https://www.sworldjournal.com/index.php/swj/article/view/swj35-01-100
DOI: 10.30888/2663-5712.2026-35-01-100

УДК 004.75

MODELS AND ALGORITHMS FOR IDENTIFYING RELATIONSHIPS IN TUNNEL TRAFFIC

МОДЕЛІ ТА АЛГОРИТМИ ІДЕНТИФІКАЦІЇ ЗАСТОСУНКІВ У ТУНЕЛЬНОМУ ТРАФІКУ

Klyushnik Vitaliy / Клюшник В.В.
postgraduate / аспірант
ORCID:0009-0004-2134-4248
Chernetsky Evgeny / Чернецький Є. В.
PhD, Associate Professor / к.т.н., доцент
ORCID:0000-0002-4197-7171
Ukrainian State University of Science and Technologies
УДУНТ ННІ УДХТУ, м. Дніпро, пр.Науки, 849107

Анотація. *У статті розглянуто задачу ідентифікації застосунків у тунельному трафіку в умовах широкого використання шифрування, VPN-технологій та анонімних мереж. Запропоновано модель ідентифікації на основі прихованих марковських моделей, що використовує статистичні характеристики потоків даних, зокрема довжини пакетів і часові інтервали між ними. Описано структуру математичної моделі, алгоритми навчання та класифікації, а також програмну реалізацію запропонованого підходу. Проведено експериментальне тестування з використанням реальних наборів мережевого трафіку, яке продемонструвало високу точність і стійкість методу навіть за умов обмеженого обсягу даних та повного шифрування трафіку. Отримані результати підтверджують можливість практичного застосування моделі в системах моніторингу, управління мережею та забезпечення інформаційної безпеки.*

Ключові слова: *тунельний трафік, ідентифікація застосунків, приховані марковські моделі, зашифрований трафік, аналіз мережевого трафіку, VPN, машинне навчання.*

Вступ.

Однією з найскладніших проблем сучасних мереж передачі даних є ідентифікація тунельного трафіку. Поширення технологій віртуальних приватних мереж (VPN), шифрованих протоколів (SSL/TLS) та анонімних сервісів (TOR) призвело до того, що значна частина інтернет-комунікацій приховується від традиційних методів аналізу. Це створює серйозні труднощі для адміністраторів мереж і систем інформаційної безпеки, адже в тунелях може передаватися як легітимний, так і небажаний або навіть шкідливий трафік [1].

Класичні підходи, зокрема аналіз номерів портів чи глибока інспекція пакетів (DPI), виявилися малоефективними в умовах масового шифрування. Більше того, використання DPI вимагає значних обчислювальних ресурсів і не забезпечує результатів, коли вміст пакетів повністю зашифрований. У зв’язку з

цим на перший план виходять методи, що ґрунтуються на статистичних та поведінкових характеристиках потоків даних, які зберігають свою інформативність навіть у зашифрованому середовищі [2].

Особливе місце серед таких підходів займає використання прихованих марковських моделей (ПММ). Цей інструмент дозволяє ефективно описувати часові залежності у послідовностях пакетів, що робить його придатним для розв'язання задачі ідентифікації застосунків у тунельному трафіку. Поєднання ПММ з алгоритмами машинного навчання створює основу для побудови гнучких і надійних моделей, здатних розрізняти типи застосунків навіть при мінімальному обсязі вхідних даних [3].

У межах дослідження було розроблено математичну модель ідентифікації застосунків у тунельному трафіку, визначено ключові параметри для аналізу, створено програмну реалізацію алгоритмів і проведено експериментальне тестування з використанням реальних наборів даних.

Мета цієї статті полягає у представленні структури розробленої моделі, описі алгоритмів і параметрів, що використовуються, а також у демонстрації результатів тестування, які підтверджують ефективність підходу.

Модель ідентифікації застосунків у тунельному трафіку. Задача ідентифікації трафіку в умовах тунелювання має низку особливостей. По-перше, інформаційне наповнення пакетів приховане, тому аналіз здійснюється виключно на основі статистичних характеристик. По-друге, трафік у тунелі може бути гетерогенним: одночасно можуть передаватися пакети від різних застосунків, що значно ускладнює розпізнавання (рисунок 1).

У дослідженні сформовано математичну модель ідентифікації, яка описує поведінку потоку даних за допомогою прихованої марковської моделі (ПММ). Основна ідея полягає в тому, що для кожного застосунку будується окрема модель, яка відображає характерні властивості його трафіку. Далі для нового потоку визначається ймовірність належності до кожної з моделей, після чого приймається рішення про класифікацію.

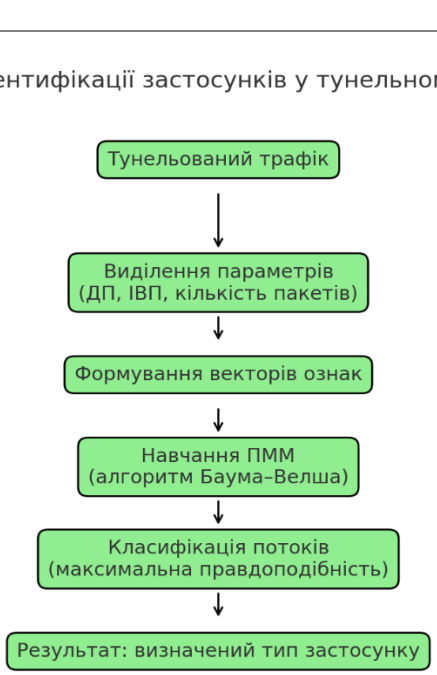

Рисунок 1 - Модель ідентифікації застосунків у тунельному графіку

Модель складається з таких елементів:

- набір станів, що відповідають прихованим характеристикам поведінки трафіку;
- матриця ймовірностей переходів між станами, яка відображає послідовність змін параметрів у часі;
- функції емісії, які задають розподіл спостережуваних величин (довжини пакетів, інтервалів часу тощо);
- початкові ймовірності станів, що відображають імовірний початок сесії.

У якості спостережуваних параметрів у моделі використовуються довжини пакетів (ДП) та інтервали між ними (ІВП). Ці характеристики обрані через їхню стійкість до шифрування і здатність відображати специфіку роботи застосунку. Математично задача ідентифікації формулюється як задача максимізації правдоподібності: для кожного потоку обчислюється ймовірність, що він згенерований певною моделлю. Клас застосунку визначається за принципом максимальної правдоподібності [4].

Таким чином, розроблена модель забезпечує можливість віднесення потоків до відповідних застосунків навіть за умови повної відсутності доступу до вмісту

пакетів. Це створює основу для розробки практичних алгоритмів аналізу тунельованого трафіку.

Параметри моделі. Якість роботи моделі ідентифікації безпосередньо залежить від правильно обраних характеристик, які описують поведінку трафіку. Оскільки в умовах тунелювання зміст пакетів є недоступним, акцент робиться на статистичних параметрах, що зберігають свою інформативність навіть при шифруванні [5].

У розробленій моделі використано такі ключові параметри:

1. Довжина пакета (ДП). Відображає кількість байтів у кожному пакеті. Для різних застосунків характерні свої закономірності: чат-програми передають здебільшого короткі пакети, тоді як протоколи обміну файлами генерують довгі й нерівномірні послідовності.

2. Інтервал часу між пакетами (ІВП). Показує часовий проміжок між двома сусідніми пакетами. VoIP-трафік має регулярні інтервали, тоді як P2P-протоколи відзначаються високою варіативністю цього показника.

3. Кількість пакетів у потоці (N). Дає змогу оцінити тривалість і інтенсивність сесії. Для веб-браузингу характерні короткі потоки, а для завантаження файлів — довгі.

4. Середня довжина пакета та середній інтервал (ДПср, ІВПср). Агреговані статистики, які використовуються для зниження впливу випадкових коливань і шумів.

5. Розподіл значень параметрів. У моделі враховується не лише середнє значення, а й характер розподілу довжин пакетів та інтервалів. Це дозволяє точніше описати поведінку застосунків.

Завдяки поєднанню цих характеристик модель формує багатовимірний вектор ознак, який використовується для навчання прихованої марковської моделі. Такий підхід дає змогу врахувати як локальні особливості потоку, так і його загальні статистичні закономірності. У результаті обрані параметри забезпечують баланс між простотою обчислень і високою інформативністю, що робить модель придатною для використання у системах реального часу.

Розроблена модель ідентифікації тунельного трафіку реалізована у вигляді комплексу алгоритмів, що забезпечують повний цикл обробки даних — від збору та підготовки до класифікації.

Алгоритм побудови моделі

1. Алгоритм роботи системи складається з кількох послідовних етапів:

2. Збір даних і відновлення потоків — трафік фіксується у вигляді окремих IP-пакетів, які групуються у цілісні сесії.

3. Виділення параметрів — для кожного потоку визначаються довжина пакетів (ДП), інтервали між ними (ІВП), кількість пакетів та агреговані характеристики.

4. Формування навчальної та тестової вибірок — дані діляться на частини: одна використовується для налаштування моделі, інша — для перевірки її ефективності.

5. Навчання прихованої марковської моделі — параметри визначаються за допомогою алгоритму Баума–Велша, що дозволяє знайти оптимальні ймовірності переходів і емісій.

6. Тестування і класифікація — для нових потоків обчислюється ймовірність належності до кожної моделі, і на основі принципу максимальної правдоподібності приймається рішення про їхній клас.

Алгоритм вирізняється ефективністю навіть при мінімальному обсязі даних: для початкової класифікації достатньо аналізу близько десяти пакетів. Це робить підхід придатним для роботи в реальних високошвидкісних мережах.

Програмна реалізація. Для реалізації алгоритмів було використано середовище MATLAB 2016, яке забезпечило можливість моделювання прихованих марковських процесів, роботи з великими наборами даних та проведення статистичного аналізу. Реалізація включала:

- модулі для формування вибірок і попередньої обробки трафіку;
- функції для навчання ПММ з використанням алгоритму Баума–Велша;
- інтерфейс для візуалізації результатів і порівняння моделей.

Програмні засоби дозволяють легко модифікувати параметри моделі,

експериментувати з кількістю станів і типами розподілів, що робить систему гнучкою і придатною для подальших досліджень.

Переваги підходу. Серед основних переваг алгоритмів і програмних засобів можна виділити:

- роботу з зашифрованим трафіком без необхідності доступу до вмісту пакетів;
- мінімальні вимоги до даних — аналіз ефективний навіть для коротких потоків;
- стабільність і точність завдяки поєднанню ПММ та оптимізаційних процедур;
- можливість інтеграції у системи моніторингу та управління мережами.

Таким чином, запропоновані алгоритми і програмні засоби не лише підтвердили наукову обґрунтованість моделі, а й показали готовність до практичного впровадження в інфраструктуру реальних мереж.

Тестування моделі. Для перевірки ефективності розробленої моделі ідентифікації застосунків у тунельному трафіку було проведено серію експериментів. Мета тестування полягала у визначенні точності, повноти, валідності та рівня помилок при класифікації потоків різних типів.

У дослідженні використовувалися набори трафіку університету Нью-Брансвіка (Канада), які широко застосовуються у науковій спільноті для аналізу мережевих протоколів. Ці набори є відкритими та містять різні типи застосунків, що дозволяє створити реалістичні умови для тестування.

Трафік був поділений на дві групи:

- навчальні дані — застосовувалися для побудови прихованих марковських моделей;
- тестові дані — використовувалися для перевірки точності класифікації нових потоків.

До складу тестових даних увійшли такі типи застосунків, як чат-застосунки, FTP, P2P, VoIP, веб-трафік та інші. Це забезпечило різноманітність експериментів і можливість перевірити універсальність моделі.

Результати класифікації. Експериментальні результати показали високу ефективність моделі (рисунок 2).

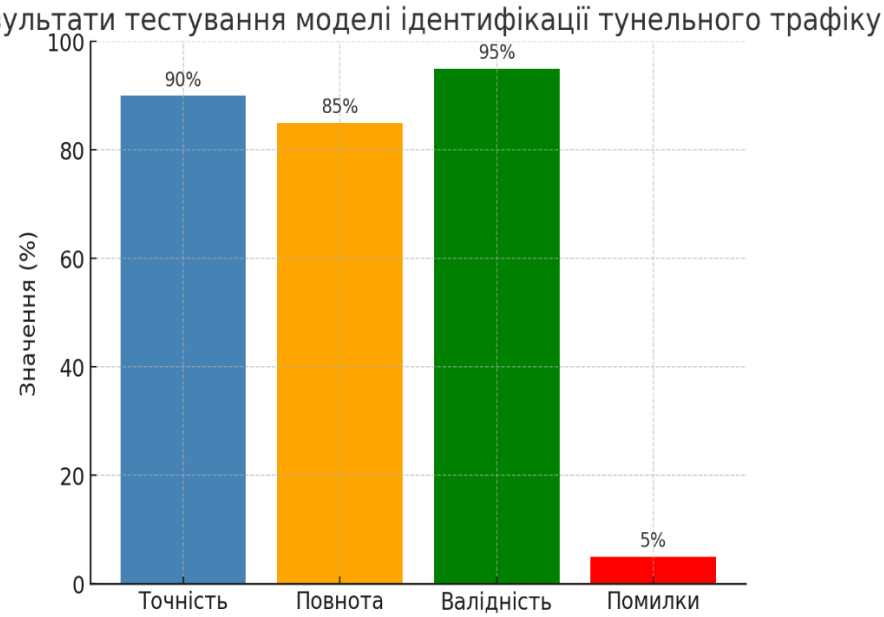

Рисунок 2 - Результати тестування моделі ідентифікації тунельного трафіку

У середньому було досягнуто такі показники:

- точність (precision) – понад 90%;
- повнота (recall) – близько 85%;
- валідність (validity) – понад 95%;
- частка помилок (error rate) – не перевищувала 5%.

Ці результати підтверджують здатність моделі правильно класифікувати більшість потоків навіть за умови обмеженого обсягу даних.

Особливо показовим є те, що для досягнення високої точності достатньо аналізу близько десяти пакетів. Це робить метод придатним для використання у високошвидкісних мережах, де критичною є швидкість прийняття рішень.

Порівняння з іншими методами. Розроблений підхід було порівняно з традиційними методами:

- класифікація за портами показала низьку точність (60–65%), що робить її малопридатною для сучасних мереж;

- глибока інспекція пакетів (DPI) досягає високої точності, але вимагає великих ресурсів і непридатна для зашифрованого трафіку;
- методи машинного навчання на багатьох ознаках забезпечують точність 75–85%, але потребують великих обсягів вхідних даних.

У порівнянні з ними метод на основі прихованих марковських моделей демонструє найкраще співвідношення між точністю, швидкодією та стійкістю до шифрування.

Практична перевірка. Тестування також підтвердило можливість інтеграції розробленої моделі в системи моніторингу та управління мережею. Алгоритми можуть працювати у реальному часі, здійснюючи автоматичну класифікацію потоків та допомагаючи адміністраторам приймати рішення щодо пріоритетів обслуговування чи блокування небажаних застосунків.

Висновки. У статті розглянуто задачу ідентифікації застосунків у тунельному трафіку та представлено підхід, що базується на використанні прихованих марковських моделей. Було описано структуру розробленої моделі, наведено основні параметри для аналізу, реалізовані алгоритми та програмні засоби, а також проведено експериментальне тестування з використанням реальних наборів даних.

Результати показали, що запропонований підхід забезпечує точність понад 90% і низьку частку помилок, навіть за умови обмеженого обсягу даних та повного шифрування трафіку. Модель виявилася стійкою, універсальною та придатною для використання в системах моніторингу та управління мережею в режимі реального часу.

Практична значущість роботи полягає у можливості інтеграції розроблених алгоритмів у сучасні системи інформаційної безпеки та QoS-контролю. Це дозволяє своєчасно виявляти небажані застосунки, оптимізувати розподіл ресурсів і підвищувати рівень захищеності мереж.

Таким чином, результати дослідження підтверджують ефективність і перспективність використання прихованих марковських моделей для ідентифікації тунельного трафіку в умовах сучасних інформаційних мереж.

Література

1. Mazel, J., Saudrais, M. and Hervieu, A. (2022) ML-based tunnel detection and tunneled application classification, arXiv:2201.10371. Available at: https://doi.org/10.48550/arXiv.2201.10371

2. Zhang, Y., Sun, W. and Zhang, S. (2023) 'Identify VPN Traffic Under HTTPS Tunnel Using Three-Dimensional Sequence Features', in Proceedings of the 2022 11th International Conference on Networks, Communication and Computing (ICNCC '22), ACM, pp. 18–23. doi: 10.1145/3579895.3579899

3. Razooqi, Y.S. and Pekár, A. (2025) VPN Traffic Analysis: A Survey on Detection and Application Identification, IEEE Access, 13, pp. 132830–132848. doi:10.1109/ACCESS.2025.3592152

4. Encrypted Network Traffic Classification Based on Machine Learning (2023) Ain Shams Engineering Journal, 15(2), Article 102361. doi:10.1016/j.asej.2023.102361

5. Жилич, В.А., Цаволик, Т.Г. (2023) Алгоритми безпеки для віртуальних приватних мереж VPN, магіст. кваліфікаційна робота, Західноукраїнський національний університет, Тернопіль. http://dspace.wunu.edu.ua/handle/316497/50201

Abstract. *The article considers the problem of identifying applications in tunnel traffic in conditions of widespread use of encryption, VPN technologies, and anonymous networks. An identification model based on hidden Markov models is proposed, which uses statistical characteristics of data flows, in particular packet lengths and time intervals between them. The structure of the mathematical model, training and classification algorithms, and software implementation of the proposed approach are described. Experimental testing using real network traffic sets demonstrated the high accuracy and robustness of the method even under conditions of limited data volume and full traffic encryption. The results confirm the possibility of practical application of the model in monitoring, network management, and information security systems.*

Keywords: *tunneled traffic; application identification; hidden Markov models; encrypted traffic; network traffic analysis; VPN; machine learning.*