**BỘ KHOA HỌC VÀ CÔNG NGHỆ**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

-----------------------------



**BÁO CÁO   
ĐỒ ÁN MÔN HỌC**

**ĐỀ TÀI: ENCRYPTED TRAFFIC CLASSIFICATION**

**Môn học: Machine Learning**

**Giảng viên hướng dẫn: Nguyễn Hồng Sơn**

**Thực hiện bởi nhóm sinh viên, bao gồm:**

1. **Hà Hữu Thắng N21DCCN174**
2. **Nguyễn Hoàng Thiện N21DCCN080**

**TP.HCM, ngày 17 tháng 05 n2025**

# MỤC LỤC

1. Cover Page
2. Introduction
3. Dataset Overview
4. Feature Engineering
5. Methodology
6. Evaluation & Result
7. Discussion
8. Conclusion & Future Work
9. References

# INTRODUCTION

Traffic classification technologies have received increased attention over the last decade due to the implementation of mechanisms for network quality of service, security, accounting, design and engineering. The networking industry as well as the research community have dedicated many efforts to the research of these technologies and came up with several classification techniques. However, the continuous expansion of Internet and mobile technologies are creating a dynamic environment where new applications and services emerge every day, and the existing ones are constantly evolving. Moreover, encryption is becoming pervasive in today’s Internet, serving as a base for secure communications. This constant creation, evolution, and securization of applications makes traffic classification a great challenge for the Internet research community. Traffic classification can be categorized based on its final purpose: associating traffic with encryption (e.g., encrypted traffic), protocol encapsulation (e.g., tunneled through VPN or HTTPS); according to specific applications, (e.g., Skype), or according to the application type (e.g., Streaming, Chat), also called traffic characterization. Some applications (e.g., Skype, Facebook) support multiple services like chat, voice call, file transfer, etc. These applications require identifying both the application itself and the specific task associated with it. Very few traffic classification techniques in the literature address this challenging trends. In early 90’s, the initial traffic classification techniques associated transport layer ports with specific applications, a simple and fast technique. But, its low accuracy and unreliability rendered the development of Deep Packet Inspection (DPI) approaches. The DPI approach analyzes packets and classifies them according to some stored signature or pattern. However, DPI techniques that require payload examination are not computationally efficient, specially over high-bandwidth network. Moreover, they are often circumvented by encapsulated, encrypted, or obfuscated traffic that precludes payload analysis. Selecting effective and reliable features for traffic analysis is still a serious challenge. Generally speaking the classification of network traffic falls mainly into two categories: flow-based classification, using properties such as flow bytes per second, duration per flow, etc. and packet-based classification, using properties such as size, inter-packet duration of the first (or n) packets, etc. In this paper, we focus on analyzing regular encrypted traffic and encrypted traffic tunneled through a Virtual Private Network (VPN). The characteriza tion of VPN traffic is a challenging task that remains to be solved. VPN tunnels are used to maintain the privacy of data shared over the physical network connection holding packet-level encryption, therefore making very difficult to identify the applications running through these VPN services.

In this report, we aim to classify network traffic into VPN and non-VPN categories using three popular machine learning models: Random Forest, Support Vector Machine (SVM), and Gradient Boosting. These models have proven to be effective in handling structured data and can offer high accuracy in classification tasks.

# DATASET OVERVIEW

In this project, we use the ISCX VPN-nonVPN Traffic Dataset, which is provided by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick.

The dataset is available publicly at: [https://www.unb.ca/cic/datasets/vpn.html](https://www.unb.ca/cic/datasets/vpn.html.)

This dataset is specifically designed for research on encrypted traffic classification and includes both VPN and non-VPN traffic collected from various applications and services.

1. Dataset Description

- The ISCX VPN-nonVPN dataset consists of network traffic that was captured and labeled across a variety of real-world applications, including:

+ Browsing: Under this label we have HTTPS traffic generated by users while browsing or performing any task that includes the use of a browser. For instance, when we captured voice-calls using hangouts, even though browsing is not the main activity, we captured several browsing flows.

+ Email: The traffic samples generated using a Thunderbird client, and Alice and Bob Gmail accounts. The clients were configured to deliver mail through SMTP/S, and receive it using POP3/SSL in one client and IMAP/SSL in the other

+ Chat: The chat label identifies instant-messaging applications. Under this label we have Facebook and Hangouts via web browser, Skype, and IAM and ICQ using an application called pidgin.

+ Streaming: The streaming label identifies multimedia applications that require a continuous and steady stream of data. We captured traffic from Youtube (HTML5 and flash versions) an

+ File Transfer: This label identifies traffic applications whose main purpose is to send or receive files and documents. For our dataset we captured Skype file transfers, FTP over SSH (SF

+ VoIP: The Voice over IP label groups all traffic generated by voice applications. Within this label we captured voice-calls using Facebook, Hangouts and Skype.

+ P2P: This label is used to identify file-sharing protocols like Bittorrent. To generate this traffic we downloaded different .torrent files from a public a repository (archive.org) and captured traffic sessions using the uTorrent and Transmission applications.

1. Feature Set

The dataset provides pre-extracted features in CSV format, where each row corresponds to a network flow/session. Key features include:

+ duration: n The duration of the flow.

+ fiat: Forward Inter Arrival Time, the time between two packets sent forward direction (mean, min, max, std).

+ biat: Backward Inter Arrival Time, the time between two packets sent backwards (mean, min, max, std).

+ flowiat: Flow Inter Arrival Time, the time between two packets sent in either direction (mean, min, max, std).

+ active: The amount of time time a flow was active before going idle (mean, min, max, std).

+ idle: The amount of time time a flow was idle before becoming active (mean, min, max, std).

+ fb\_psec: Flow Bytes per second.

+ fp\_psec: Flow packets per second.

# FEATURE ENGINEERING

# METHODOLOGY

# 

# EVALUATION AND RESULT

# **DISCUSSION**

**CONCLUSION AND FUTURE WORK**

**REFERENCES**