Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Центр післядипломної освіти_____
(повна назва)

Кафедра _____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження методів ідентифікації та класифікації трафіку для безпеки мереж
(тема)

Виконав: студент 2 курсу, групи ІПЗмзд-18-1__
спеціальності _121- Інженерія програмного забезпечення_____
(код і повна назва спеціальності)
освітньо-наукової програми _Інженерія програмного_забезпечення_ _____
(повна назва освітньої програми)
_____Воропаєв А.В._____
(прізвище, ініціали)
Керівник _____проф. Шостак І.В._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти

Кафедра програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121 – Інженерія програмного забезпечення

(код і повна назва)

Освітньо-наукова програма Інженерія програмного забезпечення

(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові Воропаєву Антону Вікторовичу

(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів ідентифікації та класифікації трафіку для безпеки мереж

затверджена наказом по університету від « »_____2020 р № ___

2. Термін подання студентом роботи до екзаменаційної комісії «21» травня 2020 р.

3. Вихідні дані до роботи *Алгоритми функціонування штучних агентів, методи взаємодії інтелектуальних агентів, методи стримінгу великих даних та пояснювальна записка. Використовувати ОС Windows, середовище об'єктно-орієнтованого проектування.*

4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, методи пошуку корисних даних, опис об'єктних моделей, використовувані. методи та алгоритми, архітектура програмної системи, опис розробленої програмної системи. , результати тестування програмної системи*

5. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Шостак І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	25 березня 2020 р.	
2.	Огляд існуючих методів	31 березня 2020 р.	
3.	Методи безпеки мережевого трафіку	15 квітня 2020 р.	
4.	Підготовка пояснювальної записки	20 квітня 2020 р.	
5.	Спецчастина	28 квітня 2020 р.	
6.	Підготовка презентації та доповіді	03 травня 2020 р.	
7.	Попередній захист	15 травня 2020 р.	
8.	Нормоконтроль, рецензування	17 травня 2020 р.	
9.	Занесення диплома в електронний архів	18 травня 2020 р.	
10.	Допуск до захисту у зав. кафедри	20 травня 2020 р.	

Дата видачі завдання _ «____»_ _____2020 р.

Студент _____
(підпис)

Керівник роботи _____ __проф. Шостак І.В._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 97 с., 4 табл., 43 рис., 3 дод., 29 джерел.

ІДЕНТИФІКАЦІЯ ТРАФІКУ, МАРКІВСЬКІ МОДЕЛІ, ТУНЕЛЬНИЙ ТРАФІК, ІНФОРМАЦІЙНА БЕЗПЕКА МЕРЕЖ

Об'єктом дослідження є методи, моделі й засоби швидкої ідентифікації корисного й тунельного трафіку в реальному часі.

Метою атестаційної роботи є підвищення рівня захисту й надійності СПД за рахунок поліпшення ідентифікації трафіку в реальному часі на рівні додатків, та в тунелях, з використанням схованої марківської моделі.

В результаті модифіковано методи й алгоритми ідентифікації для рішення завдання ідентифікації тунельних додатків, розроблено програмний засіб ідентифікації на основі запропонованих алгоритмів.

TRAFFIC IDENTIFICATION, MARKIV MODELS, TUNNEL TRAFFIC, NETWORK INFORMATION SECURITY

The object of research are methods, models and means of rapid identification of useful and tunnel traffic in real time.

As a result, the methods and algorithms of identification have been modified to solve the problem of identification of tunnel applications

The software of identification has been developed on the basis of the proposed algorithms.

ЗМІСТ

ВСТУП

Інтенсивне використання інтернет-додатків у різних аспектах життя привело до збільшення обсягу переданого трафіку й одночасно до збільшення погроз безпеки інформації. Реакцією на це стало вдосконалювання способів захисту даних і користувачів. Найбільш відомим методом захисту даних є шифрування. Але при використанні шифрування для керування трафіком і безпекою мережі стає складніше через неможливість здійснення перевірки вмісту зашифрованих пакетів [1]. Ситуація збільшується ще тим, що шифрування найчастіше використовується для обходу політики безпеки й правил використання мережних ресурсів. У цьому випадку необхідно дотримувати балансу між забезпеченням конфіденційності й забезпеченням мережної безпеки при загрозі з боку зловмисників. Політика безпеки як на рівні мереж інтернет-провайдерів, так і на рівні глобальних мережних операторів визначає правило поведінки з потенційно небезпечними додатками з погляду їх впливу на економіку, політику, моральні підвалини суспільства й т.п. Прикладом застосування політики безпеки на рівні держави може служити проект «Золотий щит» (неофіційна назва – «Великий китайський фаєрвол»), який блокує ряд сервісів та сайтів, у тому числі такі відомі сервіси, як «Facebook» і «Google» [2]. У свою чергу ефективність застосування політики безпеки багато в чому залежить від валідності класифікації трафіку, збільшення якої є нетривіальним завданням. Так, незважаючи на всі спроби блокування такого небажаного трафіку, як BiTorrent, статистика показує, що його обсяг залишається дуже високим.

Все вищесказане підтверджує, що ідентифікація й класифікація трафіку мереж передачі даних (СПД) є важливою темою магістерської роботи, оскільки визначають собою основні кроки по створенню моделі керування трафіком при рішенні завдань коректного застосування політики безпеки.

Зв'язане це ще й з тим, що кожний інтернет-провайдер визначає свою власну політику безпеки, яка включає правила заборони використання певних служб, відвідування небажаних веб-сайтів або IP-адресу. Однак іноді системи мережної безпеки, такі, як брандмауери або системи виявлення вторгнень, блокують поряд з небезпечним трафіком та додатками, трафік та додатка, які містять ознаки підозрілої діяльності, не будучи такими (наприклад, тунельний трафік). Розпізнавання таких сервісів та додатків неможливо без використання ефективного методу ідентифікації трафіку, який може виявити й заблокувати потенційні погрози в мережі.

Іншою стороною розробки ефективного методу класифікації трафіку є підвищення якості обслуговування абонента. Додатки відрізняються друг від друга вимогами до ресурсів мережі з метою одержання певного рівня якості роботи кожного додатка. Інтернет-Провайдер класифікує трафік додатків і встановлює відповідні пріоритети для кожного потоку на основі вимог додатків. Наприклад, вимоги до параметра часової затримки й числу бітових помилок при передачі трафіку IP-телефонії й відеоконференції відрізняються від вимог до аналогічних параметрів при передачі трафіку веб-сервісів. Інтернет-провайдер оптимізує свою інфраструктуру для забезпечення необхідного якості обслуговування користувача, а для цього необхідно провести правильну класифікацію.

Таким чином, дослідження, спрямовані на розробку нових методів і моделей ідентифікації трафіку (IT) СПД, що функціонують у реальному масштабі часу, є як і раніше актуальними й мають практичне значення при вирішенні проблем забезпечення політик безпеки.

Метою роботи є підвищення рівня захисту й надійності СПД за рахунок поліпшення ідентифікації трафіку в реальному часі на рівні додатків, та в тунелях, з використанням схованої марківської моделі.

Для досягнення поставленої мети атестаційної роботи були вирішені наступні завдання:

– аналіз існуючих моделей трафіку СПД, виявлення їх переваг і недоліків і особливостей організації трафіку стосовно до завдання ідентифікації;

– порівняльний аналіз існуючих методів ідентифікації трафіку СПД і розробка моделі, методу й алгоритмів ідентифікації мережного трафіку в реальному часі з використанням його статистичних характеристик і схованої марківської моделі (СММ);

– розробка алгоритму обчислення значень параметрів пропонованої моделі з використанням ітераційної процедури Баума-Велша;

– розробка алгоритму ініціалізації СММ на основі моделі гаусової суміші, що забезпечує оптимальну збіжність процедури Баума-Велша в рамках необхідного якості ідентифікації;

– розробка методики підготовки наборів даних на основі реального та модельного мережного трафіку на етапах навчання й тестування запропонованої моделі;

– розробка програмних засобів ідентифікації на основі запропонованих алгоритмів і їх експериментальне дослідження.

В відповідності з метою й завданнями атестаційної роботи об'єктом дослідження є трафік СПД, а предметом дослідження – методи, моделі й засобу швидкої ідентифікації корисного й тунельного трафіку в реальному часі.

1. АНАЛІЗ МЕТОДІВ ІДЕНТИФІКАЦІЇ ТРАФІКУ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕНЯ

1.1 Порівняльний аналіз існуючих моделей трафіку мереж передачі даних

Ідентифікація мережного трафіку є важливим завданням в області безпеки, захисту й керування трафіком мереж передачі даних. Ефективним математичним інструментом її рішення є моделювання мережного трафіку. Саме математичне моделювання широке використовується для рішення наступних мережних підзавдань [4]:

– прогнозування майбутнього трафіку з метою оцінки ресурсів, необхідних для одержання певного рівня якості обслуговування; сюди ставиться, наприклад, оцінка необхідної пропускної здатності й розміру буферів з метою досягнення прийнятних показників втрат і затримки пакетів;

– оцінка впливу алгоритмів керування мережним трафіком на характеристики мережі;

– вивчення специфічних явищ і процесів, що відбуваються в мережі (явищ фрактальності трафіку, пульсації трафіку і т.д. [5]);

– генерація трафіку для цілей імітаційного моделювання мережних взаємодій;

– ідентифікація джерела трафіку на основі різних його характеристик, наприклад, ідентифікація додатків у системах виявлення вторгнень.

В основі ряду моделей трафіку лежать стаціонарні випадкові процеси $x(t)$ з різними законами розподілу, за допомогою яких відтворюються характеристики трафіку (кількість пакетів, отриманих або відправлених протягом певного проміжку часу; інтервали між пакетами, де $\boldsymbol{i = 1, 2, \ldots}$; довжини пакетів $\{l_i\}, i = 1, 2, \ldots$, послідовність напрямків передачі пакетів $\{\boldsymbol{\delta_i}\}, \boldsymbol{i = 1, 2, \ldots}$ і т.д.).

В залежності від способу опису $X(t)$ моделі діляться на групи, найпоширенішими з яких є:

– моделі на основі законів розподілу;
– моделі на основі стохастичних часових рядів;
– моделі на основі теорії фракталів;
– моделі на основі ланцюгів Маркова.

Перша група моделей є класичною й будується на основі відомих законів розподілу [6]. Найпоширенішими з них є модель Пуассона та «On/Off»-модель. У моделі Пуассона $X(t)$ визначає кількість вхідних пакетів, причому ймовірність одержання $X(t) = k$ пакетів за інтервал часу t задається експонентним законом розподілу:

$$P\{X(t) = k\} = \frac{(\lambda t)^k e^{(\lambda t)}}{k!}, \quad (1.1)$$

де λ – інтенсивність вступу пакетів. При цьому ймовірність одержання нуль пакетів рівна

$$P\{X(t) = 0\} = e^{(-\lambda t)}.$$

Рівняння (1.1) показує, що розподіл інтервалу часу між двома пакетами є експонентним розподілом з параметром λ [7]. Перевагою цієї моделі є простота в застосуванні й той факт, що сума декількох незалежних пуасонівських процесів становить новий процес із сумарною інтенсивністю $\lambda = \sum_i \lambda_i$. Однак ця модель не має пам'ять [8], тому що в будь-який момент часу ймовірність одержання пакета не залежить від вступу пакетів у минулому. Із цієї причини ця модель не пояснює феномен пульсації трафіку.

Модель «On/Off» використовується в тих випадках, коли виділяються два стани джерела трафіку: активне й пасивне. Час настання активного стану має експонентний розподіл з параметром $1/\alpha$, а час настання пасивного стану розподіляється за експонентним законом розподілу з параметром $1/\beta$. Схема взаємного переходу станів наведено на рисунку 1.1.

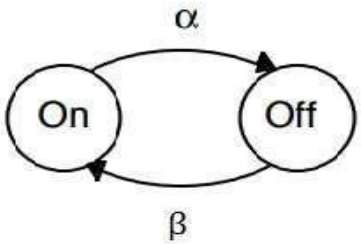

Рисунок 1.1 – Модель On/Off

В активному стані пакети генеруються з постійною інтенсивністю $1/T$, де T – час між двома послідовними пакетами, а в пасивному стані пакети не генеруються. Цей процес можна розглядати, як добуток основного процесу, що й модулює процес, що є марківським процесом (0,1) (рисунок 1.2).

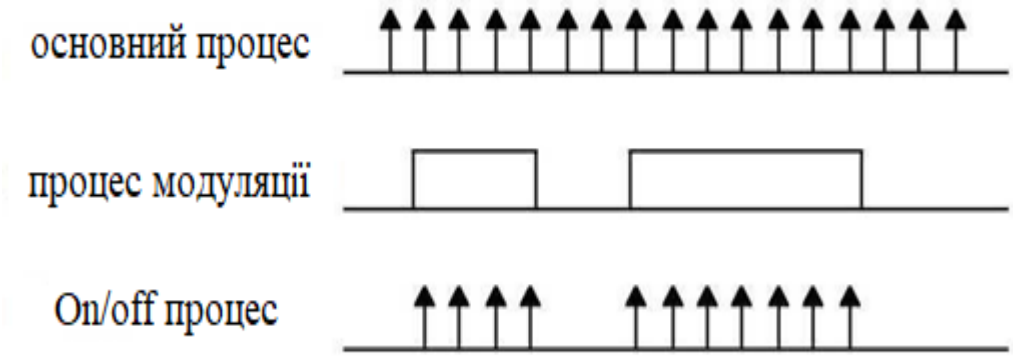

Рисунок 1.2 – Модель On/Off як модульований випадковий процес

Ця модель може бути використана для представлення поведінки декількох додатків (джерел трафіку), таких як додатки IP-телефонії (аудіо- та відеопотік).

Однак умова постійної інтенсивності пакетів вважається основним недоліком цієї моделі, тому що така умова рідко виконується в реальних мережних додатках. Якщо замість постійної інтенсивності $1/T$ генеруються в активному стані з використанням пуасонівського розподілу з параметром λ, то процес є переривчастим пуассонівським процесом (IPP – Interrupted Poisson Process). Модель On/Off

використовується для моделювання відео-трафіку, де виявляються активне (буферізація) і пасивний стану, при цьому інтенсивність в активному стані є фіксованою.

1.2 Моделі часових рядів

Класичні моделі не можуть застосовуватися для рішення завдання ідентифікації трафіку СПД, тому що не мають пам'ять, а використовувана в них проста форма розподілу рідко зустрічається на практиці опису трафіку.

Моделям трафіку на основі теорії тимчасових рядів присвячені роботи [7]. Випадковий процес X_n у моделях мережного трафіку на основі тимчасових рядів використовується для вистави інтенсивності джерела трафіку з дискретним часом $n = 1,2,...$, де інтервали між послідовними відліками рівні. Основною властивістю цих моделей є короткочасна залежність, тому що ці моделі використовують авторегресії при виставі процесу X_n , що дозволяє використовувати їх для прогнозування трафіку. Найвідомішими моделями тимчасових рядів є моделі лінійної авторегресії AR (Autoregressive), з авторегресійним ковзним середнім ARMA (Autoregressive moving average) і з авторегресійним інтегрованим ковзним середнім ARIMA (Autoregressive integrated moving average).

Найпростішою моделлю часових рядів є модель AR [8], у якій значення тимчасового ряду в конкретний момент часу лінійно залежать від попередніх значень ряду. Ця модель представляється у вигляді:

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + e_n$$

де p – порядок моделі,

$\{a_1, a_2, \dots, a_p\}$ – її коефіцієнти

e_n – білий шум.

Модель AR зручна при дослідженні відео-трафіку з повільним рухом. У цьому випадку X_n являє собою зміна для *n*-го відеокадра. Повільний рух викликає кореляції між послідовними відеокадрами [75].

Модель ARMA узагальнює дві більш прості моделі тимчасових рядів – модель AR і модель ковзного середнього (Moving Average) (MA). Модель ковзного середнього являє собою лінійну комбінацію попередніх значень шуму:

$$X_n = e_n + \sum_{i=1}^{q} \beta_1 e_{n-i}$$

де $\{e_n\}$ – білий шум,

$\{\beta_1, \beta_2, \dots, \beta_q\}$ – коефіцієнти ковзного середнього,

q порядок моделі.

Модель ARMA(p,q) складається з комбінації моделей AR і MA у такий спосіб:

$$X_n = e_n + \sum_{i=1}^{p} a_1 X_{n-i} + \sum_{i=1}^{q} \beta_1 e_{n-i}$$

Якщо в цій моделі ввести зворотний оператор B, такий, що $BX_n = X_{n-1}$, то модель ARMA можна представити у вигляді:

$$\phi(B)X_n = \theta(B)e_n$$

де:

$$\phi(B)X_n = 1 - \alpha_1 B - \dots - a_p B^p$$

$$\theta(B) = 1 + \beta_1 B - \cdots + \beta_q B^q$$

Модель ARMA може використовуватися для прогнозування трафіку додатків, що володіють сезонним характером або циклічністю, де модель ARMA застосовується для моделювання й прогнозування трафіку мережі BitTorrent. Також ця модель може використовуватися для ідентифікації мережних вторгнень і атак [9].

Модель ARIMA розглядається як окремий випадок моделі ARMA, якщо прийняти $Y_n = \nabla^d X_n$ (d порядок різниці) і ∇X_n дорівнює:

$$\nabla X_n = X_n - X_{n-1} = (1 - B)X_n$$

Тоді для ARMA (p, q):

$$\phi(B)Y_n = \theta(B)e_n$$

$$\phi(B)(1 - B)^d X_n = \theta(B)e_n$$

В цьому випадку X_n є процесом моделі ARIMA, тому що X_n є інтегралом процесу Y_n (моделі ARMA).

Дані часові ряди часто підкоряються деякому тренду, наприклад, повільно збільшуються або мають циклічні особливості, як це випливає з моделі ARMA (1.2). Для того, щоб згладити ці сезонні зміни, використовують різницю значень тимчасових рядів $(1 - B)^d X_n$. У цьому випадку ця різниця є стаціонарним тимчасовим рядом. На практиці порядок різниці звичайно рівний 1 або 2 [10].

Моделі мережного трафіку на основі моделі ARIMA використовуються для прогнозування трафіку й продуктивності мережі, також можуть використовуватися для виявлення аномалій поведінки трафіку.

Моделі мережного трафіку на основі теорії тимчасових рядів підходять для рішення завдань прогнозування трафіку й використовуються для аналізу типів трафіку із залежностями між пакетами або із сезонним їхнім характером, але вони не підходять для моделювання ряду інших типів трафіку, тому їх не можна використовувати для повного рішення завдання ідентифікації мережного трафіку довільного типу. Крім того, моделі на основі тимчасових рядів відрізняються

високою обчислювальною складністю, що різко обмежує їхнє застосування для ідентифікації в реальному масштабі часу.

Поняття фрактала (самоподібності) використовувалося для опису природнього явища збереження ряду властивостей об'єкта на різних масштабах простору або часу. Наприклад, якщо об'єкт є фрактальним, те його частина при збільшенні може бути схожа на весь об'єкт (рисунок 1.3).

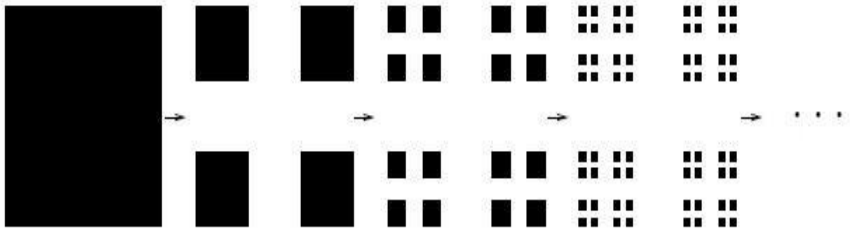

Рисунок 1.3 – Самоподібний двовимірний об'єкт

Фрактальні моделі трафіку мереж передачі даних дозволяють досліджувати феномен пульсації (Burstiness) трафіку, який виявляється на різних масштабах часу (рисунок 1.4).

Основні поняття фрактальних моделей, які використовуються для моделювання мережного трафіку. Нехай $X(n), n \in N$ – дискретний стаціонарний у широкому змісті випадковий процес із математичним очікуванням $M[X(n)] = \mu$ і дисперсією $D[X(n)] = \sigma^2$.

Фрактальний процес має властивість довгочасної залежності:

$$r(k) = \frac{1}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})$$

$$r(k) \sim ck^{-\beta}, \qquad k \to \infty$$

де $\beta = 2 - 2H$,

$c = 2H(2H-1)$.

Фрактальні моделі можна використовувати для моделювання трафіку СПД [12]. До неї ставляться фрактальний броуновський рух, фрактальний гаусовський шум , фрактальна ARIMA, фрактальний On/Off процес та фрактальний процес відновлення [9].

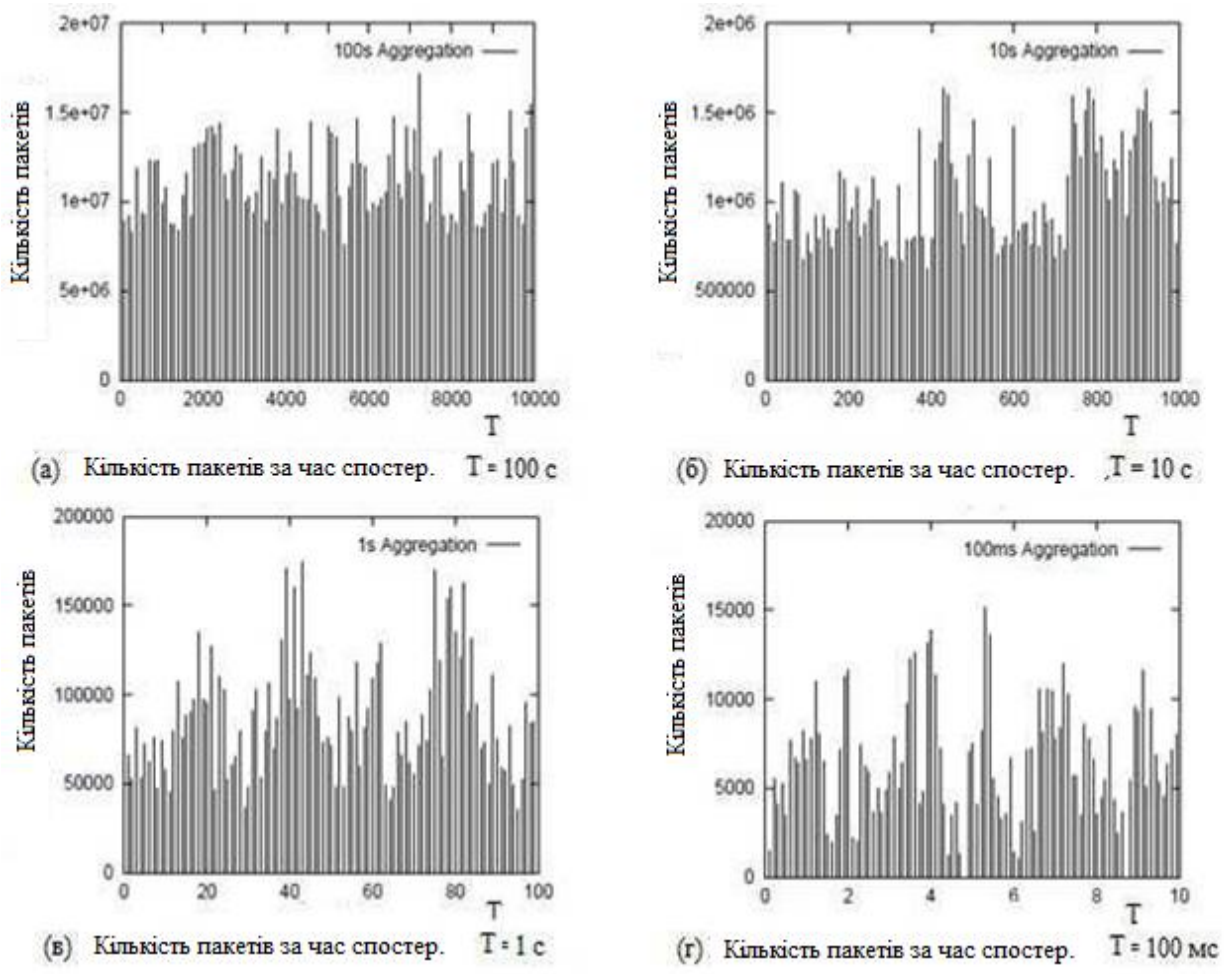

Рисунок 1.4 – Обсяг пульсації мережного трафіку в масштабах часу:
100 с (а), 10 с (б), 1 с (в) і 100 мс (г)

До найбільш відомих фрактальних моделей ставляться фрактальний броуновський рух (FBM), фрактальний гауссовський шум (FGN) і фрактальна ARIMA (FARIMA).

У моделі FBM

$$X(t) = B_H(t + t_0) - B_H(T_0)$$

де t_0– константа,

$B_H(t)$ є винеровський процес із $t \geq 0$ і коваріацією

$$M[B_H(t).B_H(s)] = \frac{\sigma^2}{2}(|t|^{2H} + |s|^{2H} - |t - s|^{2H})$$

де H – показник Херста ($H \in [0,1]$).

Модель FBM використовується для моделювання агрегації множини джерел даних (додатків або IP-адрес), тому її можна застосувати для прогнозування необхідних розмірів буферів.

Модель FGN представляється у вигляді

$$X_n(t) = (\frac{1}{\delta})(B_H(t + \delta) - B_H(t))$$

де δ – збільшення часу.

Процес $X_n(t)$ є нормально розподіленим з нормованою кореляційною функцією

$$r(t) = \frac{1}{2}(|t + 1|^{2H} - 2|st|^{2H} - |t - s|^{2H})$$

Ця модель має ті ж властивості, що й модель FBM.

Модель FARIMA являє собою частий випадок ARIMA [14] з параметром різниці $d \in \mathbb{R}$ і $d \in < [-0.5, 0.5]$. Ця модель має наступну форму запису

$$\phi(B)(1 - B)^d X_n = \theta(B)e_n$$

де $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt (x \in \mathbb{C}$ та $Re(x) > 0)$ є гамма-функция.

Фрактальні моделі використовуються для проектування мереж, особливо в завданнях, пов'язаних із прогнозуванням і розрахунком продуктивності мережі на основі довгочасної залежності [15]. Також їх можна використовувати для керування трафіком, поліпшення якості обслуговування й виявлення аномалій в інформаційних процесах. Однак фрактальні моделі не знаходять свого застосування при рішенні

завдання ідентифікації трафіку в реальному часі через обмеження на число використовуваних пакетів.

Ланцюг Маркова визначається як послідовність випадкових подій з кінцевим або рахунковим числом результатів (станів), що характеризується тем властивістю, що при фіксованому сьогоденні, майбутнє незалежне від минулого [23]. Інакше кажучи, послідовність дискретних випадкових величин $\{X_n, n \in \mathbb{N}\}$ називається простим ланцюгом Маркова, якщо

$$P(X_{n+1} = X_{n+1} | X_n = x_n, \dots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n),$$

значення випадкових величин $\{X_n\}$ утворюють простір станів. Якщо простір станів звичайно і його розмірність рівна N, то перехідна матриця ймовірностей на *n*-м кроці $A(n)$ містить наступні елементи:

$$a_{ij}(n) = P(X_{n+1} = x_i | X_n = x_j)$$

Ланцюг Маркова називається однорідним, якщо елементи матриці перехідних ймовірностей не залежать від номера кроку, тобто

$$a_{ij}(n) = a_{ij}, \qquad \forall n \in \mathbb{N}$$

Марківська модель використовується для моделювання мережного трафіку в тих випадках, коли спостережувана величина (стан) залежить тільки від попереднього стану, наприклад, від стану системи (успіх/невдача), від стану користувача (активний/неактивний) або при моделюванні одержання пакетів у плині деякого інтервалу часу.

Марківська модель із параметрами у вигляді довжини пакета (ДП), напрямку переміщення пакета й у вигляді порядку пакета в потоці. При цьому ДП розділяється на чотири частини, тому для кожного пакета в потоці задаються 8 станів (2 напрямку * 4 частини), як показано на рисунку 1.5.

Недоліком цієї моделі є неможливість ідентифікації з її допомогою додатків у тунельних або шифрованих протоколах, де може мінятися число стерпних пакетів.

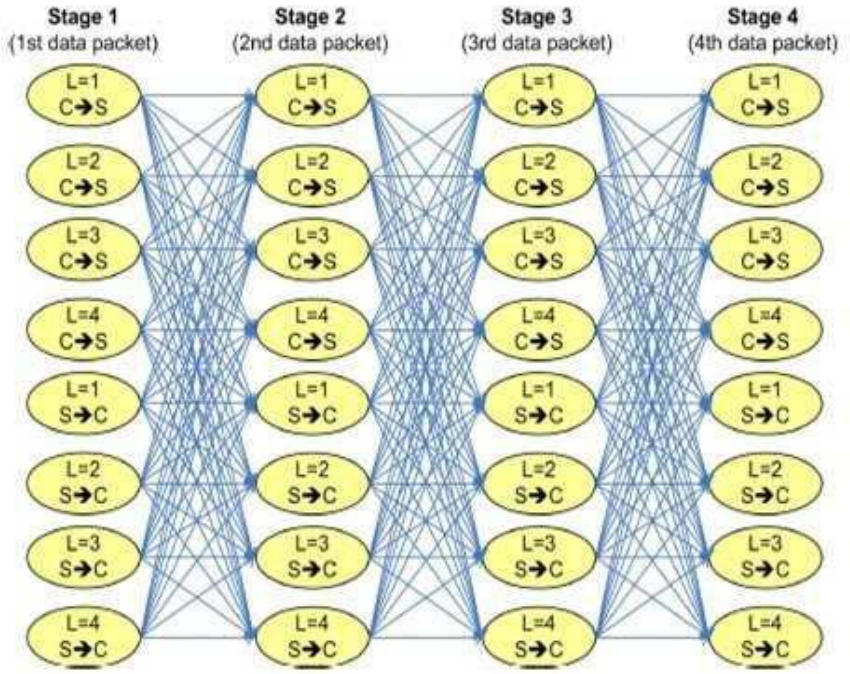

Рисунок 1.5 – Модель Мунза для ідентифікації мережних додатків на основі марківської моделі

Можлива модель, утворена з комбінації марківського й пуасонівського процесів – пуасонівський процес, керований марківським процесом (Markov Modulated Poisson Process – MMPP). Вона є агрегацією з N переривчастих пуассонівських процесів (джерел трафіку) з однаковою інтенсивністю λ в активному стані. Сумарна інтенсивність $\Lambda(t) = n(t)\lambda$, де $n(t)$ визначає кількість активних станів джерел, які є процесом народження-смерті (рисунок 1.6).

Пропонується використовувати модель MMPP для вистави процесів відправлення або одержання пакетів. При цьому станом моделі є кількість джерел трафіку в активному стані (активних додатків). Недоліком MMPP у цьому випадку є те, що умова однакової інтенсивності λ для всіх джерел трафіку рідко виконується на практиці. Тому цю модель узагальнюють на випадок з різною інтенсивністю $\lambda_i(t)$ джерел трафіку, збільшуючи, при цьому її складність.

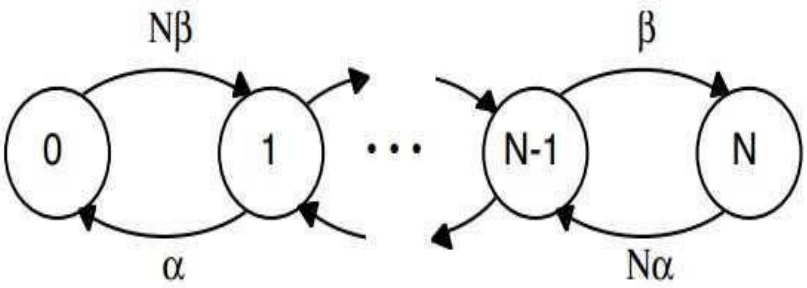

Рисунок 1.6 – Процес народження-смерті

Особливий інтерес викликають сховані марковські моделі (СММ), у яких стани не відомі заздалегідь (сховані) і формуються в процесі навчання моделі.

В [12] запропоновано використовувати дві сховані марковські моделі. У першій моделі в якості спостережуваного значення використаний інтервал часу між пакетами (ІЧП) – τ, а в другий – ДП l. У даній роботі передбачається, що вектори матриці емісії задовольняють розподіл^-розподілові-розкладу-гамма-розподілу:

$$f_i^{(\tau)}\left(d, w_i^{(\tau)}, g_i^{(\tau)}\right) = \frac{(d/w_i^{(\tau)})^{g_i^{(\tau)}-1} e^{(d/w_i^{(\tau)})}}{w_i^{(\tau)}\Gamma(g_i^{(\tau)})}$$

$$f_i^{(l)}\left(b, w_i^{(l)}, g_i^{(l)}\right) = \frac{(b/w_i^{(l)})^{g_i^{(l)}-1} e^{(b/w_i^{(l)})}}{w_i^{(l)}\Gamma(g_i^{(l)})}$$

де $\Gamma(x)$– гамма-функція,

$w_i^{(\tau)}$ і $g_i^{(\tau)}$ – параметри гама розподілу для ІЧП у стані *i*,

$w_i^{(l)}$ та $g_i^{(l)}$ – параметри гама розподілу для ДП у стані *i*.

Недоліками запропонованої моделі є, по-перше, використання тільки одного напрямку переданого трафіку, без включення другого напрямку, що приведе до збільшення часу визначення протоколу. По-друге, у моделі використовуються рівноімовірно початкові стани й рівні ймовірності переходів між станами b випадкові значень для ініціалізації матриць емісії, що, однак, вимагає окремого доказу. По-третє, у роботі не доведена доцільність використання гамма-розподілів у запропонованій моделі.

В представленої Райтом [14] модель ідентифікації також складається із двох СММ. У першій моделі використана ДП у якості спостережуваного значення, а в другий – ІЧП. У моделі передбачається розпізнавання джерела пакета (сервер або клієнт), а також розпізнавання дубльованих пакетів (рисунок 1.7).

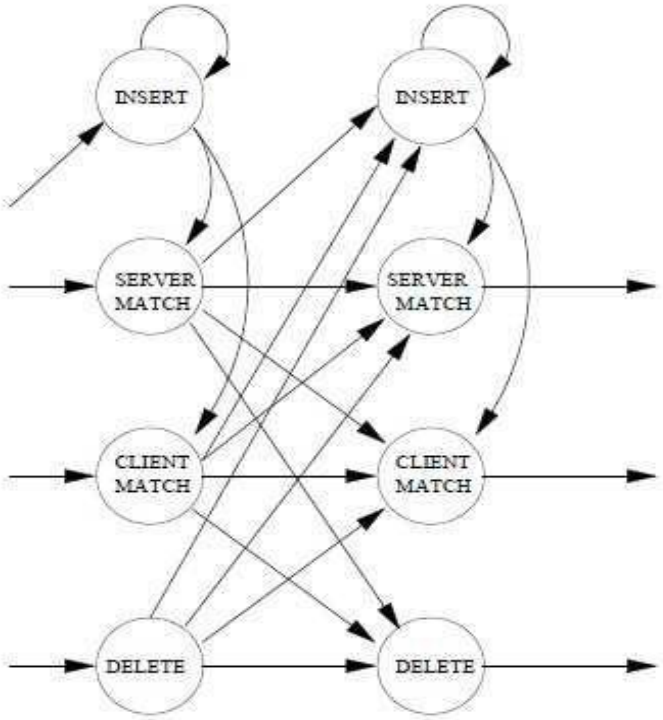

Рисунок 1.7 – Модель Райта для ідентифікації мережних додатків на основі СММ

Дана модель має два істотні недоліки. По-перше, число станів у ній є занадто більшим, оскільки містить вісім станів для вистави кожного пакета в потоці. Тобто, кількість станів моделі збільшується пропорційно кількості пакетів (наприклад, якщо таких пакетів 6, то число використаних станів стане рівним 48), що є занадто більшим у порівнянні з іншими моделями. Це число обмежує її продуктивність і, як наслідок, її використання в більших мережах. По-друге, запропонований метод не підходить уводити, увести до ладу ідентифікації інтерактивних додатків, таких як чат і telnet-сервіси.

1.3 Аналіз методів класифікації трафіку

Математична постановка завдання ідентифікації трафіку в СПД вимагає проведення порівняльного аналізу існуючих методів класифікації. Найпоширенішими методами є класифікації по номеру порту, на основі глибокого аналізу вмісту пакета (Deep Packet Inspection – DPI), з використанням машинного навчання й на основі аналізу статистичних характеристик пакета.

Класифікації трафіку по номеру порту TCP/UDP заснована на призначенні номерів портам для стандартних інтернет-додатків і послуг, певних організацією адміністрування адресного простору інтернету (IANA). Наприклад, веб-трафік класифікується за значенням TCP-порту, рівному 80, а трафік протоколу електронної пошти SMTP – за значенням TCP-порту, рівному 25.

Цей метод є простим і швидким, але його застосування обмежене, оскільки, по-перше, багато з нинішніх мережних додатків не мають зареєстрованого в IANA номера порту й, по-друге, деякі типи мережних додатків міняють номер порту або використовують номери портів інших додатків, щоб уникнути фільтрації на брандмауерові. Рисунок 1.8 ілюструє використання додатком Skype TCP-портів 80 і 443, які призначені для веб-служб (протоколи HTTP і HTTPS).

Хоча номер порту використовується на брандмауерові для блокування певних типів атак і додатків, застосування класифікації трафіку по номеру порту не приведе до рішення завдання IT, особливо при поточній тенденції використання протоколу HTTPS і хмарних IP-адрес.

Метод класифікації трафіку на основі глибокого аналізу вмісту пакетів (корисного навантаження), припускає, що кожний відомий додаток містить у переданих даних певні сигнатури, які дозволяють відрізняти один додаток від іншого. Щоб класифікувати додаток, який згенерував досліджуваний потік, необхідно шукати сигнатури всіх відомих додатків у вмісті пакетів трафіку. Цей

підхід уважається одним з найточніших способів класифікації трафіку й часто використовується при розробці простих комерційних інструментів класифікації [11].

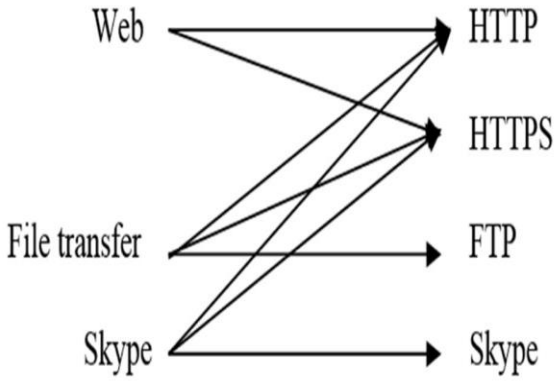

Рисунок 1.8 – Сервіси, що використовують порти стандартних інтернет-додатків і служб

Однак йому властиві наступні недоліки, що можуть вплинути на область його застосування:

– дослідження корисного навантаження мережних пакетів порушує конфіденційність користувача, тому що вміст пакетів може містити конфіденційну інформацію, використання якої третіми особами заборонене національними законами;

– цей метод не застосуємо у випадку зашифрованого трафіку;

– для виявлення конкретних сигнатур необхідно виконати велику кількість операцій порівняння, що неприпустимо в мережах з великою пропускною здатністю й навантаженням.

Машинне навчання ставиться до методів штучного інтелекту і його характерною рисою є навчання в процесі рішення завдання [17]. При використанні цього методу необхідно спочатку визначити ознаки трафіку, які будуть використовуватися в алгоритмі навчання. Ознаки трафіку зазвичай накопичуються в міру дослідження різних потоків і носять, як правило, статистичної характер.

Прикладами ознак можуть служити кількість пакетів у потоці, кількість пакетів з певними значеннями полів заголовків, тривалість потоку, середній розмір пакета, середній інтервал часу між послідовними пакетами й т.п.

Незважаючи на те, що машинне навчання достатнє широко описане, йому властивий цілої ряд істотних недоліків:

– існуючі роботи з машинного навчання присвячені, як правило, вузьким проблемам (наприклад, використання при ідентифікації трафіку певного типу додатків – Skype, Torrent або TOR [5]), більш загальні рішення відсутні;

– в багатьох робіт, що використовують машинне навчання, використовується відкладена класифікація, пов'язана з необхідністю нагромадження всіх пакетів потоку, що виключає можливість використання навчання в реальному часі;

– у багатьох роботах у навчанні використовується більш 20 ознак, що так само обмежує його використання в реальному часі;

– результати класифікації залежать від використовуваних ознак, однак теорія вибору оптимальних ознак відсутня.

Метод класифікації трафіку на основі статистичного аналізу властивостей пакетів припускає пошук статистичних сигнатур додатків, які відрізняють додатка друг від друга. У деяких роботах пропонується використовувати безпосередньо спостережувані статистичні властивості пакетів: інтенсивність пакетів, середнє значення й дисперсію розмірів пакетів і т.п. [11]. В інших роботах розглядаються параметри пакетів, статистичні характеристики яких не змінюються після шифрування. До ними ставляться розміри пакетів, інтервали часу між пакетами, та напрямок передачі пакетів.

В [12] для ідентифікації пропонуються додаткові параметри, що витягають з корисного навантаження пакета: частота байтів і кількість повторюваних байтових пар. Оскільки кожний параметр являє собою імовірнісний розподіл метрика Кульбака-Лейблера, що являє собою відстань між розподілом параметра в спостережуваному потоці й розподілом параметра в потоці відомого додатка

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)},$$

де $P(x)$ – розподіл ймовірностей параметра x у спостережуваному потоці,

$Q(x)$ – розподіл ймовірностей атрибута **e** відомого додатка.

В цьому випадку на етапі навчання розподілу тому одержують ймовірностей $Q_i(x)$ для кожного члена безлічі G відомих додатків, а також виконують оцінку впливу кожного параметра окремо з використанням F-заходу. При спостереженні потік вважається згенерованим певним додатком $g_i \in G$ якщо:

$$D_{KL}(P||g_i) = Min_{gi \in G}(D_{KL}(P||g_i)$$
$$D_{KL}(P||g_i) < T_x$$

де $P(x)$ – граничне значення параметра x.

Це метод використовується для ідентифікації додатка Skype [97] і його сервісів (чат, голос, відео й завантаження файлів). Однак його використання вимагає значних обчислювальних ресурсів.

Використаний метод класифікації мережного трафіку на основі розпізнавання сигнатур відомих додатків, у якості яких використовуються розподіли ймовірностей для всіх відомих додатків $\{p_i\} = \{l_i, \tau_i\}$, де є порядковим номером пакета в потоці, l_i − довжина i-го пакета, а τ_i інтервал часу між i-м $(i-1)$-м пакетами. Для видалення шумових складових кожний елемент p_i , пропускається через фільтр Гауса. Результати зберігаються у вигляді вектора $\vec{M} = \{M_i\}$. Для відстані між спостережуваним потоком F і потоком відомого додатки g_i використовується вектор оцінки аномалії $\vec{A}$ з елементами**:**

$$A_i(P_i M_i) = \frac{1}{\max(\varepsilon, M_i(p_i)},$$

де ε є як завгодно мала позитивна величина.

Якщо відстань S_n менша заданого порогу$T_n^{g_i}$, то потік вважається належним додатком g_i

$$S_n(F, \vec{M}) \leq T_n^{g_i},$$

де поріг $T_n^{g_i}$ визначається сумою математичного очікування й стандартного відхилення функції S_n для потоків, що містять сигнатури додатка g_i

Класифікація мережного трафіку на основі аналізу статистичних властивостей пакетів є діючим методом, тому що вирішує завдання ідентифікації трафіку декількох типів додатків з високою точністю.

Єдиним його обмеженням може служити висока обчислювальна складність, однак спільне його використання зі СММ дозволить одержати ефективний інструмент ідентифікації, що володіє високою точністю й продуктивністю.

1.4 Постановка завдання

Під класифікацією мережного трафіку розуміється процес поділу згенерованих додатками мережних потоків на групи (класи) .

Ідентифікація трафіку СПД є часткам случаємо завдання класифікації трафіку, що де цікавлять нас додатки ставляться до основних класам, а інші невідомі додатки – до додаткових класів.

При роботі мережних додатків запускаються мережні процеси, які активізують передачу даних на основі стека протоколів TCP/IP. Керування передачею даних здійснюється на рівні транспортного протоколу, де передані дані розділяються на частині й передаються в окремих TCP-пакетах. Потоком мережного трафіку є безліч

переданих пакетів, що містять усю сукупність повідомлень одного мережного додатка. Потік трафіку містить п'ять атрибутів (IP-адреса відправника, IP-адреса одержувача, порт відправника, порт одержувача, ідентифікатор транспортного протоколу), і його ідентифікація здійснюється за значеннями цих атрибутів. Виходячи з особливостей роботи додатків ідентифікація трафіку може здійснюватися як для пакетів, що передаються в одному напрямку, так і для пакетів, що передаються в обох напрямках. Останнє впливає на обсяг оброблюваних даних.

Завдання ідентифікації трафіку формулюється в такий спосіб: нехай X є простором досліджуваних потоків, які необхідно класифікувати. Кожний окремий потік $x_i \in X$ характеризується деяким вектором ознак $A_i = \{a_{i1}, a_{i2}, \dots, a_{ik}\}$, $A_i \in D_1 \times D_2 \times \dots D_k$, где D_j – простір значень ознаки a_{ij}. Нехай C є безліч класів додатків (або типів додатків) $C = \{c_1, c_2, \dots, c_{n+1}\}$, де $\{c_1, c_2, \dots, c_{n+1}\}$ – класи відомих додатків і c_{n+1} – клас невідомих додатків. Ідентифікація трафіку є методом зіставлення потоку x_i з відповідним класом додатка c_j на основі значень його вектора ознак. Завдання ідентифікації полягає із двох етапів. На першому етапі визначається сукупність ознак вектора A_i і перебуває функція відображення простору потоків у простір ознак відповідно до їхніх значень:

$$f_f : X \to D_1 \times D_2 \times \dots D_k,$$

$$f_f(x_i) = (f_1(x_i), f_2(x_i), \dots, f_k(x_i)).$$

На другому етапі перебуває функція зіставлення, що зв'язує значення ознак з відповідним класом (відображення простору ознак у простір класів):

$$f_c : D_1 \times D_2 \times \dots D_k \to C,$$

$$f_C\{a_1, a_2, \dots, a_k\} = c_i.$$

Функція класифікації f є відображенням потоку у відповідний клас і складається з композиції дві функції f_f и f_c:

$$f : X \to C$$

$$f(x_i) = (f_c \circ f_f)(x_i)$$

Якість відображення й зіставлення описується за допомогою таких показників, як точність (precision) і повнота (recall). Точність визначається як відношення числа вірно класифікованих потоків до всіх потоків, які процедура класифікації віднесла до певного класу. Повнота класифікації визначається відношенням числа знайдених класифікатором потоків, що належать даному класу, до всіх потоків цього класу в тестовій вибірці. Для більш точної класифікації використовуються три додаткові показники: частка помилкових рішень (error), валідність (accuracy) і *F*-захід (F-measure). Показники класифікації трафіку визначаються за допомогою матриці помилок по наступним формулам:

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

$$Accuacy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}$$

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

$$Error = \frac{N_{FP} + N_{FN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

де N_{TP} визначає число потоків, що належать, досліджуваного з додатком та вірно обраних класифікатором, N_{TN} – число потоків, що належать досліджуваного з додаткм та не вибраних класифікатором, N_{FP} – число потоків, які не належать досліджуваного з додатком та помилково вибраних класифікатором і N_{FN} – число потоків, які не що належать досліджуваного додатком і не вибраних класифікатором.

Класифікація трафіку здійснюється на декількох рівнях залежно від цілей класифікації, причому рівень класифікації трафіку визначає тип вихідних класів. Рівнями класифікації є:

– рівень типу додатка (сервісу): одним з найвідоміших видів класифікації мережного трафіку є класифікація на основі типу додатка, наприклад, потік мультимедіа-контенту, веб, чат, електронна пошта, надання загального доступу до файлів, пірінговий трафік та ін.

Класифікації на основі типу додатка потрібна, щоб забезпечити виконання політик безпеки використання ресурсів:

– рівень додатка: метою даного виду класифікації є ідентифікація певних додатків, наприклад, Skype, WhatsApp, Google Maps та ін.;

– рівень протоколу: даний вид класифікації необхідний для віднесення трафіку до відомого протоколу, наприклад, HTTP, SMTP, FTP та ін.;

– рівень умісту (контенту) прикладного рівня: класифікація проводиться на основі типу даних, які передаються по мережі, наприклад, текст, зображення, бінарні дані, відео контент, зашифровані дані та ін.;

– рівень характеру мережного трафіку: класифікація ґрунтується на певному характері трафіку, який визначається методом передачі даних між мережними обладнаннями, наприклад, інтерактивний трафік, що пульсує, потоковий і т.ін.

2 РОЗРОБКА МОДЕЛІ ІДЕНТИФІКАЦІЇ ТРАФІКУ МЕРЕЖ ПЕРЕДАЧІ ДАНИХ

2.1 Моделі ідентифікації трафіку мереж на основі схованої марківської моделі

Схована марківська модель є окремим випадком загальної марківської моделі і відрізняється тим, що стани моделі є заздалегідь невідомим. ряд робіт присвячений використанням СММ при рішенні різних завдань розпізнавання об'єктів у різних галузях науки й техніки [16].

СММ складається з послідовності спостережуваних значень $x_1, \dots x_T$, де $x_T \in \{0_1, 0_2 \dots 0_M\}$, у моменти часу $t = 1, \dots, T$ і послідовності схованих станів $z_1, \dots z_T$, де $z_T \in \{s_1, s_2 \dots s_n\}$, N – число станів моделі, а M – число символів у спостережуваній послідовності. На рисунку 2.1 наведений приклад такий СММ.

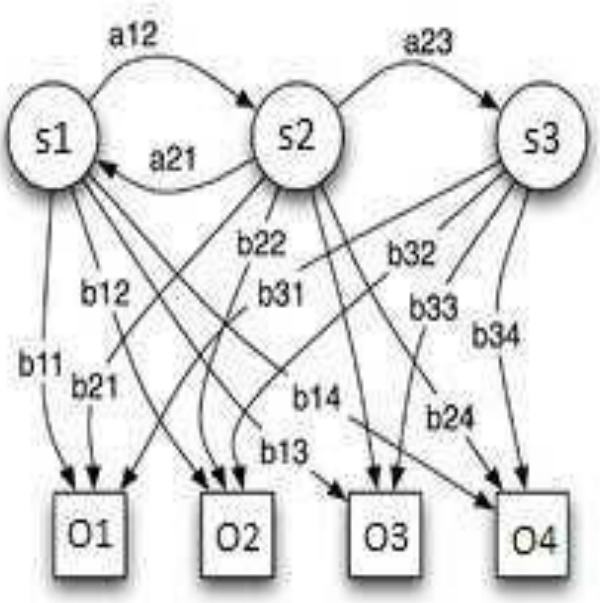

Рисунок 2.1 – Приклад схованої марківської моделі

В основі цієї моделі лежать дві гіпотези:

– у будь-який момент часу t, спостережуване значення x_t залежить тільки від схованого стану z_t;

– схований стан z_t залежить тільки від попереднього схованого стану z_{t-1}.

Параметри СММ записуються у вигляді можини $\theta = \{\pi, A, B\}$, де $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$ визначає ймовірності початкових станів, тобто кожна величина $\pi_i = P(z_1 = s_1)$ визначає ймовірність настання стану s_1 у початковий момент часу $t = 1$.

Матриця ймовірностей переходів між станами $A = \{a_{ij}\}, 1 \le i, j \le N$, складається з елементів $a_{ij} = P(z_{t+1} = s_j | z_t = s_i)$, які задають імовірності переходу зі стану s_i у стан s_j, а матриця $B = \{b_{ij}\}, 1 \le i \le N$ та $j \le j \le M$, імовірностей появи символів у спостережуваній послідовності складається з елементів $b_{ij} = P(x_t = 0_j | z_t = s_i)$, рівних імовірності появи спостереження 0_j в стані s_i.

2.2 Визначення значень параметрів моделі ідентифікації трафіку

У СММ задаються число станів і початкові значення всіх параметрів і на основі спостережуваних послідовностей обчислюються остаточні значення параметрів $\theta = \{\pi, A, B\}$моделі. Найбільше відомим способом обчислення параметрів моделі є алгоритм Баума-Велша (Baum-Welch) [17], який є часткам случаємо алгоритму EM (Expectation Maximization algorithm). Метою цього алгоритму є обчислення таких параметрів моделі, при яких виходить максимальна імовірність виявлення набору спостережень (даних тренування), тобто по алгоритму Баума-Велша перебуває θ^***,** яке визначається наступної формулою:

$$\theta^* = ARGMax_\theta(P(X, Z; \theta)),$$

де $X = \{, X^2, \dots, X^D\}$ – набір спостережень

$X^d = \{x_1^d, x_2^d, \dots, x_T^d\}$ послідовність спостережуваних значень $d \in \{1, \dots, D\}$. Передбачається, що послідовності значень спостережень є незалежними й однаково розподіленими випадковими величинами. Обчислення θ^* є нетривіальним завданням через те, що сховані змінні Z^d невідомі для кожної послідовності спостережень X^d.

Алгоритм Баума-Велша являє собою ітераційну процедуру, в якої на кожній ітерації обчислюються нові параметри моделі θ^{i+1} на основі параметрів, обчислених на попередній ітерації θ^i і набору поточних спостережень. Ітерації повторюються до виконання умови $\theta^i \to \theta^*$, де i – порядок ітерації. При цьому на кожній ітерації виконуються наступні два кроки.

E – Крок (Expectation step): На цьому кроці обчислюються ймовірності вступу спостережень на основі поточних параметрів моделі θ^i**,** для чого використовується допоміжні змінні алгоритму прямого-зворотного ходу [110]**.**

Змінна прямого ходу $a_t(i)$ визначається як імовірність настання спостережуваної послідовності $X^d = x_1, x_2, \dots, x_t$ з моменту 1 до моменту t й стану z_t момент t, рівний i

$$a_t(i) = P(x_1, x_2, \dots, x_t, z_t = i|\theta)$$

При цьому в початковий момент $t = 1$ для всіх станів $i = 1,2, \dots, N$:

$$a_t(i) = \pi_i b_{ix_1},$$

для усіх станів $i = 1,2 \dots, N$ в моментах $t = 2, \dots, T$

$$a_t(i) = \left[\sum_{j=1}^{N} a_{t-1}(j) a_{ji}\right] b_{i_{x_t}}$$

Імовірність настання X^d при заданих параметрах моделі θ дорівнює

$$P(X^d|\theta) = \sum_{j=1}^{N} P(x_1, x_2, \dots, x_t, z_t = i|\theta)$$

$$P(X^d|\theta) = \sum_{j=1}^{N} a_T(i)$$

.

Мінливість зворотного ходу $\beta_t(i)$ визначається як ймовірність настання спостерігається послідовності $X^d = x_{t+1}, \dots, x_T$ з моменту $t+1$ до моменту T за умови знаходження z_t в стані i в момент t. За аналогії змінна $\beta_t(i)$ представляється у вигляді:

$$\beta_t(i) = P(x_{t+1}, x_2, \dots, x_T | z_t = i, \theta)$$

де на початку для всіх станів $i = 1,2,\dots,N$ в момент T

$$\beta_T(i) = 1$$

Для всіх станів $i = 1,2,\dots,N$ в моментах $t = T-1, T-2,\dots,1$:

$$\beta_t(i) = \left[\sum_{j=1}^{N} a_{ij} b_{i,x_{t+1}} \beta_{t+1}(j)\right].$$

М-крок (Maximization step):

Для обчислення параметрів моделі θ, при яких виходить максимальна ймовірність настання спостережень $P(X, Z; \theta)$, використовується функція $Q(\theta, \theta^i) = \sum_{z \in Z} Log[P(X, z; \theta)] P(z|X; \theta^i)$ і обчислюється значення θ^{i+1}, яке забезпечує максимізацію $argmax_\theta(\theta, \theta^i)$ на основі ймовірностей вступу послідовності значень, які виконувалися на попередньому кроці.

2.3 Визначення параметрів спостереження

В запропонованої моделі можливе застосування трьох спостережуваних значень: довжини пакета l, напрямку передачі пакетів (dir) і інтервалу часу між

пакетами τ. Їхнє число можна скоротити, якщо напрямок передачі пакета сполучити з довжиною пакета, уважаючи, що передача від клієнта к серверу має позитивні значення довжини $l_c \in \{1, \dots, l_{max}\}$, а від сервера до клієнта – негативними значеннями $l_c \in \{-l_{max}, \dots, -1\}$, де l_{max} являє собою максимальну довжину пакета в мережі. Тоді реальне значення ДП буде належати цілочисельному інтервалу від $-l_{max}$ до l_{max} байтів. Максимальна довжина пакета у використаних наборах даних рівна 1500 (стандартний розмір кадра в мережах Ethernet). Перший пакет у потоці уважається від клієнта до сервера. Тому що напрямок пакета включений у довжину пакета, то параметрами моделі залишилися тільки величини l й τ.

Мінімальний інтервал часу між послідовно вступниками пакетами рівний 10^{-7} з, а максимальний – 10 хвилин.

Для кожного досліджуваного додатка App, проводиться етап навчання з використанням набору спостережень додатка, який включає значення ДП і ІЧП (рисунок 2.2). На цьому етапі обчислюються параметри моделей ідентифікації досліджуваного додатка

$$\theta_{App,i} = \{\pi_{App,l}, A_{App,l}, B_{App,l}\}$$
$$\theta_{App,\tau} = \{\pi_{App,\tau}, A_{App\tau}, B_{App,\tau}\}$$

На етапі тестування при появі нового потоку обчислюється F імовірність настання цього потоку для ідентифікації додатка:

Потік F вважається приналежним до додатка App, якщо задовольняє умові:

$$App = Argmax_{App(k)} P\left(O_{F,l}, O_{F,\tau} | \theta_{App(k),l}, \theta_{App(k),\tau}\right).$$

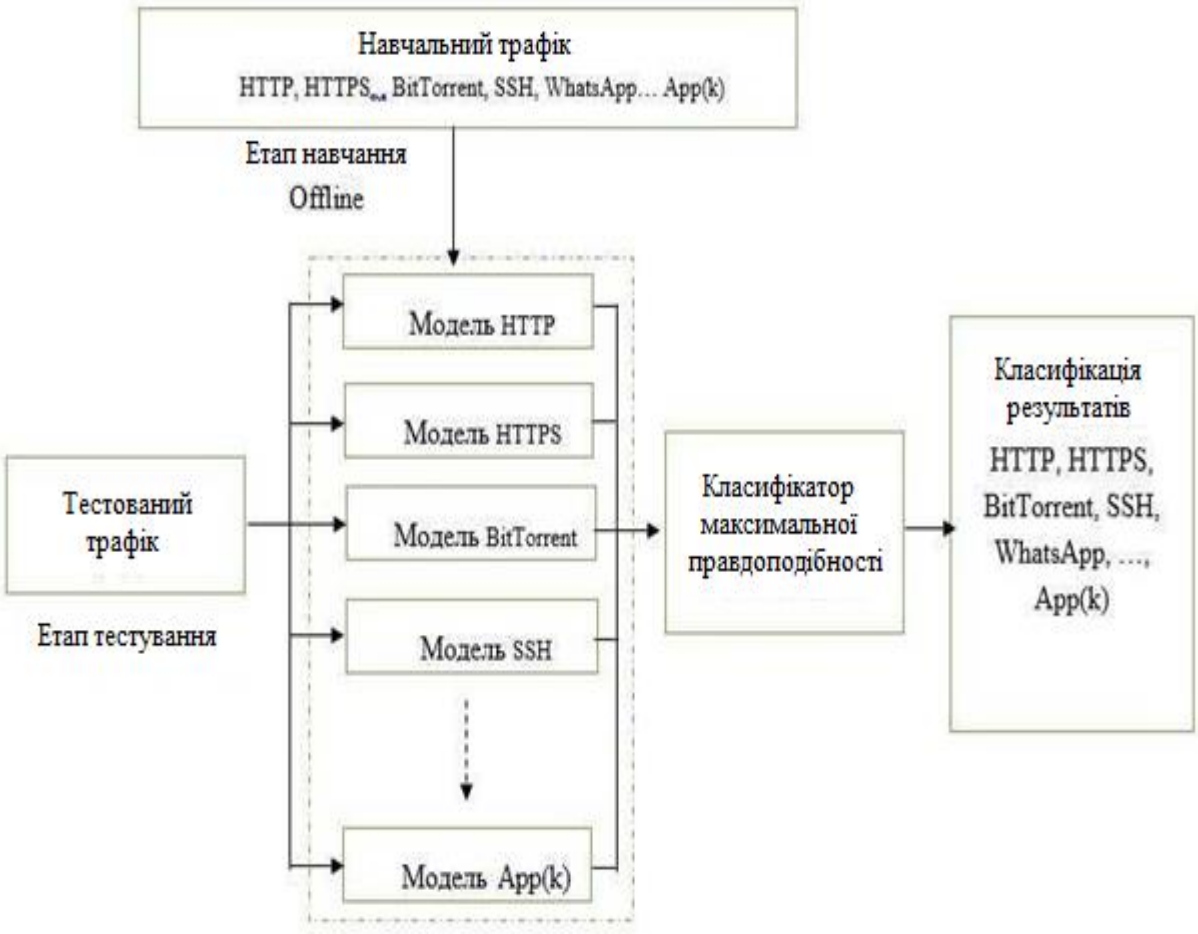

Рисунок 2.2 – Модель ідентифікації мережних додатків

2.4 Визначення початкових значень параметрів моделі

Для визначення параметрів СММ використовується алгоритм Баума-Велша, який, однак, має двома особливостями [48]:

– збіжність алгоритму залежить від початкових значень параметрів моделі;

– іноді алгоритм сходиться до локального максимуму ймовірності $P(X, z; \theta)$ отримані параметри $\theta_1^* = \{\pi, A, B\}$ не забезпечують абсолютний глобальний максимум імовірності (рисунок 2.3).

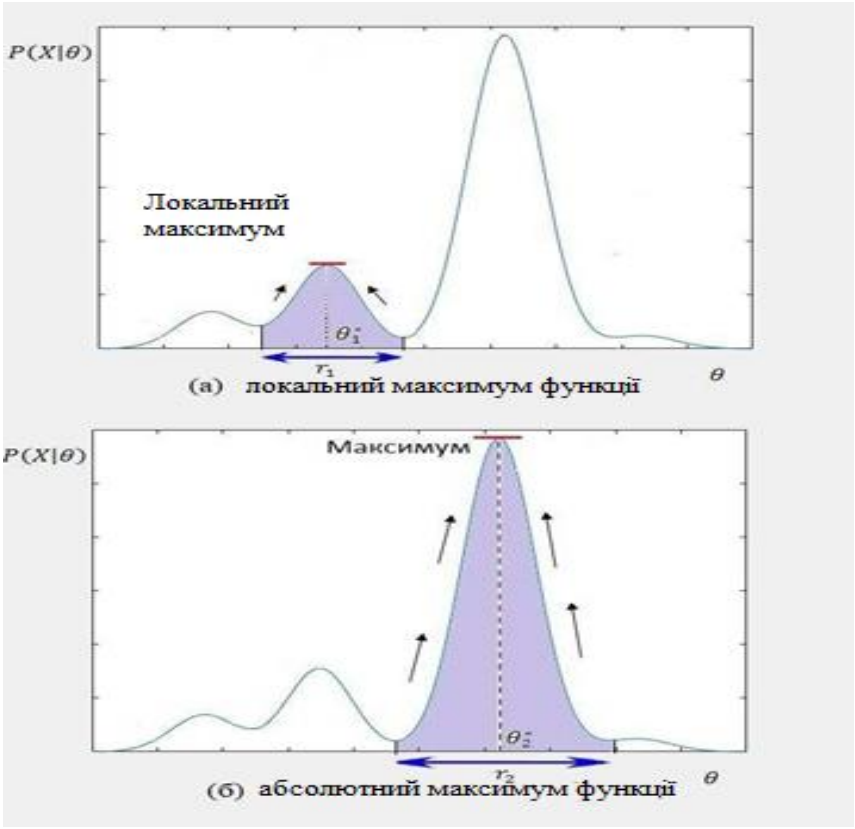

Рисунок 2.3 – Приклад збіжності алгоритму Баума-Велша:
(а) – до локального максимуму; (б) – до абсолютного максимуму

В роботах із класифікації трафіку на основі СММ використовуються або рівномірні, або випадкові значення для початкової моделі, так у роботах [18] всі ймовірності початкових станів рівні $\pi_i = \frac{1}{N}$, усі елементи матриці перехідних імовірностей $a_{ij} = \frac{1}{N}$ (де $1 \le i, j \le N$), а елементи матриці емісії представляються випадковими величинами. Однак метод ініціалізації значень параметрів моделі з використанням випадкових величин має двома недоліками: по-перше, початкові значення параметрів СММ сильно впливають на збіжність алгоритму Баума-Велша [72]; а по-друге, навіть у випадку збіжності алгоритму Баума-Велша, отримані значення параметрів можуть мінятися при повторенні експерименту при тому самому наборі. Інакше кажучи, цей метод показує нестабільні результати.

В даній роботі для рішення завдання вибору числа станів СММ і початкових значень параметрів пропонується модель гаусових сумішей (МГС), яка розглядається як окремий випадок СММ із векторами матриці емісії, що підкоряються нормальному закону розподілу.

МГС визначається наступним чином:

$$p(x) = \sum_{i=1}^{k} \varphi_i N(\mu_i, \sigma_i),$$

де k – число компонентів,

φ_i – вага компонента i,

$N(\mu_i, \sigma_i)$ – функція нормального розподілу.

Оптимізація числа компонентів МГС простіше, чим оптимізація числа станів у СММ. Можна прийняти, що число компонентів МГС дорівнює числу станів у СММ, де кожний компонент у суміші еквівалентний одному стану СММ. У даній роботі вибір числа станів моделі СММ виконаний на основі МГС, у якій число станів **:** дорівнює числу компонентів суміші K і початкове значення-го вектора матриці підкоряється розподілу Гаусса $N(\mu_i, \sigma_i)$ у суміші, тому для обох моделей θ_l і θ_τ матриця емісії

$$b_{l,ij} = N_l(, \mu_i, \sigma_i),$$

$$b_{\tau,ij} = N_\tau(, \mu_i, \sigma_i).$$

Для обчислення розподілу станів у початковий момент часу необхідно визначити розподіл довжини першого пакета кожного потоку т розподіл інтервалу часу між першим і другим пакетами кожного потоку. Імовірність кожного стану в початковий момент часу перебуває в такий спосіб:

$$\pi_{l,i} = \sum_{j=-50}^{50} P_{l,j} \cdot b_{l,ij},$$

$$\pi_{\tau,i} = \sum_{j=1}^{100} P_{\tau,j} \cdot b_{\tau,ij} .$$

На рисунку 2.4 наведений розподіл (гістограма) спостережуваних довжин пакетів протоколу HTTPS.

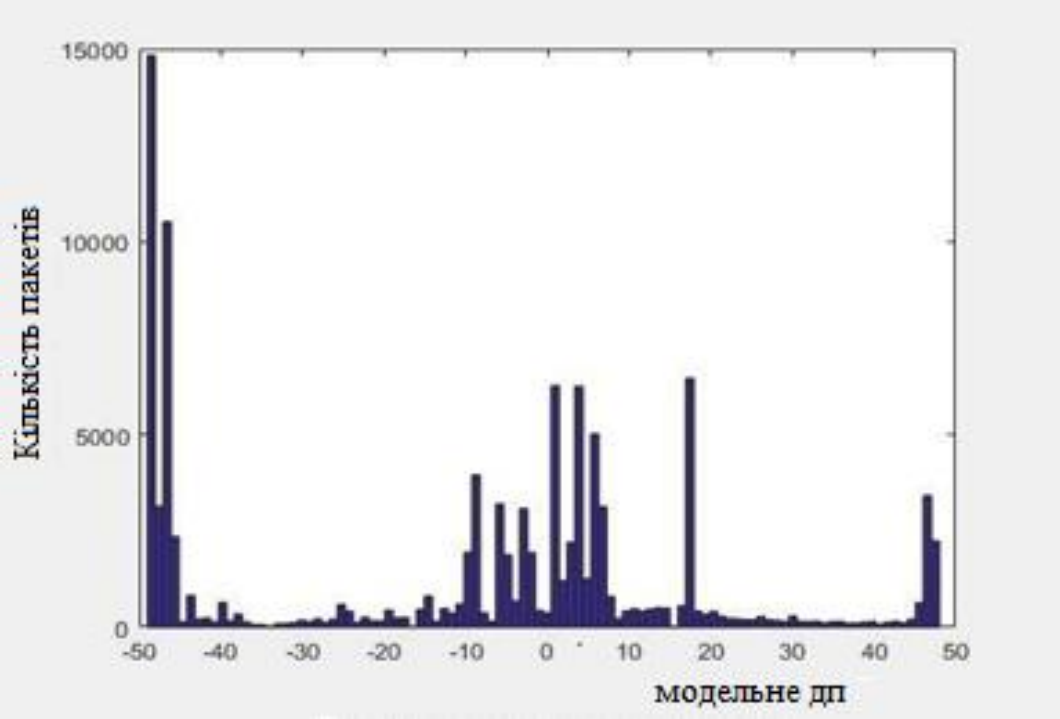

Рисунок 2.4 – Гістограма модельних значень ДП протоколу HTTPS

На рисунку 2.5 (а) наведений відповідний розподіл імовірності вступу ДП для протоколу HTTPS, а на рисунку 2.5 (б) наведені відповідні обчислені гауссові компоненти, які використовуються для ініціалізації параметрів СММ. З рисунків випливає, що шести станів досить для вистави спостережуваних значень ДП для HTTPS.

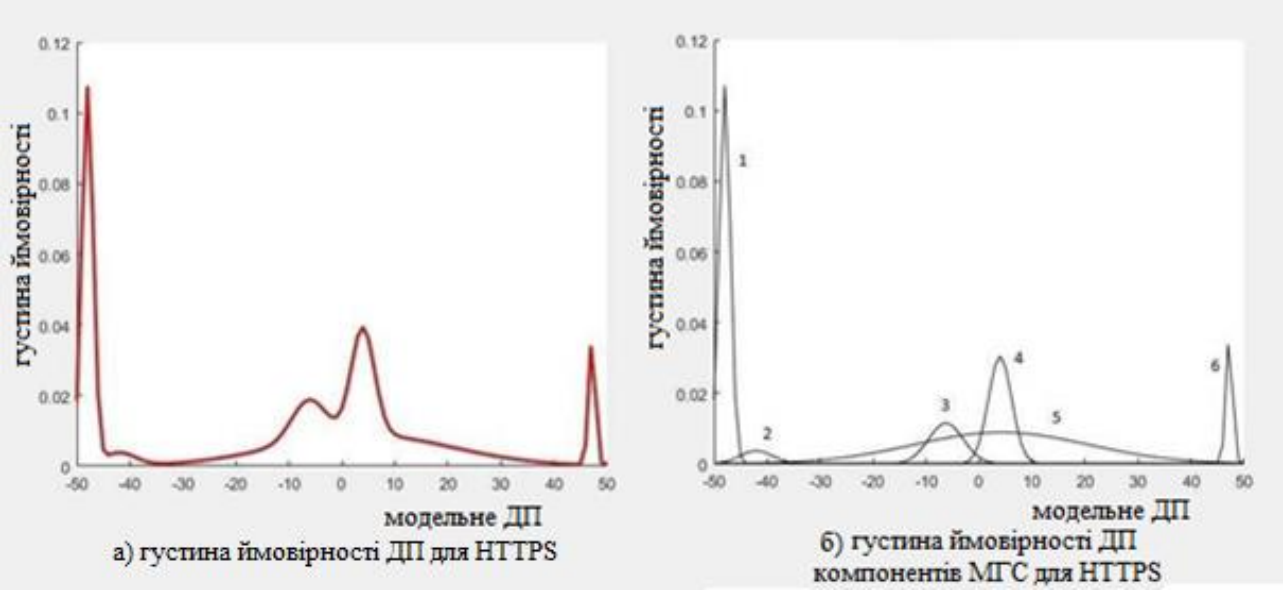

Рисунок 2.5 – Гістограми модельних значень ДП протоколу HTTPS

На рисунку 2.6 наведений розподіл (гістограма) спостережуваних інтервалів часу між пакетами протоколу HTTPS.

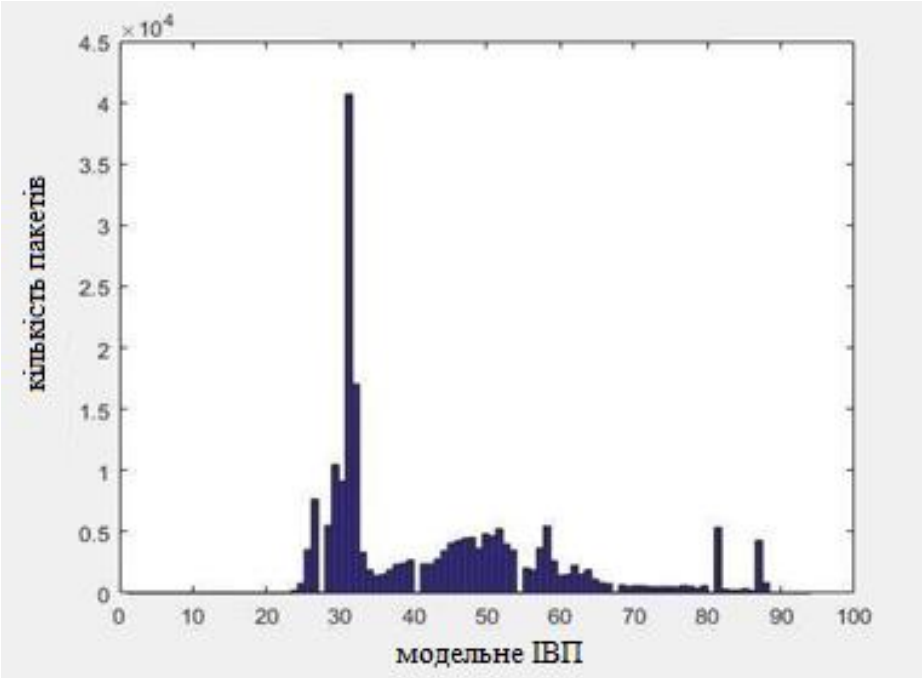

Рисунок 2.6 – Гістограми модельних значень ІЧП протоколу HTTPS

На рисунку 2.7 (а) наведений відповідний розподіл імовірності вступу ІЧП для протоколу HTTPS, а на рисунку 2.7 (б) наведено відповідні обчислені гаусові компоненти, які використовуються для ініціалізації параметрів СММ. З рисунків випливає, що шести станів досить для вистави спостережуваних значень ІЧП для HTTPS.

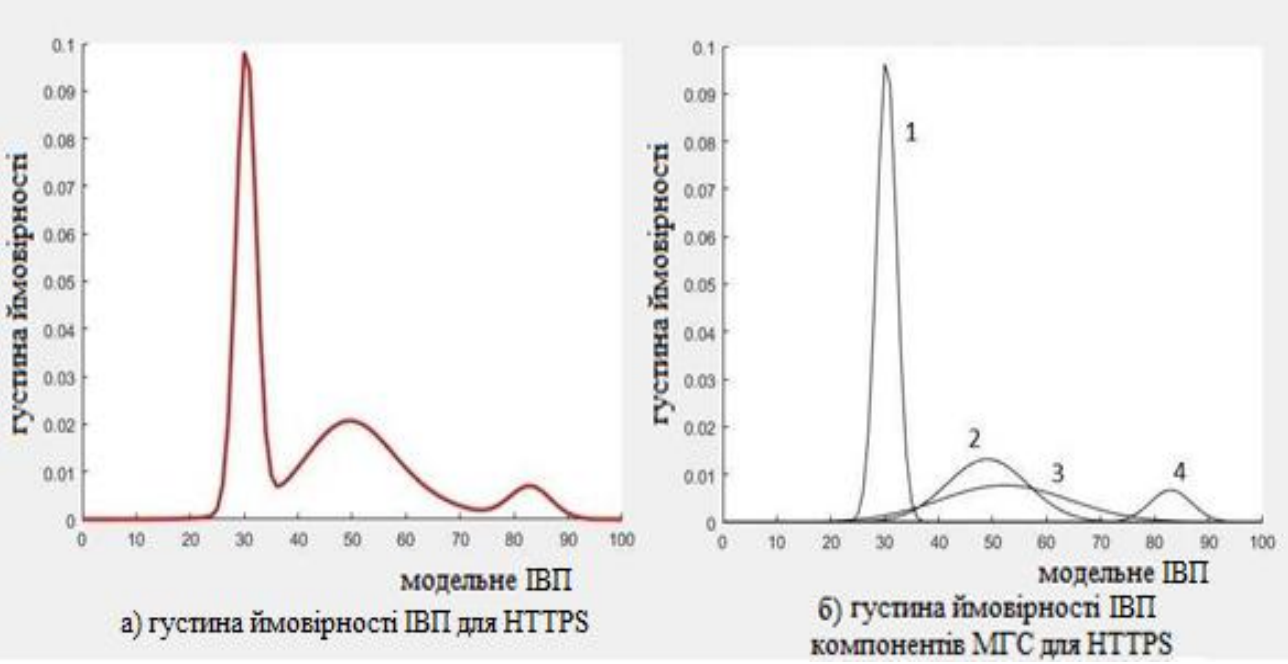

Рисунок 2.7 – Гістограми модельних значень ДП протоколу HTTPS

Для підтвердження переваги методу МГС при виборі початкової моделі в роботі проведено п'ять експериментів з використанням трьох методів вибору початкової моделі:

– повністю випадкова початкова модель, де π_l, A_l, B_l та π_τ, A_τ, B_τ обрані випадково;

– частково випадкова початкова модель, де B_l і B_τ обрані випадково, а с π_l, A_l і π_τ, A_τ мають рівно ймовірнісні розподіли:

$$A_{l,ij} = A_{\tau,ij} = \frac{1}{N} \text{ для всіх } i, j \in \{1, \dots, N\},$$

$$\pi_{l,ij} = \pi_{\tau,ij} = \frac{1}{N} \text{ для всіх } i,j \in \{1,\dots,N\};$$

– початкова модель на основі МГС.

З використанням повністю випадкової початкової моделі й частково випадкової початкової моделі були проведені по два експерименти, а для початкової моделі на основі МГС – один експеримент. Результати експериментів наведені значення математичного очікування й дисперсії логарифма ймовірності послідовностей ДП і ІЧП потоків додатка. Порівняння результатів показує, що середні значення логарифмічних метрик для всіх способів одержання початкових значень параметрів моделі практично збігаються. Крім того, при МГС значення дисперсії логарифмічної метрики мінімальні. Тобто, значення ймовірності появи спостережень при методі гаусової суміші є більш стабільними, ніж при інших моделях.

На рисунку 2.8, де наведені впливи числа станів моделі на точність ідентифікації для трьох додатків (протоколів) при використанні різного числа станів моделі від 2 до 10 для ДП і ІЧП

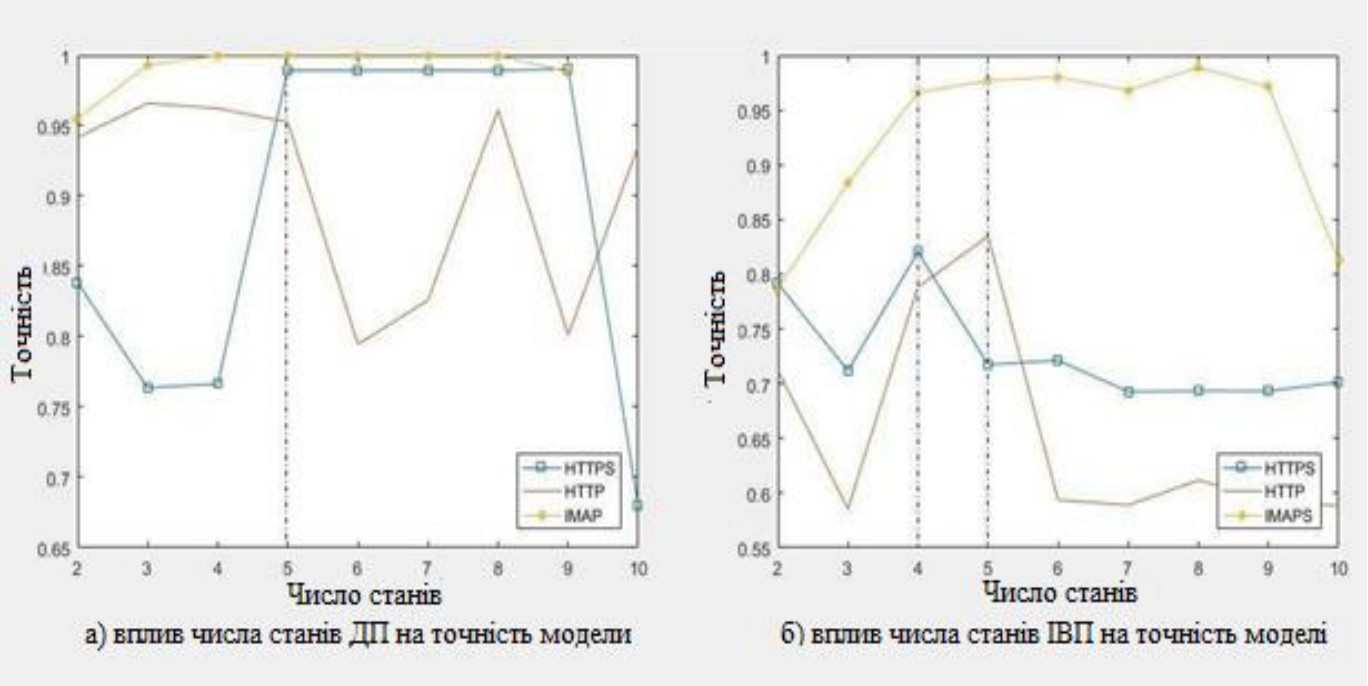

Рисунок 2.8 – Вплив числа станів моделі на точність ідентифікації додатків

Для визначення числа станів, при яких можливе одержання максимальної точності ідентифікації додатка, мають бути проведені експериментальні дослідження. результати яких представлено.

Експерименти підтверджують той факт, що оптимальне число станів СММ збігається з оптимальним числом компонентів МГС. Так, наприклад, оптимальні числа гаусових компонентів для протоколу HTTPS є рівними 6 для ДП і 4 для ІЧП.

3. ОПИС АЛГОРИТМІВ ІДЕНТИФІКАЦІЇ МЕРЕЖНОГО ТРАФІКУ

3.1 Алгоритм підготовки трафіку для ідентифікації

Представлені алгоритми ідентифікації мережного трафіку виконуються у два етапи. На першому етапі проводиться підготовка потоків трафіку до процесу ідентифікації, а на другому етапі виконується сама ідентифікація додатків, які генерують досліджувані потоки.

В свою чергу підготовка мережного трафіку включає в собі три етапи (рисунок 3.1):

– захоплення трафіку;

– відновлення потоків;

– обчислення параметрів потоків.

Існують два типи підготовки трафіку залежно від цілей етапу (навчання або тестування). У випадку навчання трафік кожного додатка знімається з журналів трафіку й зберігається в окремих файлах, на яких згодом проводиться навчання програм для ідентифікації додатків. Для маркування потоків трафіку (набору даних) використовуються різні методи, розглянуті далі в параграфі 3.3. На етапі тестування обчислюється кількість пакетів усіх потоків для всіх додатків.

Для захоплення мережного трафіку розроблена спеціальна програма, трафік для якої може бути представлений у наступних формах:

– мережний журнал (offline) – запис мережного трафіку у формі бінарного файлу, збереженого на носії;

– дані реального часу (online) – поточний трафік, одержуваний з мережного інтерфейсу.

Алгоритм, реалізований у програмі захоплення, дозволяє працювати з обома формами мережного трафіку.

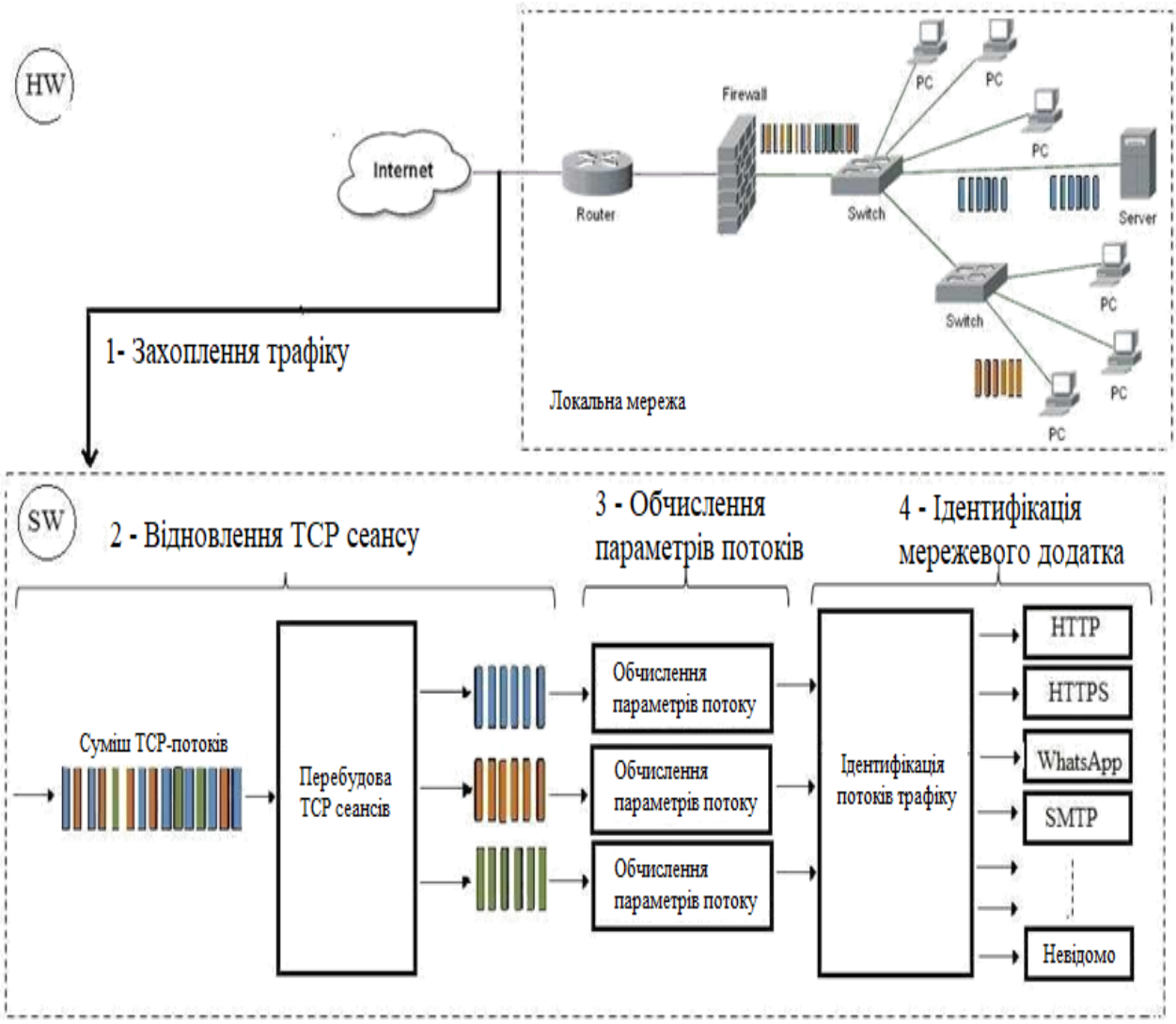

Рисунок 3.1 – Етапи захоплення й підготовки трафіку до ідентифікації

В розробленій програмі захоплення кожний сеанс з'єднання (з'єднання TCP або потік) визначається п'ятьома атрибутами (IP-адреса відправника, IP-адреса одержувача, порт відправника, порт одержувача, тип транспортного протоколу).

Алгоритм виконує відновлення потоку трафіку, тобто виявляє початок потоку по сигналах синхронізації протоколу TCP. Потім проводиться запис довжини пакетів, а також інтервалів часу між пакетами «нового» потоку. Схема цього алгоритму представлено на рисунку 3.2. Реалізація алгоритму має бути виконана мовою C# з використанням програмної платформи Visual Studio.

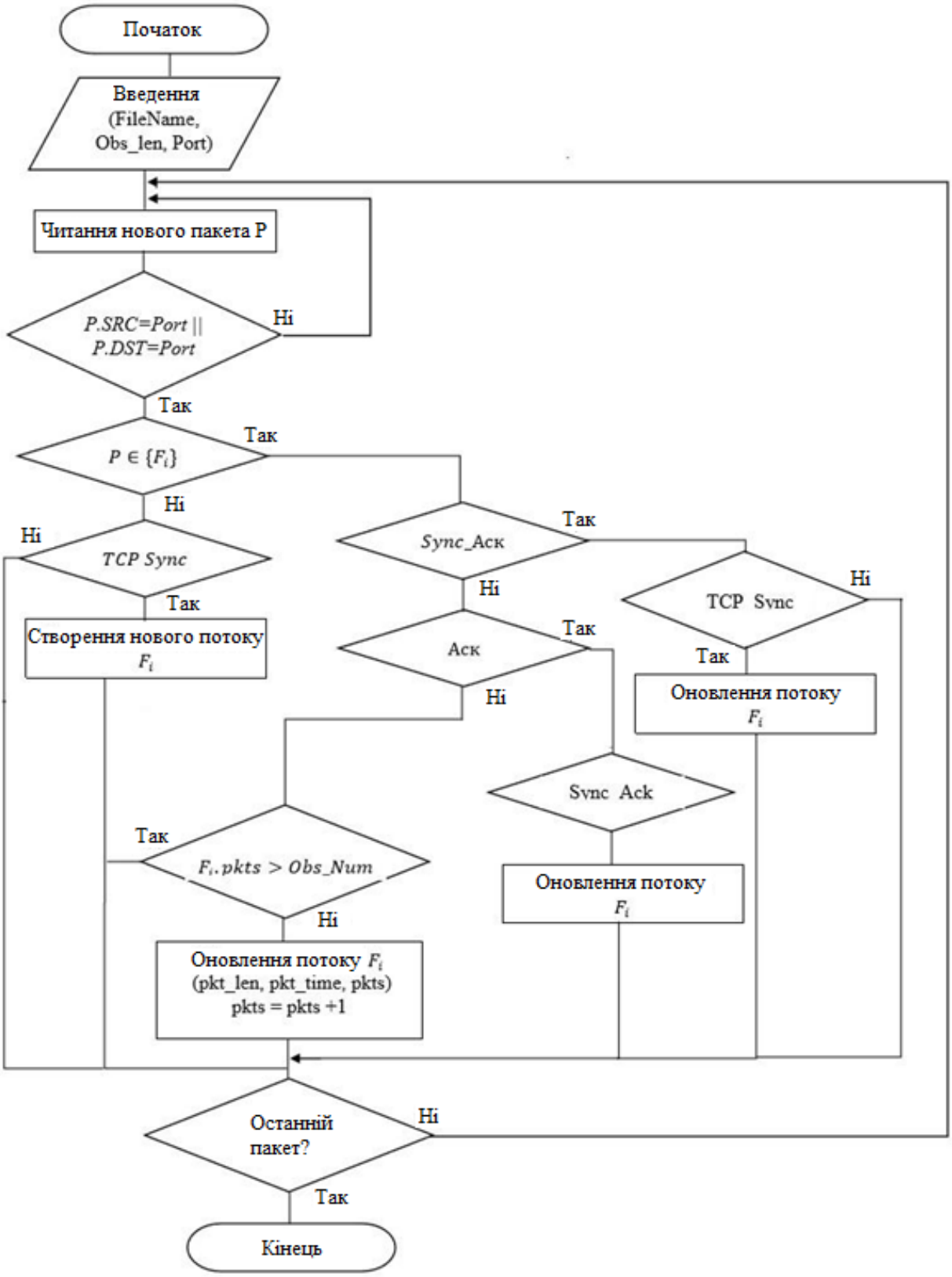

Рисунок 3.2 – Схема алгоритму поділу потоків мережного трафіку

Кількість розглянутих пакетів в експериментах становить більш 20 і всі пакети, які не несуть корисного навантаження, не ухвалюються до обробки (Sync, Ack, Reset і т.ін.). Також потоки, що полягають із менш, ніж 20 пакетів, не включаються в процес ідентифікації. У запропонованій моделі виконується виключення однакові атрибути, що мають, пакетів, тому що вони не несуть додаткової інформації й, отже, приводять у результаті до погіршення ідентифікації мережного трафіку.

3.2 Алгоритми ідентифікації на етапі навчання й тестування

Відновлення TCP сеансу проводиться після його ініціалізації, яка визначається послідовністю трьох пакетів (Sync, Sync-Ack, Ack) як показано на рисунку 3.3.

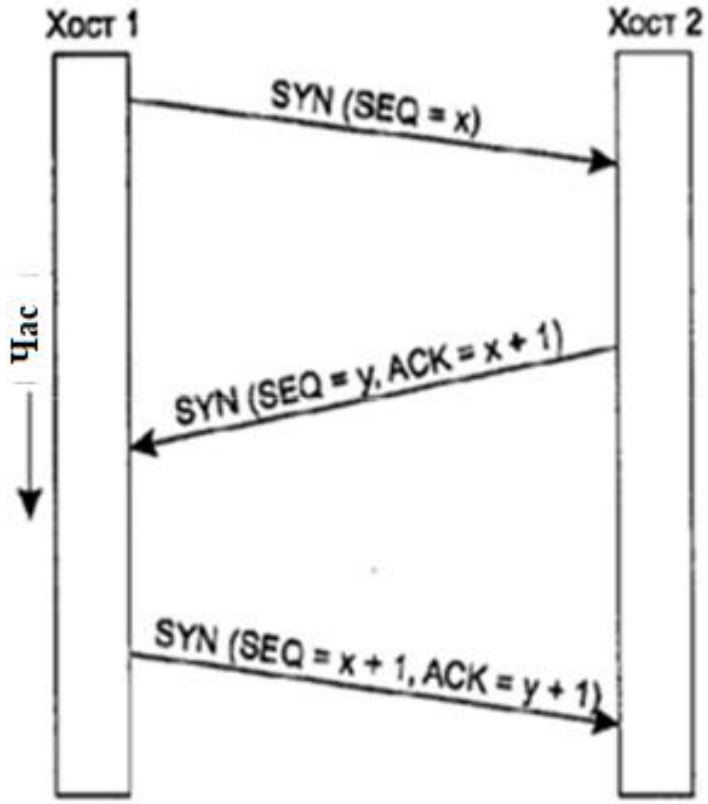

Рисунок 3.3 – Ініціалізації TCP сеансу

Схема алгоритму ініціалізації значень параметрів моделі з використанням моделі гаусової суміші показано на рисунку 3.4.

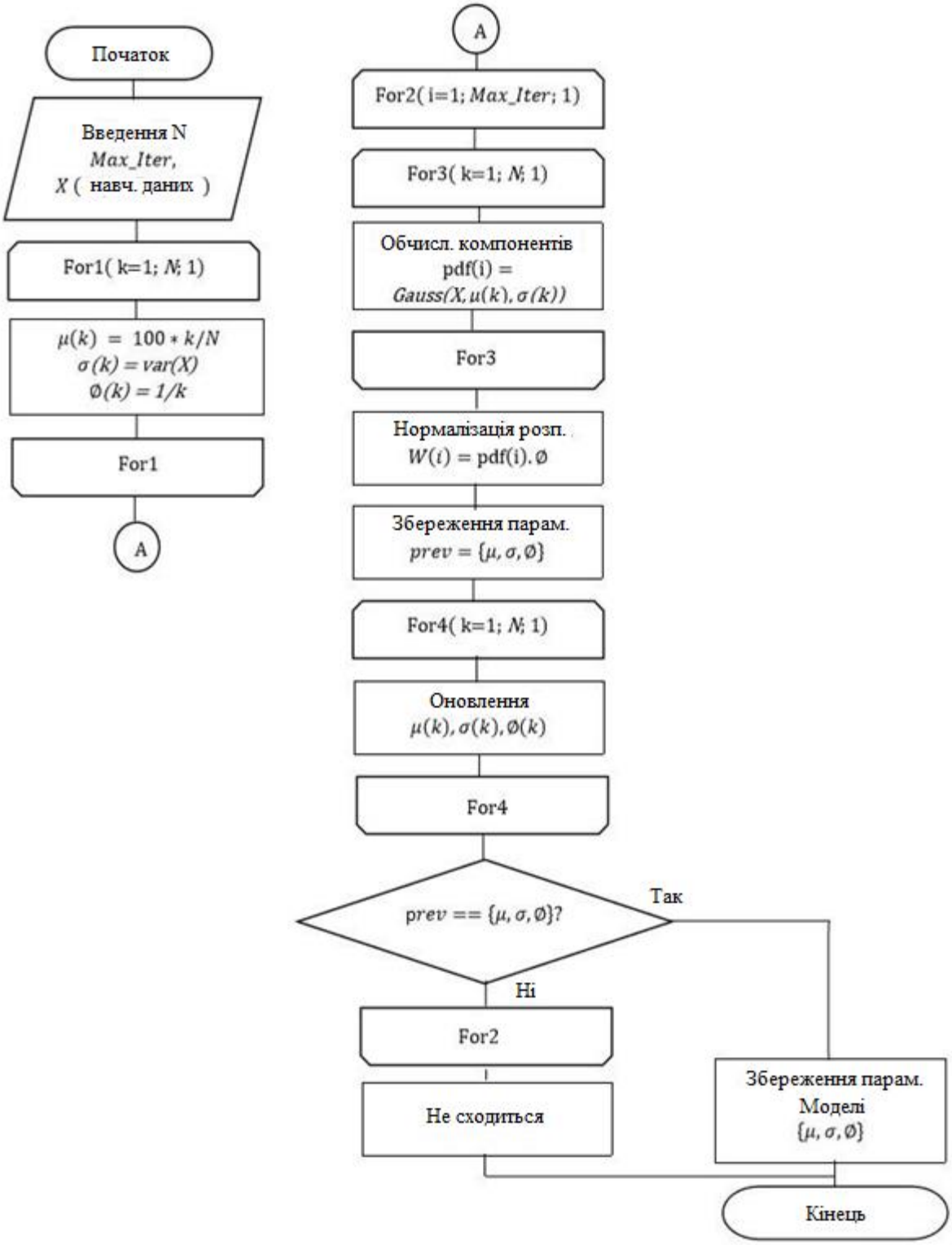

Рисунок 3.4 – Схема алгоритму ініціалізації значень параметрів моделі на основі моделі гауссової суміші

Для оцінки початкових значень параметрів моделі використаний метод гаусової суміші. У цьому методі передбачається кількість компонентів суміші відомим і за допомогою алгоритму максимальної правдоподібності обчислюються параметри кожного компонента (вага компонента, середнє значення й дисперсія). Якщо запропоноване число компонентів не підходить для вистави розподілу спостережень у вигляді гаусовських сумішей, то алгоритм вертається в початковий стан.

Кількість станів у СММ впливає на продуктивність запропонованої моделі ідентифікації трафіку СПД: чим більше кількість станів, тим більше обчислювальна складність. Тому, оптимізація компонентів здійснюється з мінімально можливого числа компонентів, рівного 2, до максимального числа компонентів, рівного 10.

Схема алгоритму обчислення параметрів моделі ідентифікації на основі процедури Баума-Велша наведено на рисунку 3.5. З використанням цього алгоритму обчислені параметри моделі ідентифікації для обраних додатків. Входом алгоритму є навчальний набір даних, а виходом – початкові значення параметрів моделі. Кількість станів моделі N визначається експериментально в алгоритмі ініціалізації значень параметрів моделі.

Поріг ідентифікації кожного досліджуваного додатка обчислений за допомогою навчального набору даних X , де кількість спостережень, що задовольняють граничному рівню, становить 99% із усіх виконуваних спостережень додатка:

$$P(x|\theta_{App}) \geq \text{Порог}(App).$$

Параметри моделі й поріг ідентифікації кожного додатка зберігаються, щоб використовувати для тестування нових наборів даних, а також для обчислення показників моделі ідентифікації додатків.

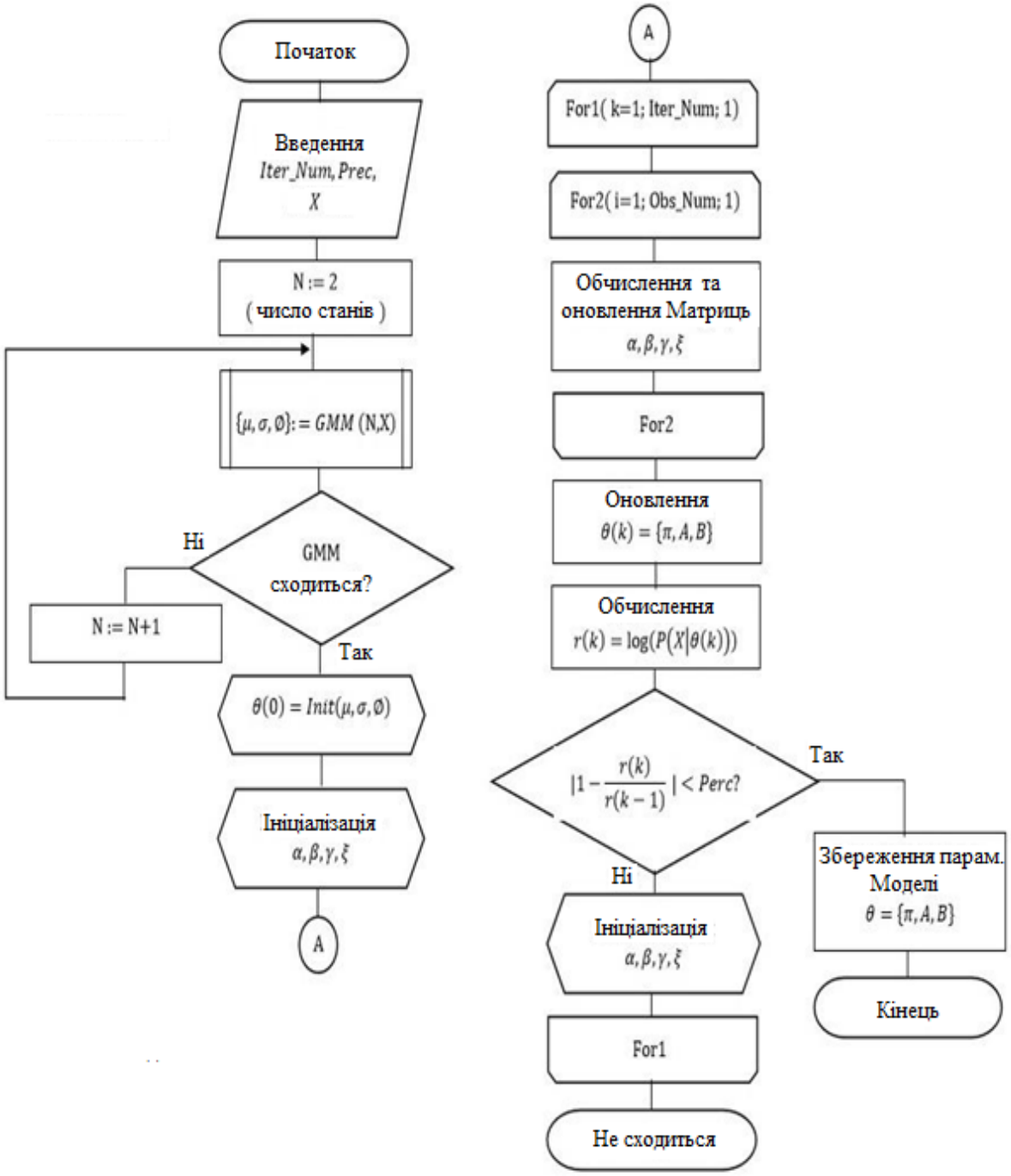

Рисунок 3.5 – Схема алгоритму обчислення параметрів моделі ідентифікації на основі процедури Баума-Велша

Для оцінки якості запропонованої моделі використовуються чотири стандартні показники якості машинного навчання: точність, повнота, валідність і частка помилок. Усі показники якості моделі ідентифікації використовують значення:

TP, TN, FP, FN. Ці значення обчислюються за допомогою алгоритму оцінки якості моделі, схема якого наведено на рисунку 3.6.

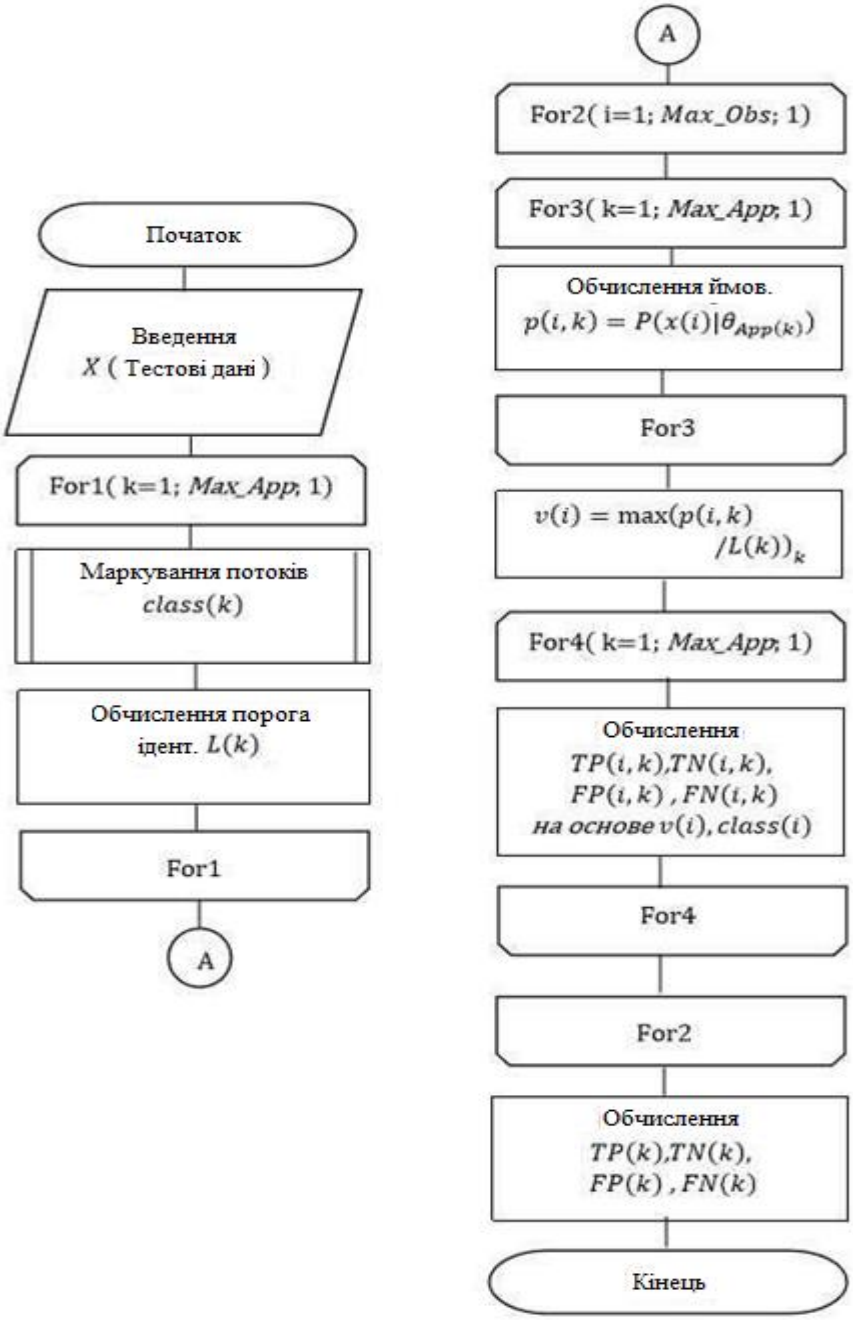

Рисунок 3.6 – Схема алгоритму оцінки якості моделі ідентифікації мережного трафіку

Показники якості ідентифікації обчислюються для кожного досліджуваного додатка. Наведений алгоритм обчислює ймовірності того, що потік був згенерований кожним з додатків. Максимальна ймовірність визначає маркування додатка (вихід моделі) і порівнюється зі справжніми додатками, що генерують відповідний потік.

3.3 Алгоритм вибору мережних додатків на етапах навчання й тестування

В роботі обрано шість додатків для тестування моделі ідентифікації мережного трафіку:

- веб-додатки HTTP (протокол передачі гіпертексту).
- безпечні веб-додатки HTTPS (безпечний протокол передачі гіпертексту).
- безпечна електронна пошта IMAPS.
- однорангові додатки P2P.
- чат-додатка WhatsApp.
- тунельний додаток TOR.

Для навчання моделі использоан мережний трафік, зібраний у МГТУ ім. Н.Є. Баумана (розмір журналів 30 ГБ), і трафік, зібраний в університеті Нью-Брансуика (Канада) (розмір жуналов 45 ГБ). Для маркування трафіку використане програмний пакет Libprotoident 2.09.

Далі даний короткий опис кожного типу трафіку, які використовуються для навчання й тестування моделі.

Трафік застосування веб-додатка (звичайний і безпечний) згенерований через навігацію різних типів веб-сайтів (новин, університетів, державних, спортивних, розваг і інших); при цьому використані веб-браузери «Google Chrome» версія і «Opera». На рисунках 3.7 і 3.8 наведені гістограми спостережень ДП і ІЧП для протоколів HTTP і HTTPS.

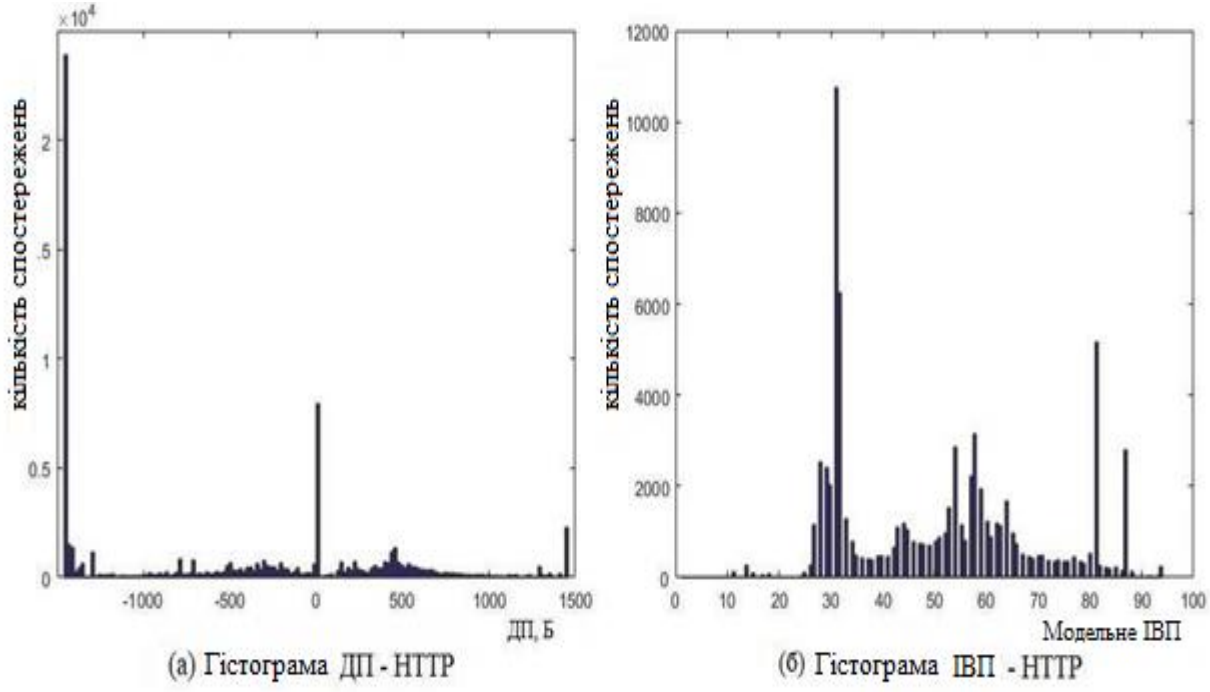

Рисунок 3.7 – Гістограми спостережень протоколу HTTP

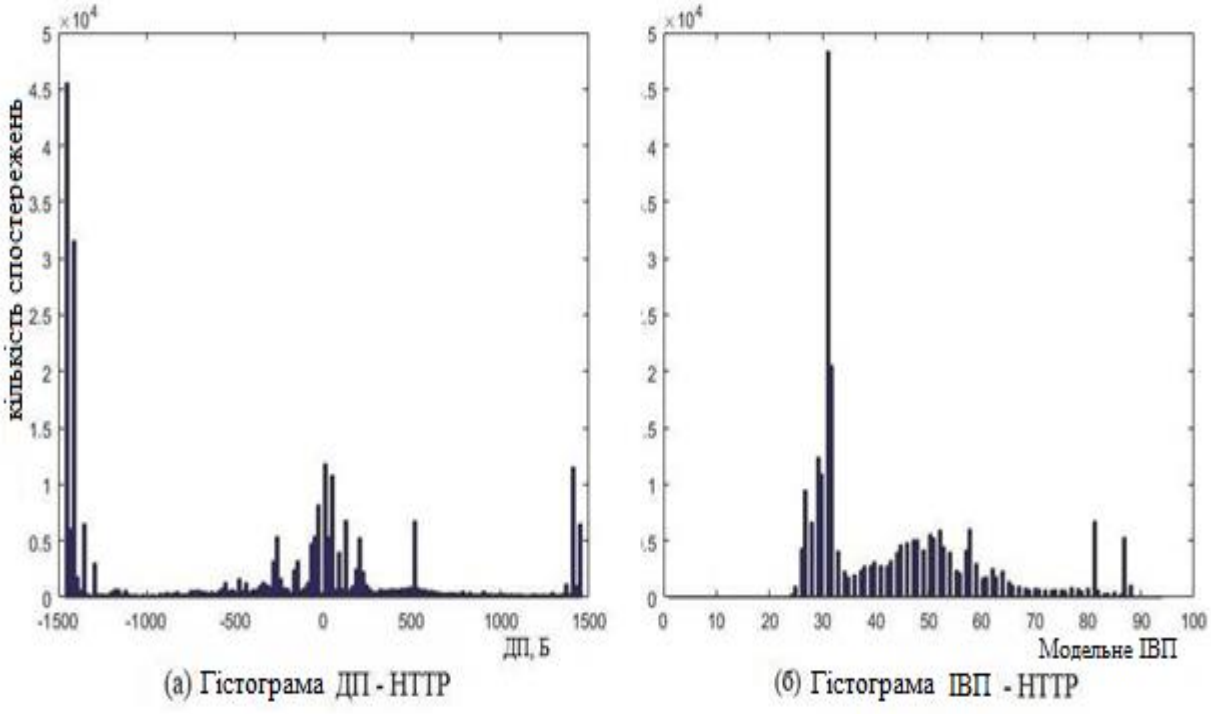

Рисунок 3.8 – Гістограми спостережень протоколу HTTPS

Трафік електронної пошти – IMAPS протокол використовується для того, щоб одержувати повідомлення електронної пошти із сервера. У процесі роботи цього протоколу можна виділити два типи сесій:

– активної, коли існує нове повідомлення,

– пасивної, коли немає нових повідомлень.

На рисунку 3.9 наведені гістограми спостережень ДП і ІЧП для протоколу IMAPS.

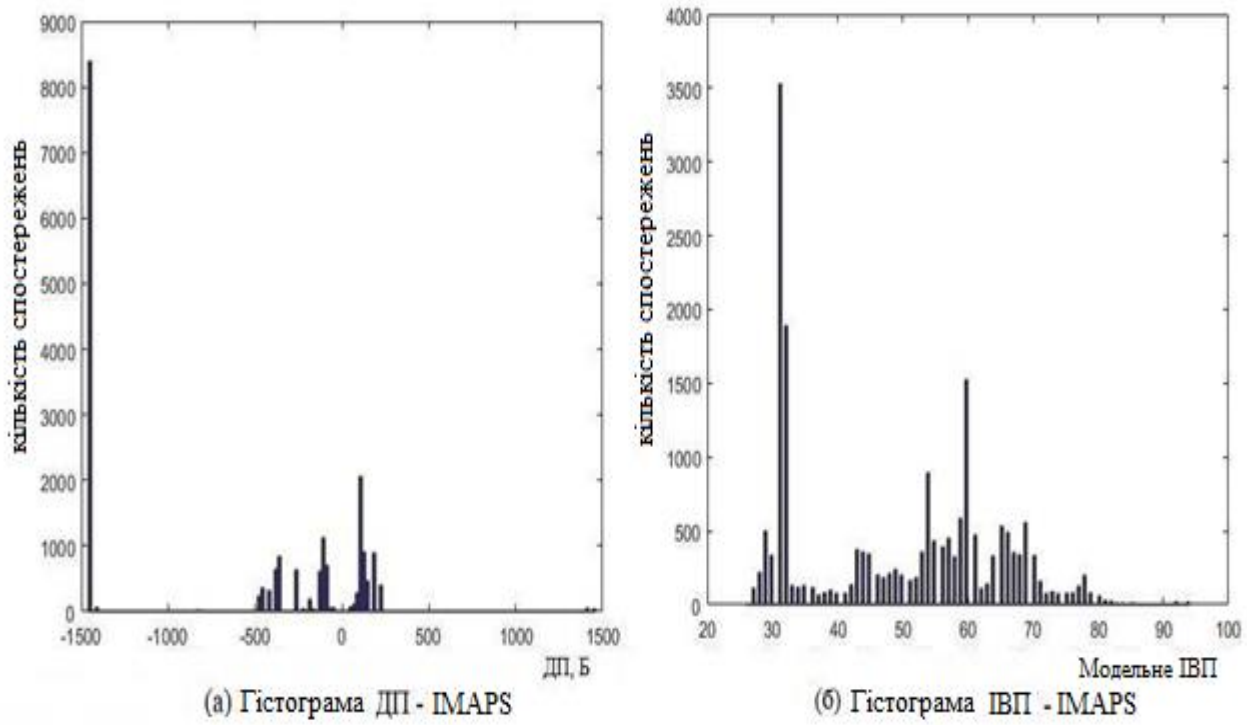

Рисунок 3.9 – Гістограми спостережень протоколу IMAPS

Для тренування й тестування запропонованої моделі створений тестової електронна адреса й використана додаток MS Outlook. При виконанні експерименту використані тільки активні сесії, оскільки пасивні сесії є дуже короткими (усього кілька пакетів) і непридатні для роботи алгоритмів.

Трафік однорангового додатка – однорангові додатки використовуються в різних областях, але основне їхнє використання полягає в забезпеченні – загального

доступу до файлів. Найвідомішим одноранговим додатком є додаток торент, що використовується для обміну різними типами даних, такими, як відео, програмне забезпечення, електронні книги й т.п. Такий тип додатків може використовуватися для обміну нелегальними даними з порушенням авторських прав, тому більшість Інтернет-Провайдерів забороняють їх, але відстежити трафік однорангових додатків (як ми вже відзначали в першому розділі) дуже складно.

Однорангові додатки мають розподілену архітектуру (рисунок 3.10). У ній вузли можуть спілкуватися один з одним безпосередньо, без використання якого-небудь центрального сервера.

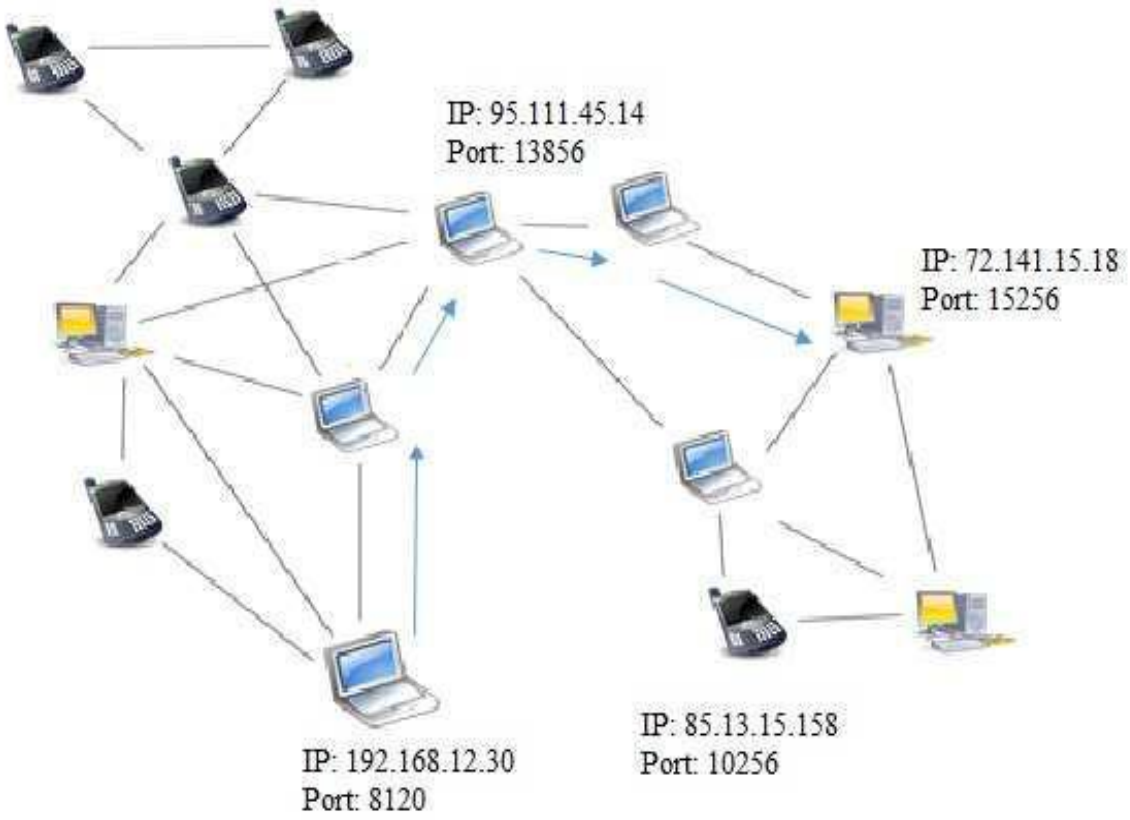

Рисунок 3.10 – Розподілена архітектура однорангового додатка

На рисунку 3.11 наведені гістограми спостережень ДП і ІЧП для протоколу P2P.

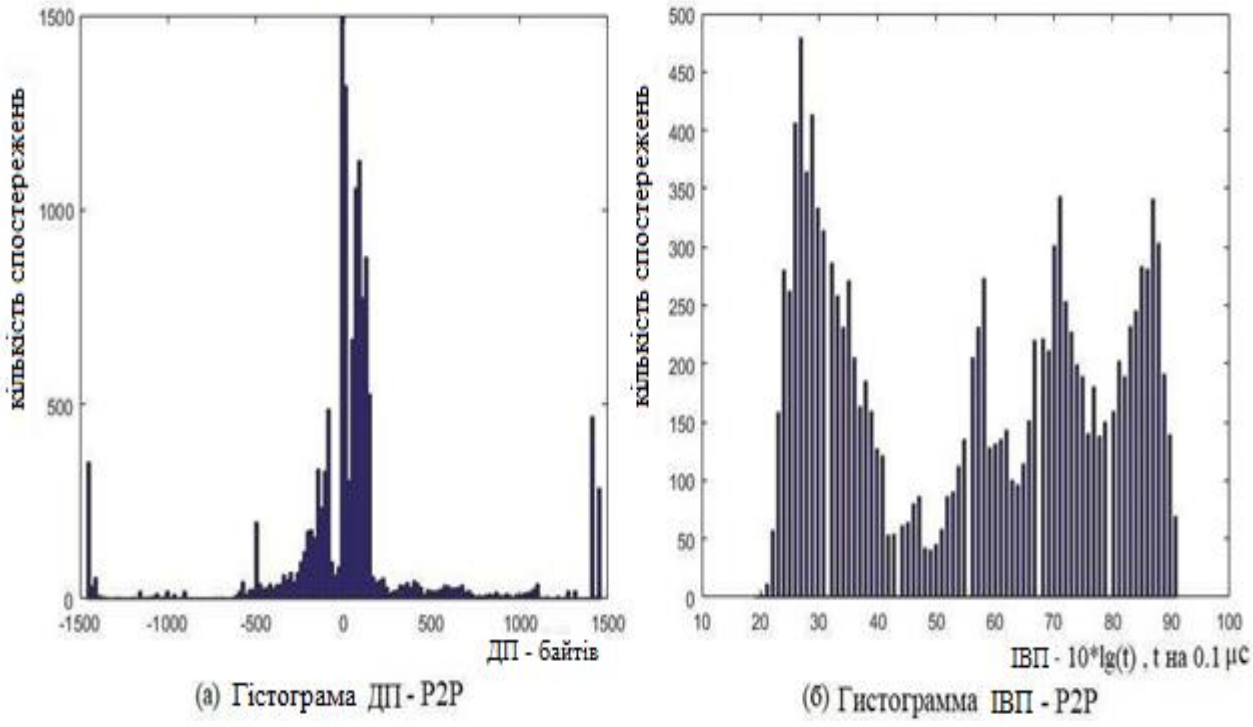

Рисунок 3.11 – Гістограми спостережень однорангового додатка P2P

Для дослідження однорангового додатка BitTorrent у даній роботі використаний набір трафіку університету Нью-Брансуика (Канада).

4 ОПИС РОЗРОБЛЕНОГО ПРОГРАМНОГО ПРОДУКТУ

4.1 Тестування розроблених алгоритмів

Трафік WhatsApp: чат-додаток WhatsApp є одним з найвідоміших чат-додатків. У наш час близько 1,5 мільярди людей використовують його, щодня через нього проходять більш 50 мільярдів повідомлень. WhatsApp має складну структуру, де чат є основним сервісом поряд з такими, як голосове повідомлення, голосовий і відео дзвінки, а також передача файлів (рисунок. 4.1).

WhatsApp використовує для різних пропонованих послуг порти 80 або 443, які пов'язані із протоколами HTTP і HTTPS. Оскільки весь трафік цього додатка зашифрований, розпізнавання трафіку цього додатка з використанням адрес IP і TCP-портів або вмісту пакетів неможливо. При запуску додаток відправляє пакет Dns-Запиту, щоб довідатися IP-адреса WhatsApp сервера, потім починається TCP-сесія керування сервером, після чого інформація й інструкції всіх нових установлених сесій передаються через цю першу сесію, із цієї причини виявити вкладені сесії практично неможливо.

Тестування трафіку для додатку TOR, що дозволяє користувачеві відвідувати веб-сайти без відстеження в брандмаурі [14].

Увесь трафік цього додатка передається між кінцями з'єднання через декілька агентів, які виконують додаткове шифрування (рисунок 4.4). Додаток TOR використовує Socks-протоколи для шифрування трафіку.

Ідентифікація трафіку додатка TOR з використанням адреси IP неможлива через те, що агенти TOR постійно змінюють використовувані TOR-порти (наприклад, можуть використовувати порт 443, який пов'язаний із протоколом HTTPS).

В цій роботі TOR-трафік згенерований з використанням власного TOR-браузера, а маркування трафіку виконане на основі TOR-порту, який був зафіксовано на значенні 9001.

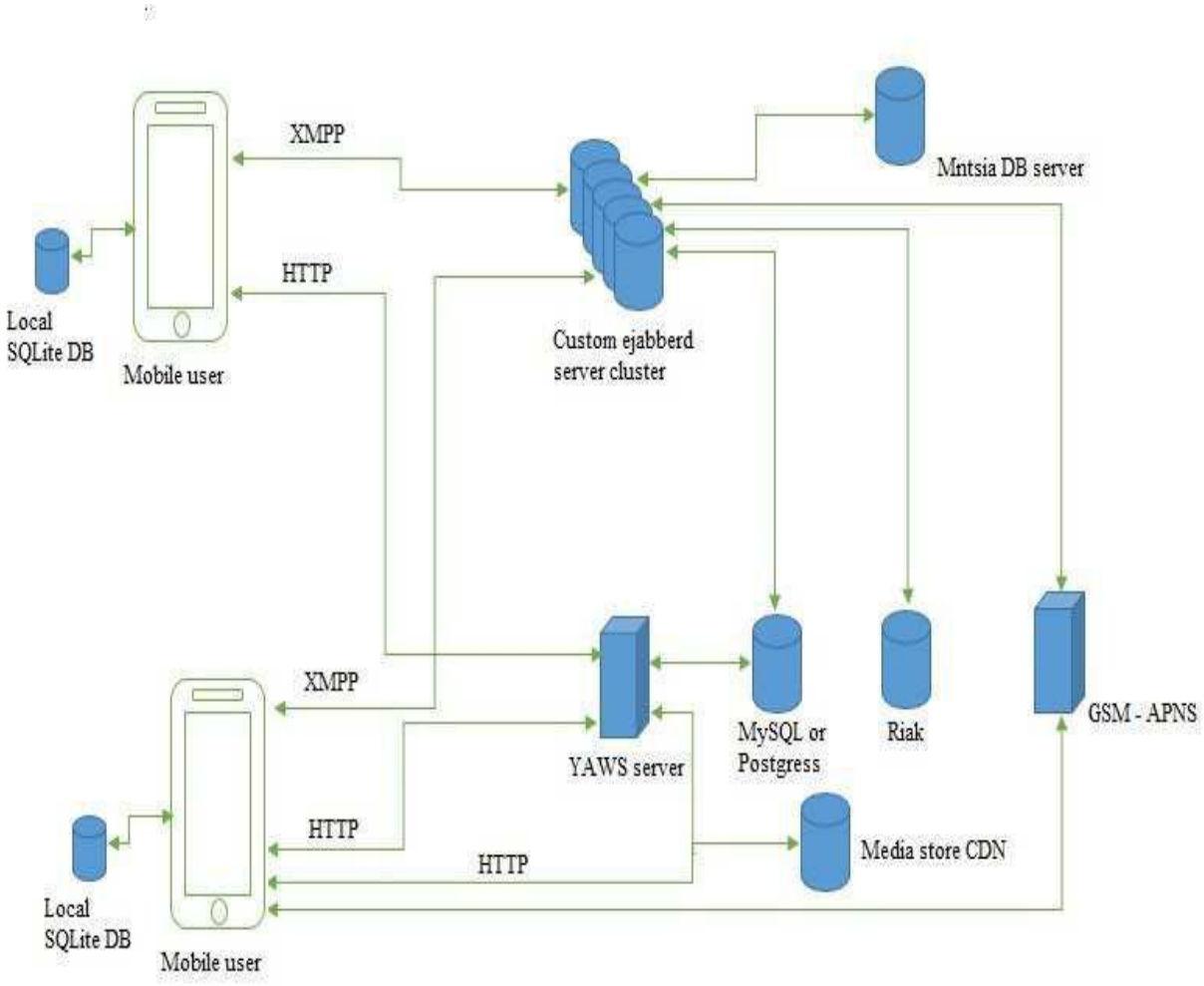

Рисунок 4.1 – Інфраструктура й сервіси додатку WhatsApp

Для виявлення сесій керування розроблені програми, що дозволяють записати пакети сесій трафіку додатка WhatsApp. На рисунку 4.2 наведені гістограми спостережень ДП і ІЧП для протоколу WhatsApp.

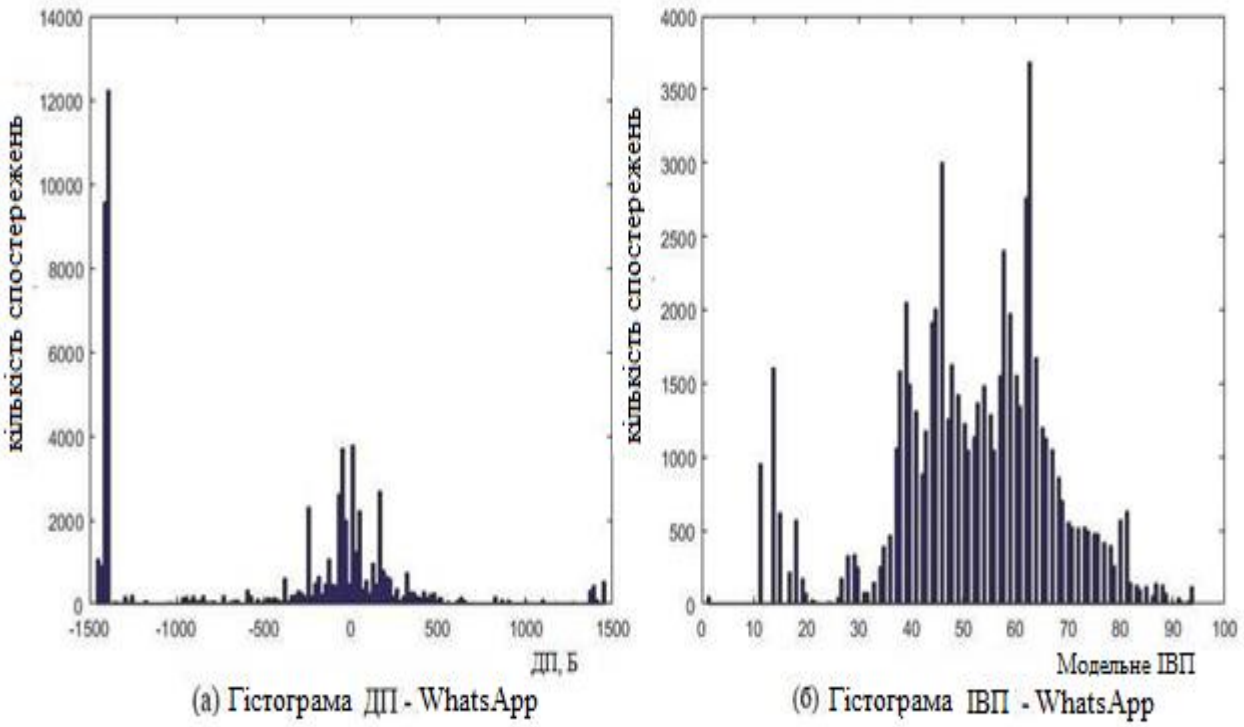

Рисунок 4.2 – Гістограми спостережень однорангового додатка WhatsApp

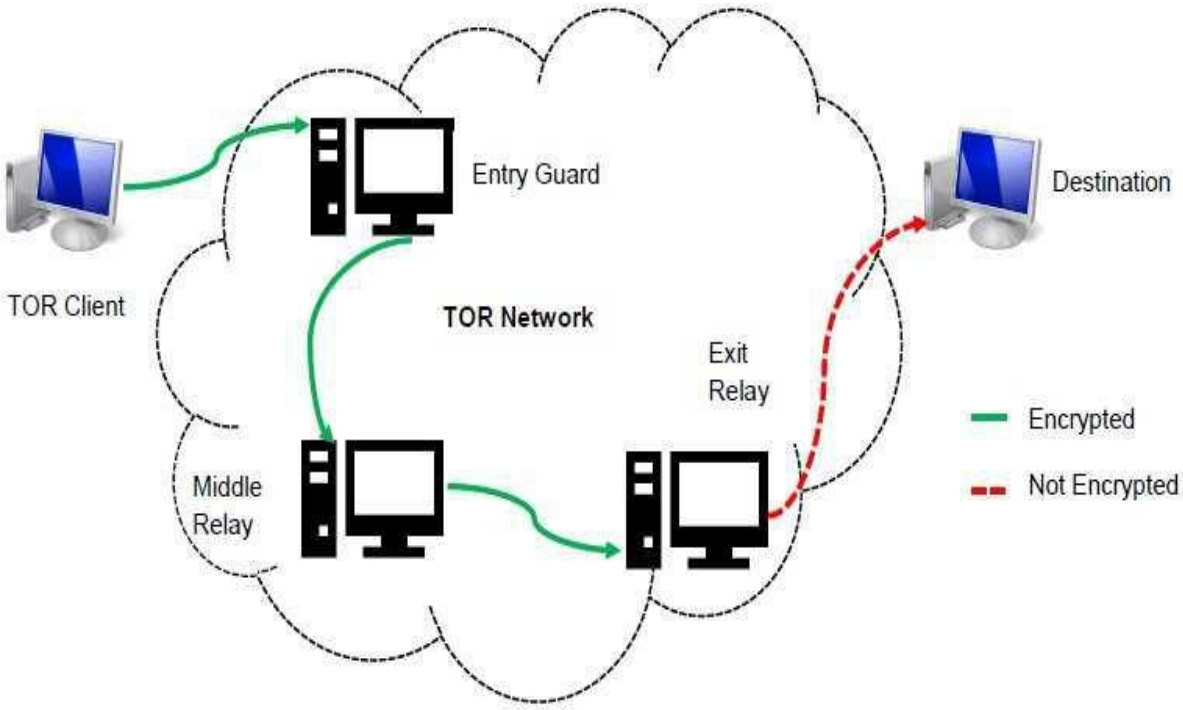

Рисунок 4.3 – Схема з'єднання з використанням додатка TOR

На рисунку 4.4 наведені гістограми спостережень ДП і ІЧП для протоколу TOR.

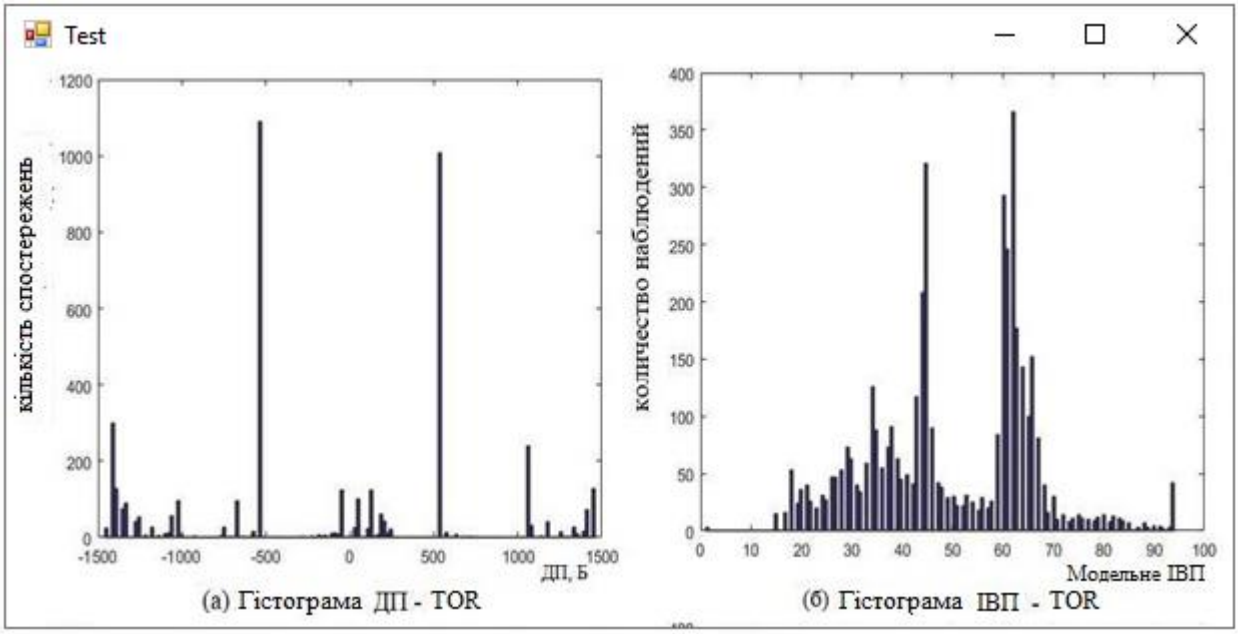

Рисунок 4.4 – Гістограми спостережень однорангового додатка TOR

Для оцінки якості моделі ідентифікації використані теж чотири показники, що відповідають рівнянням (1.3) – (1.6).

Таблиця 4.1 – Кількість потоків для навчання й тестування

Додаток	Кількість пакетів	Кількість потоків	Метод маркування
Безпечний веб (HTTPS)	267400	13267	ПЗ LibProtoIdent
Веб (HTTP)	252800	12670	ПЗ LibProtoIdent
Tor	130600	6350	Номер порту
IMAPS	85040	4252	Номер порту
P2P	160200	8010	ПЗ LibProtoIdent
WhatsApp	40700	2035	Записи DNS

На рисунку 4.5 наведено результати впливу числа пакетів на точність ідентифікації трафіку в запропонованій моделі.

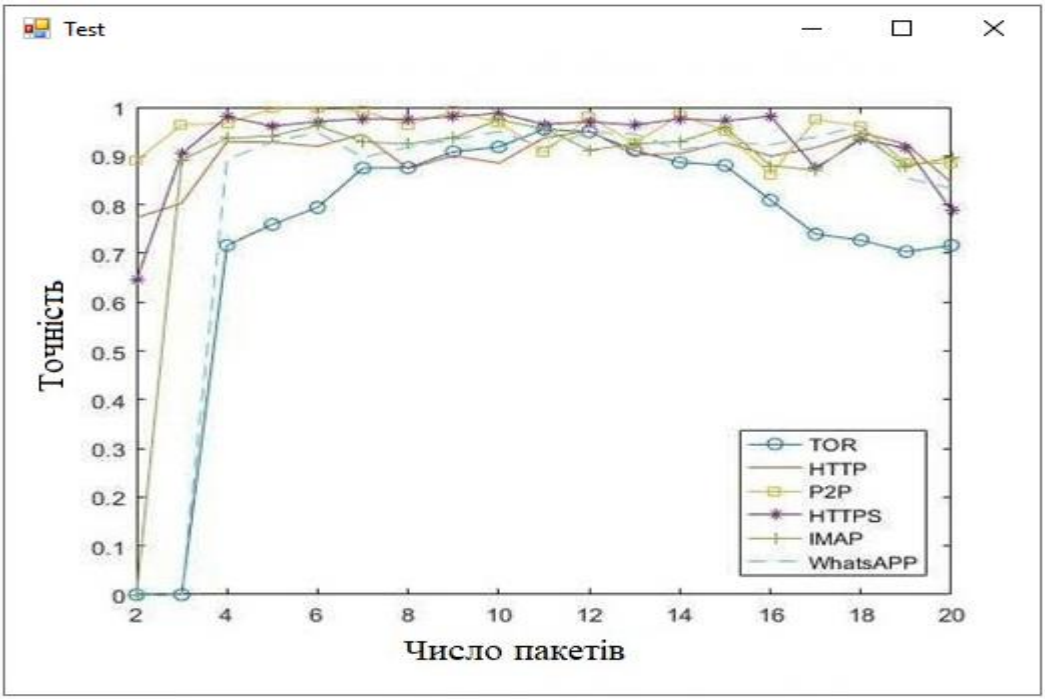

Рисунок 4.5 – Вплив числа пакетів на точність ідентифікації додатків у моделі

З рисунка випливає, що оптимальне число пакетів для ідентифікації всіх додатків з точністю не менш 90% лежить в інтервалі між 9 та 14 пакетами.

На рисунку 4.6 наведені значення показника повноти ідентифікації додатків, який показує відсоток ідентифікованого трафіку додатка стосовно всього трафіку додатка, залежно від числа пакетів. З рисунка 4.6 випливає, що повнота ідентифікації всіх додатків більше 80% при використанні числа пакетів від 4 і до 15, а при кількості пакетів від 8 до 10 повнота ідентифікації трафіку більше 90% для всіх додатків. При використанні більш 15 пакетів повнота ідентифікації стає менше через вплив пакетів ініціалізації додатка, які мають певні довжини й часовий порядок (наприклад, згідно зі стандартами роботи сервісу, протокол HTTPS починається із трьох пакетів по стандартах SSL і TLS).

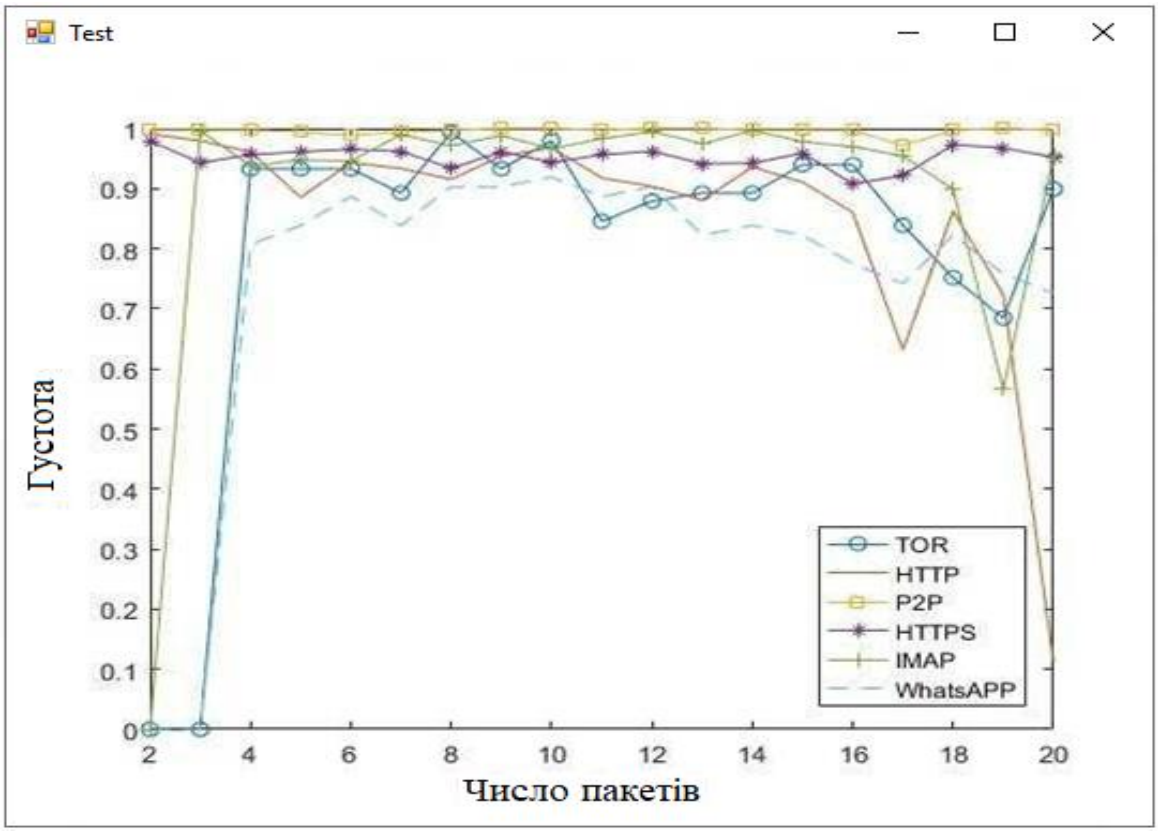

Рисунок 4.6 – Екранна форма, що відображає вплив числа пакетів на повноту ідентифікації додатків

Повнота ідентифікації додатків також зменшується через високий ступінь схожості мережних пакетів між собою. Наприклад, відмінність між трафікуми HTTP і HTTPS стає практично нерозрізнене після 4 й 5 пакетів через те, що трафік HTTPS фактично є шифрованим трафіком HTTP. Вони стають дуже схожими між собою, оскільки шифрування мале впливає на спостережувані параметри (довжини пакетів і час між пакетами).

На рисунку 4.7 наведені результати оцінки загальної валідності ідентифікації трафіку для кожного досліджуваного додатка.

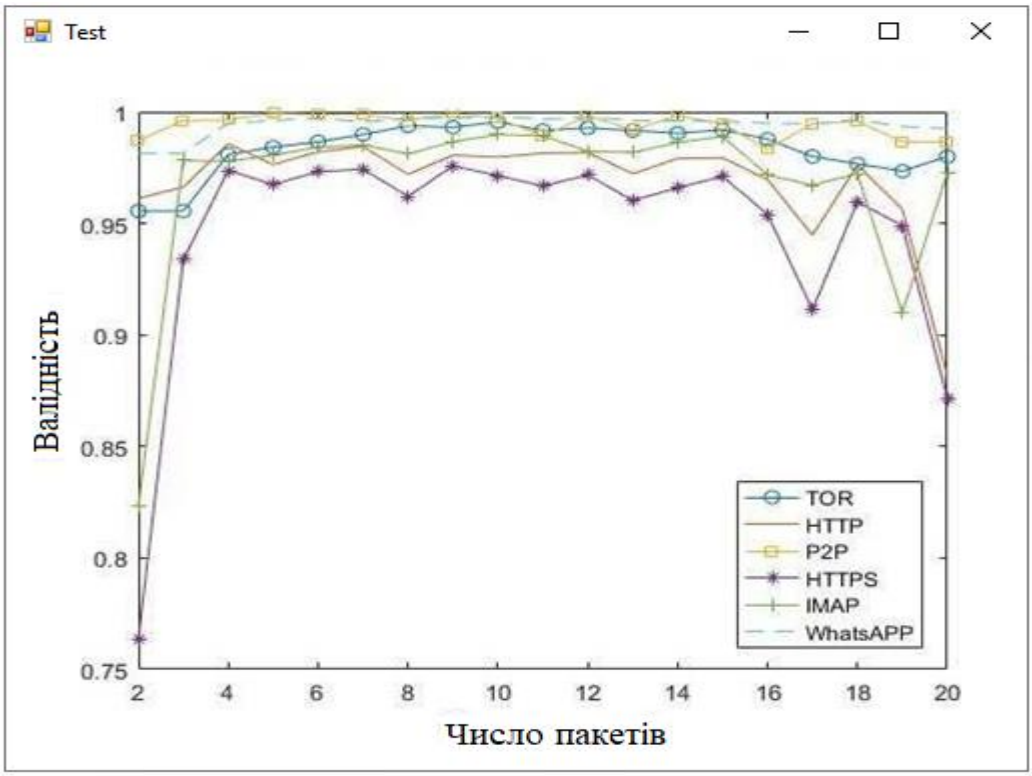

Рисунок 4.7 – Екранна форма «Вплив числа пакетів на валідність ідентифікації додатків у моделі»

З рисунка 4. 8 випливає, що валідність ідентифікації трафіку більш 95% для всіх досліджуваних додатків при використанні від 4 до 15 пакетів.

4.2 Розрахунок залежності між трафіком

Для виявлення залежності між трафіком HTTP і HTTPS в роботі була обчислена функція кореляції значень спостережень цих додатків (рисуок 4.8)

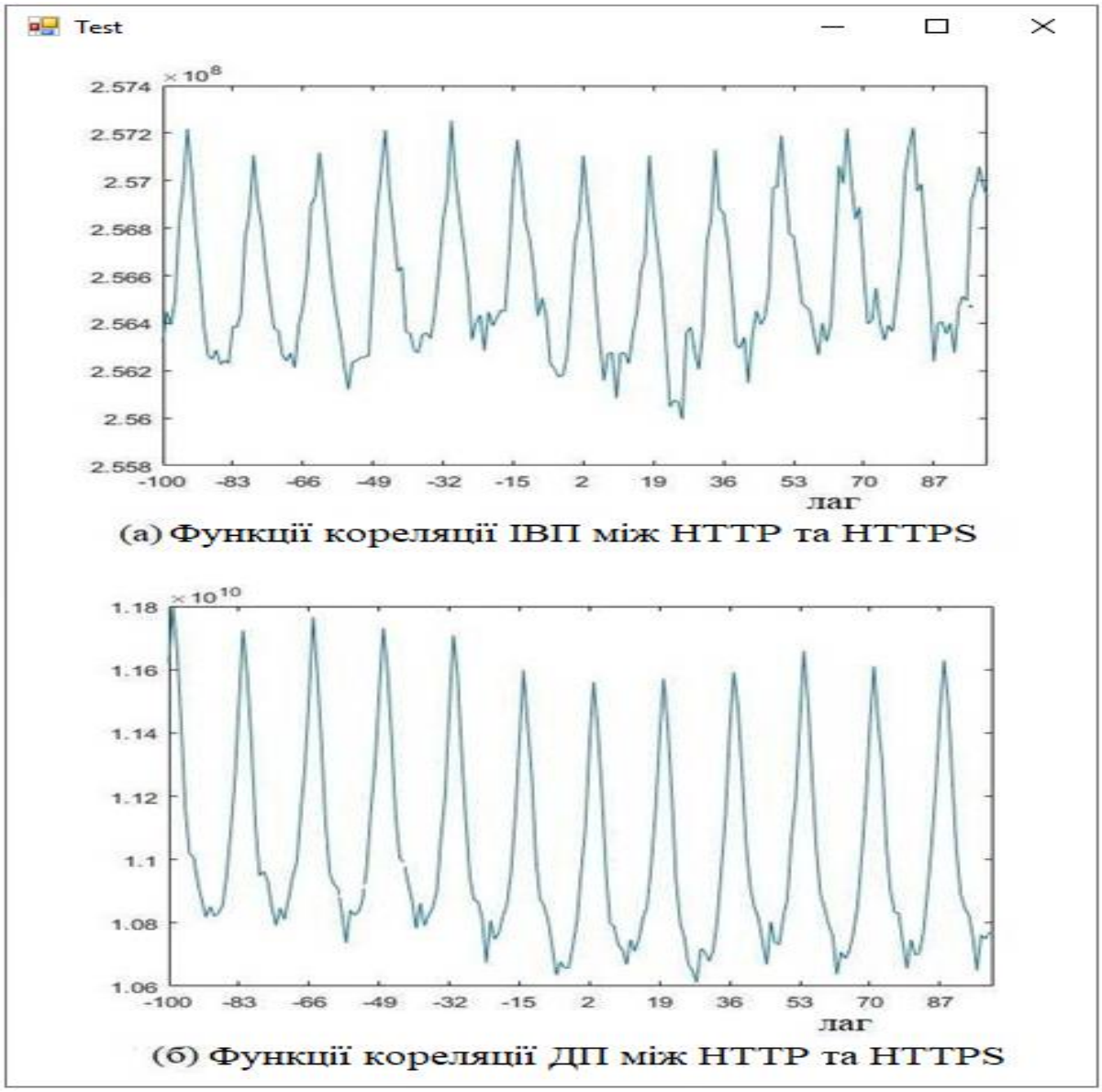

Рисунок 4.8 – Функції кореляції (а) значень часу між пакетами HTTP і HTTPS; (б) значень довжини пакетів HTTP і HTTPS

Значення перших трьох пакетів з потоків HTTPS, побудованих по протоколу SSL, не розглядаються. Не розглядаються також останні три пакети з потоків HTTP, щоб одержати однакове число значень спостережень для цих двох додатків. Таким чином, у результаті використано 17 пакетів з кожного потоку.

З рисунку 4.8 випливає, що існує висока кореляція між трафікуми додатків HTTP і HTTPS, де повторюються максимальні величини функції кореляції для

кожних 17 пакетів (довжина потоку) і ці величини мало змінюються. Проведений також аналіз кореляції між трафікуми HTTPS і TOR, результати якого показано на рисунку 4.9.

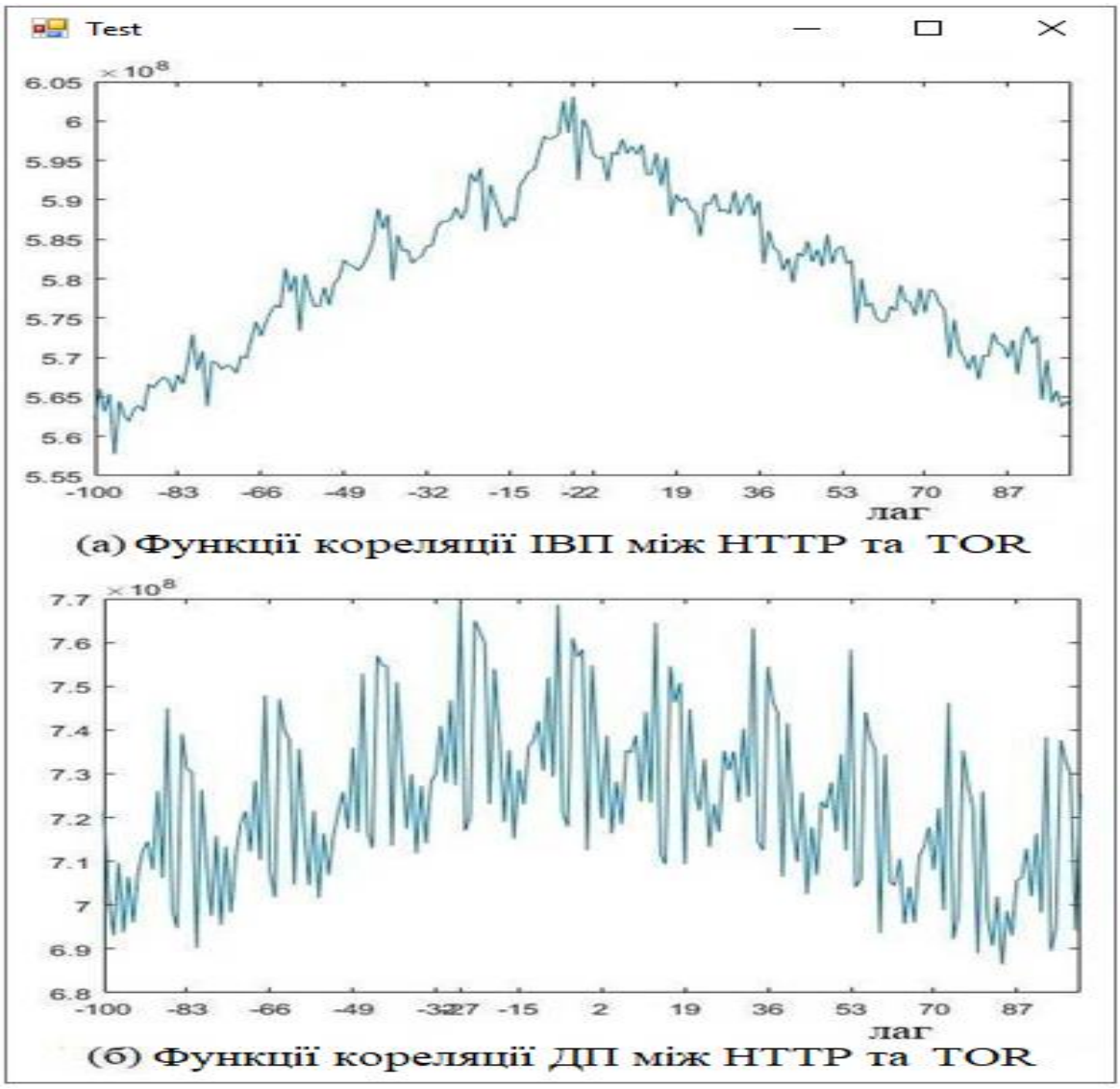

Рисунок 4.9 – Функції кореляції (а) значень часу між пакетами HTTPS і TOR; (б) значень довжини пакетів HTTPS і TOR

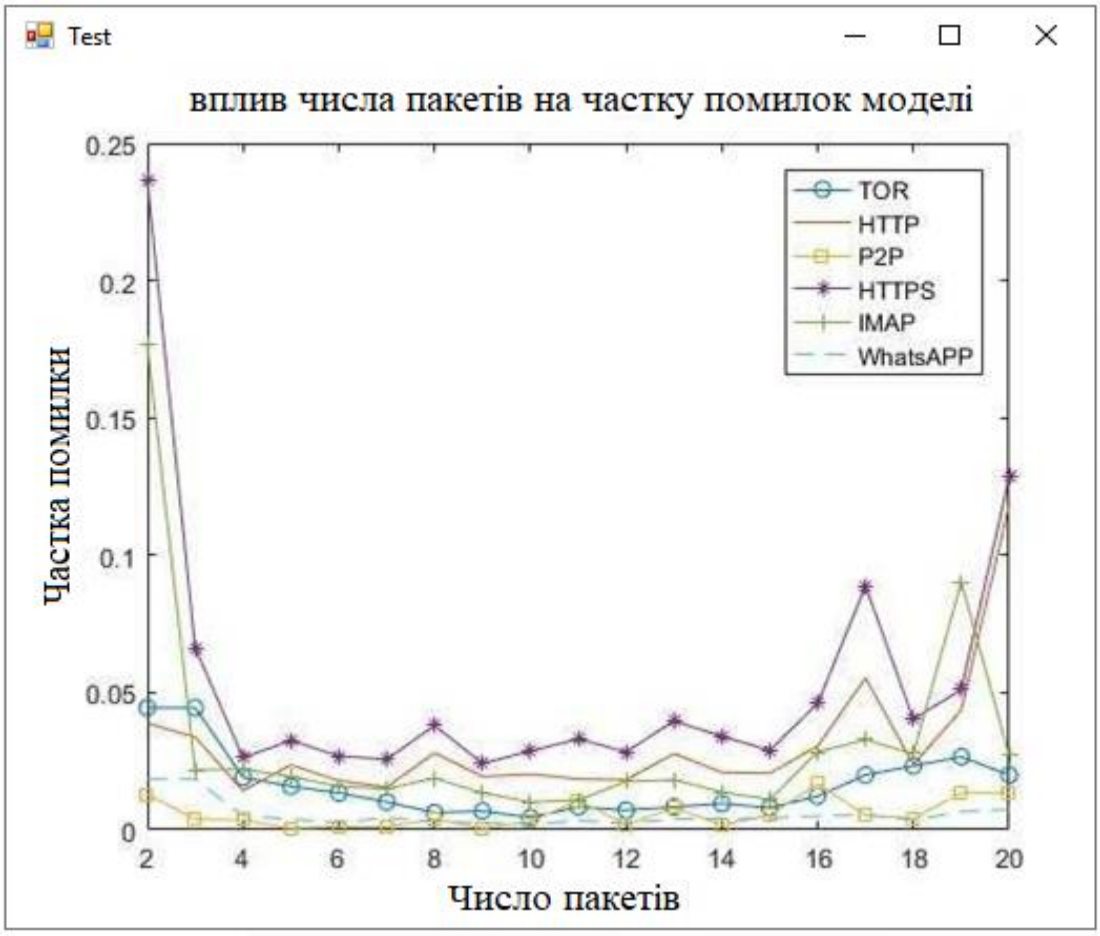

Рисунок 4.10 – Вплив числа пакетів на частка помилок ідентифікації додатків у моделі

В результаті застосування тестового програмного забезпечення випливає, що кореляція між трафіком HTTPS і TOR є дуже слабкої, хоча є локальні максимуми, які повторюються в певному порядку. У загальному кореляція сильно зменшується при видаленні від центру. За аналогією одержуємо, що залежність між трафіком різних додатків є слабкої, крім залежності між протоколами HTTP і HTTPS.

Також, виявлено вплив числа пакетів на частку помилок при ідентифікації досліджуваних додатків з використанням запропонованої моделі – частка помилок ідентифікації всіх додатків менше 5% при використанні відрізків трафіку від 4 до 14 пакетів. Очевидно, що помилки ідентифікації HTTP і HTTPS корелюються один з одним.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Для одержання коректного порівняння необхідно використовувати для аналізу отриманих і відомих методів ідентифікації однакові набори даних.

Результати ідентифікації досліджуваних додатків показано в таблиці 5.1. Спостережувані значення збігаються із середньою довжиною пакета й із середнім часом між пакетами. Проаналізовано використання 9 протоколів (aim, smtp-out, smtp-in, http, https, ftp-data, ftp, ssh, telnel).

Таблиця 5.1 – Точність ідентифікації протоколів

	Вихідний класифікатор (відсотки)									
Протокол	aim	smtpout	smtpin	http	https	ftpdata	ftp	ssh	telnet	none
aim	81.6	1.9	0.9	0.4	2.5	0.5	0.5	3.4	2.7	5.6
smtp-out	4	65.6	12.7	0	0.6	0.3	7.2	0.8	4.4	4.4
smtp-in	1.1	11.4	70.2	0	0.3	0.2	2	0.6	11.9	2.3
http	0.2	0.1	0.1	81.1	6.2	7.5	0.1	0.6	0.8	3.6
https	0.5	4.1	0.2	2.8	76.5	1.9	0.1	2.2	3.2	8.5
ftp-data	2.1	4.4	5.1	12.1	4	62.7	0.2	0.4	5.9	3.1
ftp	1	22.7	2.9	0.1	0	2	62.7	3.9	4.7	0.0
ssh	7	1.4	0.7	1.1	14.7	0.4	2.5	42.0	1.9	28.1
telnet	2.6	4.1	1	4.4	8	4.7	2.4	5.4	42.0	25.5

З аналізу результатів таблиці 5.1 випливає, що є тільки два загальні протоколи (HTTP і HTTPS), які можна використовувати для порівняння роботи Райта із роботою запропонованого тестового ПЗ. На рис 5.1 наведені результати запропонованої ідентифікації додатків з використанням 10 пакетів. Це число пакетів є оптимальним для ідентифікації додатків згідно з показниками оцінки якості моделі. У запропонованій роботі, точність ідентифікації протоколу HTTP рівна 91.5%, а в

роботі Райта – 81.1%, точність ідентифікації протоколу HTTPS у запропонованій роботі рівна 93.4%, а в роботі Райта – тільки 76.5%. Очевидно, що отримані в дисертації результати вище, чим отримані в роботі Райта при ідентифікації HTTP і HTTPS. При цьому кількість використаних пакетів для ідентифікації додатків у методі Райта набагато більше кількості використаних пакетів у запропонованій моделі.

Test

	Вихідний класифікатор (відсотки)						
Протокол	**TOR**	**HTTP**	**P2P**	**HTTPS**	**IMAPS**	**WhatsApp**	**None**
TOR	99.33	0.67	0.00	0.00	0.00	0.00	0.00
HTTP	**1.60**	**91.53**	**2.29**	**5.49**	**0.00**	**0.92**	**0.00**
P2P	0.00	0.00	99.71	0.29	0.00	0.00	0.00
HTTPS	**0.00**	**3.38**	**0.00**	**93.38**	**3.24**	**0.00**	**0.00**
IMAPS	1.86	0.34	0.34	0.68	97.13	0.00	0.00
WhatsApp	3.23	2.68	0.00	4.45	0.00	90.32	0.00

Рисунок 5.1 – Екранна форма «Точність ідентифікації досліджуваних додатків»

Було використовано два спостережувані значення: довжина пакета та час між пакетами. Число використаних пакетів для навчання й тестування методу становить 100 пакетів в одному потоці, що є набагато більшим, ніж число використаних пакетів у запропонованій роботі.

На рис. 5.2 наведено результати ідентифікації досліджуваних додатків, де відсоток правильної ідентифікації протоколу HTTP – 95.36% і додатка P2P – 97.51%, що є показниками, які дуже близькі до результатів запропонованого методу, де відсоток вірної ідентифікації протоколу P2P рівний 99.71%, а протоколу HTTP – 91.53%.

	Вихідний класифікатор (відсотки)					
Тип	**VIDEO**	**GAME**	**HTTP**	**IM**	**SMTP**	**P2P**
VIDEO	95.1	0	2.07	2.83	0	0.000
GAME	0	95.03	0	4.27	0.7	0.000
HTTP	**0.56**	**1**	**95.36**	**1.4**	**1.68**	**0.000**
IM	0.38	0.38	0.19	98.67	0.38	0.000
SMTP	0.49	5.99	3.18	4.16	82.18	4.010
P2P	**2.49**	**0**	**0**	**0**	**0**	**97.510**

Рисунок 5.2 – Екранна форма «Точність ідентифікації досліджуваних додатків»

Результати порівняння запропонованої роботи з існуючими роботами показує, що запропонована робота ідентифікує різні типи додатків з високою часткою ймовірності й відрізняється від існуючих результатів числом використовуваних для ідентифікації пакетів. У даній роботі показане, що 10 пакетів в авторській моделі досить для ідентифікації досліджуваних додатків, що є безсумнівною її перевагою (в інших же роботах це число набагато більше).

Таким чином, проведені експериментальні дослідження розроблених алгоритмів, що показали, що вони ідентифікує різні типи додатків з показниками: точність більш 90%, повнота більш 90%, загальна валідність ідентифікації більш 95% і частка помилок менш 5% при обмеженім числі пакетів (10 пакетів).

Проведений порівняльний аналіз із існуючими розробками, показав перевагу розроблених алгоритмів і можливість їх ефективного використання при ідентифікації трафіку СПД на рівні додатків в умовах реального часу.

ВИСНОВКИ

В результаті виконання магістерської атестаційної роботи вирішене й успішне апробоване науково-прикладне завдання, що полягає в розробці ефективних алгоритмів ідентифікації додатків мереж передачі даних і ідентифікації типу додатка в тунелі.

За результатами порівняльного аналізу існуючих моделей мережного трафіку запропонована нова модель, заснована на схованої марківської моделі й по своїх параметрах потенційно орієнтована на її використання у високоточних завданнях ідентифікації трафіку СПД у реальному часі.

Розроблена формальна постановка завдання ідентифікації, рішення якої дозволило сформулювати вимоги до алгоритмів ідентифікації, включаючи обов'язкову можливість класифікації відомих додатків, працездатності ідентифікації трафіку в умовах високошвидкісної передачі даних у реальному часі, збереження конфіденційності користувачів і забезпечення необхідної точності ідентифікації.

Проаналізовано метод ідентифікації трафіку, що задовольняє сформульованим вимогам, що підтвердив можливість і доцільність створення нового інструмента рішення завдання ідентифікації на основі класифікації трафіку з використанням статистичного аналізу, СММ і ітераційної процедури Баума-Велша для обчислення її параметрів.

Розроблений алгоритм ініціалізації СММ із використанням моделі гауссової суміші, що забезпечує оптимальну збіжність процедури Баума-Велша в рамках необхідного якості ідентифікації.

Застосування цього алгоритму в загальному алгоритмі ідентифікації дозволило виконати її в реальному часі з точністю більш 90%, повнотою більш 80%, загальної валідності більш 95% і часток помилок менш 5% при обмеженім числі пакетів (10 пакетів).

Розроблено тестове ПЗ підготовки наборів даних на основі реального й модельного мережного трафіку на етапах навчання й тестування запропонованої моделі, що складає основу її функціонування й експериментального дослідження.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Суперкомпьютерные технологии в науке, образовании и промышленности / Под ред. В. А. Садовничего, Г. И. Савина, чл.-корр. РАН В. В. Воеводина. – М. : МГУ, 2012. – 232 с.

2. Bailey, D. H. High-Precision Arithmetic: Progress and Challenges / D. H. Bailey, J. M. Borwein. – Electronic text data. – 2013. – 15 p. – Mode of access: http://www.davidhbailey.com/dhbpapers/hp-arith.pdf. – The title from the screen.

3. Bailey D. H. High-Precision Computation: Mathematical Physics and Dynamics / D. H. Bailey, R. Barrio, J. M. Borwein // Applied Mathematics and Computation. – 2012. – Vol. 218, Issue 20. – P. 10106–10121.

4. Robey R. W. In Search of Numerical Consistency in Parallel Programming / R. W. Robey, J. M. Robey, R. Aulwes // Parallel Computing. – 2011. – Vol. 37, Issue 4–5. – P. 217–229.

5. Толстых, С. С., Подольский В. Е. Оценка сложности крупноблочных облачных вычислений, использующих арифметику повышенной точности / С. С. Толстых, В. Е. Подольский // Труды ИСП РАН, том 26. – 2014. – № 5. – С. 29-64.

6. Complexity evaluation of the large-block parallel computing tasks using precision-trust arithmetics / A. M. Babichev, V. E. Podolskiy, S. S. Tolstyh, S. G. Tolstyh // Proceedings of the International Multidisciplinary Scientific GeoConferences SGEM. – Albena Resort, Bulgaria, 2016. – P. 141-148.

7. Воеводин В. В. Вычислительная математика и структура алгоритмов [Текст] / В. В. Воеводин. – М. : Изд-во Московского ун-та, 2010. – 168 с. – (Серия «Суперкомпьютерное образование»).

8. Варениця В. В. Проблемы вычисления метрик сложности программного обеспечения при проведении аудита безопасности кода методом ручного

рецензирования / В. В. Вареница // Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». – 2011. – № SPEC. – P. 79-84.

9. Fenton N. E. Software Metrics: A Rigorous and Practical Approach / N. E. Fenton, S. L. Pfleeger // International Thomson Computer Press, 2nd ed. PWS Publishing, 1996. – P. 36-39.

10. Липаев В. В. Обеспечение качества программных средств. Методы и стандарты / В. В. Липаев. – М.: Синтег, 2001. – 380 с.

11. Марков А. С. Оценка динамической сложности программного обеспечения на ПЭВМ / А. С. Марков // Методы и средства совершенствования сложных управляющих систем и комплексов. - СПб.: БГТУ им. Д.Ф. Устинова (Военмех), 1992. – С. 14-27.

12. Современные проблемы вычислительной математики и математического моделирования. В 2 т. Т. 2. Математическое моделирование / отв. ред. В. П. Дымников. – М. : Наука, 2005. – 405 с.

13. Котов В. Е., Сети Петри / В. Е. Котов – М.: Наука, 1984. – 160 с.

14. Питерсон Дж. Теория сетей Петри и моделирование систем / Дж. Питерсон ; пер. с англ. – М.: Мир, 1984. – 264 с.

15. Федотов И. Е. Некоторые приёмы параллельного программирования. Учебное пособие / И. Е. Федотов. – М.: Изд-во МГИРЭА(ТУ), 2008. - 188 с.

16. Bailey D. H. Experimental Mathematics: Examples, Methods and Implications / D. H. Bailey, J. M. Borwein // Notices of the AMS. – 2005. – Vol. 52. – P. 502–514.

17. Анализ вычислительной сложности крупноблочного итерационного процесса в условиях повышенной точности арифметических операций с плавающей точкой / С. С. Толстых, С. В.Мищенко, В. Е. Подольский, С. Г. Толстых // Научно-методический журнал "Информатизация образования и науки". – 2016. – №2(30). – С. 159-169.

18. Hurwitz J. What Is Platform as a Service (PaaS) in Cloud Computing? / J. Hurwitz, M. Kaufman, F. Halper, D. Kirsh // Hybrid Cloud For Dummies, Hoboken. – NJ: John Wiley & Sons, 2012. – 360 p.

19. Thomas E. B. Computer Concepts and Terminology: Types of Computers / E. B. Thomas. – Electronic text data. – Los Alamos : University of New Mexico, 2012. – Mode of access: http://www.unm.edu/~tbeach/terms/ types.html. – The title from the screen.

20. Howard A. R. The Surprising Technology Economics of Mainframe vs. Distributed Servers: Understanding the Impact of Your Strategy in Real Business Terms [Electronic resource] / A. R. Howard. – Electronic text data. – US : Rubin Worldwide, 2011. – 7 p. – Mode of access: http://www-03.ibm.com/systems/es/ resources/ZSL03135USEN.pdf. – The title from the screen.

21. Chapman B. Using OpenMP: portable shared memory parallel programming Scientific and Engineering Computation / B. Chapman, G. Jost, Ruud van der Pas // Cambridge, Massachusetts: The MIT Press., 2008. - 353 pp.

22. Характеристики качества программного обеспечения [Текст] / Б. Боэм, Дж. Браун, Х. Каспар, М. Липов, Г. Мак-Леод, М. Мерит ; пер. с англ. – М.: Мир, 1981. – 208 с.

23. Пышкин Е. В. Структурное проектирование: основание и развитие методов. С примерами на языке С++: Учеб. Пособие. / Е. В. Пышкин. – СПб.: Политехнический университет, 2005. – 324 с.

24. Брукс П. Метрики для управления ИТ-услугами / П. Брукс – М.: Альпина Бизнес Брукс, 2008. – 25-31, 49-58, 99-115 с.

25. Рыжков, Е. А. Программный код и его метрики / Е. А. Рыжков. – 2010. – Режим доступа: https://habrahabr.ru/company/intel/blog/ 106082/. СПб.: ГУАП, 2000. – 210 с.

26. Изосимов А. В. Метрическая оценка качества программ [Текст] / А. В. Изосимов, А. Л. Рыжко. – М.: Изд. МАИ, 1989. – 96 с.

27. Кулаков А. Ю. Оценка качества программ ЭВМ / А. Ю. Кулаков. – Киев: Техніка, 1984. – 167с.

28. Холстед М. Х. Начало науки о программах / М. Х. Холстед. – М.: Финансы и Статистика, 1981. – 128 с.

29. Маевский Д. А. Оценка количества дефектов программного обеспечения на основе метрик сложности / Д. А. Маевский, С. А. Яремчук // Електротехнічні та комп'ютерні системи. – 2012. – № 7. – С. 113-120.