволяє зберігати чутливість до різких порушень, але водночас зменшує кількість хибних спрацювань у разі плавної еволюції нормального режиму.

У прикладному сенсі аномальний стан у межах моделі визначається не лише за величиною статистики відхилення, а й за внеском окремих компонент вектора стану. Якщо поточне вікно класифіковане як аномальне, модель повинна дозволяти ідентифікувати, які саме ознаки найбільше відхилилися від профілю норми: чи це різке зменшення ентропії адрес призначення, чи зростання дивергенції за портами, чи нетипова комбінація об’ємних характеристик.

Хоча таке пояснення є вже передумовою для методичної реалізації третього розділу, воно логічно випливає саме з математичної моделі, де багатовимірне відхилення розкладається на складові (рисунок 2.3).

Математична модель опису станів мережного трафіку подає кожне часове вікно у вигляді вектора інформативних ознак, формує багатовимірний профіль нормального режиму та визначає аномалію як статистично значуще відхилення від цього профілю. Модель є достатньо загальною, щоб працювати з різними джерелами даних, і водночас достатньо конкретною для подальшої алгоритмічної реалізації у вигляді послідовності процедур виявлення аномальних станів.

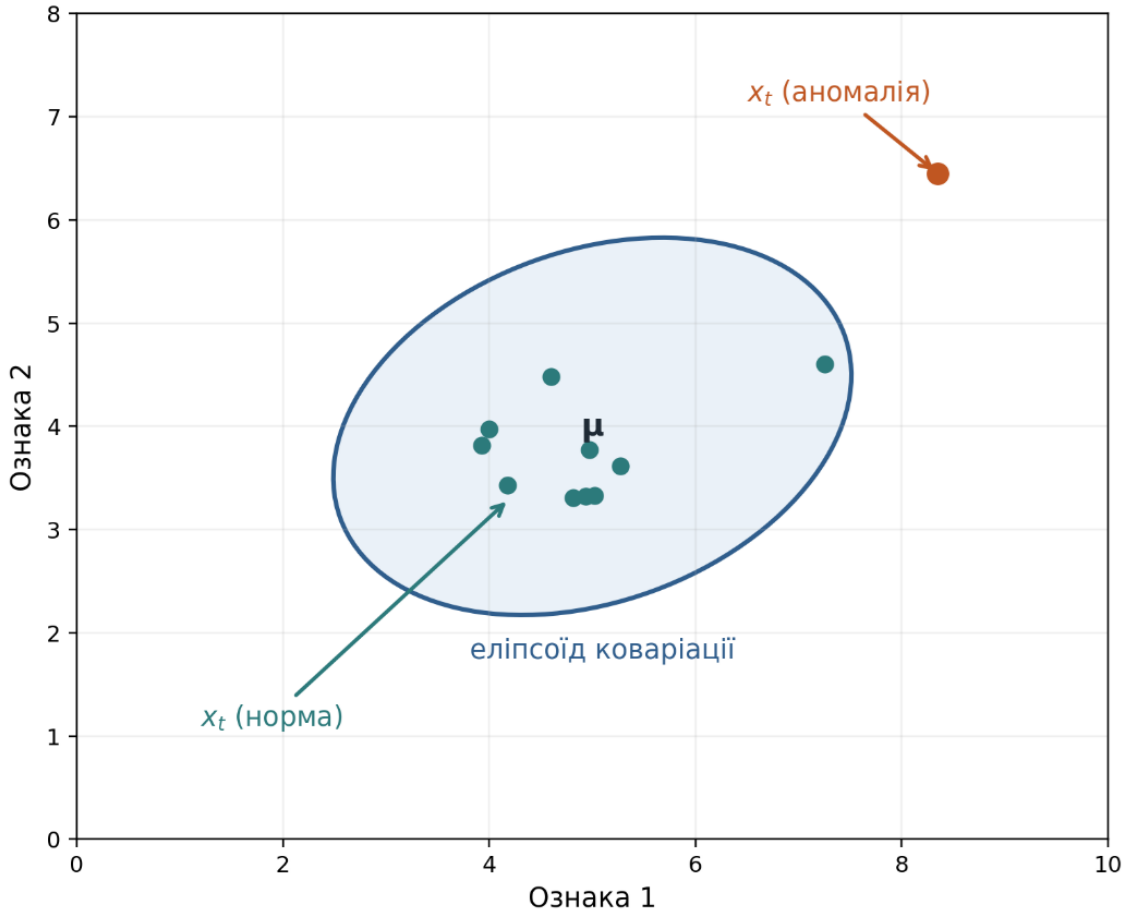


Рисунок 2.3 – Геометрична інтерпретація багатовимірного детектування у просторі ознак

2.5 Розроблення структурної моделі аналізу трафіку комп'ютерних мереж

Побудована математична модель описує стан трафіку у формальному вигляді, однак для її практичного застосування необхідно визначити структурну логіку всієї системи аналізу. Структурна модель відображає не окремі формули, а функціональні блоки, що забезпечують перехід від сирих мережевих записів до підсумкового рішення про стан мережі. Її призначення полягає в тому, щоб показати місце кожного етапу в загальній архітектурі та забезпечити узгодженість між даними, ознаками, статистичною оцінкою і модулем прийняття рішення.

Першим блоком структурної моделі є підсистема збору та підготовки даних. На цьому етапі мережеві записи надходять із джерела спостереження, яким може бути файл пакетного захоплення, система потокового експорту або інший

телеметричний модуль. Далі вони нормалізуються, очищуються від очевидних помилок, впорядковуються за часом і приводяться до єдиного формату опису. Значення цього блоку полягає в тому, що саме тут забезпечується коректність подальшої агрегації й усуваються невідповідності, які могли б спотворити статистичний опис трафіку.

Другий блок відповідає за часову агрегацію. Потік подій розбивається на послідовність вікон однакової або адаптивної тривалості, після чого кожне вікно розглядається як окремий об’єкт аналізу. В межах цього блоку визначається масштаб спостереження, а також баланс між швидкістю реакції та статистичною стійкістю оцінок. Надто короткі вікна роблять систему чутливою до випадкових коливань, тоді як надто довгі - зменшують здатність швидко локалізувати початок аномального процесу. Тому віконування виконує не допоміжну, а концептуально важливу функцію в усій структурній моделі.

Третій блок призначений для побудови розподілів ознак усередині кожного вікна. На цьому етапі формуються гістограми категоріальних атрибутів, оцінюється кількість унікальних значень і обчислюються агреговані числові характеристики. Цей блок готує статистичну основу для подальшого обчислення ентропійних і дивергентних показників. Його значення полягає в тому, що він здійснює перехід від сирих мережевих записів до компактних описів структури трафіку, придатних для подальшого аналізу на рівні станів, а не окремих подій (рисунок 2.4).

Четвертий блок реалізує обчислення інформативних ознак. Тут на основі побудованих розподілів визначаються ентропії адрес, портів і протоколів, обчислюються дивергенції відносно еталонного профілю, а також формуються об’ємні та допоміжні агреговані показники. Результатом роботи цього блоку є вектор ознак X(t), який подає стан трафіку в конкретному вікні. Функціонально цей блок є центральним для всієї моделі, оскільки саме він перетворює статистичний опис трафіку на компактне багатовимірне представлення, зручне для прийняття рішення.

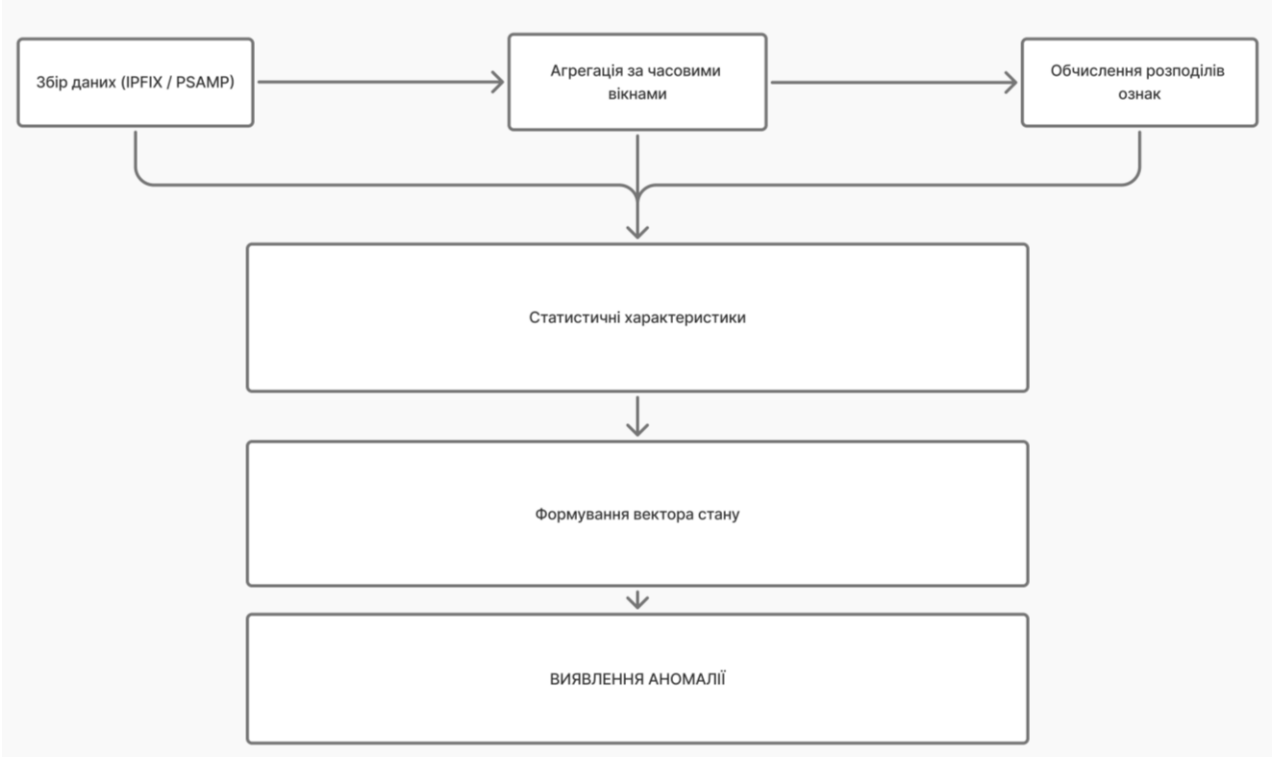


Рисунок 2.4 – Структурна модель аналізу трафіку комп’ютерних мереж

П’ятий блок становить підсистему багатовимірного статистичного аналізу. У ньому виконується стандартизація ознак, оцінювання параметрів моделі нормального режиму, а за потреби - перетворення простору ознак методом головних компонент. Після цього обчислюється інтегральна статистика відхилення, яка показує, наскільки поточний стан X(t) віддалений від профілю норми. Фактично цей блок реалізує математичне ядро моделі, в якому окремі характеристики трафіку інтегруються у єдину оцінку ризику аномального стану.

Шостий блок виконує прийняття рішення та формування інтерпретації. На його вході знаходиться інтегральна статистика відхилення або набір таких статистик, а на виході - класифікація поточного вікна як нормального чи аномального, а також супровідна інформація щодо внеску окремих ознак. За практичної реалізації цей блок може бути доповнений механізмами фільтрації одиничних сплесків, адаптивним налаштуванням порогів, веденням журналу спрацювань і передаванням сигналу зовнішній системі моніторингу. Проте навіть у базовій постановці його функція полягає в перетворенні числової статистики на управлінське рішення.

Структурна модель повинна бути не просто лінійною послідовністю дій, а узгодженою системою взаємопов’язаних блоків із чітко визначеними входами та виходами. Її важливою властивістю є модульність: зміна способу збору даних, типу віконування чи конкретної багатовимірної статистики не повинна руйнувати загальну логіку системи. Завдяки цьому модель залишається придатною для різних практичних сценаріїв, від офлайн-аналізу архівних трас до потокового моніторингу мережевої інфраструктури в реальному часі. Модульна побудова створює основу для подальшої алгоритмічної та програмної реалізації методу.

2.6 Аналіз відповідності розроблених моделей задачам аналізу мережного трафіку

Після побудови математичної і структурної моделі необхідно оцінити, наскільки вони відповідають реальним задачам аналізу мережного трафіку. Відповідність визначається не лише формальною коректністю моделі, а й її здатністю відображати властивості об’єкта спостереження та забезпечувати практично корисні рішення. Для задач мережевого моніторингу це означає, що модель повинна працювати з доступними даними, бути чутливою до основних типів аномалій, враховувати багатофакторний характер трафіку і зберігати інтерпретованість результатів.

Розроблена модель відповідає першій із цих вимог, оскільки базується на заголовкових і потокових характеристиках, які можуть бути отримані стандартними засобами спостереження без аналізу payload. Це робить її придатною для широкого кола мережевих середовищ, у тому числі для високонавантажених сегментів, де повна інспекція вмісту пакетів є або надто дорогою, або взагалі неможливою через шифрування. Орієнтація на практично доступні поля даних є важливою перевагою моделі, оскільки забезпечує можливість її реального використання, а не лише теоретичного опрацювання.

Другою важливою вимогою є чутливість до різних типів аномалій. Поєднання об’ємних, ентропійних і дивергентних ознак дозволяє виявляти як

волюметричні відхилення, пов’язані з різким зростанням навантаження, так і структурні порушення, що проявляються у зміні розподілів адрес, портів або протоколів. Багатовимірна інтеграція цих ознак у єдину статистику відхилення додатково підвищує здатність моделі реагувати на комбіновані сценарії, де аномальний стан не зводиться до однієї окремої характеристики. Таким чином, модель є достатньо універсальною щодо спектра очікуваних порушень у мережевому середовищі.

Третьою вимогою є врахування багатовимірної природи трафіку. Нормальний стан мережі не можна задати простим набором незалежних інтервалів для кожної ознаки, оскільки реальні характеристики трафіку взаємопов’язані. Запропонована модель усуває це обмеження через використання коваріаційної структури і, за потреби, підпросторового аналізу. У результаті область норми формується не як сукупність окремих порогів, а як багатовимірний профіль, що більш адекватно відображає реальну поведінку мережі. Ця властивість є принциповою для зменшення кількості хибних спрацювань у практичному застосуванні.

Важливою перевагою розробленої моделі є також її інтерпретованість. Ентропійні ознаки мають зрозумілий зміст і дозволяють якісно пояснити напрям зміни структури трафіку. Багатовимірна статистика, у свою чергу, дає можливість оцінити інтегральний характер відхилення та розкласти його на внески окремих ознак. Завдяки цьому модель може не лише сигналізувати про наявність проблеми, а й надавати вихідну інформацію для подальшого аналізу причин аномального стану. Для задач кібербезпеки та мережевого адміністрування така властивість є не менш важливою, ніж сам факт виявлення відхилення.

Разом з тим модель має певні обмеження, які необхідно враховувати ще на етапі її оцінювання. Вона залежить від вибору параметрів віконування, складу вектора ознак і якості базового профілю нормального стану. Якщо навчальна вибірка містить аномальні фрагменти або не охоплює типову варіативність робочого режиму, це може призвести до зміщення меж норми. Крім того, модель не аналізує зміст корисного навантаження, а орієнтована насамперед на

поведінкові та структурні прояви аномалій. Проте зазначені обмеження є прийнятними для задачі побудови масштабованої системи мережевого моніторингу і не нівелюють переваг запропонованого підходу.

Розроблені математична і структурна моделі в цілому відповідають задачам аналізу мережного трафіку. Вони забезпечують формалізований опис станів мережі, спираються на доступні дані, враховують структурні та об’ємні прояви аномалій, дають можливість інтегрального оцінювання відхилення від норми та зберігають інтерпретованість результатів. Тому ці моделі можуть бути використані як теоретична основа для подальшого розроблення методу аналізу трафіку, його алгоритмічної реалізації та експериментальної перевірки.

2.7 Висновки

У другому розділі виконано формалізацію задачі аналізу мережного трафіку та побудовано модель, яка пов’язує часові вікна спостереження, систему інформативних ознак і статистичне оцінювання стану мережі. Показано, що для адекватного опису трафіку недостатньо окремих об’ємних показників і доцільно використовувати поєднання ентропійних, дивергентних та агрегованих характеристик.

Обґрунтовано доцільність використання ентропійних показників для відображення структурних змін трафіку та багатовимірних статистичних методів для інтегрального оцінювання відхилень від профілю норми. Сформовано систему інформативних ознак, описано математичну модель нормального й аномального станів та побудовано структурну модель системи аналізу трафіку від етапу збору даних до прийняття рішення.

Отримані результати другого розділу становлять теоретичну основу для розроблення методу аналізу мережного трафіку. На цій основі в третьому розділі буде сформовано послідовність процедур, за допомогою яких розроблена модель реалізується у вигляді цілісного методу виявлення аномальних станів комп’ютерних мереж.

3 УДОСКОНАЛЕНИЙ МЕТОД АНАЛІЗУ ТРАФІКУ КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ТА БАГАТОВИМІРНОЇ МАТЕМАТИЧНОЇ СТАТИСТИКИ

3.1 Концепція побудови методу виявлення аномальних станів мережного трафіку

У другому розділі було побудовано математичну й структурну модель аналізу мережного трафіку, у якій кожне часове вікно подається у вигляді вектора інформативних ознак, а стан мережі визначається за величиною статистичного відхилення від профілю нормального режиму. Подальшим кроком є перехід від цієї моделі до цілісного методу, тобто до узгодженої послідовності процедур, що переводять сирі мережеві записи у практичне рішення про нормальний або аномальний стан трафіку. Розробленню такого методу присвячено третій розділ.

Запропонований метод належить до класу методів аномалійного виявлення, однак його ключова особливість полягає у поєднанні двох різних аналітичних рівнів. На першому рівні оцінюється структура трафіку через ентропійні характеристики та міри відхилення поточного розподілу від еталонного. На другому рівні окремі ознаки інтегруються у багатовимірному просторі, де визначається узагальнена відстань від поточного стану до профілю нормального режиму. Завдяки такому поєднанню метод є чутливим як до волюметричних відхилень, так і до структурних змін розподілів, які не завжди проявляються у різкому зростанні навантаження.

Концептуально метод спирається на віконне подання трафіку. Потік мережевих подій не аналізується покадрово, оскільки одиничний пакет або окремий flow-запис містить недостатньо інформації для надійного висновку про стан мережі. Натомість записи агрегуються у часові вікна, усередині яких будуються емпіричні розподіли адрес, портів, протоколів та інших характеристик. Для кожного вікна формується багатовимірний опис стану, придатний для

статистичного оцінювання відхилення від норми. У такий спосіб метод забезпечує перехід від рівня окремих подій до рівня станів мережі (рисунок 3.1).

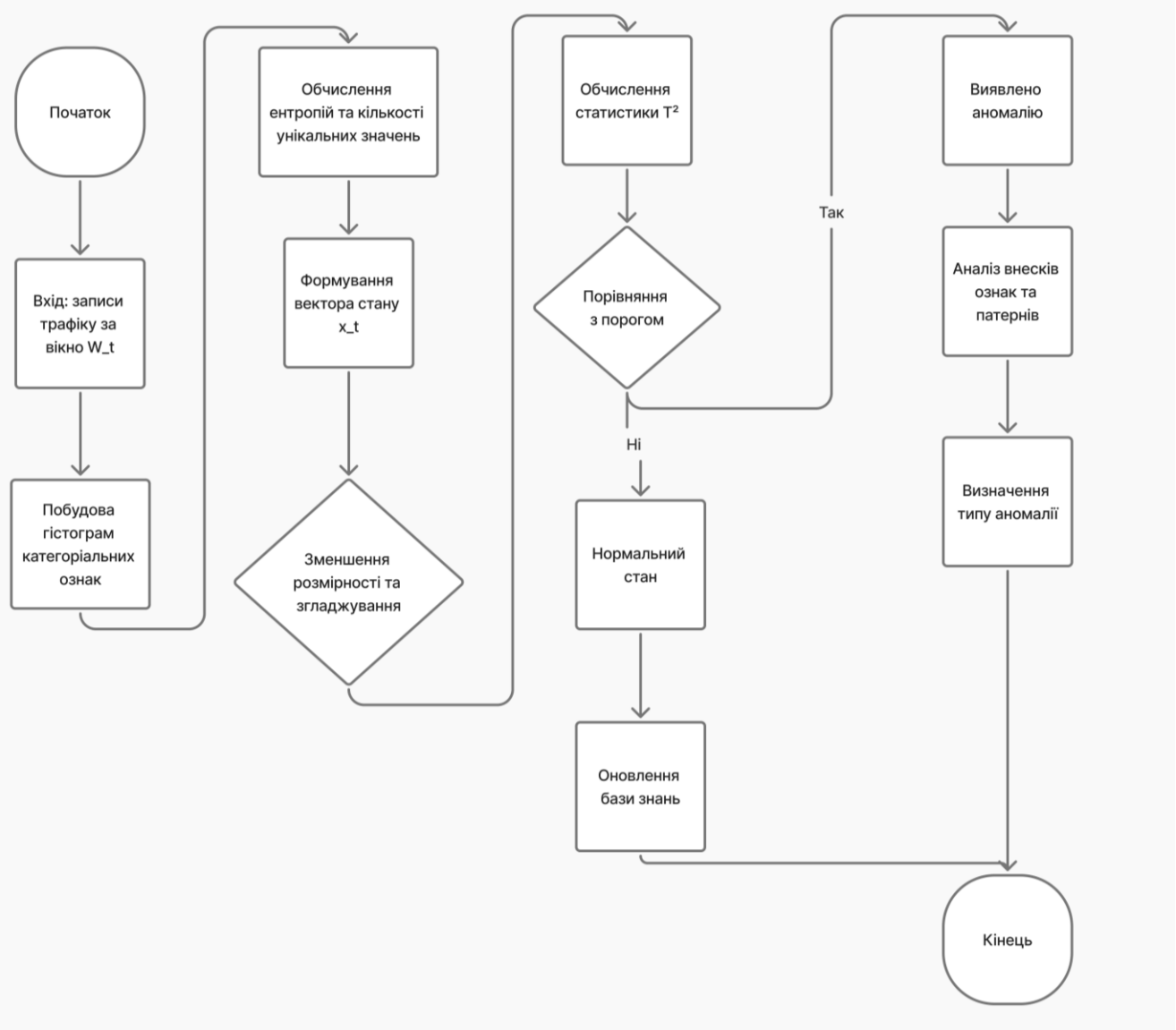


Рисунок 3.1 – Блок-схема детектування аномальних станів мережного трафіку

Важливою складовою концепції є використання базового профілю нормального режиму. Метод не порівнює поточне вікно з абстрактним уявленням про норму, а формує еталон на основі навчальної або опорної вибірки, що представляє штатну роботу мережі. Такий профіль містить середні значення ознак, характерну мінливість і взаємозв’язки між компонентами вектора стану. Відносно

цього профілю надалі виконується обчислення ентропійних відхилень, стандартизація ознак і побудова інтегральної статистики аномальності.

Метод орієнтований на практично доступні мережеві дані та не вимагає повного аналізу корисного навантаження пакетів. Це дає змогу застосовувати його як до архівних файлів захоплення трафіку, так і до потокових телеметричних даних, одержаних із мережевих сенсорів, маршрутизаторів або систем експорту потоків. При цьому сама логіка методу залишається незмінною: змінюється лише формат джерела, тоді як послідовність перетворення даних на вектор ознак і правила статистичного оцінювання зберігаються.

Ще однією принциповою вимогою до методу є інтерпретованість результатів. На відміну від суто «чорних» моделей, де система видає лише кінцеву мітку класу, запропонований підхід дозволяє простежити, які саме ознаки сформували високу статистику відхилення. Це важливо для задач кібербезпеки і мережевого адміністрування, оскільки дає змогу не лише зафіксувати факт аномального стану, а й встановити його можливу природу: концентрацію трафіку на одній цілі, різке розширення множини адрес, нетипову динаміку портів або комбіновану зміну кількох характеристик.

3.2 Загальна послідовність кроків методу аналізу мережного трафіку

Узагальнена послідовність реалізації методу складається з двох взаємопов’язаних фаз. Перша фаза є підготовчою і призначена для формування профілю нормального режиму. На цьому етапі з опорної вибірки нормального трафіку виконуються попередня обробка даних, побудова часових вікон, обчислення інформативних ознак, оцінювання їхніх середніх значень, масштабів зміни й коваріаційної структури. Результатом підготовчої фази є параметризований опис нормального стану, який використовується далі як еталон.

Друга фаза є робочою й виконується для кожного нового вікна спостереження. Поточні записи трафіку нормалізуються, агрегуються у часове вікно і перетворюються на вектор ознак. Далі для цього вектора визначаються

ентропійні характеристики, міри відхилення поточних розподілів від базового профілю, а також інтегральні багатовимірні статистики. На основі цих оцінок формується рішення про належність вікна до нормального або аномального режиму.

Послідовність реалізації методу повинна забезпечувати узгодженість усіх етапів. Це означає, що правила формування опорної вибірки і правила аналізу робочих вікон мають бути ідентичними за способом нормалізації записів, побудови розподілів і обчислення ознак.

Лише за цієї умови статистичне порівняння поточного вікна з еталоном буде коректним. Якщо ж на різних етапах використовуються різні схеми агрегації або різні набори ознак, значення статистики відхилення втрачає однозначну інтерпретацію (табл. 3.1).

В узагальненому вигляді реалізація методу передбачає перехід від сирих мережевих подій до рішення через послідовність функціональних перетворень.

Спочатку формується часове вікно спостереження, далі з нього будуються емпіричні розподіли вибраних атрибутів і обчислюються агреговані кількісні показники.

Після цього формується вектор стану, який стандартизується відносно базового профілю та подається на модуль багатовимірного статистичного аналізу. Отримана статистика відхилення надалі інтерпретується з використанням порогового правила та, за потреби, механізму згладжування короткочасних флуктуацій (рисунок 3.2).

Окремого акценту потребує питання часової організації методу. Оскільки трафік мережі є нестаціонарним і має добові та поведінкові цикли, метод не може розглядати всі спостереження як незалежні точки без часового контексту (рисунок 3.3).

Таблиця 3.1 – Вхідні дані та результати основних етапів методу

Етап методу	Вхідні дані	Результат
Формування базового-періоду	Нормалізовані записи трафіку початкового інтервалу	Базовий профіль нормального режиму та параметри масштабу ознак
Віконування трафіку	Послідовність мережевих записів із часовими мітками	Часові вікна спостереження однакової структури
Обчислення ознак	Записи в межах окремого вікна	Вектор об’ємних, структурних і порівняльних характеристик
Ентропійний аналіз	Дискретні розподіли атрибутів у вікні	Набір ентропійних та дивергентних показників
Багатовимірне оцінювання	Стандартизований вектор ознак	Статистика відхилення від профілю нормального стану
Прийняття рішення	Інтегральна статистика та порогове правило	Висновок про нормальний або аномальний стан трафіку

Тому послідовність реалізації методу враховує не лише внутрішню структуру окремого вікна, а й зміну оцінок між сусідніми вікнами.

Завдяки цьому стає можливим виділення стійких відхилень, які тривають у часі, та відокремлення їх від випадкових одиничних коливань.

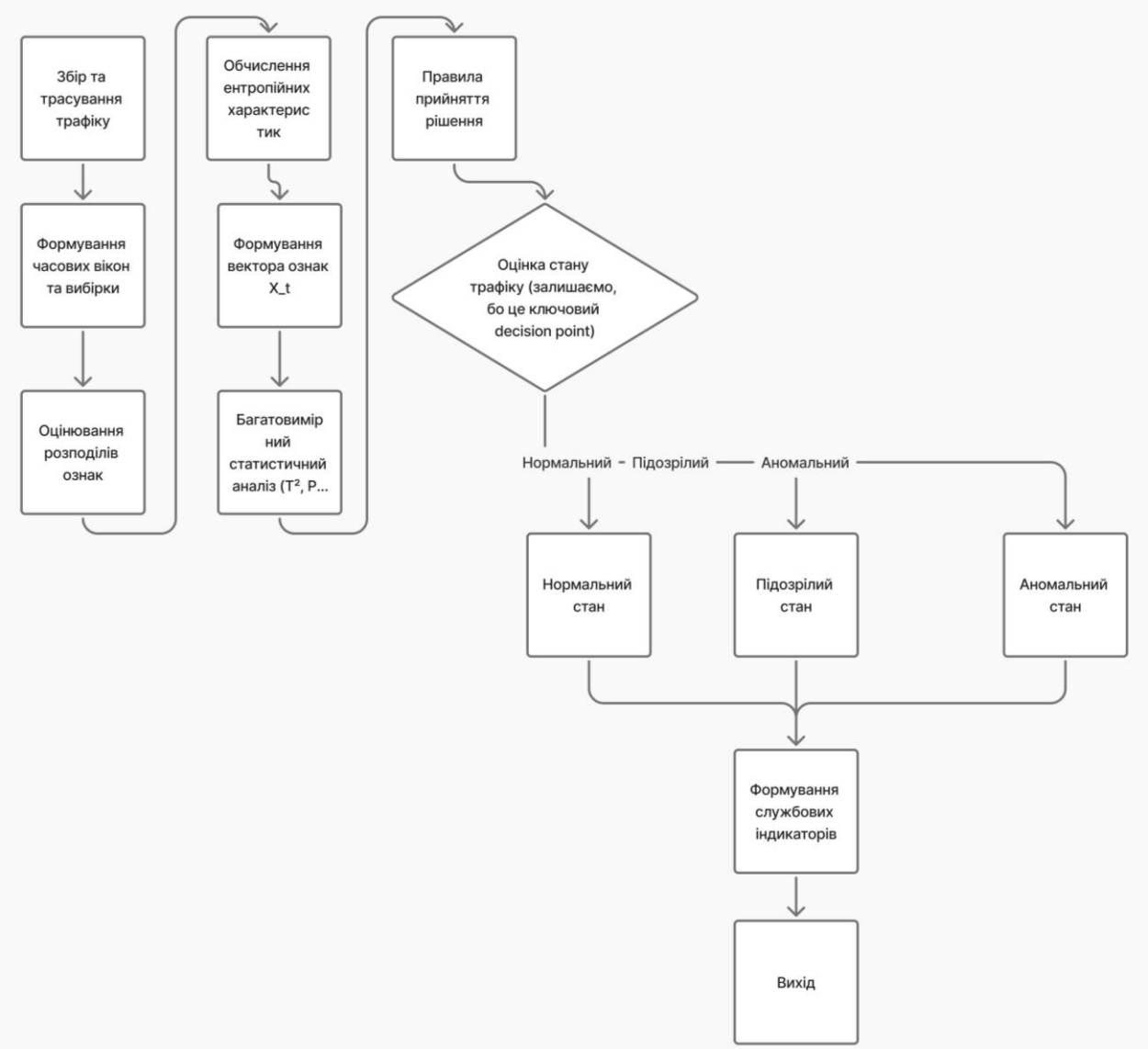


Рисунок 3.2 – Загальна схема реалізації методу аналізу мережного трафіку

Рисунок 3.3 – Часова організація методу: базовий інтервал, робоча ділянка моніторингу та ковзні вікна спостереження

3.3 Формування часових вікон і підготовки вибірки мережного трафіку

Метод формування часових вікон і підготовки вибірки мережного трафіку є початковим етапом алгоритмічної реалізації розробленого підходу. Його призначення полягає у перетворенні неструктурованої послідовності мережевих подій на впорядкований набір інтервалів спостереження, для кожного з яких можна коректно обчислити систему ознак. На цьому етапі вирішуються питання синхронізації часових міток, уніфікації формату записів, відсікання очевидно некоректних спостережень та забезпечення однакового правила агрегації як для опорної вибірки, так і для робочого потоку даних.

Нехай кожний запис трафіку описується кортежем r = (t, as, ad, ps, pd, pr, b, n), де t є часовою міткою, as і ad - адресами джерела та призначення, ps і pd - портами, pr - типом протоколу, b - обсягом байтів, а n - кількістю пакетів. Після початкової нормалізації всі записи впорядковуються за часом і розбиваються на часові вікна однакової тривалості Tw. Якщо крок зсуву позначити через Δ, то множина записів у k-му вікні може бути подана як $W_k = \{r : t_k <= t(r) < t_k + Tw\}$. При Δ = Tw отримуємо неперекривні вікна, а при Δ < Tw - ковзне віконування з частковим перекриттям.

Вибір параметрів Tw і Δ істотно впливає на властивості методу. Короткі вікна підвищують швидкість реакції на аномальний процес, але зменшують статистичну стійкість оцінок, оскільки кількість спостережень усередині інтервалу може бути недостатньою для надійного обчислення розподілів. Надто довгі вікна, навпаки, згладжують локальні відхилення і можуть приховувати короткочасні атаки або імпульсні збої. Тому в загальному випадку доцільно використовувати фіксоване вікно помірної тривалості з кроком, що забезпечує компроміс між оперативністю та стійкістю. Для мереж із вираженою спалаховою активністю доцільним є часткове перекриття сусідніх вікон.

Підготовка вибірки не обмежується розбиттям на часові інтервали. Для кожного вікна необхідно забезпечити узгоджене подання мережевих подій, зокрема привести адресні та портові поля до дискретного формату, відокремити

відсутні значення, обробити службові записи та усунути дублікати, якщо вони виникають унаслідок особливостей захоплення трафіку. У разі аналізу flow-записів важливо також уніфікувати напрям потоку, щоб уникнути штучного подвоєння ознак. Результатом цього етапу є очищений і впорядкований набір вікон, придатний для побудови розподілів і подальшого статистичного аналізу.

Окремим завданням є формування базової вибірки нормального трафіку. Вона повинна містити вікна, що репрезентують штатний режим функціонування мережі, і водночас не включати фрагменти, в яких уже присутні аномальні події. Якщо такий контроль неможливо забезпечити повністю, доцільно використовувати робастні процедури оцінювання параметрів норми, що зменшують вплив поодиноких забруднених вікон на підсумковий профіль. Базова вибірка має охоплювати природну варіативність навантаження, інакше модель нормального стану виявиться надто вузькою та спричинить зростання кількості хибних спрацювань.

Таким чином, метод формування часових вікон і підготовки вибірки забезпечує перехід від потоку подій до впорядкованої множини інтервалів спостереження, у межах яких можливо коректно обчислити систему ентропійних і агрегованих ознак. Якість цього етапу значною мірою визначає стійкість усього подальшого аналізу, оскільки похибки нормалізації, некоректне віконування або нерепрезентативна базова вибірка безпосередньо впливають на точність статистичного оцінювання стану мережного трафіку.

3.4 Обчислення ентропійних характеристик мережного трафіку

Після формування часових вікон для кожного інтервалу спостереження необхідно побудувати ентропійний опис структури трафіку. З цією метою всередині вікна Wk формуються емпіричні розподіли за вибраними атрибутами, такими як адреси джерела, адреси призначення, порти джерела, порти призначення та типи протоколів. Якщо для j-го атрибута у вікні зафіксовано nj різних значень, а кількість появ i-го значення дорівнює cij(k), то ймовірність цього значення

визначається як $pij(k) = \frac{cij(k)}{\Sigma cij(k)}$.На основі цих ймовірностей формується ентропійна характеристика відповідного розподілу.

Базовою ентропійною оцінкою є ентропія Шеннона, яка для j-го атрибута у k-му вікні обчислюється за формулою:

$$Hj(k) = -\Sigma\, pij(k)\, log2\, pij(k). \quad (3.1)$$

Отримане значення характеризує ступінь невизначеності або різноманітності розподілу. Чим рівномірніше розподілені елементи, тим вищою є ентропія; чим сильніше домінує невелика кількість значень, тим ентропія нижча. Оскільки кількість унікальних значень у різних вікнах може відрізнятися, доцільно додатково використовувати нормовану ентропію, що визначається як відношення Hj(k) до максимально можливого значення log2(nj). Нормалізація переводить показник у порівнянний масштаб і спрощує подальшу багатовимірну інтеграцію ознак.

Окрім абсолютних ентропійних оцінок, метод передбачає аналіз відхилення поточних розподілів від профілю нормального режиму. Для цього з базової вибірки формується еталонний розподіл Pref,j для кожного атрибута, а поточний розподіл у вікні Wk позначається як Pk,j. Ступінь їхньої відмінності доцільно оцінювати за симетричною та стійкою мірою, наприклад дивергенцією Дженсена-Шеннона. У загальному вигляді ця оцінка задається співвідношенням:

$$DJS(Pk,j,Pref,j) = \frac{1}{2}\, DKL(Pk,jM) + \frac{1}{2}\, DKL(Pref,jM), \quad (3.2)$$

$$\text{де M} = \frac{1}{2}\,(\text{Pk,j} + \text{Pref,j}).$$

Використання дивергенційного показника дає можливість врахувати не лише ступінь концентрації розподілу, а й зміну його форми відносно нормального режиму. Це особливо важливо в ситуаціях, коли сама ентропія залишається

близькою до типових значень, але перерозподіл імовірностей між категоріями вже свідчить про структурну зміну трафіку.

Під час практичного обчислення ентропійних характеристик необхідно враховувати проблему нульових частот і нестабільності оцінок на малих вибірках. Якщо у поточному вікні або в еталонному профілі деякі значення атрибута відсутні, то для коректного обчислення дивергенцій доцільно виконувати узгодження підтримок розподілів і застосовувати слабке згладжування. Крім того, для дуже коротких вікон корисно встановлювати мінімальну допустиму кількість спостережень, нижче якої ентропійні оцінки не використовуються окремо, а інтерпретуються разом з агрегованими показниками обсягу трафіку.

Результатом цього етапу є набір ентропійних і дивергентних характеристик, який описує структуру трафіку всередині кожного вікна. У подальшому ці показники об’єднуються з агрегованими кількісними ознаками в єдиний вектор стану. На цій основі виконується багатовимірний статистичний аналіз, що дозволяє перейти від локальних характеристик окремих атрибутів до інтегральної оцінки аномальності поточного стану мережі.

3.5 Багатовимірний статистичний аналіз ознак трафіку

Багатовимірний статистичний аналіз у запропонованому методі виконує функцію інтеграції окремих ознак у цілісну оцінку стану мережного трафіку. Після обчислення ентропійних, дивергентних та агрегованих кількісних характеристик для кожного вікна формується вектор $X(k) = [x1(k), x2(k), ..., xm(k)]^T$. Оскільки ознаки мають різні масштаби та фізичний зміст, безпосереднє порівняння таких векторів є некоректним. Тому першим кроком багатовимірного аналізу є стандартизація кожної компоненти відносно параметрів базового профілю нормального режиму.

Якщо для i-ї ознаки за базовою вибіркою оцінено середнє значення μi та стандартне відхилення σi, то стандартизована ознака визначається співвідношенням $zi(k) = (xi(k) - \mu i) / \sigma i$. Стандартизований вектор $Z(k)$ дозволяє

привести всі компоненти до узгодженого масштабу і водночас зберегти інформацію про напрям та величину відхилення від норми. На основі множини таких векторів, побудованих для вікон базової вибірки, оцінюється коваріаційна матриця, яка відображає взаємозв'язки між ознаками. Для зменшення впливу поодиноких аномальних фрагментів доцільно використовувати робастну оцінку коваріації, наприклад підхід Minimum Covariance Determinant.

Інтегральна статистика відхилення може бути побудована на основі квадрата відстані Махаланобіса або статистики Хотеллінга T^2. У компактному вигляді вона задається як:

$$T^2(k) = Z(k)^T \Sigma z^{-1} Z(k).$$

де Σz є коваріаційною матрицею стандартизованих ознак у нормальному режимі. Зміст цієї статистики полягає в тому, що вона враховує не лише величину відхилення кожної ознаки окремо, а й кореляції між ними. Завдяки цьому навіть відносно помірні зміни кількох взаємопов'язаних характеристик можуть бути правильно інтерпретовані як суттєве порушення нормального профілю.

Для підвищення стійкості методу до надлишкових або сильно корельованих ознак доцільно додатково використовувати аналіз головних компонент. У цьому випадку стандартизований простір ознак розкладається на нормальний підпростір, що описує основну варіативність штатного режиму, і залишковий підпростір, у якому накопичуються відхилення, не пояснені моделлю норми. Для кожного вікна поряд зі статистикою T^2 може обчислюватися залишкова статистика Q, яка відображає енергію компоненти, що не відтворюється обраною кількістю головних компонент.

Поєднання T^2 і Q забезпечує більш гнучке виявлення аномалій. Якщо T^2 зростає, це означає, що стан трафіку зміщується в межах структури, уже представленої у моделі нормального режиму, але виходить за її характерні межі. Якщо ж зростає статистика Q, це свідчить про появу нового типу варіації, яка не описується нормальним підпростором. Інтерпретація є корисною для розрізнення різних типів відхилень і підвищує пояснювальну здатність методу (рисунок 3.4).

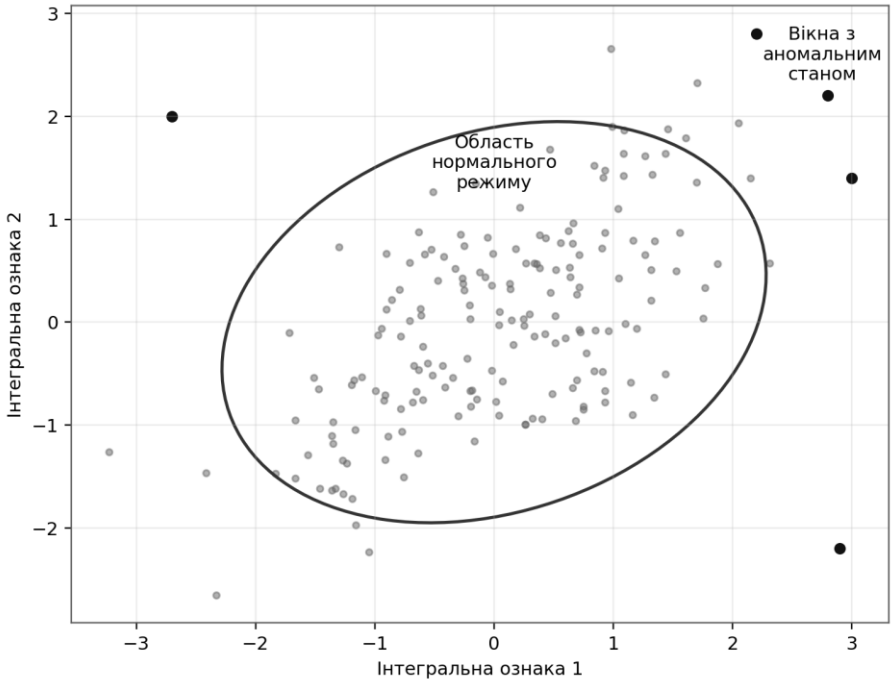


Рисунок 3.4 – Інтерпретація нормального та аномального станів у багатовимірному просторі ознак

Багатовимірний статистичний аналіз забезпечує перехід від набору окремих характеристик до інтегрального оцінювання стану трафіку. В цьому модулі закладається механізм виявлення складних аномалій, що проявляються не у значенні однієї ознаки, а в нетиповій комбінації кількох взаємопов’язаних характеристик. Подальший етап методу полягає у перетворенні одержаних статистик на однозначне рішення про стан мережного трафіку.

3.6 Прийняття рішення щодо стану мережного трафіку

Метод прийняття рішення щодо стану мережного трафіку є завершальним етапом запропонованого підходу. Його призначення полягає в перетворенні інтегральних статистик, отриманих у багатовимірному просторі ознак, на інтерпретовану класифікацію поточного вікна. На цьому етапі необхідно

забезпечити одночасно дві властивості: високу чутливість до реальних аномалій і стійкість до короткочасних флуктуацій, що не мають характеру інциденту.

У найпростішому випадку рішення може прийматися окремо за статистикою T^2 або за парою статистик T^2 і Q, кожна з яких порівнюється зі своїм граничним значенням. Якщо хоча б одна зі статистик перевищує допустимий рівень, поточне вікно відноситься до аномального стану. Проте така схема є чутливою до випадкових одиничних сплесків. Тому для підвищення надійності доцільно використовувати згладжену інтегральну оцінку аномальності, що враховує як поточне, так і попередні значення статистик.

Нехай A(k) є інтегральною аномальною оцінкою, сформованою на основі нормованих статистик $T^2(k)$, Q(k) та, за потреби, узагальненого дивергентного показника. Для подавлення високочастотних випадкових коливань використовується експоненційно зважене згладжування за правилом $Y(k) = \rho\, A(k) + (1 - \rho)\, Y(k - 1), \quad 0 < \rho \leq 1$, де параметр ρ визначає чутливість системи до швидких змін. Якщо значення Y(k) перевищує поріг τ, поточне вікно позначається як аномальне. Додатково може застосовуватися умова стійкості, за якої сигнал вважається підтвердженим лише тоді, коли перевищення порога спостерігається у кількох сусідніх вікнах або коли одне перевищення має критично велику амплітуду. Такий підхід дозволяє поєднати оперативність реакції з прийнятним рівнем хибних тривог.

Важливим елементом прийняття рішення є інтерпретація внеску окремих ознак у підсумкову оцінку аномальності. Якщо для поточного вікна зафіксовано перевищення порогу, система не повинна обмежуватися лише бінарною міткою. Доцільно також аналізувати, які саме компоненти стандартизованого вектора зробили найбільший внесок у статистику відхилення. Це дозволяє встановити, чи зумовлена аномалія концентрацією трафіку на одному напрямі, різким зростанням різноманітності адрес, нетиповим перерозподілом портів або комбінованою зміною кількох характеристик.

Таким чином, метод прийняття рішення в запропонованому підході не зводиться до простого порогового контролю одного показника. Він поєднує

багатовимірну статистичну оцінку, часове згладжування та аналіз внеску ознак, що дозволяє більш надійно відокремлювати істотні аномальні стани від випадкових локальних коливань трафіку. Ця завершальна частина робить розроблений підхід придатним до практичного використання у системах моніторингу комп’ютерних мереж.

3.7 Висновки

У третьому розділі на основі побудованої раніше моделі розроблено цілісний метод аналізу мережного трафіку, орієнтований на виявлення аномальних станів комп’ютерних мереж. Показано, що запропонований метод ґрунтується на часовій агрегації подій, формуванні ентропійних, дивергентних та агрегованих ознак, їх багатовимірному статистичному аналізі та подальшому прийнятті рішення за інтегральною оцінкою відхилення від профілю норми.

Детально розглянуто процедури підготовки вибірки, формування часових вікон, побудови емпіричних розподілів, обчислення ентропійних характеристик, стандартизації ознак, оцінювання коваріаційної структури та формування статистик аномальності. Обґрунтовано доцільність використання поєднання ентропійного опису структури трафіку та багатовимірного статистичного оцінювання для виявлення як явних волюметричних, так і прихованих структурних відхилень.

Отриманий метод створює алгоритмічну основу для подальшої програмної реалізації й експериментальної перевірки.

4 СИСТЕМА АНАЛІЗУ ТРАФІКУ КОМП’ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ТА БАГАТОВИМІРНОЇ МАТЕМАТИЧНОЇ СТАТИСТИКИ

4.1 Опис засобів програмної реалізації методу

Розроблений у попередньому розділі метод аналізу мережного трафіку потребує такої програмної реалізації, яка одночасно забезпечує відтворюваність

обчислень, достатню швидкодію під час роботи з часовими вікнами, коректну підтримку статистичних процедур і можливість зручного представлення результатів. На відміну від спрощених навчальних прикладів, у практичній реалізації методу необхідно послідовно виконати кілька різнорідних дій: прийняти вхідні записи трафіку, привести їх до уніфікованого формату, сформувати часові вікна, обчислити ентропійні та об’ємні ознаки, оцінити багатовимірні статистики, порівняти їх з пороговими значеннями та зберегти підсумкові артефакти експерименту.

З огляду на наведені вимоги доцільним є використання середовища, орієнтованого на наукові та інженерні обчислення. Для цього найкраще підходить Python, оскільки він дозволяє поєднати обробку мережних даних, матричні операції, статистичний аналіз і побудову графічних матеріалів у межах єдиного відтворюваного конвеєра. Важливою перевагою такого підходу є не лише наявність готових бібліотек, а й можливість фіксувати весь ланцюг перетворень від вхідних даних до підсумкового висновку про стан трафіку, що є критичним для експериментальної перевірки методу.

Для попередньої підготовки даних доцільно використовувати інструменти, які підтримують роботу з пакетними трасами та flow-записами, а також дозволяють коректно обробляти часові мітки, мережеві адреси, порти та службові поля. На етапі статистичної обробки потрібні засоби для лінійної алгебри, стандартизації ознак, оцінювання коваріаційної структури, побудови підпростору головних компонент і обчислення інтегральних статистик відхилення. Для завершального етапу, пов’язаного з інтерпретацією результатів, необхідні засоби візуалізації часових рядів, порогових перевищень, кривих якості детекції та службових журналів експерименту.

Окремо слід підкреслити, що програмна реалізація методу повинна бути модульною. Вимога зумовлена тим, що різні джерела вхідних даних можуть відрізнятися за форматом, а вибір ознак, параметрів часових вікон або способу статистичного оцінювання може змінюватися залежно від сценарію перевірки. Тому реалізація не повинна бути жорстко прив’язаною до одного набору трас або

одного типу аномалій. Модульний підхід дозволяє окремо перевіряти кожний етап обробки, а також повторно використовувати сформовані компоненти під час нових експериментів.

Таким чином, вибір засобів програмної реалізації визначається не лише зручністю розробки, а передусім відповідністю структурі запропонованого методу. Програмне середовище має підтримувати обробку часових вікон, формування векторів ознак, обчислення ентропійних характеристик, багатовимірний статистичний аналіз та формування зрозумілого підсумкового висновку. За таких умов програмна реалізація може розглядатися як коректне продовження математичної моделі й алгоритмічного опису, поданих у попередніх розділах.

4.2 Структура програмної реалізації розробленого методу

Структура програмної реалізації повинна відтворювати логіку методу, розробленого у третьому розділі, але в термінах взаємодії функціональних компонентів. У центрі такої структури перебуває послідовний перехід від мережних записів до інтегральної статистики аномальності. На першому рівні здійснюється зчитування даних і приведення їх до уніфікованої схеми. На другому рівні формується набір часових вікон, для яких обчислюються ознаки стану. На третьому рівні виконується багатовимірний статистичний аналіз і прийняття рішення. На завершальному рівні результати подаються у вигляді часових рядів, журналів спрацювань та узагальнених показників якості.

Функціональне розділення реалізації на модулі дозволяє зменшити зв’язаність компонентів і спростити перевірку коректності окремих етапів. Модуль джерела даних відповідає за приймання пакетних або потокових записів і контролює цілісність схеми вхідних полів. Модуль попередньої підготовки виконує нормалізацію часових міток, перевірку ознак, усунення дублювань та формування єдиного опису спостереження. Після цього модуль віконування перетворює суцільний потік подій у впорядковану послідовність інтервалів спостереження, кожний з яких далі розглядається як окрема одиниця аналізу.

Модуль обчислення ентропійних ознак формує частотні розподіли для вибраних категоріальних полів та перетворює їх на компактний опис структури трафіку всередині вікна. Поряд із ним можуть обчислюватися додаткові об’ємні або поведінкові ознаки, якщо вони включені до вектора стану. Далі модуль багатовимірного аналізу одержує стандартизований вектор ознак, оцінює його відхилення від базового профілю та формує числову статистику, яка використовується для ухвалення рішення про нормальний або аномальний режим роботи мережі (рисунок 4.1).

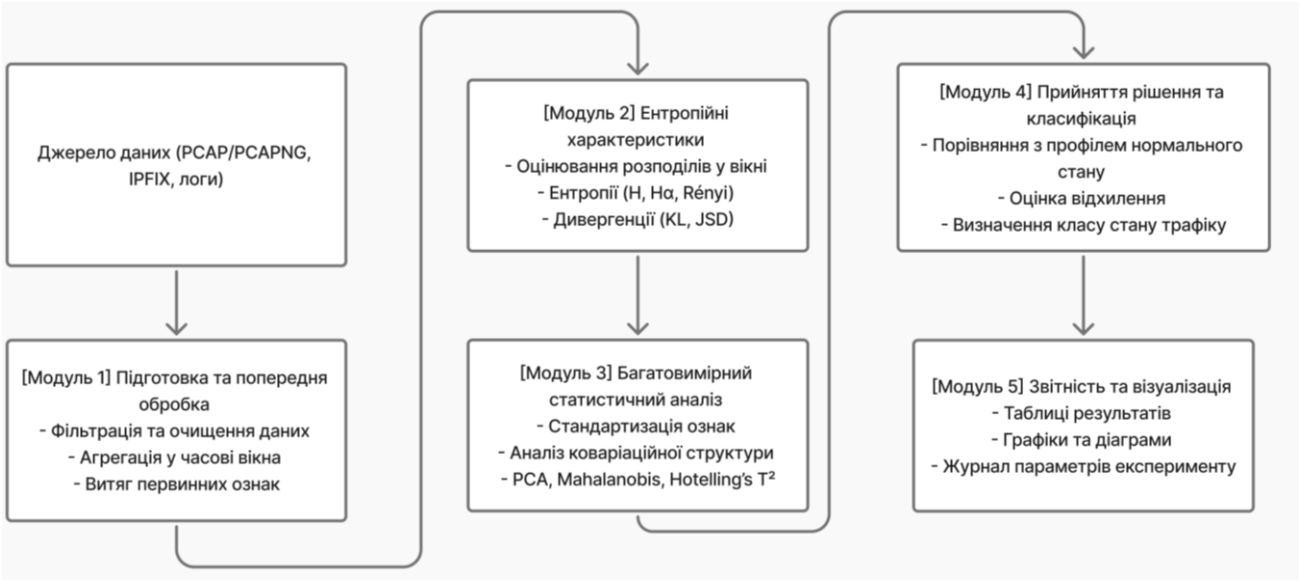


Рисунок 4.1 – Функціональна схема реалізації методу

Підсумковий блок реалізації призначений для накопичення та представлення результатів. На цьому етапі зберігаються часові ряди інтегральної статистики, фіксуються моменти перевищення порога, формуються графіки для інтерпретації та створюються артефакти, необхідні для повторного запуску експерименту. Побудова дозволяє зробити реалізацію не лише придатною для одиничного тестування, а й придатною для серії відтворюваних порівнянь між різними конфігураціями методу.

4.3 Основні етапи програмної реалізації та проведення експерименту

Програмна реалізація методу доцільно будується як послідовність взаємопов’язаних етапів, кожний з яких завершується формуванням проміжного, але перевірюваного результату. На початковому етапі виконується зчитування мережних записів і перевірка їхньої придатності для аналізу. Для пакетних трас це означає коректне вилучення часових міток і службових полів заголовків, а для flow-записів – перевірку цілісності ключових ідентифікаторів потоку, часу початку та завершення, а також службових числових атрибутів. У разі виявлення пропусків, конфліктних позначень або нетипових значень ці записи повинні або виправлятися за встановленими правилами, або виключатися з подальшого аналізу.

Наступний етап пов’язаний з формуванням часових вікон. Після того як записи впорядковано за часом, обирається тривалість вікна та крок його зсуву. У межах кожного вікна формується локальний зріз поточного стану трафіку. Вв такому зрізі обчислюються частотні розподіли для обраних полів, а також агрегуються числові показники.

Коректність цього етапу визначає весь подальший результат, оскільки надто короткі вікна збільшують варіативність оцінок, а надто довгі згладжують короткочасні аномальні події.

Після формування вікон виконується побудова векторів ознак. Для кожного вікна визначаються ентропійні характеристики адресних, портових чи протокольних розподілів, а також додаткові узагальнені показники, що характеризують інтенсивність і структуру трафіку. На цьому ж етапі відбувається стандартизація ознак та підготовка базового-сегмента, який описує нормальний режим роботи мережі. Базовий рівень використовується для оцінювання параметрів моделі нормального стану, з якими потім порівнюється кожне нове вікно спостереження.

Далі реалізація переходить до багатовимірного статистичного аналізу. Отримані вектори ознак перетворюються у простір, де можна коректно врахувати взаємозв’язки між компонентами. Для цього оцінюється центр нормального

профілю, коваріаційна структура, а за потреби – підпростір зниженої розмірності. Після цього для кожного нового вікна обчислюється статистика відхилення, яка набуває ролі інтегральної оцінки аномальності. Якщо її значення перевищує заданий поріг, вікно позначається як аномальне (рисунок 4.2).

Завершальний етап стосується безпосередньої організації експерименту. Після того як програмна реалізація пройшла перевірку на коректність обробки даних, вона застосовується до наборів, у яких відомі нормальні та аномальні сегменти.

Для кожного запуску фіксуються параметри віконування, набір ознак, спосіб формування базового рівня, поріг спрацювання й склад тестового сегмента. Фіксація є принципово важливою, тому що без неї результати різних конфігурацій методу неможливо порівнювати між собою (рисунок 4.3).

Програмна реалізація не зводиться до окремого скрипта для обчислення ентропії або побудови графіків. Вона являє собою повний конвеєр обробки, у якому кожний наступний етап спирається на результат попереднього.

Організація дозволяє перейти від суто алгоритмічного опису до реальної процедури експериментальної перевірки, яку можна повторити на інших даних без зміни загальної логіки роботи методу.

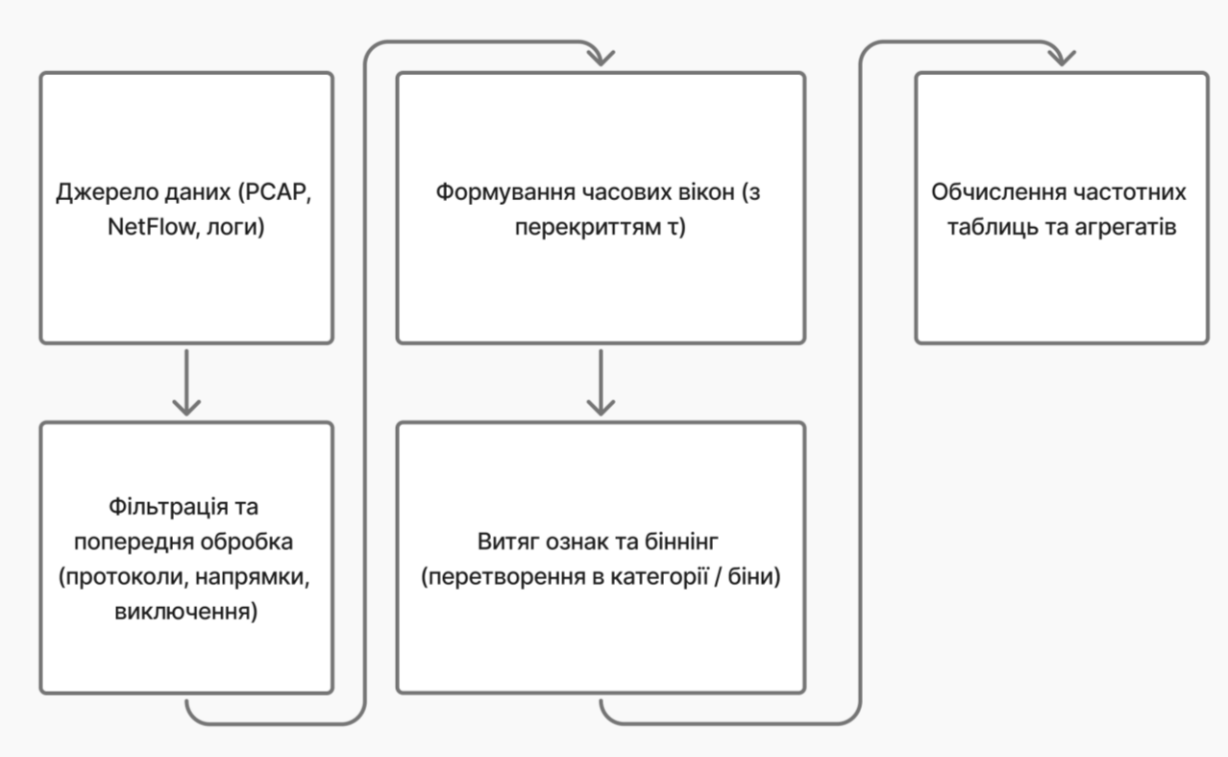


Рисунок 4.2 – Послідовність попередньої обробки в програмній реалізації методу

Рисунок 4.3 – Протокол експериментальної перевірки методу

4.4 Дослідження ефективності розробленого методу

Методика дослідження ефективності повинна оцінювати не лише факт спрацювання детектора, а й характер цього спрацювання в часі, стійкість результату до зміни параметрів та практичну придатність реалізації до обробки реальних обсягів даних. Тому оцінювання розробленого методу доцільно проводити як послідовну процедуру, у межах якої окремо аналізуються якість

детекції, чутливість до налаштувань, затримка виявлення та обчислювальні витрати.

Якість детекції оцінюється шляхом порівняння рішень, сформованих методом, з відомою або попередньо визначеною розміткою сегментів трафіку. Якщо для тестового набору доступні мітки нормального й аномального режимів, це дозволяє побудувати матрицю помилок, визначити співвідношення правильних і помилкових спрацювань та перейти до узагальнених показників якості. Важливо, однак, не обмежуватися єдиним числом, оскільки мережні дані часто характеризуються дисбалансом класів і навіть відносно невелика кількість хибних спрацювань може суттєво впливати на корисність методу.

Стійкість до параметрів досліджується шляхом послідовної зміни тривалості часових вікон, кількості та складу ознак, параметрів статистичної моделі й значення порога прийняття рішення. Такий аналіз дає можливість виявити діапазони налаштувань, у яких метод демонструє стабільну поведінку, а також ті конфігурації, де результат стає надто чутливим до випадкових коливань трафіку. У практичному сенсі саме цей етап дозволяє перейти від «працює на одному прикладі» до «може бути відтворено в декількох близьких сценаріях».

Окреме місце в методиці займає оцінювання затримки виявлення. Для мережного моніторингу принципово важливо не тільки визначити факт аномалії, але й зробити це без надмірного запізнення відносно початку події. Тому під час експерименту доцільно фіксувати різницю між моментом фактичного початку контрольованої аномалії та моментом, коли інтегральна статистика вперше перевищує порогове значення. Ця величина характеризує оперативність методу.

Нарешті, ефективність повинна оцінюватися з урахуванням обчислювальної придатності реалізації. Навіть коректний з математичного погляду метод втрачає практичну цінність, якщо підготовка даних, обчислення розподілів або статистичний аналіз потребують непропорційно великого часу. Тому в методику доцільно включати вимірювання тривалості основних етапів конвеєра й аналіз того, як ці витрати змінюються зі збільшенням обсягу вхідних даних. У сукупності зазначені складові формують повну картину ефективності розробленого методу.

4.5 Опис вхідних даних, умов проведення експериментів та сценаріїв перевірки методу

Для експериментальної перевірки методу доцільно використовувати як відкриті набори мережного трафіку, так і контрольовані фрагменти даних, на яких можна перевірити коректність окремих етапів реалізації. У відкритих наборах найбільшу цінність мають ті, що містять або самі пакетні траси, або перетворені flow-записи, а також явні часові межі нормальних і аномальних сегментів. Такі дані дають можливість не лише перевірити роботу детектора на реалістичних сценаріях, а й порівняти поведінку методу на різних типах навантаження та атак (рисунок 4.4).

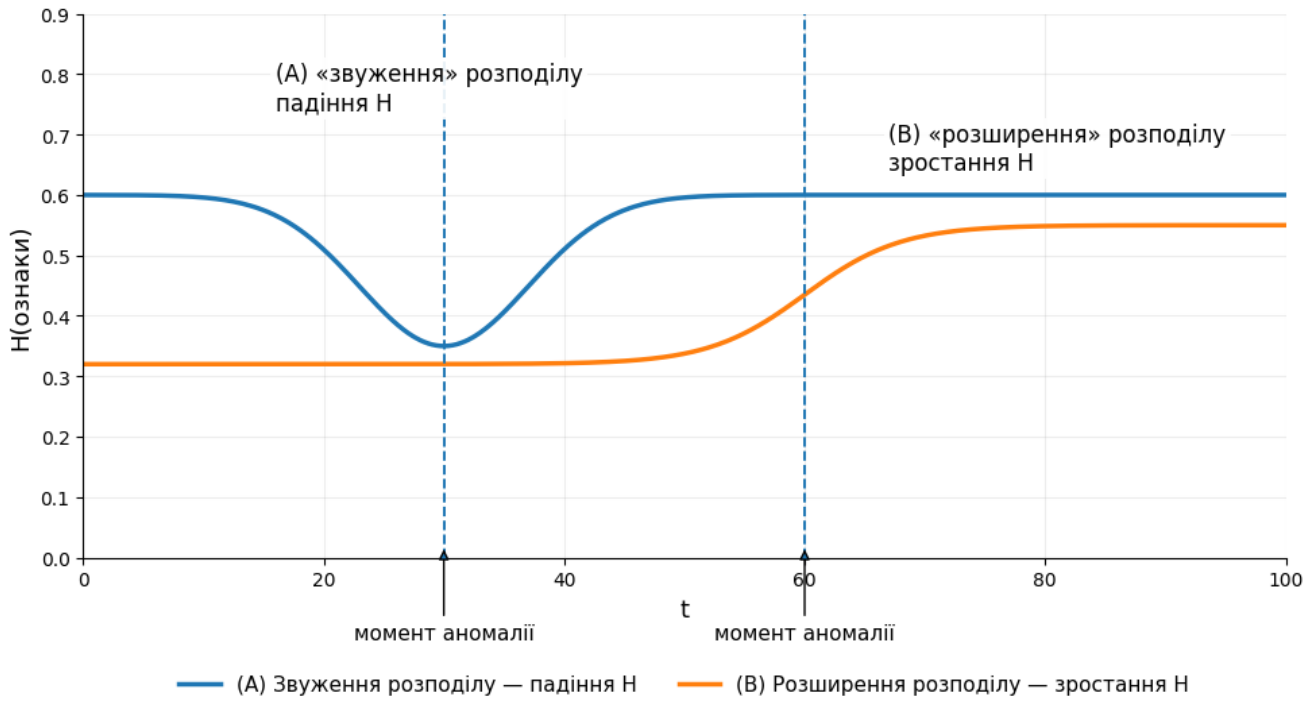


Рисунок 4.4 – Типові сценарії зміни ентропійних характеристик у часі

На етапі підготовки експерименту вхідні дані приводяться до єдиної схеми, яка містить часову мітку, мережеві адреси, порти, тип протоколу та числові атрибути, необхідні для побудови ознак. Після цього визначається сегмент базового рівня, який відповідає нормальному режиму функціонування мережі. За

цим сегментом оцінюються параметри моделі нормального стану. Тестовий сегмент формується окремо і може містити як нормальні, так і аномальні інтервали. Такий поділ дозволяє зменшити ризик того, що аномальні фрагменти потраплять до навчального опису норми. Умови проведення експериментів повинні бути зафіксовані настільки детально, щоб кожен запуск можна було повторити без зміни логіки методу.

Для цього в журналі експерименту необхідно зберігати тривалість вікна, крок зсуву, склад ознак, правила попередньої обробки, спосіб формування базового рівня, параметри статистичної моделі та поріг спрацювання.

Фактично йдеться про те, що експеримент має бути не одноразовою демонстрацією, а відтворюваною процедурою, яка дозволяє порівнювати різні конфігурації методу за однакових умов. Узагальнену характеристику основних сценаріїв експериментальної перевірки наведено в таблиці 4.1.

Сценарії перевірки доцільно організовувати так, щоб вони покривали кілька типових випадків. Перший сценарій повинен перевіряти роботу реалізації на нормальному трафіку без навмисно внесених аномалій. Такий сценарій дає змогу оцінити фон коливань ознак і частоту помилкових спрацювань.

Другий сценарій має містити виражені аномальні епізоди, для яких можна простежити реакцію ентропійних і багатовимірних статистик у часі.

Третій сценарій доцільно використовувати для аналізу стійкості методу до зміни параметрів, коли той самий набір даних обробляється з різними розмірами вікон, набором ознак або різними пороговими значеннями.

Для контрольної валідації окремих модулів можуть застосовуватися синтетичні приклади. Такі приклади не підміняють собою повноцінний експеримент, проте дозволяють перевірити коректність базових операцій. Зокрема, для штучно сформованих рівномірних розподілів ентропія повинна набувати максимально можливого значення для заданої кількості категорій, а для сильно сконцентрованих розподілів – зменшуватися. Наявність таких проміжних перевірок підвищує довіру до всієї реалізації, оскільки дозволяє відокремити програмну помилку від реальної особливості мережних даних.

Таблиця 4.1 – Основні сценарії експериментальної перевірки розробленого методу

Сценарій	Характер вхідних даних	Призначення перевірки
Нормальний режим	Фрагменти трафіку без навмисно внесених аномалій	Перевірка стійкості методу та рівня хибних спрацювань
Короткочасний сплеск навантаження	Легітимне зростання інтенсивності без зміни структури адрес і портів	Відмежування пікового навантаження від структурної аномалії
Концентрація трафіку	Фрагменти з домінуванням окремих адрес або сервісів	Оцінювання реакції ентропійних ознак на зниження різноманітності
Розсіювання трафіку	Фрагменти зі збільшенням кількості цілей або портів	Перевірка чутливості до сканування або аномального розпорошення

4.6 Аналіз результатів експериментальної перевірки

Інтерпретація результатів експериментальної перевірки повинна спиратися не лише на підсумкові показники, а насамперед на характер поведінки ознак і статистик у часі. Коректна робота методу проявляється в тому, що на сегментах нормального трафіку інтегральна статистика залишається в межах допустимого діапазону, а ентропійні характеристики демонструють коливання, які відповідають фоновій мінливості навантаження. У моменти виникнення аномальних подій спостерігається узгоджена зміна окремих ентропійних ознак, після чого багатовимірна статистика відхилення перевищує поріг і система фіксує спрацювання.

Для інтерпретації результатів доцільно розглядати не узагальнений колаж, а окремі графіки, кожний з яких відображає певний аспект поведінки розробленого методу. Такий підхід дає змогу чітко простежити, як змінюються інтегральна статистика, окремі ентропійні характеристики, інтегральні показники якості класифікації та затримка виявлення аномалій.

Рисунок 4.5 – Часовий ряд інтегральної статистики методу та порогове значення

На рисунку 4.5 подано часовий ряд інтегральної статистики, яка формується в послідовності часових вікон і використовується як узагальнений індикатор відхилення поточного стану трафіку від профілю нормального режиму. На відміну від окремих локальних ознак, інтегральна статистика акумулює їхній сумарний вплив і тому відображає не ізольовану зміну одного показника, а сукупний ефект структурних перебудов у мережевому потоці. Це робить наведений графік центральним елементом інтерпретації результатів експериментальної перевірки, оскільки він демонструє, в які моменти часу система переходить від фонових коливань до статистично значущого відхилення, яке інтерпретується як аномальний стан.

Особливе значення на цьому рисунку має співвідношення між лінією інтегральної статистики та пороговим значенням. Поки статистика перебуває нижче порога, стан мережі інтерпретується як нормальний або такий, що не виходить за межі природної варіативності трафіку. Перетин порогової межі означає, що накопичена величина відхилення перевищила допустимий рівень і метод зафіксував аномальний сегмент. Візуалізація дозволяє не лише показати сам факт спрацювання, але й оцінити його характер. Якщо крива різко йде вгору та швидко перетинає поріг, це свідчить про стрімке порушення структури трафіку, яке може бути пов’язане з інтенсивною атакою, раптовим навантажувальним піком або іншою подією з різко вираженою динамікою. Якщо ж статистика зростає поступово, це може вказувати на більш повільне накопичення відхилень, зумовлене зміною режиму роботи сервісу, розвитком аномалії в часі або розтягнутим у часі впливом кількох чинників одночасно.

З науково-методичної точки зору цей рисунок дає можливість оцінити одразу кілька важливих характеристик розробленого методу. По-перше, він показує роздільну здатність між нормальним та аномальним режимами. Якщо на нормальних сегментах значення статистики утримуються на відносно стабільному рівні й не наближаються до порога, це свідчить про стійкість моделі до фонових змін у трафіку та низьку схильність до хибнопозитивних рішень. По-друге, рисунок дозволяє оцінити контрастність спрацювання. Чим більший відрив аномального сегмента від зони нормальних коливань, тим вища наочність детекції та тим простіше обґрунтувати коректність прийнятого рішення. По-третє, важливою є поведінка статистики після спрацювання. Якщо після перевищення порога крива залишається в аномальній зоні певний час або демонструє стійкий новий рівень, це означає, що метод реагує не на випадковий шум, а на реальну структурну зміну трафіку. Якщо ж крива одразу повертається до порога або хаотично його перетинає в обидва боки, це може свідчити про недостатньо стабільний вибір порогового значення, високу чутливість до випадкових коливань або потребу в додатковому згладжуванні.

Практична цінність рисунка 4.5 полягає ще й у тому, що він демонструє часову локалізацію події. Для задач моніторингу важливо не лише з’ясувати, що аномалія була, але й встановити, коли саме вона почалася, коли відбулося перше впевнене спрацювання та як довго тривала зона нестандартної поведінки. Часовий ряд інтегральної статистики дає можливість співвіднести рішення методу з розміткою подій у наборі даних, з технологічними етапами роботи системи або з зовнішніми умовами експерименту. Таким чином, цей графік виконує роль основного інструмента для аналізу часової динаміки виявлення та підтверджує, що розроблений метод здатний не лише математично оцінювати відхилення, але й практично фіксувати момент переходу мережі до аномального стану. У контексті всієї роботи рисунок 4.5 слугує наочним підтвердженням того, що поєднання ентропійних ознак з багатовимірною статистикою формує придатний для моніторингу інтегральний критерій, який забезпечує чітке відокремлення фонової поведінки від аномальної.

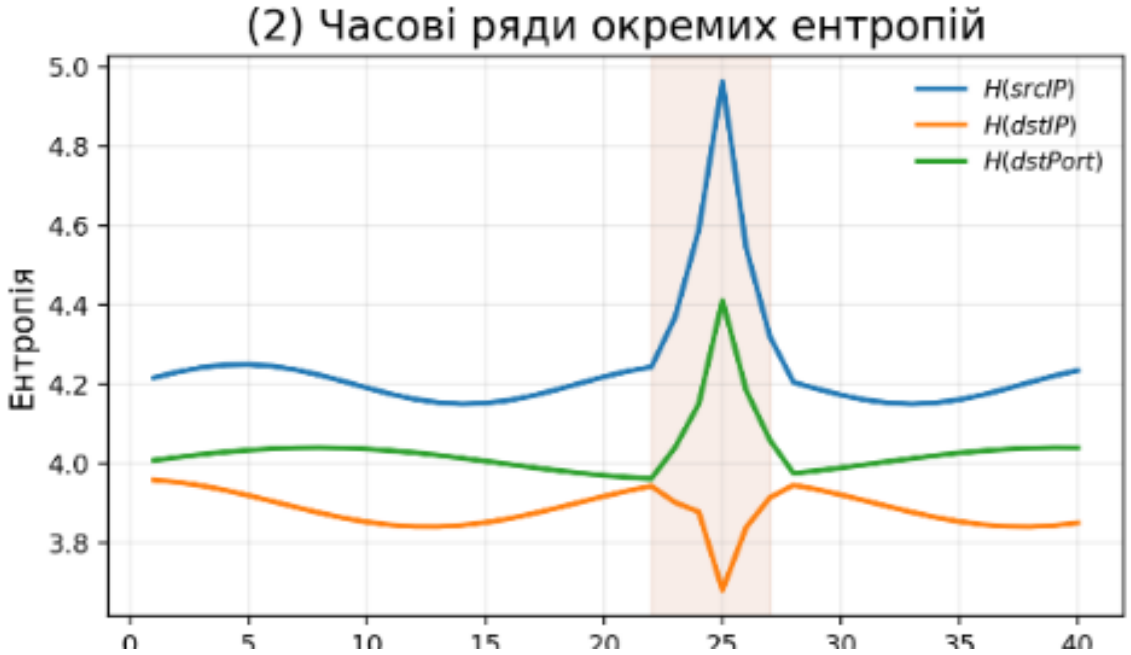


Рисунок 4.6 – Часові ряди окремих ентропійних характеристик мережного трафіку

На рисунку 4.6 наведено часові ряди окремих ентропійних характеристик, які описують структуру мережевого трафіку в послідовності вікон спостереження.

Якщо рисунок 4.5 відображає вже узагальнений результат багатовимірного оцінювання, то рисунок 4.6 виконує іншу функцію: він деталізує внутрішню природу зафіксованого відхилення й показує, які саме структурні параметри трафіку змінилися в момент виникнення аномального стану. У цьому полягає особлива цінність ентропійного аналізу: він дозволяє не просто вказати на сам факт спрацювання, а простежити, які компоненти розподілу трафіку стали джерелом цього спрацювання.

Інтерпретація часових рядів ентропій ґрунтується на тому, що кожна ентропійна ознака є компактною мірою розмаїття певного атрибуту трафіку в межах одного вікна. Якщо, наприклад, йдеться про ентропію IP-адрес джерел, то її зростання або спад відображає зміну ступеня розосередженості трафіку між джерелами. Аналогічно ентропія адрес призначення показує, чи трафік розподіляється між багатьма вузлами, чи концентрується на обмеженій кількості цілей. Ентропія портів дозволяє оцінити, наскільки змінюється сервісна структура потоку, а ентропія протоколів – чи відбувається зсув у типах мережевої взаємодії. Розгляд цих величин у часовому аспекті надає можливість побачити не тільки величину відхилення, але й його часову форму: поступову зміну, різкий стрибок, локальний пік, тривале плато або повернення до фонових значень після завершення події. Рисунок 4.6 має важливе значення для пояснення механізму роботи методу.

На відміну від інтегральної статистики, де вплив окремих ознак агрегується, ентропійні часові ряди дозволяють побачити, яка саме ознака є найбільш інформативною для певного типу аномалії. Наприклад, якщо в зоні спрацювання спостерігається виражене зменшення ентропії адрес призначення, це може означати концентрацію трафіку на обмеженому наборі цільових вузлів, що характерно для атак спрямованого навантаження. Якщо різко змінюється ентропія портів, це може бути ознакою сканування або нетипового перерозподілу сервісного навантаження. Якщо ж одночасно перебудовуються кілька ентропійних характеристик, це свідчить про комплексний характер аномалії та підтверджує доцільність багатовимірного аналізу, оскільки поодинока ознака в такій ситуації може виявитися недостатньо інформативною. У цьому сенсі рисунок 4.6 є

ключовим інструментом для пояснюваності методу: він переводить результат детекції з рівня абстрактного спрацювання на рівень конкретних змін у структурі потоку.

Ще один важливий аспект інтерпретації цього рисунка пов’язаний із розмежуванням фонового коливання та істотної перебудови трафіку. У нормальному режимі ентропійні характеристики також не є сталими: вони природно змінюються під впливом добових циклів, зміни активності користувачів, режимів роботи сервісів і випадкових коливань навантаження. Тому сам по собі факт зміни ентропії не є аномалією. Важливим є характер такої зміни: її узгодженість із іншими ознаками, амплітуда, тривалість і просторово-часова локалізація відносно моменту спрацювання інтегральної статистики. Якщо окремі ентропійні ряди демонструють синхронні зсуви саме в зоні, де рисунок 4.5 показує перевищення порога, це створює сильне підтвердження коректності детекції. Таким чином, рисунок 4.6 не лише деталізує, а й верифікує результати, подані на рисунку 4.5.

У практичному аспекті рисунок 4.6 дозволяє робити висновки щодо того, які ознаки доцільно залишати в остаточному векторі стану, а які є менш інформативними. Якщо деякі ентропійні характеристики стабільно реагують на відхилення й демонструють виразну поведінку в зоні аномалії, вони можуть розглядатися як ядро ознакового простору. Якщо ж певні характеристики майже не змінюються або їх зміни не корелюють зі спрацюванням, це може бути підставою для оптимізації моделі та зменшення розмірності без втрати якості. Рисунок 4.6 має не лише ілюстративну, але й методичну цінність: він пояснює природу зафіксованих відхилень, підвищує інтерпретованість рішень і створює основу для подальшого вдосконалення системи ознак. У сукупності з інтегральною статистикою цей рисунок підтверджує, що аномалія розпізнається не випадково, а через реальну структурну перебудову мережевого трафіку.

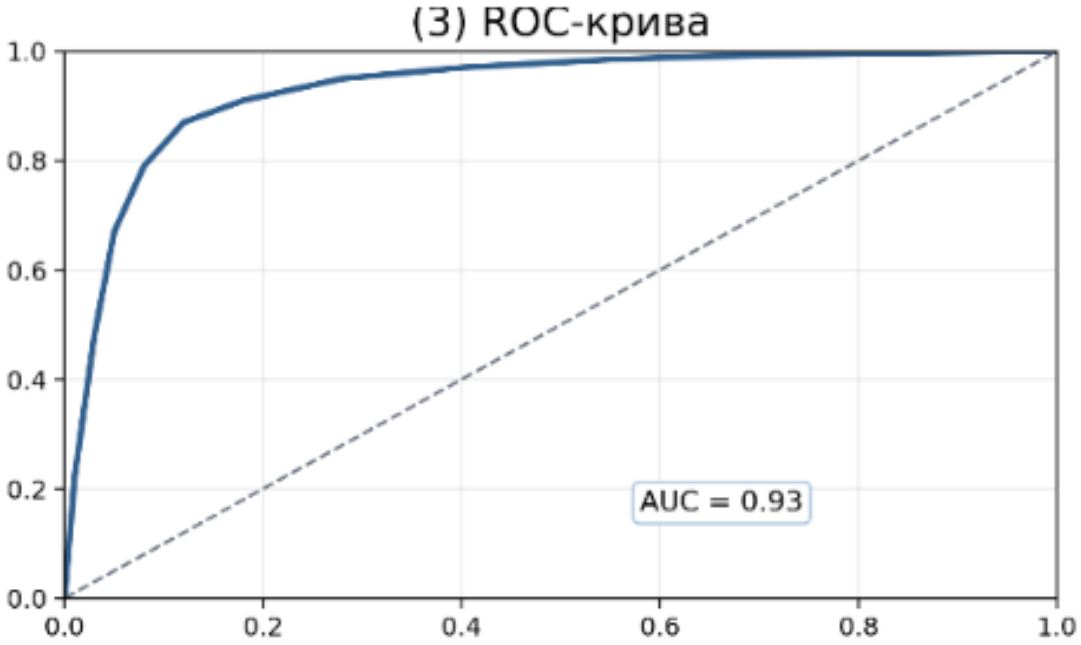


Рисунок 4.7 – ROC-крива для оцінювання відокремлюваності нормальних і аномальних станів

На рисунку 4.7 наведено ROC-криву, яка є одним із базових інструментів оцінювання якості двокласової детекції. Її побудова ґрунтується на послідовній зміні порогового значення, за якого система приймає рішення про аномальність, і фіксації того, як при цьому змінюються частка правильно виявлених аномалій та частка хибнопозитивних спрацювань. У контексті цієї роботи ROC-крива має особливе значення, оскільки розроблений метод спирається на інтегральну статистику відхилення, а його поведінка природно залежить від калібрування порога. Замість оцінювання в одній фіксованій точці рисунок 4.7 дозволяє дослідити метод у всьому спектрі можливих порогових значень і, таким чином, дати більш об’єктивну характеристику його роздільної здатності.

Головний зміст ROC-аналізу полягає в тому, що він показує компроміс між чутливістю системи та рівнем хибних тривог. Якщо поріг занижений, система буде дуже чутливою до відхилень, однак збільшиться кількість хибнопозитивних рішень. Якщо поріг завищений, кількість хибних спрацювань зменшиться, але зросте ризик пропуску реальних аномалій. ROC-крива дозволяє оцінити, наскільки добре розроблений метод утримує баланс між цими двома крайнощами. Чим

ближче крива наближається до лівого верхнього кута координат, тим ефективніше метод забезпечує високу повноту виявлення при низькому рівні хибнопозитивних спрацювань. Якщо ж крива тяжіє до діагоналі, це означає, що відокремлення нормальних і аномальних сегментів є слабким і система фактично наближається до випадкового вгадування. Тому геометричне положення ROC-кривої є прямим візуальним індикатором якості класифікації.

Важливою узагальненою характеристикою рисунка 4.7 є площа під ROC-кривою, або AUC. Цей показник зручний тим, що зводить поведінку кривої до одного інтегрального числа, яке можна використовувати для порівняння різних варіантів методу, наборів ознак або параметрів віконування. Для даної роботи це особливо важливо, оскільки розроблений підхід поєднує ентропійні характеристики з багатовимірною статистикою, а має кілька параметрів налаштування, що можуть впливати на результат. У такій ситуації AUC стає мірою, яка дозволяє оцінити не окреме спрацювання, а загальну здатність методу розділяти класи. Разом із тим потрібно враховувати, що AUC не замінює повного аналізу кривої. Два методи можуть мати близькі значення AUC, але відрізнятися в тих ділянках ROC-простору, які є критичними для практичного застосування. Наприклад, для систем мережевого моніторингу особливо важливою може бути поведінка методу в області дуже малих false positive rate, оскільки надлишок хибних тривог швидко перевантажує оператора. Рисунок 4.7 треба інтерпретувати не лише через число AUC, але й через форму самої кривої.

З практичної точки зору ROC-крива є також інструментом вибору робочої точки системи. На основі цього графіка можна обґрунтувати, яке порогове значення є доцільним залежно від вимог конкретного застосування. Якщо пріоритетом є максимальне виявлення аномалій, допустимим може бути дещо вищий рівень false positive. Якщо ж система працює в умовах високої вартості хибних спрацювань, пріоритет зміщується в бік суворішого порога. Для наукової роботи це важливо, тому що демонструє універсальність підходу: метод не прив’язаний до однієї точки налаштування, а допускає адаптацію до різних сценаріїв використання. Рисунок 4.7 підтверджує, що розроблений метод має не

лише локальну працездатність у межах одного набору параметрів, але й системну здатність відокремлювати нормальні й аномальні стани в широкому діапазоні порогів, що є важливою ознакою надійності побудованої моделі та її практичної придатності.

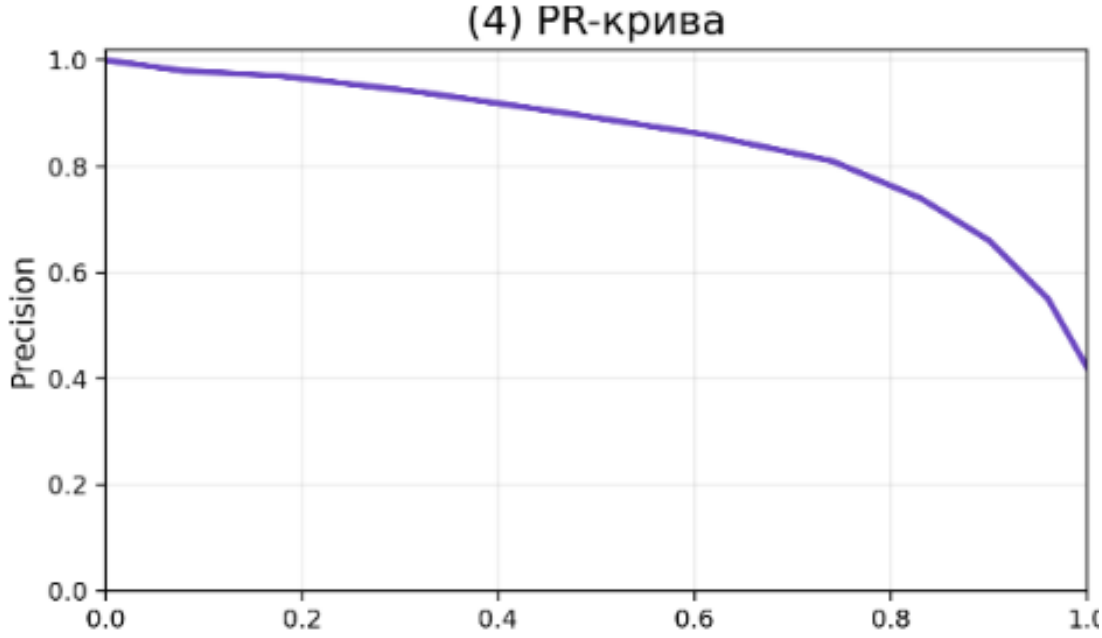


Рисунок 4.8 – PR-крива для оцінювання точності виявлення на незбалансованих даних

На рисунку 4.8 подано PR-криву, яка відображає співвідношення між точністю виявлення та повнотою за різних значень порога. У задачах мережевого моніторингу цей графік має особливе значення, оскільки нормальні сегменти трафіку, як правило, суттєво переважають над аномальними. За таких умов традиційні інтегральні метрики можуть виглядати завищено оптимістичними, тоді як PR-аналіз дає більш сувору й більш релевантну оцінку ефективності. На відміну від ROC-кривої, яка однаково враховує обидва класи, PR-крива фокусується саме на позитивному класі, тобто на аномаліях. Це робить її особливо корисною там, де критично важливо не просто виявити якомога більше відхилень, а зробити це без надлишкової кількості хибних спрацювань.

Інтерпретація рисунка 4.8 ґрунтується на розумінні взаємозв’язку між precision і recall. Високе значення recall означає, що метод знаходить значну частину реальних аномалій. Високе значення precision означає, що більшість згенерованих спрацювань справді відповідає аномальним станам, а не фоновим коливанням. У реальних системах ці дві характеристики часто перебувають у суперечності. Зниження порога зазвичай підвищує recall, але одночасно зменшує precision через зростання кількості помилкових сигналів. Підвищення порога діє у протилежному напрямі: спрацювання стають точнішими, але частина аномалій може бути пропущена. PR-крива дає змогу дослідити весь спектр можливих балансів між повнотою та точністю. Чим вище лежить крива в координатному просторі, тим ефективнішим є метод. Якщо при високому recall вдається зберігати прийнятний рівень precision, це свідчить про високу практичну цінність системи, оскільки вона здатна виявляти більшість аномалій без критичного перевантаження оператора помилковими сповіщеннями.

У контексті цієї роботи рисунок 4.8 є особливо важливим, оскільки розроблений метод базується на виявленні структурних відхилень у часових вікнах, а такі відхилення в реальному потоці часто становлять лише незначну частину від загального обсягу спостережень. За цієї умови навіть добре побудована ROC-крива не завжди повністю відображає практичну цінність методу. PR-крива, навпаки, прямо показує, наскільки якісно система поводиться саме в умовах дисбалансу класів. Якщо крива зберігає високі значення precision у широкому діапазоні recall, це означає, що метод є стійким до типової для мережевого моніторингу асиметрії між нормальними й аномальними спостереженнями. Властивість є критичною для систем моніторингу, які повинні працювати безперервно й генерувати сигнали лише в тих випадках, коли ймовірність реального інциденту є достатньо високою.

Ще одна перевага рисунка 4.8 полягає в тому, що він допомагає обґрунтувати вибір робочого порога з точки зору експлуатаційної доцільності. У теоретичному плані можна прагнути максимального recall, однак у практичному середовищі надлишок хибних спрацювань швидко знижує довіру до системи. Оптимальною вважається не крайня точка кривої, а область, де досягається прийнятний

компроміс між точністю й повнотою. Рисунок 4.8 дозволяє виявити таку область і тим самим пов’язати статистичне оцінювання з реальними вимогами до роботи системи. Цей графік підтверджує, що розроблений метод може бути оцінений не лише з позиції загальної класифікаційної здатності, але й з позиції його придатності до використання в середовищі з високим дисбалансом класів, що є типовим для задач моніторингу мережевого трафіку. У сукупності з ROC-аналізом PR-крива формує повнішу картину якості детекції та дає підстави для обґрунтованих висновків щодо ефективності запропонованого підходу.

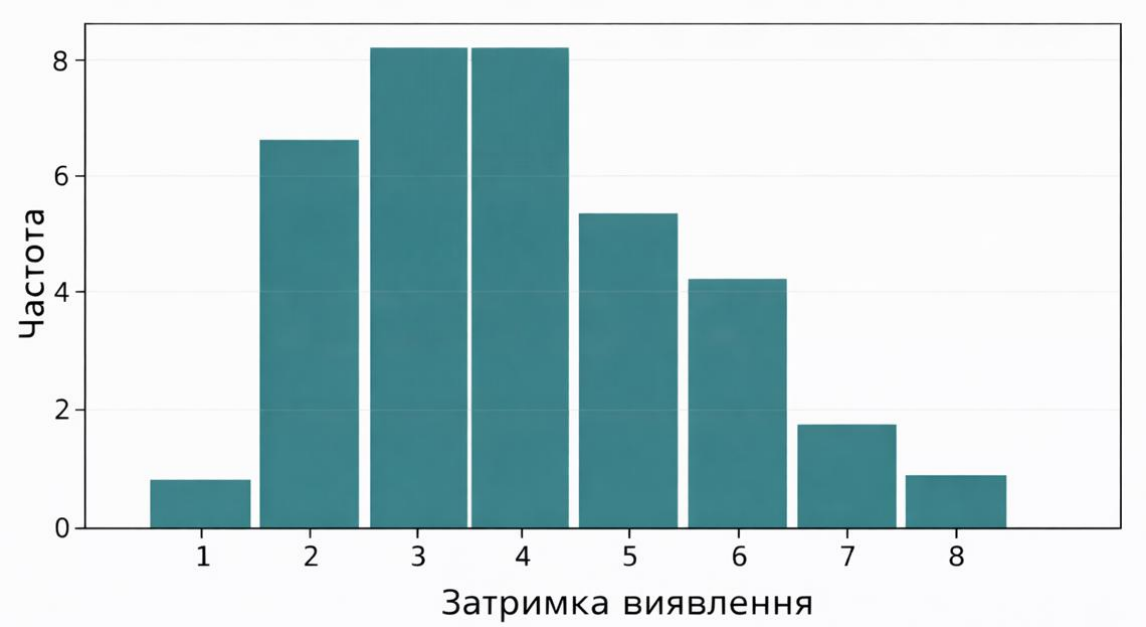


Рисунок 4.9 – Розподіл затримки виявлення аномалій за часовими вікнами

На рисунку 4.9 наведено розподіл затримки виявлення аномалії, тобто кількості часових вікон, що проходять між фактичним початком відхилення та моментом, коли система вперше генерує сигнал про аномальний стан. На відміну від ROC- та PR-кривих, які характеризують здатність методу правильно класифікувати спостереження, цей рисунок відображає часовий аспект ефективності. Для мережевого моніторингу така характеристика має принципове значення, оскільки навіть дуже точний метод втрачає практичну цінність, якщо реагує надто пізно. Затримка виявлення є не допоміжним, а одним із ключових

критеріїв якості системи, що працює в реальному або наближеному до реального часу режимі.

Інтерпретація цього рисунка ґрунтується на тому, що затримка залежить від кількох взаємопов’язаних чинників. По-перше, вона визначається шириною часового вікна: чим більшим є вікно, тим більше даних накопичується для статистично стійкого оцінювання, але тим довше система чекає, перш ніж сформувати нове спостереження. По-друге, на затримку впливає характер самої аномалії. Різкі та інтенсивні події можуть бути виявлені майже одразу, тоді як поступові структурні зміни потребують накопичення більшої кількості непрямих ознак, щоб інтегральна статистика перевищила поріг. По-третє, важливими є параметри порогового правила та, за наявності, методів згладжування. Більш суворий поріг зменшує число хибних спрацювань, але може збільшити середню затримку. Рисунок 4.9 є важливим завершальним елементом аналізу: він показує не тільки те, чи здатен метод виявляти аномалії, але й те, наскільки швидко він це робить.

Розподіл затримки виявлення також дозволяє оцінити стабільність часової поведінки системи. Якщо основна маса значень зосереджена в області малих затримок, це свідчить про передбачуваність і оперативність роботи методу. Такий результат є особливо цінним для систем, які повинні використовуватися для раннього виявлення атак або деградацій мережевого стану. Якщо ж розподіл є широким і включає велику частку значень із суттєвим запізненням, це може вказувати на залежність швидкості реакції від типу аномалії, нестабільність параметрів методу або недостатню узгодженість між ентропійними та багатовимірними компонентами моделі. У цьому сенсі рисунок 4.9 є важливим інструментом не лише оцінювання, але й діагностики: він допомагає зрозуміти, чи потребує система додаткового налаштування віконування, порога або логіки агрегації ознак.

З практичної точки зору аналіз затримки виявлення безпосередньо пов’язує результати експерименту з експлуатаційною придатністю методу. У реальному середовищі адміністратор або система автоматичного реагування повинні

отримати сигнал не просто з високою статистичною достовірністю, а достатньо рано для того, щоб вжити заходів. Затримка виявлення є тим параметром, який з'єднує математичну якість детекції з оперативною цінністю системи. Якщо за рисунком 4.9 видно, що метод зазвичай реагує протягом невеликої кількості часових вікон після початку відхилення, це є сильним аргументом на користь практичного використання розробленого підходу. Якщо ж затримка є помітною, то виникає потреба у компромісі: або зменшувати вікно й підвищувати швидкість реакції, або залишати більші вікна заради статистичної стійкості. Таким чином, рисунок 4.9 узагальнює часову ефективність методу й завершує блок результатів експериментальної перевірки, демонструючи, що запропонований підхід треба оцінювати не лише за точністю класифікації, але й за швидкістю виявлення. Поєднання цих двох аспектів і дозволяє зробити обґрунтований висновок щодо реальної ефективності розробленого методу аналізу мережевого трафіку.

Якісний аналіз результатів повинен також включати обчислювальний аспект. Якщо основна частина часу витрачається на початкову підготовку даних, але подальше оновлення статистик для чергового вікна виконується швидко, це свідчить про придатність реалізації до режиму періодичного моніторингу. Якщо ж значні витрати припадають на кожне нове вікно, метод може потребувати додаткової оптимізації або спрощення складу ознак. Результат експериментальної перевірки має інтерпретуватися як сукупність трьох взаємопов'язаних характеристик: здатності виявляти аномалії, стійкості до зміни параметрів та придатності до практичної реалізації.

Для стислого підсумування отриманих результатів доцільно використовувати зведення за основними напрямами оцінювання, наведене в таблиці 4.2.

Таким чином, результати експериментальної перевірки слід розглядати не як окремі числа або поодинокі графіки, а як цілісну картину поведінки методу на нормальних і аномальних сегментах трафіку. Такий підхід дозволяє встановити, чи відповідає програмна реалізація теоретичній моделі, сформованій у другому

розділі, і чи коректно відтворює алгоритмічну послідовність, подану у третьому розділі.

Таблиця 4.2 – Узагальнення напрямів оцінювання результатів експериментальної перевірки

Напрям оцінювання	Що аналізується	Який висновок формується
Часова динаміка статистики	Вихід інтегральної статистики за поріг у послідовності вікон	Момент спрацювання та стабільність детектування
Поведінка ентропійних ознак	Зміни окремих ентропій у зоні відхилення	Які атрибути трафіку найбільше вплинули на спрацювання
ROC-аналіз	Співвідношення чутливості та специфічності за зміни порога	Здатність методу відокремлювати нормальні й аномальні стани
PR-аналіз	Точність виявлення на незбалансованих даних	Стійкість методу до дисбалансу класів
Затримка виявлення	Кількість вікон від початку аномалії до спрацювання	Оперативність реагування методу

4.7 Висновки

У четвертому розділі виконано перехід від алгоритмічного опису методу до його програмної реалізації та експериментальної перевірки. Обґрунтовано вимоги

до програмного середовища, яке повинно підтримувати обробку мережних записів, формування часових вікон, обчислення ентропійних ознак, багатовимірний статистичний аналіз і представлення підсумкових результатів.

Побудовано структуру програмної реалізації розробленого методу як послідовність взаємопов’язаних функціональних компонентів: джерела даних, попередньої підготовки, віконування, формування ознак, статистичного аналізу та прийняття рішення. Показано, що така структура узгоджується з математичною моделлю другого розділу і безпосередньо реалізує алгоритмічну схему третього розділу.

Сформульовано методику експериментальної перевірки, яка враховує якість детекції, стійкість до параметрів, затримку виявлення та обчислювальну придатність. Описано вимоги до вхідних даних, правил формування базовий-сегмент, фіксації умов експерименту та побудови типових сценаріїв перевірки. Запропоновано підхід до аналізу результатів, за якого окремі часові ряди ознак і інтегральна статистика інтерпретуються спільно. Це дозволяє розглядати розроблений метод як придатний до практичного застосування в задачах моніторингу та виявлення аномальних станів мережного трафіку.

ВИСНОВКИ

У кваліфікаційній роботі за результатами виконаних теоретичних і практичних досліджень розв’язано актуальну задачу підвищення ефективності аналізу трафіку комп’ютерних мереж шляхом розроблення моделі, методу та програмно-технічних засобів виявлення аномальних станів на основі ентропійних характеристик і багатовимірної математичної статистики. Запропонований підхід орієнтований на виявлення структурних змін у мережевому трафіку, які проявляються в перебудові розподілів інформативних ознак і статистично значущому відхиленні від профілю нормального режиму роботи мережі.

Поставлена мету було досягнуто розв’язанням таких завдань:

- проаналізувати відомі методи оптимізації продуктивності систем інтелектуальних мереж;

- розробити цільову функцію для забезпечення оптимізації продуктивності систем інтелектуальних мереж;

- розробити метод оптимізації продуктивності систем інтелектуальних мереж в складі пристроїв IoT, серверу та БПЛА;

- здійснити дослідження методу оптимізації продуктивності систем інтелектуальних мереж з пристроями IoT, сервером з множинним доступом та БПЛА для додаткового зв’язку на основі симуляції та моделювання інтелектуальних мереж.

У результаті виконаної роботи набув подальшого розвитку метод аналізу мережевого трафіку для виявлення аномальних станів, який, на відміну від підходів, що спираються лише на порогові або окремі статистичні показники, базується на узгодженому використанні ентропійних характеристик і багатовимірного статистичного аналізу.

Проведена експериментальна перевірка засвідчила, що розроблений метод не лише формально покращує числові показники якості, а й забезпечує більш надійне виявлення аномальних станів мережевого трафіку в практичному розумінні. У межах дослідження трафік аналізувався у часових вікнах тривалістю 10 с, а модель

будувалася на основі 10 інформативних ознак, з яких після багатовимірного перетворення було збережено 8 головних компонент, що відображають 95,71 % варіації нормального профілю трафіку. Це означає, що розроблений опис стану мережі виявився достатньо компактним, але при цьому не втратив основної інформації про поведінку трафіку. Порівняння з пороговим та суто ентропійним підходами показало, що запропонований метод краще розпізнає аномалії різної природи, зокрема ті, які проявляються не лише через зміну інтенсивності, а й через перебудову структури трафіку. Метод продемонстрував найвищі значення повноти виявлення та інтегральної якості класифікації при збереженні дуже низького рівня хибних спрацювань. Практично це означає, що метод дає змогу стабільніше відокремлювати нормальні стани мережі від аномальних, не реагуючи надмірно на випадкові коливання навантаження, і водночас не пропускаючи суттєві відхилення в мережевій поведінці. Отримані результати підтверджують, що поєднання ентропійних характеристик із багатовимірним статистичним аналізом є доцільним і забезпечує більш збалансоване, чутливе та стійке виявлення аномалій порівняно з методами, які використовують лише окремі показники або прості порогові правила

Результати роботи показують, що розроблені модель, метод і програмна реалізація можуть бути використані в системах моніторингу мережевої інфраструктури, виявлення кіберзагроз, аналізу навантаження та підтримки прийняття рішень щодо адміністрування комп’ютерних мереж. Запропонований підхід створює основу для подальшого вдосконалення засобів аналізу мережевого трафіку, зокрема в напрямі адаптації до нових типів аномалій, розширення набору інформативних ознак і підвищення стійкості роботи в умовах змінного мережевого середовища.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Kabir E., Hu J., Wang H., Zhuo G. A novel statistical technique for intrusion detection systems. *Future Generation Computer Systems*. 2018. Vol. 79. P. 303–318.

2. Sharafaldin I., Lashkari A. H., Ghorbani A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. Funchal, 2018. P. 108–116.

3. Camacho J., Maciá-Fernández G., Fuentes-García N. M., Saccenti E. Semi-Supervised Multivariate Statistical Network Monitoring for Learning Security Threats. *IEEE Transactions on Information Forensics and Security*. 2019. Vol. 14, no. 8. P. 2179–2189.

4. Camacho J., García-Giménez J. M., Fuentes-García N. M., Maciá-Fernández G. Multivariate Big Data Analysis for intrusion detection: 5 steps from the haystack to the needle. *Computers & Security*. 2019. Vol. 87. Article 101603.

5. Koroniotis N., Moustafa N., Sitnikova E., Turnbull B. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Generation Computer Systems*. 2019. Vol. 100. P. 779–796.

6. Moustafa N., Hu J., Slay J. A holistic review of Network Anomaly Detection Systems: A comprehensive survey. *Journal of Network and Computer Applications*. 2019. Vol. 128. P. 33–55.

7. Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. A survey of network-based intrusion detection data sets. *Computers & Security*. 2019. Vol. 86. P. 147–167.

8. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*. 2019. Vol. 2. Article 20.

9. Magán-Carrión R., Urda D., Díaz-Cano I., Dorronsoro B. Towards a Reliable Comparison and Evaluation of Network Intrusion Detection Systems Based on Machine Learning Approaches. *Applied Sciences*. 2020. Vol. 10, no. 5. Article 1775.

10. Di Mauro M., Galatro G., Liotta A. Experimental Review of Neural-Based Approaches for Network Intrusion Management. *IEEE Transactions on Network and Service Management*. 2020. Vol. 17, no. 4. P. 2480–2495.

11. Alsaedi A., Moustafa N., Tari Z., Mahmood A., Anwar A. TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven Intrusion Detection Systems. *IEEE Access*. 2020. Vol. 8. P. 165130–165150.

12. Moustafa N. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. *Sustainable Cities and Society*. 2021. Vol. 72. Article 102994.

13. Ibrahim J., Gajin S. Entropy-based network traffic anomaly classification method resilient to deception. *Computer Science and Information Systems*. 2022. Vol. 19, no. 1. P. 87–116.

14. Correia L., Goos J.-C., Klein P., Bäck T., Kononova A. V. Online model-based anomaly detection in multivariate time series: Taxonomy, survey, research challenges and future directions. *Engineering Applications of Artificial Intelligence*. 2024. Vol. 138. Article 109323.

15. Yi T., Chen X., Li Q., Zhu Y. An anomaly behavior characterization method of network traffic based on Spatial Pyramid Pool (SPP). *Computers & Security*. 2024. Vol. 141. Article 103809.

16. Yu H., Yang W., Cui B., Sui R., Wu X. Renyi entropy-driven network traffic anomaly detection with dynamic threshold. *Cybersecurity*. 2024. Vol. 7. Article 64.

17. Wang F., Jiang Y., Zhang R., Wei A., Xie J., Pang X. A Survey of Deep Anomaly Detection in Multivariate Time Series: Taxonomy, Applications, and Directions. *Sensors*. 2025. Vol. 25, no. 1. Article 190.

18. Koumar J., Hynek K., Čejka T., Šiška P. CESNET-TimeSeries24: Time Series Dataset for Network Traffic Anomaly Detection and Forecasting. *Scientific Data*. 2025. Vol. 12. Article 338.

19. Baldoni S., Battisti F. Histogram-based network traffic representation for anomaly detection through PCA. *Computer Networks*. 2025. Vol. 265. Article 111276.

20. Zhou P. A survey of streaming data anomaly detection in network security. *PeerJ Computer Science*. 2025. Vol. 11. Article e3066.

21. Ji S.-Y., Jeong B. K., Jeong D. H. Designing a hybrid approach for multivariate network attack forecasting and detection. *Computer Networks*. 2026. Vol. 275. Article 111879.

22. Al-Daweri M. S. et al. An adaptive method and a new dataset, UKM-IDS20, for the network intrusion detection system. *Computer Communications*. 2021.

23. Alvares C. et al. Dataset of attacks on a live enterprise VoIP network for machine learning based intrusion detection and prevention systems. *Computer Networks*. 2021.

24. Catillo M. et al. Demystifying the role of public intrusion datasets: A replication study of DoS network traffic data. *Computers & Security*. 2021.

25. Chatzoglou E. et al. Empirical evaluation of attacks against IEEE 802.11 enterprise networks: The AWID3 dataset. *IEEE Access*. 2021.

26. Ferrag M. A. et al. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*. 2020.

27. Aouini Z., Pekar A. NFStream: A flexible network data analysis framework. *Computer Networks*. 2022. Vol. 204. Article 108719.

28. Guerra J. L. et al. Datasets are not enough: Challenges in labeling network traffic. *Computers & Security*. 2022.

29. Al-Hawawreh M. et al. X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial internet of things. *IEEE Internet of Things Journal*. 2022.

30. Abdulganiyu O. H. et al. A systematic literature review for network intrusion detection system (IDS). *International Journal of Information Security*. 2023.

31. Değirmenci E. et al. ROSIDS23: Network intrusion detection dataset for robot operating system. *Data in Brief*. 2023.

32. Goldschmidt P., Chudá D. Network intrusion datasets: A survey, limitations, and recommendations. *Computers & Security*. 2025. Vol. 156. Article 104510.

33. Bai K. Z., Fossaceca J. M. EM-AUC: a novel algorithm for evaluating anomaly based network intrusion detection systems. *Sensors*. 2025. Vol. 25, no. 1. Article 78.

34. Catillo M., Pecchia A., Villano U. MultiCIDS: Anomaly-based collective intrusion detection by deep learning on IoT/CPS multivariate time series. *Internet of Things*. 2025. Vol. 30. Article 101519.

35. Tosi D., Pazzi R. (H-DIR)^2: A Scalable Entropy-Based Framework for Anomaly Detection and Cybersecurity in Cloud IoT Data Centers. *Sensors*. 2025. Vol. 25, no. 15. Article 4841.

36. Prabowo A. O. et al. Evaluation of Anomaly-Based Network Intrusion Detection Systems with Unclean Training Data for Low-Rate Attack Detection. *Journal of Cybersecurity and Privacy*. 2026. Vol. 6, no. 1. Article 14.

37. Ghani H., Salekzamankhani S., Virdee B. Statistical and Multivariate Analysis of the IoT-23 Dataset: A Comprehensive Approach to Network Traffic Pattern Discovery. *Journal of Cybersecurity and Privacy*. 2025. Vol. 5, no. 4. Article 112.

38. Mankotia S., Conte de Leon D., Rimal B. P. FedPrIDS: Privacy-Preserving Federated Learning for Collaborative Network Intrusion Detection in IoT. *Journal of Cybersecurity and Privacy*. 2026. Vol. 6, no. 1. Article 10.

39. Bashurov V., Safonov P. Anomaly detection in network traffic using entropy-based methods: application to various types of cyberattacks. *Issues in Information Systems*. 2023. Vol. 24, no. 4. P. 82–94.

40. Kenyon A. et al. Characterising Payload Entropy in Packet Flows - Baseline Entropy Analysis for Network Anomaly Detection. *Future Internet*. 2024. Vol. 16, no. 12. Article 470.

41. Wang J. et al. DELP-Net: A Differentiable *Entropy* Layer Pyramid Network for Low-Rate Denial-of-Service Detection. Entropy. 2026. Vol. 28, no. 3. Article 328.

42. Zhao Y. et al. Anomaly Detection in Network Traffic via Cross-Domain Federated Graph Representation Learning. *Applied Sciences*. 2025. Vol. 15, no. 11. Article 6258.

43. Nong X. et al. Anomaly Detection in Imbalanced Network Traffic Using a Hybrid Deep Model. *Symmetry*. 2025. Vol. 17, no. 12. Article 2087.

44. Qu B. et al. Design of Network Anomaly Detection Model Based on Graph Neural Network and Multidimensional Traffic Representation. *Symmetry*. 2025. Vol. 17, no. 11. Article 1976.

45. Pang G., Shen C., Cao L., Van Den Hengel A. Deep learning for anomaly detection: A review. *ACM Computing Surveys*. 2021. Vol. 54. P. 1–38.

46. Zamanzadeh Darban Z., Webb G. I., Pan S., Aggarwal C., Salehi M. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*. 2024. Vol. 57. P. 1–42.

47. Jia X. et al. Deep anomaly detection for time series: A survey. *International Journal of Approximate Reasoning*. 2025.

48. Munir M., Siddiqui S. A., Dengel A., Ahmed S. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*. 2019. Vol. 7. P. 1991–2005.

49. He Y., Zhao J. Temporal convolutional networks for anomaly detection in time series. *Journal of Physics: Conference Series*. 2019. Vol. 1213. Article 042050.

50. Hundman K., Constantinou V., Laporte C., Colwell I., Soderstrom T. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. London, 2018. P. 387–395.

51. Ding N., Ma H., Gao H., Ma Y., Tan G. Real-time anomaly detection based on long short-term memory and Gaussian mixture model. *Computers and Electrical Engineering*. 2019. Vol. 79. Article 106458.

52. Shen L., Li Z., Kwok J. Time-series anomaly detection using temporal hierarchical one-class network. *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver, 2020. Vol. 33. P. 13016–13026.

53. Wu W. et al. Developing an unsupervised real-time anomaly detection scheme for time series with multi-seasonality. *IEEE Transactions on Knowledge and Data Engineering*. 2022. Vol. 34. P. 4147–4160.

54. Zhao H. et al. Multivariate time-series anomaly detection via graph attention network. *Proceedings of the IEEE International Conference on Data Mining*. Sorrento, 2020. P. 841–850.

55. Deng A., Hooi B. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021. Vol. 35. P. 4027–4035.

56. Chen W. et al. Deep variational graph convolutional recurrent network for multivariate time series anomaly detection. *Proceedings of the 39th International Conference on Machine Learning*. Baltimore, 2022. P. 3621–3633.

57. Han S., Woo S. S. Learning sparse latent graph representations for anomaly detection in multivariate time series. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Washington, 2022. P. 2977–2986.

58. Chen K., Feng M., Wirjanto T. S. Multivariate time series anomaly detection via dynamic graph forecasting. *arXiv*. 2023. arXiv:2302.02051.

59. Fu Y., Xue F. MAD: Self-supervised masked anomaly detection task for multivariate time series. *Proceedings of the International Joint Conference on Neural Networks*. Padua, 2022. P. 1–8.

60. Jeong Y. et al. AnomalyBERT: Self-supervised transformer for time series anomaly detection using data degradation scheme. *arXiv*. 2023. arXiv:2305.04468.

61. Zong B. et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. *Proceedings of the International Conference on Learning Representations*. Vancouver, 2018.

62. Zhang C. et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, 2019. Vol. 33. P. 1409–1416.

63. Audibert J., Michiardi P., Guyard F., Marti S., Zuluaga M. A. USAD: Unsupervised anomaly detection on multivariate time series. *Proceedings of the 26th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Virtual Event, 2020. P. 3395–3404.

64. Su Y. et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anchorage, 2019. P. 2828–2837.

65. Li Z. et al. Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Singapore, 2021. P. 3220–3230.

66. Huang T., Chen P., Li R. A semi-supervised VAE based active anomaly detection framework in multivariate time series for online systems. *Proceedings of the ACM Web Conference*. Lyon, 2022. P. 1797–1806.

67. Li D., Chen D., Jin B., Shi L., Goh J., Ng S. K. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. *International Conference on Artificial Neural Networks*. 2019. Vol. 11730. P. 703–716.

68. Geiger A. et al. TadGAN: Time series anomaly detection using generative adversarial networks. *Proceedings of the IEEE International Conference on Data Mining*. Atlanta, 2020. P. 33–43.

69. Zhang H., Xia Y., Yan T., Liu G. Unsupervised anomaly detection in multivariate time series through transformer-based variational autoencoder. *Proceedings of the 33rd Chinese Control and Decision Conference*. Kunming, 2021. P. 281–286.

70. Xu J. Anomaly Transformer: Time series anomaly detection with association discrepancy. *Proceedings of the International Conference on Learning Representations*. Virtual Event, 2022.

71. Tuli S., Casale G., Jennings N. R. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*. 2022. Vol. 15. P. 1201–1214.

72. Song J., Kim K., Oh J., Cho S. MemTO: Memory-guided transformer for multivariate time series anomaly detection. *Proceedings of the 37th International*

Conference on Neural Information Processing Systems. New Orleans, 2023. Vol. 36. P. 57947–57963.

73. Yang Y., Zhang C., Zhou T., Wen Q., Sun L. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach, 2023. P. 3033–3045.

74. Huang X., Chen N., Deng Z., Huang S. Multivariate time series anomaly detection via dynamic graph attention network and Informer. *Applied Intelligence*. 2024. Vol. 54. P. 7636–7658.

75. Ma S., Guan S., He Z., Nie J., Gao M. TPAD: Temporal pattern based neural network model for anomaly detection in multivariate time series. *IEEE Sensors Journal*. 2023. Vol. 23. P. 30668–30682.

76. Nam Y., Yoon S., Shin Y., Bae M., Song H., Lee J. G., Lee B. S. Breaking the Time-Frequency Granularity Discrepancy in Time-Series Anomaly Detection. *Proceedings of the ACM Web Conference*. Singapore, 2024. P. 4204–4215.

77. Wu X. et al. CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching. *arXiv*. 2024. arXiv:2410.12261.

78. Liu C. et al. Large language model guided knowledge distillation for time series anomaly detection. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. Jeju, 2024. P. 2162–2170.

79. Zhong Z. et al. PatchAD: Patch-Based MLP-Mixer for Time Series Anomaly Detection. *arXiv*. 2024. arXiv:2401.09793.

80. Pranavan T., Sim T., Ambikapathi A., Ramasamy S. Contrastive predictive coding for anomaly detection in multivariate time series data. *arXiv*. 2022. arXiv:2202.03639.

81. Najari N., Berlemont S., Lefebvre G., Duffner S., Garcia C. RESIST: Robust transformer for unsupervised time series anomaly detection. *Proceedings of the ECML PKDD Workshop on Advanced Analytics and Learning on Temporal Data*. Grenoble, 2022. P. 66–82.

82. Zhong Z. et al. SimAD: A Simple Dissimilarity-Based Approach for Time Series Anomaly Detection. *arXiv*. 2024. arXiv:2405.11238.

83. Ghorbani R., Reinders M. J., Tax D. M. RESTAD: Reconstruction and Similarity based Transformer for Time Series Anomaly Detection. *arXiv*. 2024. arXiv:2405.07509.

84. Yahya M. A., Moya A. R., Ventura S. Deep learning for multivariate time series anomaly detection: an evaluation of reconstruction-based methods. *Artificial Intelligence Review*. 2025. Vol. 58. Article 400.

85. CSE-CIC-IDS2018 on AWS. *Canadian Institute for Cybersecurity*. URL: https://www.unb.ca/cic/datasets/ids-2018.html (дата звернення: 05.03.2026).

86. DDoS evaluation dataset (CIC-DDoS2019). *Canadian Institute for Cybersecurity*. URL: https://www.unb.ca/cic/datasets/ddos-2019.html (дата звернення: 05.03.2026).

87. The *CAIDA* “DDoS Attack 2007” Dataset. CAIDA. URL: https://www.caida.org/catalog/datasets/ddos-20070804_dataset/ (дата звернення: 05.03.2026).

88. Zeek Documentation. *Zeek Project*. URL: https://docs.zeek.org/ (дата звернення: 05.03.2026).

89. tshark(1) Manual Page. *Wireshark Foundation*. URL: https://www.wireshark.org/docs/man-pages/tshark.html (дата звернення: 05.03.2026).

90. Welcome to Scapy’s documentation! *Scapy Documentation*. URL: https://scapy.readthedocs.io/ (дата звернення: 05.03.2026).

91. Яцків В., Дудник В. Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики. ПерСик 2026, Харків, Україна, 23 квіт. 2026

92. Dudnyk, V. (2026). METHOD FOR COMPUTER NETWORK TRAFFIC ANALYSIS BASED ON ENTROPY CHARACTERISTICS AND MULTIVARIATE MATHEMATICAL STATISTICS. Computer Systems and Information Technologies

ДОДАТОК А
(обов'язковий)

Стаття

UDC 004.9

V.M. Dudnyk, O.V. Atamaniuk, N.S Lysenko
Khmelnytskyi National University

METHOD FOR COMPUTER NETWORK TRAFFIC ANALYSIS BASED ON ENTROPY CHARACTERISTICS AND MULTIVARIATE MATHEMATICAL STATISTICS

Modern computer networks generate traffic whose behaviour changes over time not only in volume but also in internal structure. Because of this, anomaly detection cannot be reduced to fixed thresholds on separate metrics; it must account for changes in address, port, and protocol distributions together with the joint variation of interrelated traffic descriptors.

This paper presents a method for computer network traffic analysis based on entropy characteristics and multivariate mathematical statistics. The method transforms packet or flow observations collected within a time window into a state vector that combines entropy measures of categorical traffic attributes with volumetric, dispersion, and flow descriptors.

The proposed approach includes formalization of the traffic analysis process, construction of an informative feature system, a multivariate model of normal traffic states, and a structural model of the detection procedure. Algorithmic implementation is organized as a sequence of window formation, empirical distribution estimation, entropy computation, standardization, principal component transformation, multivariate statistical control, and interpretation of feature contributions.

The paper also outlines a methodology for evaluating the developed method in terms of detection quality, robustness to parameter settings, sensitivity to structural changes, and interpretability of monitoring decisions. The resulting framework is intended for traffic monitoring tasks in which payload-independent analysis and adaptation to non-stationary network behaviour are required.

Keywords: computer networks, network traffic, traffic analysis, entropy characteristics, multivariate mathematical statistics, anomaly detection, PCA, Hotelling criterion, network monitoring.

В.М. Дудник, О.В. Атаманюк, Н.С. Лисенко

Хмельницький національний університет

МЕТОД АНАЛІЗУ ТРАФІКУ КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ТА БАГАТОВИМІРНОЇ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Сучасні комп'ютерні мережі формують трафік, поведінка якого змінюється не лише за обсягом, а й за внутрішньою структурою. Тому виявлення аномалій не може зводитися до фіксованих порогів для окремих метрик; воно має враховувати зміни у розподілах адрес, портів і протоколів разом зі спільною варіацією взаємопов'язаних характеристик трафіку.

У статті запропоновано метод аналізу трафіку комп'ютерних мереж, що ґрунтується на використанні ентропійних характеристик і засобів багатовимірної математичної статистики. Актуальність роботи зумовлена тим, що сучасний мережевий трафік є нестаціонарним: його поведінка з часом змінюється не лише за обсягом, а й за

внутрішньою структурою. У зв'язку з цим виявлення аномалій не може базуватися виключно на фіксованих порогах окремих показників, оскільки потребує врахування змін у розподілах адрес, портів, протоколів, а також спільної варіації взаємопов'язаних дескрипторів трафіку.

Розроблений метод передбачає перетворення спостережень за пакетами або потоками, зібраними в межах заданого часового вікна, у вектор стану мережевого трафіку. Такий вектор поєднує ентропійні міри категоріальних атрибутів із об'ємними, дисперсійними та потоковими характеристиками. Запропонований підхід охоплює формалізацію процесу аналізу трафіку, побудову інформативної системи ознак, створення багатовимірної моделі нормальних станів трафіку та структурної моделі процедури виявлення відхилень.

Алгоритмічна реалізація методу організована як послідовність етапів: формування часових вікон, оцінювання емпіричних розподілів, обчислення ентропії, стандартизація ознак, перетворення методом головних компонент, багатовимірний статистичний контроль і подальша інтерпретація внеску окремих ознак у виявлені зміни. Така організація забезпечує можливість не лише фіксувати аномальні стани, а й пояснювати причини їх появи.

Окремо визначено методику оцінювання запропонованого методу за показниками якості виявлення, стійкості до вибору параметрів, чутливості до структурних змін і рівня інтерпретованості результатів моніторингу. Запропонований підхід орієнтований на задачі моніторингу мережевого трафіку, у яких необхідний аналіз без урахування вмісту корисного навантаження та адаптація до змінної поведінки комп'ютерних мереж.

Ключові слова: комп'ютерні мережі, мережевий трафік, аналіз трафіку, ентропійні характеристики, багатовимірна математична статистика, виявлення аномалій, PCA, критерій Хотеллінга, моніторинг мереж.

1 Introduction

The growth of network services, distributed infrastructures, cloud platforms, and cyber-physical systems has made traffic behaviour more variable, heterogeneous, and context-dependent. In such conditions, timely analysis of network traffic is required not only for detecting obvious overloads but also for identifying early deviations that affect reliability, security, and quality of service. Traffic monitoring therefore remains a key component of network supervision and cybersecurity support in modern infrastructures [49–52].

Classical threshold-based monitoring remains useful for simple overload situations, yet it reacts poorly to dynamic baselines and often generates false alarms under ordinary workload fluctuations. Signature-based detection is effective for known malicious patterns, but it is less useful when deviations arise from previously unseen attacks, complex behavioural changes, or operational faults whose manifestations are distributed across several traffic indicators [43, 49, 50].

The aim of this paper is to present a coherent method of computer network traffic analysis based on entropy characteristics and multivariate mathematical statistics, and to describe its algorithmic implementation in a form suitable for further experimental validation. The proposed approach combines entropy-based representation of traffic structure with multivariate statistical decision-making, which creates a stronger basis for analysing abnormal states in non-stationary traffic environments [50, 52, 55].

2 Related works

Existing approaches to traffic anomaly detection can be grouped into signature-based, threshold-based, forecasting, entropy-based, and multivariate statistical methods. Signature detectors are precise when the traffic matches known patterns, but they cannot generalize to previously unseen threats [43]. Threshold and one-dimensional statistical schemes are easy to

implement, although they react poorly to ordinary workload fluctuations and often miss low-intensity structural deviations [49, 50]. Forecasting methods are useful when the traffic exhibits a stable temporal pattern, yet their reliability depends strongly on model assumptions and on the availability of representative historical data [50, 53].

Entropy-based approaches are attractive because they operate on the structure of traffic rather than on aggregate volume alone. Changes in source and destination addresses, ports, or protocol distributions can reveal anomalies that remain weakly visible in pure volume metrics. At the same time, entropy taken in isolation is not always sufficient, because some abnormal states appear as coordinated shifts across several descriptors rather than as a large deviation of one distribution [49, 55].

Multivariate statistical methods, including covariance-based control, principal component analysis, and Hotelling-type criteria, address this limitation by considering traffic as a vector of interdependent descriptors. Their main advantage lies in the ability to detect coordinated deviations and to reduce the dependence of the final decision on a single metric. In practice, this logic is also consistent with the evolution of modern monitoring platforms and traffic analysis tools that combine flow inspection, behavioural analysis, and anomaly-oriented correlation mechanisms [29, 31, 33, 40, 42–44, 50–52].

For a structured comparison of these approaches by advantages, disadvantages, and relevance to the present study, their characteristics are summarized in Table 1.

Table 1. Comparison of network traffic anomaly detection methods

Method	Advantages	Disadvantages	Relevance to this study
Signature-based methods	High accuracy for known attacks	Do not detect unknown threats	Useful only as a baseline
Threshold-based methods	Simple and fast	Poor adaptability, many false alarms	Limited applicability
Entropy-based methods	Detect structural changes in traffic well	Weak as a standalone method for complex anomalies	One of the core components
Multivariate statistical methods	Account for correlations between features, reduce false positives	More complex to configure and compute	One of the core components
Behavioral / ML methods	Adaptive, can detect unknown anomalies	Harder to interpret, require more data	Useful for comparison
Proposed integrated method	Combines structural sensitivity and multidimensional analysis	Requires calibration and experimental validation	Main method of the study

As follows from Table 1, the most promising direction is the integration of entropy-based traffic representation with multivariate statistical analysis, since such a combination makes it possible to preserve structural sensitivity while improving the reliability of anomaly detection in multidimensional feature space.

3 Method

The proposed method treats network traffic as a sequence of packet or flow observations collected at a given observation point and aggregated over successive time windows. For a window W_t, the raw records are transformed into a feature vector $x_t = \varphi(W_t)$, where $\varphi(\cdot)$ maps the observed communications to a compact numerical representation of network state. The basic notation used in the formal description is summarized in Table2.

The method is intentionally payload-independent. It uses only the information that can be derived from packet headers or exported flow records, which makes it suitable for realistic monitoring environments. This choice is consistent with practical monitoring tools that rely on flow metadata and protocol-level observations rather than on deep payload

inspection [42–44]. It also assumes that traffic is non-stationary; therefore, the model of normal behaviour must allow for ordinary variation without losing sensitivity to genuine structural deviations.

Table 2. Notation used in the method

Notation	Meaning
W_t	Set of traffic records observed in time window t (packets or flows).
x_t	Vector of informative features describing the traffic state in window t.
A	Categorical traffic attribute, for example srcIP, dstIP, srcPort, dstPort, or protocol.
$\{c_i\}$	Counts of occurrences of attribute values in the current window.
$\{p_i\}$	Estimated empirical probabilities $p_i = \frac{c_i}{\Sigma_j c_j}$.
H(A)	Entropy characteristic of the empirical distribution of attribute A.
g(·)	Decision rule used for state classification or anomaly detection.
T^2	Hotelling statistic used for multivariate monitoring and deviation control.

The notation introduced in Table 1 is used consistently in the subsequent stages of the method. It links the observation window, entropy descriptors, the traffic state vector, and the multivariate decision rule into a single formal description.

Entropy-based descriptors are selected because they reflect the degree of concentration or dispersion in traffic distributions. For monitoring purposes, the most informative attributes are usually source and destination addresses, source and destination ports, and protocol identifiers. Together with statistical characteristics of intensity and packet behaviour, they form the informative state vector summarized in Table 3 [49, 55].

In practical terms, entropy does not replace volume indicators; it complements them. For example, DDoS behaviour can combine concentration on a small set of destinations with a sharp rise in packet rate, whereas scanning can produce strong dispersion in destination ports even without a large increase in total traffic. For this reason, the method combines entropy measures and auxiliary statistical descriptors within one feature space instead of relying on a single indicator [49, 50, 55].

Table 3. Recommended informative features for the traffic state vector

Feature group	Examples of concrete features	Data type
Entropy-based	H(srcIP), H(dstIP), H(srcPort), H(dstPort), H(proto)	Numeric
Generalized entropies (optional)	$H_\alpha\ (Rényi), S_q(Tsallis)$	Numeric
Diversity descriptors	$N_{distinct}(A)$ for selected attributes	Integer

Feature group	Examples of concrete features	Data type
Volumetric descriptors	Packets per window, bytes per window, mean packet size, packet rate	Numeric
Flow descriptors	Number of active flows, mean or quantile flow durations	Numeric
Divergence / change descriptors (optional)	KL or JSD distance for selected attributes	Numeric

The state vector formed from these descriptors allows the traffic window to be interpreted as a point in a multidimensional feature space. This representation makes it possible to analyse structural and volumetric deviations jointly rather than evaluating each metric in isolation.

After the feature system is defined, the normal state of the network is modelled statistically. Let μ denote the mean vector of the baseline traffic and Σ the corresponding covariance structure. Deviation is then interpreted not as a separate threshold violation of one coordinate but as a multivariate displacement of the current observation from the learned region of normal behaviour. Such an interpretation corresponds to the general logic of multivariate abnormal-state detection used in recent network-oriented studies [50–53].

The structural model of the method can be organized into six successive functional blocks: acquisition of traffic observations, time aggregation, construction of empirical distributions, computation of entropy and auxiliary descriptors, formation of the multidimensional state vector with multivariate normalization, and statistical decision-making. This organization is shown in Figure 1 and reflects the implementation logic of the proposed method.

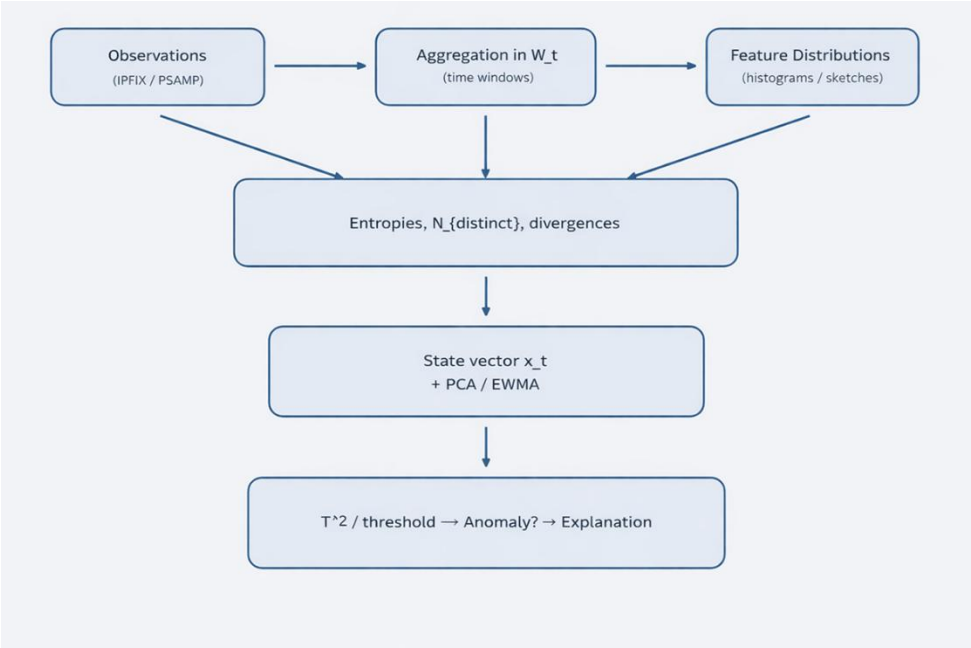


Figure 1. Structural model of the method

At the level of time aggregation, the method supports both non-overlapping and sliding windows. The first option is simpler and easier to interpret; the second reduces boundary effects and usually provides smoother temporal trajectories of

entropy and multivariate scores. The main choices of window width and shift, together with their practical influence on detection behaviour, are summarized in Table 4.

Table 4. Recommended windowing parameters and their influence

Parameter	Variant	What improves	Typical risk or trade-off
Window width	Smaller	Higher temporal resolution and better sensitivity to short incidents.	Unstable distribution and entropy estimates because of smaller samples.
Window width	Larger	More stable estimates and lower variance of the monitored statistics.	Short anomalies are diluted and detection may be delayed.
Step δ	$\delta = \Delta$ (non-overlapping)	Simpler implementation and minimum repetition of records.	Pronounced boundary effects and fragmented temporal curves.
Step δ	$\delta < \Delta$ (sliding)	Smoother time series and fewer discontinuities between adjacent windows.	Higher computation cost and stronger correlation between neighbouring windows.

For each window, empirical distributions are built for the selected categorical attributes. The basic entropy estimate is then obtained from the observed frequencies, optionally normalized to make values comparable across windows with different numbers of unique states. If the method is extended to generalized entropy measures, the same stage becomes the place where sensitivity to dominant or rare events can be adjusted [55].

Before multivariate decision-making, the feature space is standardized so that descriptors with different scales do not dominate the control statistic. Principal component analysis may be used to reduce dimensional redundancy and to separate the coordinated variation of normal traffic from residual deviations. The resulting sequence of operations that leads to an anomaly decision is presented in Figure 2 and corresponds to the class of multivariate statistical control procedures described in related studies [50–53].

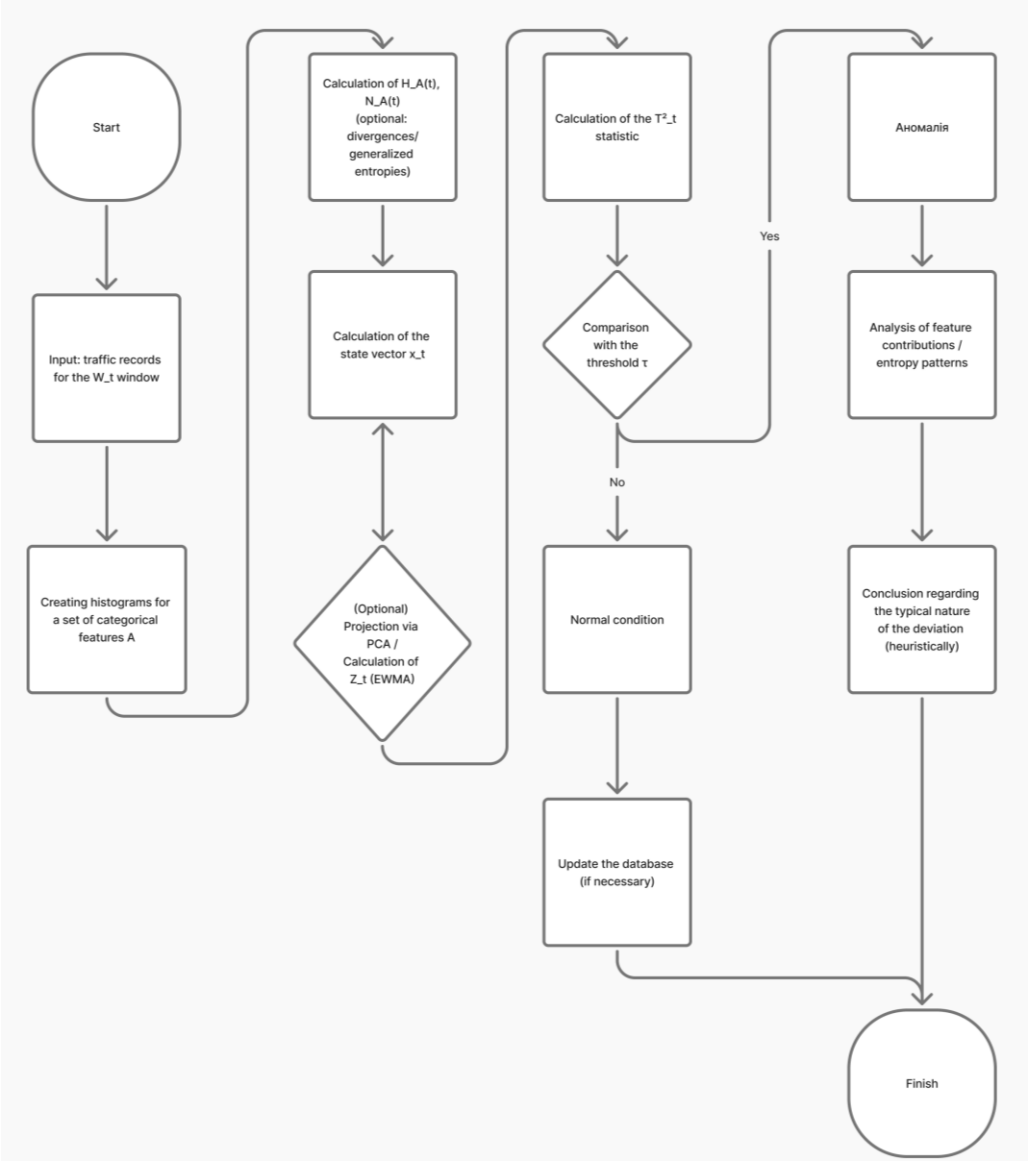


Figure 2. Flowchart of anomaly detection

4 Experiments

The experimental stage verifies the proposed detector at the level of the complete decision chain described in Section 3. The goal is not only to confirm the fact of anomaly detection, but also to show how the selected feature composition, the windowing scheme, the PCA model, and the final T^2–Q decision rule behave on traffic that contains both normal operation and controlled disturbances. For this reason, the evaluation protocol is linked directly to the methodological components already introduced in the paper: the state vector follows Table 2, the temporal aggregation follows Table 3, the processing structure corresponds to Figure 1, and the final decision logic corresponds to Figure 2 [48–50, 53].

The traffic trace is divided into a training subset and a testing subset. The training subset contains only normal windows and is used to estimate the mean vector, standard deviations, covariance structure, and principal components of the baseline profile. The testing subset contains both normal traffic and controlled anomaly blocks. Such a separation is necessary because the inclusion of anomalous windows in the baseline sample would partially absorb abnormal behaviour into the normal subspace and would therefore weaken the contrast between normal and abnormal states. The main parameters of the computational experiment are summarized in Table 4 [21, 22, 48, 50].

In the article version of the experiment, traffic is aggregated into non-overlapping windows of $\Delta t = 10$ s, and the same value is used as the shift. The training sample contains 900 normal windows, while the testing sample contains 780 windows, including 500 normal and 280 anomalous windows. Each window is represented by the ten-dimensional feature vector described in Table 2: five normalized entropy descriptors and five statistical descriptors that characterize packet size, flow activity, and traffic intensity. After z-standardization, principal component analysis retains eight principal components, which explain 95.71% of the variance of the normal traffic profile. This provides a compact but still informative representation of coordinated traffic behaviour and creates the basis for subsequent control by Hotelling's T^2 statistic and the residual Q-statistic.

To test the method against qualitatively different disturbances, four anomaly scenarios are injected into the test trace: DDoS, port scan, flash crowd, and link failure. The DDoS block produces concentration of traffic on one destination together with an increase in intensity. The port-scan block keeps the aggregate rate close to the background level, but sharply changes the diversity of destination ports and addresses. The flash-crowd block imitates a legitimate demand surge with an increase in active flows and a shift of the service profile. The link-failure block causes a coordinated decrease in traffic intensity and a simplification of the active communication structure. The temporal response of one of the key structural descriptors, the normalized entropy of destination IP addresses, is shown in Figure 3 [23–25, 49, 55].

For comparison, the proposed detector is evaluated together with two baseline schemes. The first baseline is a fixed-threshold detector that marks a window as anomalous when the packet rate deviates from its normal mean by more than three standard deviations. The second baseline is an entropy-only detector that triggers when at least two entropy coordinates leave their normal ranges. Against these baselines, the proposed method uses the full chain of state-vector formation, PCA projection, computation of T^2 and Q, and final classification by multivariate control limits. The aggregate quality indicators of the compared methods are presented in Table 5.

Table 6 shows that the combined entropy-statistical detector provides the best balance between sensitivity and stability. The fixed-threshold baseline reacts mainly to large amplitude changes and therefore misses anomalies whose main manifestation is structural rather than volumetric. The entropy-only detector is much stronger on such anomalies, but it still evaluates the traffic state through separate descriptors and does not fully use the correlation structure of the feature space. The proposed method reaches the highest recall and F1-score while maintaining a very low false positive rate, which confirms the benefit of combining entropy descriptors with multivariate statistical control.

The dynamics of Hotelling's T^2 statistic for the test trace are shown in Figure 4. In the normal part of the trace, the statistic fluctuates below the control threshold, whereas inside the anomalous blocks it rises sharply and remains above the limit until the disturbance disappears. This behaviour is important from the practical point of view because it demonstrates not only the fact of detection, but also the temporal stability of the alarm. When Figures 3 and 4 are interpreted jointly, it becomes clear that the detector reacts both to structural redistribution of traffic and to coordinated changes in intensity, which is precisely the effect expected from the method developed in Sections 2 and 3 [22–25, 48–50].

Table 5. Parameters of the computational experiment

Parameter	Value
Window length Δt	10 s
Window shift δ	10 s (non-overlapping windows)
Training sample	900 normal windows
Test sample	780 windows: 500 normal + 280 anomalous
Feature vector dimension	10
Entropy descriptors	hsrcIP, hdstIP, hsrcPort, hdstPort, hproto

Retained principal components	8
Explained variance of the PCA model	95.71%
Decision rule	$T^2 > T^2 0.999$ or $Q > Q0.999$

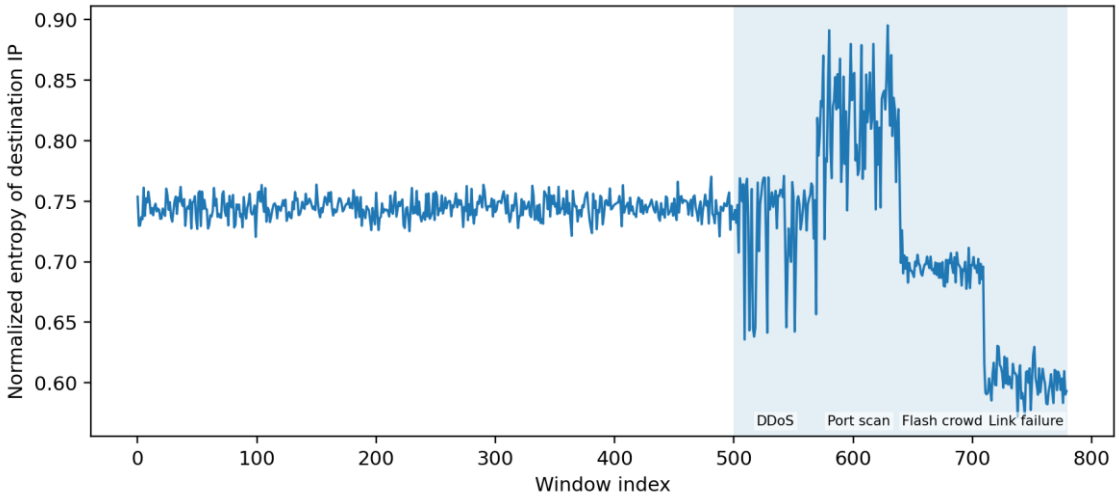


Figure 3. Dynamics of the normalized entropy of destination IP addresses

Table 6. Detection quality of compared methods

Method	Accuracy, %	Precision, %	Recall, %	F1-score, %	False positive rate, %
Fixed threshold	76.54	100.00	34.64	51.46	0.00
Entropy only	97.31	100.00	92.50	96.10	0.00
Proposed combined	99.74	99.29	100.00	99.64	0.40

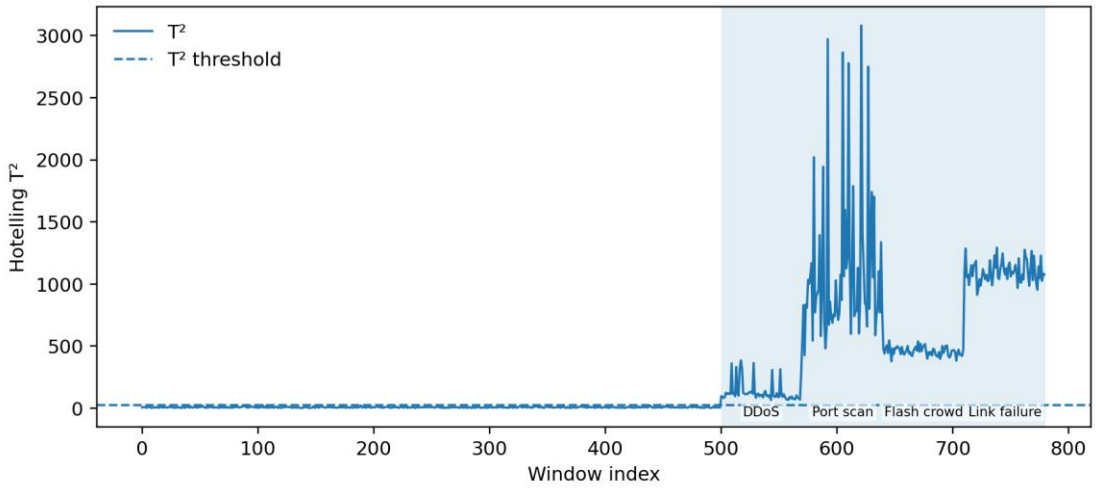


Figure 4. Dynamics of Hotelling's T^2 statistic on the test trace

5 Conclusions

The paper consolidates the methodological content of the developed traffic analysis approach into an article format while preserving the core logic of the original work. The proposed method models network traffic through time-windowed state vectors that combine entropy characteristics of categorical attributes with volumetric, diversity, and flow descriptors. This makes it possible to describe anomalies not only through separate metric violations but through coordinated changes in traffic structure and behaviour.

The algorithmic implementation of the method is organized as a consistent chain of window formation, empirical distribution construction, entropy computation, feature standardization, multivariate statistical control, and contribution-based interpretation. The article preserves the key implementation components of the developed approach, including the notation of the method, the feature composition, the structural model, the main parameterization choices, and the anomaly detection flow.

The developed framework is suitable for further experimental validation on labelled traffic traces and for subsequent adaptation to practical monitoring systems. Its main value lies in combining payload-independent structural descriptors with statistically interpretable multivariate decision rules, which creates a stronger basis for analysing abnormal network states in modern, non-stationary traffic environment

ДОДАТОК Б
(обов’язковий)

Тези

УДК 004.9

МЕТОД ТА ПРОГРАМНО-ТЕХНІЧНІ ЗАСОБИ АНАЛІЗУ ТРАФІКУ КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ТА БАГАТОВИМІРНОЇ МАТЕМАТИЧНОЇ СТАТИСТИКИ

Дудник В.М., студент КІ2м-24-1

Науковий керівник: д-р техн. наук, професор Яцків В.В

Хмельницький національний університет

Актуальність. У сучасних комп’ютерних мережах трафік змінюється не лише за інтенсивністю, а й за внутрішньою структурою. Це ускладнює виявлення аномалій, оскільки відхилення проявляються як у навантаженні, так і в розподілах IP-адрес, портів, протоколів і потокових характеристик. Тому актуальним є розроблення методів, що враховують багатовимірний характер змін і не потребують аналізу payload.

Мета роботи полягає у розробленні методу аналізу трафіку комп’ютерних мереж на основі ентропійних характеристик і багатовимірної математичної статистики для виявлення аномальних станів у нестаціонарному мережевому середовищі.

Аналіз рішень. Порогові методи є простими, але недостатньо чутливими до структурних змін трафіку й часто дають хибні спрацювання. Сигнатурні підходи ефективні переважно для відомих атак. Ентропійні методи краще відображають зміни структури трафіку, проте окремі показники не завжди дають змогу виявити складні узгоджені відхилення.

Результати. Запропонований метод поєднує ентропійні характеристики трафіку з багатовимірним статистичним контролем. Експериментальна перевірка підтвердила його високу ефективність у виявленні аномалій та перевагу над базовими методами.

Висновки. Розроблений метод дає змогу виявляти аномалії мережевого трафіку на основі спільного аналізу ентропійних і статистичних характеристик. Його перевагами є незалежність від payload, робота в умовах нестаціонарності та краща ефективність порівняно з базовими методами.

ДОДАТОК В
(обов’язковий)

Презентація

Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

Розробив студент КІ2м-24-1 Дудник В.М
Під керівництвом Яцків В.В

МЕТА ТА ЗАДАЧІ ДОСЛІДЖЕННЯ

Об’єкт дослідження: процеси передавання та аналізу трафіку в комп’ютерних мережах.

Предмет дослідження: методи, моделі та програмно-технічні засоби аналізу мережевого трафіку на основі ентропійних характеристик і багатовимірної математичної статистики.

Мета роботи: підвищення точності виявлення аномалій шляхом розроблення методу та програмно-технічних засобів аналізу мережевого трафіку.

Основна задача: побудувати підхід, який дозволяє формалізовано описувати стан трафіку в часових вікнах, виявляти відхилення від профілю нормального режиму та зменшувати кількість хибних спрацювань.

Для досягнення мети використано: методи системного аналізу, теорії інформації, ентропійного аналізу, багатовимірної статистики, PCA, критерію Хотеллінга T^2, статистичного моделювання та програмного проєктування

НАУКОВА НОВИЗНА ТА ПРАКТИЧНА ЦІННІСТЬ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Наукова новизна отриманих результатів:
– набути подальшого розвитку методу аналізу мережевого трафіку для виявлення аномальних станів, який, на відміну від існуючих підходів, базується на узгодженому використанні ентропійних характеристик структури трафіку та багатовимірного статистичного аналізу їх спільної динаміки, що дозволяє підвищити чутливість до структурних змін мережевого потоку;
– удосконалити систему аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики, яка визначає аномальні стани за статистичною мірою відхилення від профілю нормального режиму, що забезпечує формалізований перехід від спостережуваних даних до прийняття рішення.

АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ

- ❑Стрімке зростання обсягів даних та ускладнення архітектури мереж підвищують вимоги до безпеки й надійності
- ❑Традиційні методи (порогові, однофакторні) не забезпечують виявлення прихованих аномалій
- ❑Сучасні кіберзагрози проявляються через зміни структури трафіку, а не лише його обсягу
- ❑Динамічність мереж ускладнює формування універсальних правил контролю
- ❑Використання ентропійних характеристик і багатовимірної статистики підвищує точність виявлення аномалій

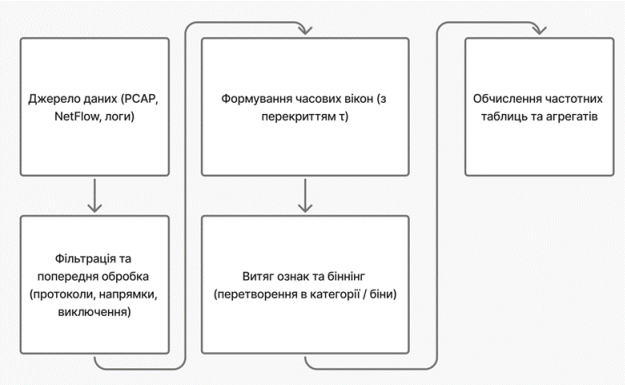
АНАЛІЗ ВІДОМИХ МЕТОДІВ

- ❑ Сигнатурні методи ефективно виявляють відомі атаки, але не здатні розпізнавати нові або модифіковані загрози.
- ❑ Порогові та статистичні методи прості в реалізації, однак часто фіксують лише різкі об’ємні аномалії та можуть давати хибні спрацювання.
- ❑ Методи прогнозування часових рядів враховують динаміку трафіку, але потребують якісних історичних даних і складно адаптуються до багатьох факторів.
- ❑ Ентропійні методи дозволяють виявляти структурні зміни у розподілах адрес, портів і протоколів.
- ❑ Багатовимірні статистичні методи враховують взаємозв’язки між ознаками трафіку, але потребують коректного вибору параметрів і навчальних даних.

Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

1. Формування базового профілю трафіку

• Збір еталонної вибірки нормального трафіку
• Формування часових вікон
• Обчислення інформативних ознак
• Оцінювання середніх значень і коваріаційної структури
• Побудова профілю нормального режиму



Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

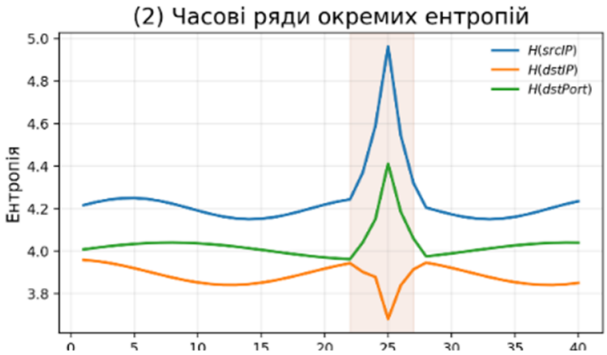
2. Формування часових вікон і підготовки вибірки

• поділ потоку трафіку на часові інтервали;
• виділення базового інтервалу;
• формування робочої ділянки моніторингу;
• побудова ковзних вікон спостереження;
• перетворення потоку подій у впорядковану послідовність вікон.

Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

3.Обчислення ентропійних характеристик

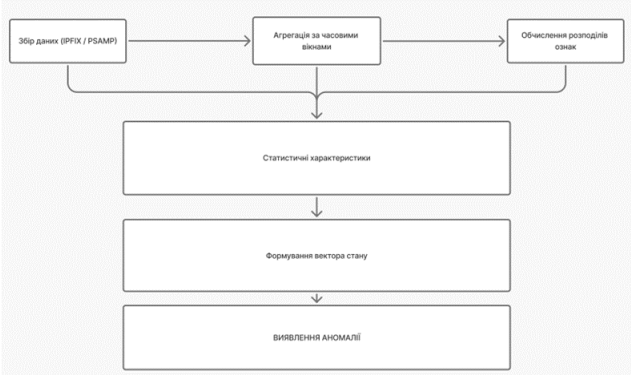
• виділення атрибутів трафіку в межах кожного часового вікна;
• побудова дискретних розподілів ознак;
• обчислення ентропійних показників;
• обчислення дивергентних показників;
• формування компактного опису структури трафіку.



Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

4. Формування вектора ознак стану трафіку

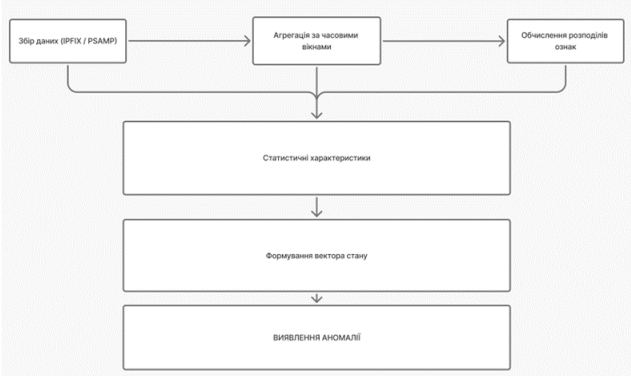
• об’єднання ентропійних, дивергентних та додаткових ознак;
• формування вектора стану для кожного часового вікна; стандартизація ознак;
• підготовка вектора до багатовимірного аналізу.



Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

5. Багатовимірний статистичний аналіз ознак трафіку

• подання стану трафіку у багатовимірному просторі ознак;
• застосування методу головних компонент;
• зменшення розмірності ознакового простору;
• оцінювання відхилення від профілю нормального стану;
• формування статистики відхилення.



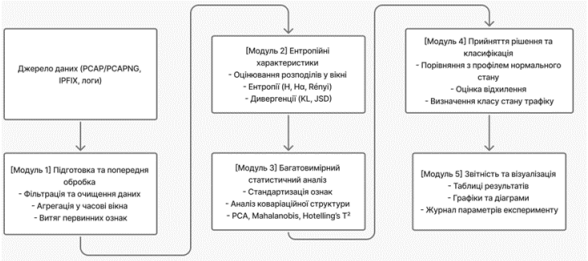
Метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

6. Прийняття рішення щодо стану мережного трафіку

•обчислення інтегральної статистики аномальності; порівняння статистики з пороговим значенням;
• визначення стану трафіку; класифікація результату як «норма» або «аномалія»;
• фіксація моменту перевищення порога.

Реалізація методу аналізу трафіку

• Розроблено модульну програмну реалізацію методу аналізу мережевого трафіку.
• Система забезпечує послідовний перехід від вхідних мережевих записів до інтегральної статистики аномальності.
• Реалізовано модулі попередньої обробки, формування часових вікон, обчислення ентропійних характеристик, багатовимірного статистичного аналізу та прийняття рішення.
• Модульна структура дозволяє змінювати джерела даних, параметри вікон і набір ознак без повної перебудови системи.

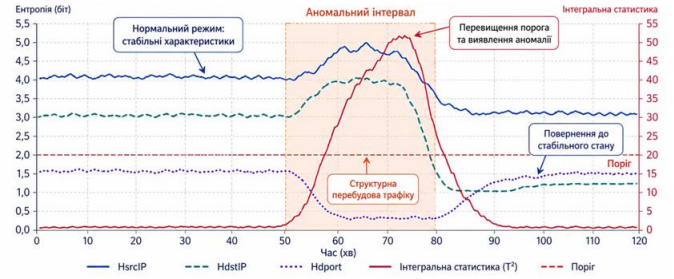


Експериментальна перевірка методу

• Експериментальна перевірка виконувалася для оцінювання стійкості методу та його здатності виявляти різні типи аномальних станів.
• Розглядалися сценарії нормального режиму, короткочасного сплеску навантаження, концентрації трафіку та розсіювання трафіку.
• Перевірка дозволила оцінити реакцію ентропійних ознак, інтегральної статистики та порогового правила на зміну структури мережевого потоку.
• Основна увага приділялася не лише факту спрацювання, а й стабільності роботи методу та рівню хибних спрацювань.

Результати експериментальних досліджень

• У нормальному режимі інтегральна статистика залишається в межах допустимого діапазону.
• У моменти появи аномалій ентропійні характеристики змінюються узгоджено, після чого багатовимірна статистика перевищує порогове значення.
• Метод дозволяє виявляти не лише різкі сплески інтенсивності, а й приховані структурні зміни трафіку.
• Результати підтверджують доцільність поєднання ентропійного аналізу з багатовимірною статистикою.



ВИСНОВКИ

У роботі за результатами виконаних теоретичних та практичних досліджень:

- проаналізовано сучасні методи аналізу мережевого трафіку та виявлення аномалій;
- розроблено метод аналізу трафіку на основі ентропійних характеристик і багатовимірної математичної статистики;
- реалізовано програмно-технічну систему аналізу мережевого трафіку;
- проведено експериментальну перевірку ефективності запропонованого методу.

Запропонований метод базується на аналізі трафіку у часових вікнах тривалістю 10 с та використанні 10 інформативних ознак, які після застосування методу головних компонент зводяться до 8 головних компонент, що пояснюють 95,71% варіації нормального трафіку.

Порівняно з пороговими та ентропійними підходами метод забезпечує більш високу повноту виявлення аномалій, підвищення інтегральної якості класифікації та зниження рівня хибних спрацювань.Отримані результати підтверджують ефективність запропонованого підходу та доцільність використання поєднання ентропійного аналізу з багатовимірною статистикою для задач моніторингу мереж і кібербезпеки.

ПУБЛІКАЦІЇ

1. Яцків В., Дудник В. Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики. ПерСик 2026, Харків, Україна, 23 квіт. 2026
2. Dudnyk, V. M (2026). METHOD FOR COMPUTER NETWORK TRAFFIC ANALYSIS BASED ON ENTROPY CHARACTERISTICS AND MULTIVARIATE MATHEMATICAL STATISTICS. Computer Systems and Information Technologies

Дякую за увагу!

ДОДАТОК Г

Лістинг програмного забезпечення реалізації методу аналізу трафіку комп’ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

А.1 Конфігурація параметрів та службові структури

У цьому підрозділі наведено конфігураційні структури та базові параметри, які задають часове віконування, ентропійні обчислення, багатовимірний статистичний аналіз і правило прийняття рішення.

```
from __future__ import annotations
from dataclasses import dataclass
from pathlib import Path
from typing import Dict, Iterator, List, Optional, Tuple, Union
import math
import json
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import entropy as scipy_entropy
from sklearn.decomposition import PCA
from sklearn.covariance import MinCovDet
from sklearn.metrics import (
    roc_curve,
    precision_recall_curve,
    confusion_matrix,
    classification_report,
)
try:
    from scapy.utils import RawPcapReader
    from scapy.layers.l2 import Ether
    from scapy.layers.inet import IP, TCP, UDP
except Exception:
    RawPcapReader = None
    Ether = None
    IP = None
    TCP = None
    UDP = None
@dataclass(frozen=True)
class WindowConfig:
    window_size_s: float
    step_s: float

@dataclass(frozen=True)
class EntropyConfig:
    log_base: float = 2.0
    smoothing_eps: float = 1e-12
@dataclass(frozen=True)
class MSPCConfig:
    use_pca: bool = True
    n_components: int = 5
    use_robust_cov: bool = False
    regularization: float = 1e-6

@dataclass(frozen=True)
class DecisionConfig:
    w_t2: float = 1.0
    w_q: float = 1.0
    w_e: float = 1.0
    use_ewma: bool = True
    ewma_lambda: float = 0.2
    threshold: Optional[float] = None
def is_tumbling(cfg: WindowConfig) -> bool:
    return abs(cfg.window_size_s - cfg.step_s) < 1e-12
```

А.2 Зчитування та попередня обробка даних

Лістинг містить функції для зчитування CSV- і PCAP-даних, а також для нормалізації подій мережевого трафіку перед формуванням часових вікон.

```
def read_flow_csv(path: Union[str, Path]) -> pd.DataFrame:
    df = pd.read_csv(path)
    required = {"ts", "src_ip", "dst_ip", "src_port", "dst_port", "proto", "bytes"}
    missing = required - set(df.columns)
    if missing:
        raise ValueError(f"CSV не містить обов'язкових колонок: {sorted(missing)}")

    df = df.copy()
    df["ts"] = pd.to_numeric(df["ts"], errors="coerce")
    df = df.dropna(subset=["ts"])

    for c in ["src_port", "dst_port", "proto", "bytes"]:
        df[c] = pd.to_numeric(df[c], errors="coerce")

    df["src_ip"] = df["src_ip"].astype(str)
    df["dst_ip"] = df["dst_ip"].astype(str)

    if "packets" in df.columns:
        df["packets"] = pd.to_numeric(df["packets"], errors="coerce")

    df = df.sort_values("ts").reset_index(drop=True)
    return df

def read_pcap(path: Union[str, Path], limit_packets: Optional[int] = None) -> pd.DataFrame:
    if RawPcapReader is None:
        raise RuntimeError(
            "Scapy не доступний. Встановіть scapy або використайте CSV з flow records."
        )

    rows: List[Dict[str, object]] = []
    n = 0

    for pkt_bytes, meta in RawPcapReader(str(path)):
        n += 1
        if limit_packets is not None and n > limit_packets:
            break

        ts = float(meta.sec) + float(meta.usec) / 1e6

        try:
            eth = Ether(pkt_bytes)
        except Exception:
            continue

        if IP is None or IP not in eth:
            continue

        ip = eth[IP]
        proto = int(ip.proto)
        src_ip = str(ip.src)
        dst_ip = str(ip.dst)

        src_port = np.nan
        dst_port = np.nan

        if TCP is not None and TCP in ip:
            src_port = int(ip[TCP].sport)
            dst_port = int(ip[TCP].dport)
        elif UDP is not None and UDP in ip:
            src_port = int(ip[UDP].sport)
            dst_port = int(ip[UDP].dport)

        b = int(getattr(meta, "wirelen", len(pkt_bytes)))

        rows.append(
            {
                "ts": ts,
                "src_ip": src_ip,
                "dst_ip": dst_ip,
                "src_port": src_port,
                "dst_port": dst_port,
```

```
                "proto": proto,
                "bytes": b,
            }
        )

    df = pd.DataFrame(rows)
    if df.empty:
        raise ValueError("PCAP не дав придатних IPv4-пакетів або файл порожній.")

    df = df.sort_values("ts").reset_index(drop=True)
    return df


def normalize_events(df: pd.DataFrame) -> pd.DataFrame:
    df = df.copy()

    required = ["ts", "src_ip", "dst_ip", "src_port", "dst_port", "proto", "bytes"]
    missing = [c for c in required if c not in df.columns]
    if missing:
        raise ValueError(f"Вхідні події не містять колонок: {missing}")

    df["ts"] = pd.to_numeric(df["ts"], errors="coerce")
    df["bytes"] = pd.to_numeric(df["bytes"], errors="coerce")

    for c in ["src_port", "dst_port", "proto"]:
        df[c] = pd.to_numeric(df[c], errors="coerce")

    df["src_ip"] = df["src_ip"].astype(str)
    df["dst_ip"] = df["dst_ip"].astype(str)

    df = df.dropna(subset=["ts", "bytes"])
    df = df.sort_values("ts").reset_index(drop=True)
    return df
```

А.3 Формування часових вікон

У цьому підрозділі подано реалізацію генератора часових вікон, який забезпечує послідовну обробку подій трафіку у фіксованих інтервалах спостереження.

```
def iter_time_windows(
    df: pd.DataFrame,
    cfg: WindowConfig,
    *,
    ts_col: str = "ts",
) -> Iterator[Tuple[float, float, pd.DataFrame]]:
    if df.empty:
        return
    df = df.sort_values(ts_col).reset_index(drop=True)
    ts = df[ts_col].to_numpy(dtype=float)
    n = len(ts)
    win_size = float(cfg.window_size_s)
    step = float(cfg.step_s)
    t0 = ts[0]
    t_last = ts[-1]

    start_idx = 0
    end_idx = 0
    win_start = t0
    while win_start <= t_last:
        win_end = win_start + win_size

        while start_idx < n and ts[start_idx] < win_start:
            start_idx += 1

        if end_idx < start_idx:
            end_idx = start_idx

        while end_idx < n and ts[end_idx] < win_end:
            end_idx += 1

        yield (win_start, win_end, df.iloc[start_idx:end_idx])
        win_start += step
```

А.4 Обчислення ентропійних характеристик і дивергентних мір

Наведений код реалізує побудову розподілів, обчислення ентропії Шеннона, нормованої ентропії, а також дивергентних мір KL і JSD для порівняння поточного та базового профілів.

```
def _counts_to_prob(counts: pd.Series) -> Tuple[np.ndarray, np.ndarray]:
    cats = counts.index.to_numpy()
    c = counts.to_numpy(dtype=float)
    s = c.sum()
    if s <= 0:
        return cats, np.array([], dtype=float)
    return cats, c / s
def shannon_entropy_from_counts(counts: pd.Series, cfg: EntropyConfig) -> float:
    _, p = _counts_to_prob(counts)
    if p.size == 0:
        return 0.0
    return float(scipy_entropy(p, base=cfg.log_base))
def normalized_entropy_from_counts(counts: pd.Series, cfg: EntropyConfig) -> float:
    k = int(counts.shape[0])
    if k <= 1:
        return 0.0

    h = shannon_entropy_from_counts(counts, cfg)
    denom = math.log(k, cfg.log_base)
    return float(h / denom) if denom > 0 else 0.0
def _align_distributions(
    p_counts: pd.Series,
    q_counts: pd.Series,
    eps: float,
) -> Tuple[np.ndarray, np.ndarray]:
    all_cats = p_counts.index.union(q_counts.index)
    p = p_counts.reindex(all_cats, fill_value=0.0).to_numpy(dtype=float)
    q = q_counts.reindex(all_cats, fill_value=0.0).to_numpy(dtype=float)
    p = p + eps
    q = q + eps
    p = p / p.sum()
    q = q / q.sum()
    return p, q
def kl_divergence_from_counts(
    p_counts: pd.Series,
    q_counts: pd.Series,
    cfg: EntropyConfig,
) -> float:
    p, q = _align_distributions(p_counts, q_counts, cfg.smoothing_eps)
    return float(scipy_entropy(p, qk=q, base=cfg.log_base))

def js_divergence_from_counts(
    p_counts: pd.Series,
    q_counts: pd.Series,
    cfg: EntropyConfig,
    weight: float = 0.5,
) -> float:
    p, q = _align_distributions(p_counts, q_counts, cfg.smoothing_eps)
    w = float(weight)
    m = w * p + (1.0 - w) * q
    d1 = scipy_entropy(p, qk=m, base=cfg.log_base)
    d2 = scipy_entropy(q, qk=m, base=cfg.log_base)

    return float(w * d1 + (1.0 - w) * d2)
def window_distributions(df_win: pd.DataFrame) -> Dict[str, pd.Series]:
    d: Dict[str, pd.Series] = {}
    d["src_ip"] = df_win["src_ip"].value_counts(dropna=True)
    d["dst_ip"] = df_win["dst_ip"].value_counts(dropna=True)
    d["src_port"] = df_win["src_port"].dropna().astype(int).value_counts()
    d["dst_port"] = df_win["dst_port"].dropna().astype(int).value_counts()
    d["proto"] = df_win["proto"].dropna().astype(int).value_counts()
    return d
```

А.5 Формування вектора ознак трафіку

У цьому фрагменті формується узагальнений вектор ознак одного часового вікна, який поєднує об’ємні показники, ентропійні характеристики та дивергентні міри.

```
def compute_window_features(
    df_win: pd.DataFrame,
    baseline_dists: Optional[Dict[str, pd.Series]],
    e_cfg: EntropyConfig,
) -> Dict[str, float]:
    feats: Dict[str, float] = {}
    feats["n_events"] = float(len(df_win))
    feats["bytes_sum"] = float(df_win["bytes"].sum()) if "bytes" in df_win.columns else 0.0
    feats["bytes_mean"] = float(df_win["bytes"].mean()) if "bytes" in df_win.columns else 0.0
    if "packets" in df_win.columns:
        feats["packets_sum"] = float(df_win["packets"].sum())
        feats["packets_mean"] = float(df_win["packets"].mean())
    dists = window_distributions(df_win)
    for key, counts in dists.items():
        feats[f"H_{key}"] = shannon_entropy_from_counts(counts, e_cfg)
        feats[f"Hn_{key}"] = normalized_entropy_from_counts(counts, e_cfg)
        feats[f"uniq_{key}"] = float(counts.shape[0])
        if baseline_dists is not None and key in baseline_dists and baseline_dists[key].shape[0] >
0:
            feats[f"KL_{key}"] = kl_divergence_from_counts(counts, baseline_dists[key], e_cfg)
            feats[f"JS_{key}"] = js_divergence_from_counts(counts, baseline_dists[key], e_cfg)
        else:
            feats[f"KL_{key}"] = 0.0
            feats[f"JS_{key}"] = 0.0
    feats["H_srcdst_diff"] = feats.get("H_src_ip", 0.0) - feats.get("H_dst_ip", 0.0)
    feats["H_ports_diff"] = feats.get("H_src_port", 0.0) - feats.get("H_dst_port", 0.0)
    return feats
```

А.6 Багатовимірний статистичний аналіз

Підрозділ містить реалізацію навчання базової моделі нормального трафіку та обчислення статистик багатовимірного відхилення для подальшого виявлення аномальних станів.

```
@dataclass
class BaselineModel:
    feature_names: List[str]
    mean_: np.ndarray
    std_: np.ndarray
    cov_: np.ndarray
    cov_inv_: np.ndarray
    pca_: Optional[PCA] = None
    robust_cov_: Optional[MinCovDet] = None
    t2_mean: float = 0.0
    t2_std: float = 1.0
    q_mean: float = 0.0
    q_std: float = 1.0
    e_mean: float = 0.0
    e_std: float = 1.0
def fit_baseline_model(
    X_train: pd.DataFrame,
    m_cfg: MSPCConfig,
) -> BaselineModel:
    feature_names = list(X_train.columns)
    X = X_train.to_numpy(dtype=float)

    mean_ = np.nanmean(X, axis=0)
    std_ = np.nanstd(X, axis=0)
    std_[std_ == 0] = 1.0

    Z = (X - mean_) / std_

    if m_cfg.use_robust_cov:
```

```
            rc = MinCovDet().fit(Z)
            cov_ = rc.covariance_
            robust_cov = rc
        else:
            cov_ = np.cov(Z, rowvar=False)
            robust_cov = None

        cov_reg = cov_ + m_cfg.regularization * np.eye(cov_.shape[0])
        cov_inv_ = np.linalg.pinv(cov_reg)

        if m_cfg.use_pca:
            pca = PCA(n_components=min(m_cfg.n_components, Z.shape[1]))
            pca.fit(Z)
        else:
            pca = None

        model = BaselineModel(
            feature_names=feature_names,
            mean_=mean_,
            std_=std_,
            cov_=cov_reg,
            cov_inv_=cov_inv_,
            pca_=pca,
            robust_cov_=robust_cov,
        )

        t2_list = []
        q_list = []

        for z in Z:
            t2, q = mspc_scores(z, model)
            t2_list.append(t2)
            q_list.append(q)
        t2_arr = np.array(t2_list, dtype=float)
        q_arr = np.array(q_list, dtype=float)
        model.t2_mean = float(np.mean(t2_arr))
        model.t2_std = float(np.std(t2_arr) if np.std(t2_arr) > 0 else 1.0)
        model.q_mean = float(np.mean(q_arr))
        model.q_std = float(np.std(q_arr) if np.std(q_arr) > 0 else 1.0)
        return model
    def mspc_scores(z: np.ndarray, model: BaselineModel) -> Tuple[float, float]:
        z = z.reshape(-1)
        t2 = float(z.T @ model.cov_inv_ @ z)
        q = 0.0
        if model.pca_ is not None:
            z_proj = model.pca_.inverse_transform(model.pca_.transform([z]))[0]
            resid = z - z_proj
            q = float(np.dot(resid, resid))
        return t2, q
```

A.7 Прийняття рішення щодо аномального стану

Нижче наведено функції нормалізації статистик, згладжування та формування інтегрального бала аномальності з подальшим пороговим прийняттям рішення.

```
    def normalize_score(x: float, mu: float, sigma: float) -> float:
        sigma = sigma if sigma > 0 else 1.0
        return (x - mu) / sigma
    def ewma_update(prev: float, x: float, lam: float) -> float:
        return lam * x + (1.0 - lam) * prev
    def compute_anomaly_score(
        z: np.ndarray,
        entropy_scalar: float,
        model: BaselineModel,
        d_cfg: DecisionConfig,
        prev_ewma: Optional[float] = None,
    ) -> Tuple[float, float, float, float]:
        t2, q = mspc_scores(z, model)
        t2n = normalize_score(t2, model.t2_mean, model.t2_std)
        qn = normalize_score(q, model.q_mean, model.q_std)
        en = normalize_score(entropy_scalar, model.e_mean, model.e_std)
        s_raw = d_cfg.w_t2 * t2n + d_cfg.w_q * qn + d_cfg.w_e * en
```

```
    if d_cfg.use_ewma:
        if prev_ewma is None:
            s_sm = s_raw
        else:
            s_sm = ewma_update(prev_ewma, s_raw, d_cfg.ewma_lambda)
    else:
        s_sm = s_raw
    return float(s_raw), float(s_sm), float(t2), float(q)
def choose_threshold_by_roc(scores: np.ndarray, y_true: np.ndarray) -> float:
    fpr, tpr, thr = roc_curve(y_true, scores, pos_label=1)
    j = tpr - fpr
    idx = int(np.argmax(j))
    return float(thr[idx])
def apply_threshold(scores: np.ndarray, thr: float) -> np.ndarray:
    return (scores >= thr).astype(int)
def top_feature_contributions(
    z: np.ndarray,
    model: BaselineModel,
    k: int = 5,
) -> List[Tuple[str, float]]:
    idx = np.argsort(-np.abs(z))[:k]
    return [(model.feature_names[i], float(z[i])) for i in idx]
```

A.8 Експериментальна перевірка та оцінювання результатів

У цьому підрозділі представлено реалізацію наскрізного конвеєра обробки, побудови ROC- і PR-кривих, формування підсумкових графіків і збереження результатів експериментальної перевірки.

```
def _entropy_scalar_from_row(row: pd.Series) -> float:
    return float(
        row.get("JS_src_ip", 0.0)
        + row.get("JS_dst_ip", 0.0)
        + row.get("JS_src_port", 0.0)
        + row.get("JS_dst_port", 0.0)
    )
def run_pipeline(
    df_events: pd.DataFrame,
    w_cfg: WindowConfig,
    e_cfg: EntropyConfig,
    m_cfg: MSPCConfig,
    d_cfg: DecisionConfig,
    *,
    train_end_ts: Optional[float] = None,
    labels: Optional[pd.Series] = None,
) -> Dict[str, object]:
    df_events = normalize_events(df_events)
    baseline_dists: Optional[Dict[str, pd.Series]] = None
    feature_rows: List[Dict[str, float]] = []
    win_meta: List[Tuple[float, float]] = []
    for (t_start, t_end, df_win) in iter_time_windows(df_events, w_cfg):
        if len(df_win) == 0:
            continue
        feats = compute_window_features(df_win, baseline_dists, e_cfg)
        feature_rows.append(feats)
        win_meta.append((t_start, t_end))
    X = pd.DataFrame(feature_rows).fillna(0.0)
    meta = pd.DataFrame(win_meta, columns=["t_start", "t_end"])
    if X.empty:
        raise ValueError("Не вдалося сформувати жодного вікна з ознаками.")
    if train_end_ts is None:
        n_train = max(1, int(0.3 * len(X)))
        train_mask = np.zeros(len(X), dtype=bool)
        train_mask[:n_train] = True
    else:
        train_mask = meta["t_end"].to_numpy() <= float(train_end_ts)

    if train_mask.sum() == 0:
        raise ValueError("Навчальний період порожній. Перевір параметр train_end_ts.")
    baseline_dists: Dict[str, pd.Series] = {}
    train_end_value = float(meta.loc[train_mask, "t_end"].max())
    df_train_events = df_events[df_events["ts"] <= train_end_value]
```

```
baseline_dists.update(window_distributions(df_train_events))
feature_rows2: List[Dict[str, float]] = []
win_meta2: List[Tuple[float, float]] = []
for (t_start, t_end, df_win) in iter_time_windows(df_events, w_cfg):
    if len(df_win) == 0:
        continue
    feature_rows2.append(compute_window_features(df_win, baseline_dists, e_cfg))
    win_meta2.append((t_start, t_end))
X = pd.DataFrame(feature_rows2).fillna(0.0)
meta = pd.DataFrame(win_meta2, columns=["t_start", "t_end"])
X_train = X.loc[train_mask].copy()
model = fit_baseline_model(X_train, m_cfg)
e_train = X.loc[train_mask].apply(_entropy_scalar_from_row, axis=1).to_numpy(dtype=float)
model.e_mean = float(np.mean(e_train))
model.e_std = float(np.std(e_train) if np.std(e_train) > 0 else 1.0)
scores_raw = []
scores_sm = []
t2_list = []
q_list = []
contribs = []
prev = None
for i in range(len(X)):
    x = X.iloc[i].to_numpy(dtype=float)
    z = (x - model.mean_) / model.std_
    e_val = _entropy_scalar_from_row(X.iloc[i])
    s_raw, s_sm, t2, q = compute_anomaly_score(
        z,
        e_val,
        model,
        d_cfg,
        prev_ewma=prev,
    )
    scores_raw.append(s_raw)
    scores_sm.append(s_sm)
    t2_list.append(t2)
    q_list.append(q)
    contribs.append(top_feature_contributions(z, model, k=5))

    prev = s_sm
scores_raw = np.array(scores_raw, dtype=float)
scores_sm = np.array(scores_sm, dtype=float)
t2_arr = np.array(t2_list, dtype=float)
q_arr = np.array(q_list, dtype=float)
y_true = None
y_pred = None
threshold = d_cfg.threshold
if labels is not None:
    labels = pd.Series(labels).astype(int).reset_index(drop=True)
    if len(labels) != len(X):
        raise ValueError(
            "Кількість міток labels має відповідати кількості сформованих вікон."
        )
    y_true = labels.to_numpy(dtype=int)
    if threshold is None:
        threshold = choose_threshold_by_roc(scores_sm, y_true)
    y_pred = apply_threshold(scores_sm, threshold)
else:
    if threshold is None:
        threshold = float(np.mean(scores_sm[train_mask]) + 3.0 * np.std(scores_sm[train_mask]))
    y_pred = apply_threshold(scores_sm, threshold)
out = {
    "X": X,
    "meta": meta,
    "train_mask": train_mask,
    "model": model,
    "scores_raw": scores_raw,
    "scores_smoothed": scores_sm,
    "t2": t2_arr,
    "q": q_arr,
    "threshold": threshold,
    "y_pred": y_pred,
    "feature_contributions": contribs,
}
if y_true is not None:
    fpr, tpr, roc_thr = roc_curve(y_true, scores_sm, pos_label=1)
    precision, recall, pr_thr = precision_recall_curve(y_true, scores_sm)
    cm = confusion_matrix(y_true, y_pred)
    report = classification_report(y_true, y_pred, output_dict=True, zero_division=0)
    out.update(
```

```
            {
                "y_true": y_true,
                "roc_curve": {
                    "fpr": fpr,
                    "tpr": tpr,
                    "thresholds": roc_thr,
                },
                "pr_curve": {
                    "precision": precision,
                    "recall": recall,
                    "thresholds": pr_thr,
                },
                "confusion_matrix": cm,
                "classification_report": report,
            }
        )
    return out
def save_roc_plot(results: Dict[str, object], path: Union[str, Path]) -> None:
    if "roc_curve" not in results:
        raise ValueError("ROC-крива недоступна: для run_pipeline не передано labels.")
    roc = results["roc_curve"]
    fpr = roc["fpr"]
    tpr = roc["tpr"]
    plt.figure(figsize=(7, 5))
    plt.plot(fpr, tpr, label="ROC")
    plt.plot([0, 1], [0, 1], linestyle="--")
    plt.xlabel("False Positive Rate")
    plt.ylabel("True Positive Rate")
    plt.title("ROC-крива методу")
    plt.grid(True, alpha=0.3)
    plt.legend()
    plt.tight_layout()
    plt.savefig(path, dpi=200)
    plt.close()
def save_score_plot(results: Dict[str, object], path: Union[str, Path]) -> None:
    scores = results["scores_smoothed"]
    scores = results["scores_smoothed"]
    thr = results["threshold"]
    x = np.arange(len(scores))
    plt.figure(figsize=(10, 5))
    plt.plot(x, scores, label="S(t), згладжений")
    plt.axhline(thr, linestyle="--", label=f"Поріг τ = {thr:.3f}")
    plt.xlabel("Номер вікна")
    plt.ylabel("Інтегральний бал аномальності")
    plt.title("Динаміка бала аномальності по часових вікнах")
    plt.grid(True, alpha=0.3)
    plt.legend()
    plt.tight_layout()
    plt.savefig(path, dpi=200)
    plt.close()
def save_results_summary(results: Dict[str, object], path: Union[str, Path]) -> None:
    summary = {"threshold": float(results["threshold"]),
        "n_windows": int(len(results["X"])),
        "n_train_windows": int(np.sum(results["train_mask"])),
    }
    if "classification_report" in results:
        summary["classification_report"] = results["classification_report"]
        summary["confusion_matrix"] = np.asarray(results["confusion_matrix"]).tolist()

    with open(path, "w", encoding="utf-8") as f:
        json.dump(summary, f, ensure_ascii=False, indent=2)
```

Протокол аналізу звіту подібності експертом

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Володимир ДУДНИК

Співавтор:

Назва: Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

Експерт: Сергій ЛИСЕНКО

Підрозділ: Кафедра комп'ютерної інженерії та інформаційних систем

Коефіцієнт подібності 1:6.91%

Коефіцієнт подібності 2:1.85%

Мікропробіли: 3

Заміна букв: 10

Інтервали: 0

Білі знаки: 6

Дата створення звіту: 2026-04-21 18:44:52.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обґрунтування:

2026-04-21 Доцент Андрій Нічепорук

Дата експерт

Tue Apr 21 19:06:18 EEST 2026, Медзатий Дмитро Миколайович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (http://ap.km.ua) v-15.701

Максимальне співпадіння з одним документом 0.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 11%**

	Документ		Сумарний збіг по Базі Даних	
ID: 270584 Назва: МКР Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики Додано в БД: 2026-04-21 Автора: Володимир ДУДНИК Керівники: Сергій ЛИСЕНКО Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	148997	1066	2487 (2%)	33 (3%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

РЕЦЕНЗІЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

Здобувач: Володимир ДУДНИК

Тема: Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

Спеціальність: 123 «Комп'ютерна інженерія»

Обсяг кваліфікаційної роботи магістра:

Кількість листів креслень _–_; кількість сторінок записки _81_

1. Короткий зміст роботи та прийнятих рішень _У роботі запропоновано метод аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики

2. Висновок про відповідність роботи дипломному завданню _____ Кваліфікаційна робота магістра відповідає виданому завданню_____

3. Характеристика виконання кожного розділу, ступінь використання останніх досягнень науки і техніки і передових методів роботи: _ У першому розділі проаналізовано мережевий трафік як об'єкт моніторингу, класи деградацій та аномалій, відомі методи й засоби виявлення відхилень, а також обґрунтовано доцільність використання ентропійних характеристик і багатовимірної статистики для аналізу трафіку. У другому розділі виконано постановку задачі дослідження, формалізацію процесу аналізу мережевого трафіку, обґрунтовано вибір ентропійних і багатовимірних статистичних методів. У третьому розділі розроблено алгоритмічну реалізацію методу, зокрема алгоритми формування часових вікон і підготовки вибірки, обчислення ентропійних характеристик. У четвертому розділі обґрунтовано вибір засобів програмної реалізації, реалізовано модулі підготовки даних, формування часових вікон та обчислення ентропійних характеристик, модулі багатовимірного статистичного аналізу і прийняття рішення, а також описано вхідні дані, умови проведення експериментів та сценарії перевірки методу.

4. Позитивні сторони роботи: Удосконалена система аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної

математичної статистики, яка визначає аномальні стани за статистичною мірою відхилення від профілю нормального режиму, що забезпечує формалізований перехід від спостережуваних даних до прийняття рішення.

5. Негативні сторони роботи: _____

6. Оцінка графічного оформлення та пояснювальної записки роботи: _–_____

_____

_____

_____

_____

_____

7. Відгук про роботу в цілому: В загальному робота виконана на високому рівні. _____

_____

_____

_____

_____

_____

8. Інші зауваження: _–_____

_____

_____

_____

_____

_____

9. Оцінка кваліфікаційної роботи магістра:

Розглянувши позитивні та негативні сторони представленої кваліфікаційної роботи магістра вважаю, що робота заслуговує оцінки «відмінно» 90.00 (A) _____

Рецензент (прізвище, ім'я, по батькові, посада, місце роботи) _____ д.т.н., професор, Бармак Олександр Володимирович, завідувач кафедри комп'ютерних наук _____

"__1 травня__" _____2026р.

_____

Зав. кафедри КІІС
д-р. філософії Ользі ПАВЛОВІЙ

Володимир ДУДНИК
ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КІ2м-24-1

ЗАЯВА

З правилами чинного Положення про систему забезпечення академічної доброчесності у Хмельницькому національному університеті, згідно з яким виявлення академічного плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту і застосування заходів академічної відповідальності, ознайомлений (а). Про використання спеціалізованих програмних засобів (СПЗ) StrikePlagiarism та Anti-Plagiarism для перевірки кваліфікаційних робіт здобувачів вищої освіти на наявність академічного плагіату оповіщений (а). Надаю університету право на передачу моєї роботи для обробки та збереження в базах даних СПЗ і використання роботи для виявлення академічного плагіату в інших роботах, які перевіряються СПЗ.

Також надаю свою згоду на обробку й збереження університетом моєї роботи в Інституційному репозитарії Хмельницького національного університету.

Робота надається для перевірки в електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

1 травня 2026 року

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи Метод та програмно-технічні засоби аналізу трафіку комп'ютерних мереж на основі ентропійних характеристик та багатовимірної математичної статистики
Автор Володимир ДУДНИК
Освітня програма Комп'ютерна інженерія та програмування
Рівень вищої освіти другий (магістерський)
Спеціальність 123 Комп'ютерна інженерія
Науковий керівник: д-р техн. наук, професор, Василь ЯЦКІВ

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмним засобом(ами) комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	відповідає
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) усі запозичення фрагментарні, або мають належним чином оформленні посилання;

2) окремі виявлені збіги є загальновживаними фразами або виразами, про що свідчить посилання системи на збіг з джерелами на один фрагмент речення;

3) всі зафіксовані системою ознаки модифікації тексту відносяться до комбінування латинських символів зі україномовними скороченнями індексів в формулах, що не є модифікацією тексту.

4) значна частина знайденого плагіату відноситься до списку використаних джерел

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості StrikePlagiarism, складає 6.91% і адресується до 10 першоджерел; та системою Anti-Plagiarism складає 0%, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

21.04.2026

Завідувач кафедри _____ Підпис — Ольга ПАВЛОВА (Ім'я, ПРІЗВИЩЕ)

Гарант освітньої програми _____ Підпис — Олег САВЕНКО (Ім'я, ПРІЗВИЩЕ)

Керівник кваліфікаційної роботи _____ Підпис — Василь ЯЦКІВ (Ім'я, ПРІЗВИЩЕ)