ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NIÊN LUẬN NGÀNH

AN TOÀN THÔNG TIN

ĐỀ TÀI

TĂNG CƯỜNG KHẢ NĂNG GIÁM SÁT VÀ PHÂN TÍCH GIAO THÔNG TRONG MẠNG (E.G., VPN) VỚI CÁC MÔ HÌNH HỌC SÂU

Sinh viên: Phạm Thái Khiêm

Mã số sinh viên: B2203725

Khóa: 48

Cần Thơ, 12/2025

ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NIÊN LUẬN NGÀNH

AN TOÀN THÔNG TIN

KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG

ĐỀ TÀI

TĂNG CƯỜNG KHẢ NĂNG GIÁM SÁT VÀ PHÂN TÍCH GIAO THÔNG TRONG MẠNG (E.G., VPN) VỚI CÁC MÔ HÌNH HỌC SÂU

Giảng viên hướng dẫn:
TS. Nguyễn Hữu Vân Long

Sinh viên thực hiện:
Phạm Thái Khiêm
MSSV: B2203725
Khóa 48

Cần Thơ, 11/2025

LỜI CẢM ƠN

Trước hết, em xin gửi lời cảm ơn chân thành và sâu sắc đến ***Tiến sĩ Nguyễn Hữu Vân Long*** *- người đã hướng dẫn em trong quá trình thực hiện niên luận ngành với đề tài " Tăng cường khả năng giám sát và phân tích giao thông trong mạng (e.g., vpn) với các mô hình học sâu ". Trong thời gian làm việc với thầy, em đã nhận được rất nhiều sự quan tâm và định hướng quý báu. Thầy đã giải đáp từng thắc mắc nhỏ, đào sâu vấn đề để hiểu bản chất của các khái niệm, nguyên lý trong lĩnh vực mạng máy tính và máy học.*

Trong quá trình nghiên cứu đề tài, em nhận thấy kiến thức về hệ thống mạng khá rộng và bao gồm nhiều khía cạnh từ lý thuyết đến ứng dụng thực tế. Nếu không có sự định hướng cụ thể từ thầy, em sẽ rất dễ đi sai hướng hoặc tập trung vào những phần không cần thiết. Thầy đã giúp em xác định rõ phạm vi nghiên cứu, ưu tiên những nội dung cốt lõi, đồng thời gợi ý các nguồn tài liệu chất lượng để tham khảo. Nhờ vậy, em có thể xây dựng được một khung nội dung mạch lạc và phù hợp với mục tiêu của niên luận.

Bên cạnh việc hướng dẫn kiến thức chuyên môn, thầy còn truyền đạt cho em nhiều kinh nghiệm quý báu về phương pháp làm việc khoa học. Từ cách phân chia thời gian hợp lý, lập kế hoạch nghiên cứu, đến việc trình bày kết quả sao cho rõ ràng, thầy đều tận tình chỉ bảo. Điều này không chỉ giúp em hoàn thành tốt bài niên luận mà còn là hành trang quan trọng cho các dự án học tập và công việc sau này.

Dù công việc giảng dạy và nghiên cứu rất bận rộn, thầy vẫn dành thời gian xem xét kỹ lưỡng từng phần nội dung em gửi, góp ý chi tiết và kịp thời. Ngoài ra, em xin gửi lời cảm ơn đến các bạn bè xung quanh đã chia sẻ kinh nghiệm và hỗ trợ em trong việc tìm kiếm tài liệu, kiểm tra kết quả thử nghiệm khi em gặp khó khăn. Sự giúp đỡ và đồng hành của mọi người đã góp phần không nhỏ vào việc hoàn thiện đề tài này.

Cuối cùng, em xin chân thành cảm ơn thầy một lần nữa vì sự tận tâm, kiên nhẫn và nhiệt huyết trong suốt quá trình hướng dẫn. Thành quả của đề tài này không chỉ là kết quả học tập, mà còn là minh chứng cho những gì em đã học hỏi được từ thầy.

Cuối lời, em xin chúc Thầy sức khỏe, thành công và hạnh phúc.

Em xin chân thành cảm ơn!

Cần Thơ, ngày 15 tháng 12 năm 2025
Sinh viên thực hiện

Phạm Thái Khiêm

MỤC LỤC

DANH MỤC HÌNH

TÓM LƯỢC

Sự gia tăng nhanh chóng của lưu lượng mạng được mã hóa, đặc biệt thông qua các dịch vụ Mạng Riêng Ảo (VPN), đã đặt ra nhiều thách thức cho công tác giám sát mạng, quản lý lưu lượng và đảm bảo an ninh hệ thống. Mặc dù công nghệ VPN có khả năng bảo vệ nội dung truyền tải thông qua cơ chế mã hóa, các đặc trưng thống kê và hành vi của lưu lượng mạng vẫn không thể bị che giấu hoàn toàn. Nghiên cứu này tập trung đánh giá hiệu quả của các phương pháp học máy và học sâu trong bài toán phân loại lưu lượng VPN và Non-VPN, cũng như nhận dạng loại ứng dụng trong môi trường lưu lượng được mã hóa.

Một khung thực nghiệm toàn diện được xây dựng dựa trên các tập dữ liệu lưu lượng mạng quy mô lớn với kích thước lần lượt là 500K, 800K và 1,1 triệu mẫu. Hai kịch bản phân loại chính được xem xét bao gồm: (i) phân loại nhị phân giữa lưu lượng VPN và Non-VPN và (ii) phân loại đa lớp với 14 loại ứng dụng khác nhau, bao gồm cả lưu lượng VPN và Non-VPN. Các mô hình học máy truyền thống (K-Nearest Neighbors và cây quyết định C4.5) được so sánh với các mô hình học sâu (Multi-Layer Perceptron, Convolutional Neural Network và Long Short-Term Memory). Hiệu năng mô hình được đánh giá thông qua các chỉ số Precision, Recall, F1-score, ma trận nhầm lẫn và quá trình hội tụ trong huấn luyện.

Kết quả thực nghiệm cho thấy các mô hình học sâu vượt trội hơn rõ rệt so với các phương pháp truyền thống, đặc biệt khi được huấn luyện trên các tập dữ liệu lớn. Trong số các mô hình được khảo sát, LSTM đạt hiệu năng tổng thể cao nhất nhờ khả năng khai thác hiệu quả các đặc trưng theo chuỗi thời gian của luồng lưu lượng. Mô hình CNN thể hiện hiệu năng cao với tốc độ hội tụ nhanh và chi phí tính toán thấp hơn, trong khi mô hình C4.5 vẫn đạt kết quả cạnh tranh trong bài toán phân loại nhị phân với hiệu quả tính toán tốt. Các kết quả cũng cho thấy rằng mặc dù VPN che giấu nội dung gói tin, các đặc trưng hành vi của lưu lượng mạng vẫn đủ rõ ràng để cho phép phân loại chính xác cả việc sử dụng VPN và loại ứng dụng.

Nghiên cứu khẳng định rằng việc kết hợp phân tích lưu lượng quy mô lớn với các kỹ thuật học sâu là một hướng tiếp cận hiệu quả cho bài toán phân loại lưu lượng mạng được mã hóa, đồng thời mang lại những đóng góp quan trọng cho các hệ thống an ninh mạng và quản lý lưu lượng trong tương lai.

ABSTRACT

The rapid growth of encrypted network traffic, especially through Virtual Private Network (VPN) services, poses significant challenges for network monitoring, traffic management, and security enforcement. Although VPN technology effectively protects user content by encryption, it cannot completely conceal statistical and behavioral characteristics of network traffic. This study investigates the effectiveness of machine learning and deep learning approaches in classifying VPN and non-VPN traffic, as well as identifying application types within encrypted traffic.

A comprehensive experimental framework is designed using large-scale traffic datasets with different sizes, including 500K, 800K, and 1.1 million samples. Two classification scenarios are considered: binary classification (VPN versus Non-VPN) and multi-class classification with fourteen application categories, covering both VPN and non-VPN traffic. Traditional machine learning models (K-Nearest Neighbors and C4.5 decision tree) are compared with deep learning models (Multi-Layer Perceptron, Convolutional Neural Network, and Long Short-Term Memory network). Performance is evaluated using Precision, Recall, F1-score, confusion matrices, and training convergence analysis.

Experimental results demonstrate that deep learning models significantly outperform traditional approaches, particularly when trained on large-scale datasets. Among all evaluated models, the LSTM achieves the best overall performance, benefiting from its ability to capture temporal dependencies in traffic flows. CNN also shows strong performance with faster convergence and lower computational cost, while C4.5 delivers competitive results in binary classification with high efficiency. The results further indicate that although VPN encryption obscures payload content, distinctive traffic behavior patterns remain observable, enabling accurate classification of both VPN usage and application types.

The findings confirm that large-scale traffic analysis combined with deep learning techniques provides an effective solution for encrypted traffic classification and offers valuable insights for future network security and traffic management systems.

PHẦN GIỚI THIỆU

I. Đặt vấn đề

Trong những năm gần đây, lưu lượng mạng Internet ngày càng tăng trưởng mạnh mẽ, đi kèm với sự đa dạng của các loại ứng dụng như trình duyệt web, VoIP, truyền tệp, streaming, mạng ngang hàng (P2P), trò chuyện trực tuyến... Đặc biệt, các dịch vụ VPN ngày càng được sử dụng rộng rãi để bảo mật kết nối và ẩn danh người dùng.

Mặc dù VPN mang lại lợi ích về quyền riêng tư, nhưng nó cũng gây khó khăn cho các hệ thống giám sát, phân tích và phát hiện bất thường trong mạng. Do cơ chế mã hóa và ẩn danh, các gói tin VPN có xu hướng che giấu thông tin về loại ứng dụng thực sự đang được sử dụng. Điều này làm suy giảm khả năng phân loại lưu lượng mạng bằng các phương pháp truyền thống.

Việc phân loại chính xác lưu lượng VPN và Non-VPN, cũng như phân biệt 14 loại traffic ứng dụng (Browsing, VoIP, P2P, FTP, Chat, Mail, Streaming và tương ứng các biến thể VPN) đóng vai trò quan trọng trong:

+ Giám sát an toàn mạng.
+ Phát hiện và ngăn chặn tấn công.
+ Phân tích hành vi người dùng.
+ Tối ưu QoS và quản lý băng thông.

Tuy nhiên, các phương pháp máy học truyền thống mặc dù đã cho kết quả tương đối tốt trên dataset ISCXVPN2016, nhưng vẫn còn nhiều hạn chế trong việc mô hình hóa các quan hệ phi tuyến và phụ thuộc thời gian. Sự phát triển của Deep Learning mở ra cơ hội cải thiện đáng kể độ chính xác và khả năng khái quát hóa trong bài toán phân tích lưu lượng mạng.

II. Tính cấp thiết của đề tài

– Sự gia tăng mạnh mẽ của các loại dịch vụ Internet và ứng dụng trực tuyến kéo theo sự đa dạng và phức tạp của lưu lượng mạng. Đồng thời, việc sử dụng VPN ngày càng phổ biến do nhu cầu bảo vệ quyền riêng tư và vượt tường lửa. VPN mã hóa toàn bộ lưu lượng, khiến cho:

+ Các hệ thống giám sát truyền thống khó nhận diện loại ứng dụng thực sự nằm bên trong VPN
+ Các giải pháp an ninh mạng gặp khó khăn khi cần phát hiện hành vi bất thường hoặc tấn công
+ Các nhà quản trị không thể phân bổ tài nguyên mạng hợp lý dựa trên loại ứng dụng.

– Trong khi đó, các kỹ thuật phân loại lưu lượng truyền thống dựa trên chữ ký (signature) hoặc port-based gần như mất hiệu lực do:

+ Nhiều ứng dụng sử dụng cổng ngẫu nhiên
+ Kỹ thuật mã hóa và ngụy trang ngày càng tinh vi
+ Lưu lượng của các ứng dụng khác nhau có thể trở nên tương đồng trong môi trường VPN.

– Các phương pháp Machine Learning truyền thống như KNN hay Decision Tree đã cho thấy hiệu quả nhất định, nhưng vẫn còn hạn chế trong việc:

+ Mô hình hóa các quan hệ phi tuyến phức tạp
+ Nhận diện mẫu dựa trên tính trừu tượng cao
+ Phân loại multi-class với số lượng lớp lớn.

– Sự phát triển mạnh mẽ của Deep Learning trong lĩnh vực nhận dạng mẫu, xử lý tín hiệu và phát hiện hành vi bất thường cho thấy tiềm năng vượt trội trong:

+ Khai thác cấu trúc ẩn trong dữ liệu lưu lượng mạng
+ Cải thiện độ chính xác phân loại
+ Cãng cao khả năng tổng quát hóa cho mô hình.

– Trong bối cảnh các hệ thống mạng hiện đại cần giám sát thông minh, tự động và hiệu quả, việc nghiên cứu ứng dụng mô hình học sâu để phân tích lưu lượng (đặc biệt là lưu lượng VPN) trở nên cấp thiết và có giá trị thực tiễn cao, phục vụ cho:

+ An toàn mạng
+ Quản lý chất lượng dịch vụ
+ Phân tích hành vi người dùng
+ Điều phối tài nguyên mạng.

III. Mục tiêu nghiên cứu

Mục tiêu tổng quát: Tăng cường khả năng giám sát và phân tích lưu lượng mạng (đặc biệt là lưu lượng VPN) bằng các mô hình học sâu và dữ liệu synthetic.

Mục tiêu cụ thể:

– Phân loại lưu lượng mạng thành VPN và Non-VPN (binary classification).
– Phân loại 14 loại traffic gồm nhiều nhóm ứng dụng và các biến thể VPN tương ứng.
– Sinh dữ liệu bằng CTGAN nhằm giảm mất cân bằng và kiểm tra nguy cơ overfitting.
– So sánh liệu năng giữa mô hình Machine Learning (KNN, C4.5) và Deep Learning (MLP, CNN, LSTM).
– Đánh giá các mô hình theo:
+ Precision, Recall, F1-score
+ Ma trận nhầm lẫn
+ Thời gian huấn luyện và dự đoán
– Đề xuất mô hình phù hợp nhất cho bài toán phân tích lưu lượng VPN.

IV. Phạm vi và giới hạn nghiên cứu

Phạm vi:

– Sử dụng dataset ISCXVPN2016 (unb.ca/cic/datasets/vpn.html) và các phiên bản synthetic (500k, 800k, 1.1M record).
– Làm việc trên flow-based features (23 đặc trưng).
– Phân loại supervised learning.
– So sánh mô hình ML và DL trên cùng dữ liệu.

Giới hạn:

– Không phân tích payload (do dữ liệu đã được trích xuất thành flow).
– Không triển khai hệ thống giám sát thời gian thực.
– Không xem xét phương pháp semi-supervised hoặc unsupervised.

- Không nghiên cứu kỹ thuật VPN obfuscation phức tạp hơn.

V. Phương pháp nghiên cứu

1. Thu thập dữ liệu: sử dụng ISCXVPN2016.
2. Tiền xử lý: loại dữ liệu trùng, chuẩn hóa, mã hóa nhãn.
3. Tạo dữ liệu cân bằng: sinh bằng CTGAN ở ba mức dung lượng.
4. Huấn luyện mô hình Machine Learning: KNN, Decision Tree (C4.5)
5. Huấn luyện mô hình Deep Learning: MLP, CNN, LSTM
6. Đánh giá mô hình: Precision, Recall, F1-score, Confusion Matrix, Train/Test time.
7. So sánh ML và DL trên cả binary và multi-class.
8. Kết luận mô hình tối ưu và ảnh hưởng của dữ liệu synthetic.

PHẦN NỘI DUNG
CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1 Tổng quan về phân tích lưu lượng mạng

1.1.1 Khái niệm lưu lượng mạng (Network Traffic)

Lưu lượng mạng là tập hợp các gói tin được truyền trong mạng giữa các thiết bị. Mỗi phiên truyền thông có thể được biểu diễn thành các network flows chứa thông tin thống kê như: thời lượng, số gói tin, kích thước gói, tốc độ byte/packet, thời gian giữa các gói (IAT), trạng thái idle/active...

1.1.2 Phân loại lưu lượng mạng (Traffic Classification)

Phân loại lưu lượng là quá trình xác định loại ứng dụng hoặc tính chất của kết nối dựa trên đặc trưng hành vi của nó.

Trong luận văn này, bài toán phân loại bao gồm:

- Phân loại nhị phân: VPN vs Non-VPN
- Phân loại đa lớp (14 nhãn): Browsing, VoIP, P2P, FTP, Chat, Mail, Streaming và các biến thể tương ứng khi truyền qua VPN

1.1.3 Khó khăn trong phân loại lưu lượng VPN

- Toàn bộ payload bị mã hóa, không thể dùng DPI (Deep Packet Inspection).
- VPN che giấu hành vi thật của ứng dụng.
- Các ứng dụng khác nhau có thể có đặc trưng phân bố tương đồng khi chạy qua VPN.
- Lưu lượng multi-class thường mất cân bằng nghiêm trọng, gây khó cho các mô hình học máy.

1.2 Phân loại bằng phương pháp truyền thống

1.2.1 KNN (K-Nearest Neighbors)

KNN là thuật toán phân loại dựa trên cơ chế tìm k điểm dữ liệu gần nhất với mẫu cần dự đoán trong không gian đặc trưng. Khoảng cách phổ biến được sử dụng là Euclidean, Manhattan hoặc Minkowski. Mẫu mới được gán nhãn theo đa số nhãn của k điểm gần nhất này.

Ưu điểm:

- Dễ triển khai, không cần huấn luyện phức tạp.
- Hiệu quả với dữ liệu ít nhiễu.

Nhược điểm:

- Rất chậm khi dự đoán trên dataset lớn (vì phải tính khoảng cách với toàn bộ dữ liệu train).
- Nhạy cảm với nhiễu và mất cân bằng lớp.
- Không phù hợp với dữ liệu high-dimensional.

1.2.2 Decision Tree (C4.5)

Decision Tree là mô hình phân loại dựa trên cấu trúc cây, gồm các nút quyết định dựa trên giá trị thuộc tính. Phương pháp C4.5 lựa chọn thuộc tính phân chia bằng cách tối đa hóa Information Gain Ratio, giúp giảm chênh lệch khi thuộc tính có nhiều giá trị.

Ưu điểm:
- Thời gian dự đoán rất nhanh.
- Dễ diễn giải, trực quan.

Nhược điểm:
- Dễ overfitting.
- Khó mô hình hóa quan hệ phi tuyến phức tạp trong dữ liệu VPN.

1.3 Học sâu trong phân tích lưu lượng mạng

Học sâu đã chứng minh hiệu quả vượt trội trong các bài toán phân loại dữ liệu có cấu trúc phức tạp, nhờ khả năng tự động học đặc trưng và mô hình hóa quan hệ phi tuyến. Trong đề tài này, ba mô hình được sử dụng gồm MLP, CNN, và LSTM.

1.3.1 MLP (Multi-Layer Perceptron)

MLP là dạng mạng neuron truyền thẳng (feed-forward neural network) gồm nhiều lớp tuyến tính kết hợp với các hàm kích hoạt phi tuyến như ReLU, sigmoid hoặc tanh.

Đặc điểm chính:
- Bao gồm input layer, một hoặc nhiều hidden layer, và output layer.
- Mỗi neuron trong lớp này kết nối đầy đủ với neuron của lớp sau.
- Học thông qua thuật toán backpropagation và tối ưu gradient.

Ưu điểm:
- Mô hình hóa được các quan hệ phi tuyến.
- Dễ triển khai và điều chỉnh.
- Hoạt động tốt trên dữ liệu dạng vector như các flow thống kê.

Nhược điểm:
- Không khai thác được thông tin cục bộ hay tính tuần tự.
- Dễ overfitting nếu mô hình quá lớn hoặc dữ liệu không đủ.

1.3.2 CNN (Convolutional Neural Network)

CNN là mô hình đặc biệt hiệu quả trong việc trích xuất đặc trưng cục bộ bằng cơ chế tích chập (convolution). Trong phân loại lưu lượng mạng, CNN có thể xử lý vector đặc trưng như một chuỗi 1D.

Đặc điểm chính:

- Convolution layers dùng kernel quét qua dữ liệu, học ra đặc trưng cục bộ.
- Pooling layers giảm chiều và chống overfitting.
- Fully connected layers dùng để phân loại.

Ưu điểm:
- Học được đặc trưng quan trọng mà không cần thiết kế thủ công.
- Tổng quát hóa tốt hơn MLP.
- Phù hợp với dữ liệu có cấu trúc cục bộ hoặc tương quan giữa các đặc trưng.

Nhược điểm:
- Cần reshape dữ liệu đúng định dạng.
- Đòi hỏi nhiều tài nguyên tính toán hơn mô hình ML.

1.3.3 LSTM (Long Short-Term Memory)

LSTM là mạng nơ-ron tuần tự (Recurrent Neural Network) được thiết kế để xử lý dữ liệu có quan hệ theo thời gian.

Đặc điểm chính:
- Sử dụng các cơ chế input gate, forget gate, output gate để duy trì trạng thái.
- Học được các phụ thuộc dài hạn mà RNN thông thường không thể xử lý.
- Thích hợp cho dữ liệu gồm chuỗi hoặc các biến có tương quan thời gian cao.

Ưu điểm:
- Mô hình hóa tốt sự thay đổi theo thời gian của traffic.
- Nhận diện được các pattern phức tạp trong dữ liệu lưu lượng.
- Được xem là mô hình mạnh nhất trong số ba mô hình khi phân tích flow.

Nhược điểm:
- Huấn luyện chậm hơn CNN và MLP.
- Dễ overfitting nếu dữ liệu ít hoặc mất cân bằng.

1.4 Vấn đề mất cân bằng dữ liệu và tăng cường dữ liệu:

1.4.1 Mất cân bằng trong bài toán 14 lớp

Dataset ISCXVPN2016 gốc mất cân bằng nặng:
- Lớp Streaming, Mail, VPN-VoIP rất ít mẫu
- Lớp Browsing, Chat có số lượng lớn hơn nhiều

Điều này khiến mô hình:
- Nghiêng về lớp lớn,
- F1-score thấp,
- Không học được đặc trưng của lớp hiếm.

1.4.2 Tổng quan về CTGAN

CTGAN (Conditional Tabular GAN) là mô hình GAN được thiết kế cho dữ liệu bảng, như các flow trong mạng.

CTGAN có thể:

- Sinh dữ liệu synthetic sát với phân phối thật,
- Điều khiển phân phối class bằng "conditional sampling",
- Giúp cân bằng dataset,
- Tạo thêm dữ liệu để tránh overfitting.

1.5 Đánh giá mô hình phân loại

Các chỉ số dùng để đánh giá mô hình gồm:

- Precision
- Recall
- F1-score
- Ma trận nhầm lẫn
- Training time & Testing time

CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Quy trình nghiên cứu tổng quát

Quy trình nghiên cứu của luận văn được thiết kế theo các bước sau:

1. Thu thập và phân tích dữ liệu
2. Sử dụng dataset ISCXVPN2016.
3. Tiền xử lý dữ liệu
4. Loại bỏ lưu lượng trùng.
5. Chuẩn hóa (StandardScaler).
6. Shuffle dữ liệu.
7. Mã hóa nhãn cho bài toán đa lớp.
8. Sinh dữ liệu synthetic bằng CTGAN
9. Giải quyết mất cân bằng dữ liệu multi-class.
10. Sinh các bộ 500k, 800k, 1.1M mẫu để kiểm tra nguy cơ overfitting.
11. Huấn luyện mô hình Machine Learning
- KNN
- Decision Tree (C4.5)
12. Huấn luyện mô hình Deep Learning
- MLP
- CNN
- LSTM
13. Đánh giá mô hình
- Precision, Recall, F1-score
- Confusion Matrix
- Thời gian train / test
- So sánh giữa ML và DL
14. Kết luận và đề xuất mô hình tối ưu

2.2 Dữ liệu sử dụng

Tên dataset: VPN-nonVPN dataset (ISCXVPN2016)
Nguồn: Đại học New Brunswick (Canada)
Lưu lượng gốc: 59706 dòng
Lưu lượng sau làm sạch: 57596 dòng
Dataset chứa các lưu lượng mạng từ nhiều loại ứng dụng bao gồm:

- **BROWSING:** Nhãn này bao gồm lưu lượng HTTPS do người dùng tạo ra khi duyệt web hoặc thực hiện bất kỳ tác vụ nào liên quan đến việc sử dụng trình duyệt. Ví dụ, khi họ ghi lại các cuộc gọi thoại bằng Hangouts, mặc dù duyệt web không phải là hoạt động chính, họ vẫn ghi lại được một số luồng duyệt web.

- **MAIL:** Các mẫu lưu lượng truy cập được tạo ra bằng cách sử dụng ứng dụng Thunderbird và tài khoản Gmail của Alice và Bob. Các ứng dụng này được cấu hình để gửi thư qua SMTP/S và nhận thư bằng POP3/SSL trên một ứng dụng và IMAP/SSL trên ứng dụng còn lại.

- **CHAT:** Nhãn trò chuyện xác định các ứng dụng nhắn tin tức thời. Dưới nhãn này, chúng ta có Facebook và Hangouts thông qua trình duyệt web, Skype, và IAM và ICQ sử dụng một ứng dụng có tên là pidgin.

- **STREAMING:** Nhãn này xác định các ứng dụng đa phương tiện yêu cầu luồng dữ liệu liên tục và ổn định. Thu thập lưu lượng truy cập từ YouTube (phiên bản HTML5 và Flash) và Vimeo bằng trình duyệt Chrome và Firefox.

- **FT:** Nhãn này xác định các ứng dụng lưu lượng truy cập mà mục đích chính là gửi hoặc nhận tập tin và tài liệu. Đối với tập dữ liệu này đã thu thập các phiên truyền tải tập tin Skype, FTP qua SSH (SFTP) và FTP qua SSL (FTPS).

- **VOIP:** Nhãn này nhóm tất cả lưu lượng truy cập được tạo ra bởi các ứng dụng thoại. Trong nhãn này, đã thu thập các cuộc gọi thoại sử dụng Facebook, Hangouts và Skype.

- **P2P:** Nhãn này được sử dụng để xác định các giao thức chia sẻ tệp như BitTorrent. Để tạo ra lưu lượng truy cập này, họ đã tải xuống các tệp .torrent khác nhau từ một kho lưu trữ công cộng và ghi lại các phiên truy cập bằng ứng dụng uTorrent và Transmission.

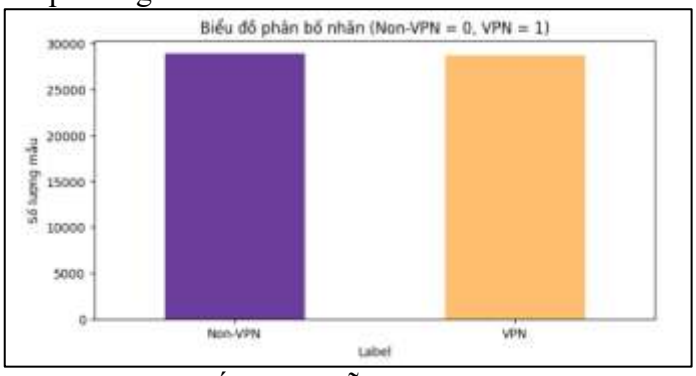
Mỗi loại ứng dụng đều có hai dạng:
- Non-VPN Traffic
- VPN Traffic (được mã hóa qua OpenVPN)

Số lượng đặc trưng:
- Bộ dữ liệu bao gồm 23 đặc trưng thống kê của mỗi flow:
    + duration: Tổng thời gian lưu lượng tính bằng micro giây
    + total_fiat: Tổng thời gian giữa các lần đến (thời gian giữa hai gói tin liên tiếp theo hướng truyền đi)
    + total_biat: Tổng thời gian giữa các lần đến ngược chiều (thời gian giữa hai gói tin liên tiếp theo chiều ngược lại)
    + min_fiat: Thời gian đến giữa các lần đến tối thiểu
    + min_biat: Thời gian đến ngược tối thiểu
    + max_fiat: Thời gian đến giữa các lần đến tối đa
    + max_biat: Thời gian đến ngược tối đa
    + mean_fiat: Thời gian trung bình giữa các lần đến
    + mean_biat: Thời gian trung bình giữa các lần đến ngược chiều

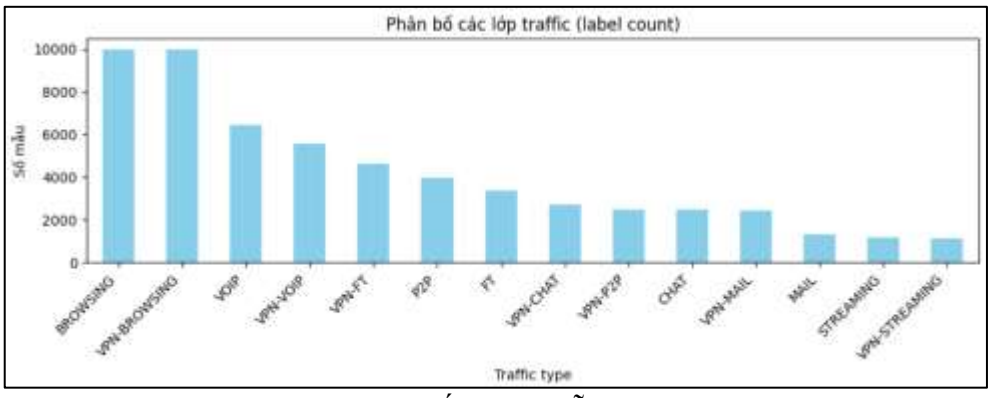
+ flowPktsPerSecond: Số lượng gói dữ liệu mỗi giây
+ flowBytesPerSecond: Số byte truyền tải mỗi giây
+ min_flowiat: Thời gian giữa các lần đến của lưu lượng tối thiểu
+ max_flowiat: Thời gian giữa các lần đến của lưu lượng tối đa
+ mean_flowiat: Thời gian giữa các lần đến của lưu lượng trung bình
+ std_flowiat: Độ lệch chuẩn của thời gian giữa các lần đến của lưu lượng
+ min_active: Thời gian tối thiểu của một giai đoạn hoạt động trong lưu lượng
+ mean_active: Thời gian trung bình của một giai đoạn hoạt động trong lưu lượng
+ max_active: Thời gian tối đa của một giai đoạn hoạt động trong lưu lượng
+ std_active: Độ lệch chuẩn của thời gian hoạt động
+ min_idle: Thời gian tối thiểu của giai đoạn không hoạt động trong lưu lượng
+ mean_idle: Thời gian trung bình của giai đoạn nhàn rỗi trong lưu lượng
+ max_idle: Thời gian tối đa của giai đoạn không hoạt động trong lưu lượng
+ std_idle: Độ lệch chuẩn của thời gian nhàn rỗi

– 1 cột nhãn ứng dụng (traffic_type) bao gồm 14 lớp: BROWSING, VOIP, P2P, FT, CHAT, MAIL, STREAMING, VPN-BROWSING, VPN-VOIP, VPN-P2P, VPN-FT, VPN-CHAT, VPN-MAIL, VPN-STREAMING.

Vấn đề của dataset gốc:
– Bài toán binary classification (VPN vs Non-VPN): dữ liệu khá cân bằng.
– Bài toán multi-class (14 lớp): mất cân bằng nghiêm trọng - một số lớp chỉ vài trăm mẫu, một số lớp vài nghìn.



Hình 2.1 Số lượng mẫu VPN & Non-VPN

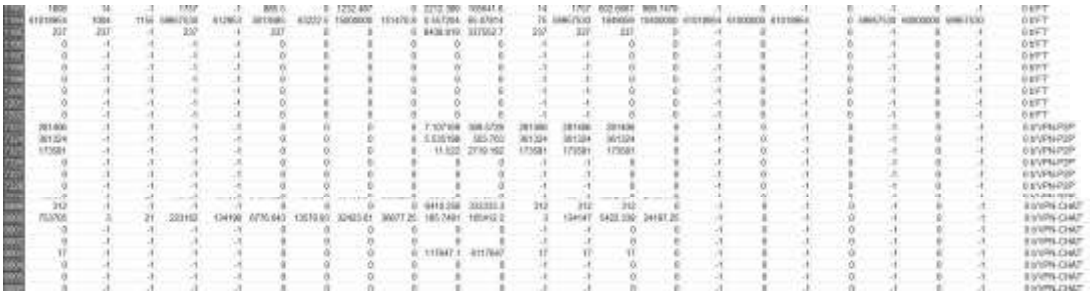


Hình 2.2 Số lượng mẫu đa lớp

2.3 Tiền xử lý dữ liệu

Các bước tiền xử lý bao gồm:

- Loại bỏ các lưu lượng bị trùng:



Hình 2.3 Các lưu lượng bị trùng

- Xáo trộn dữ liệu
- Chuẩn hóa đặc trưng:
  Tất cả 23 đặc trưng được chuẩn hóa về phân phối chuẩn:

$$x' = \frac{x - \mu}{\sigma}$$

  Mục đích:
  + Giúp mô hình hội tụ nhanh hơn,
  + Tránh đặc trưng có biên độ lớn lấn át đặc trưng nhỏ.
- Mã hóa nhãn
- One-hot encoding (cho Deep Learning multi-class)

2.4 Sinh dữ liệu cân bằng bằng CTGAN

2.4.1 Mục đích sinh dữ liệu tổng hợp:

Tập dữ liệu ISCXVPN2016 ở bài toán phân loại đa lớp (14 nhãn) tồn tại mất cân bằng nghiêm trọng giữa các lớp, trong đó một số loại lưu lượng (ví dụ: MAIL, CHAT, VPN-CHAT, VPN-MAIL) có số lượng mẫu rất nhỏ so với các lớp phổ biến như BROWSING VPN- BROWSING.

Hiện tượng mất cân bằng này gây ảnh hưởng tiêu cực đến quá trình huấn luyện mô hình học máy và học sâu, dẫn đến:

- Mô hình thiên lệch về các lớp chiếm đa số
- Hiệu năng kém trên các lớp thiểu số
- Giá trị F1-score macro thấp, phản ánh khả năng phân loại đa lớp không đồng đều

Do đó, trong nghiên cứu này, CTGAN được sử dụng để sinh thêm dữ liệu tổng hợp nhằm:

- Cân bằng số lượng mẫu giữa các lớp
- Giữ nguyên phân bố thống kê của các đặc trưng mạng

2.4.2 Lý do lựa chọn CTGAN

CTGAN là một mô hình GAN được thiết kế chuyên biệt cho dữ liệu dạng bảng (tabular data), phù hợp với dữ liệu lưu lượng mạng có đặc điểm:

- Bao gồm nhiều đặc trưng số liên tục
- Có cột rời rạc (traffic_type)

– Phân bố dữ liệu phức tạp và không tuyến tính

So với các phương pháp sinh dữ liệu truyền thống (SMOTE, ADASYN), CTGAN có ưu điểm:

– Học được mối quan hệ phi tuyến giữa các đặc trưng
– Sinh dữ liệu đa chiều có tính thực tế cao
– Hạn chế việc sao chép dữ liệu gốc gây overfitting

2.4.3 Cấu hình và tham số sinh dữ liệu

Quá trình sinh dữ liệu được thực hiện riêng biệt cho từng nhãn, nhằm đảm bảo mỗi lớp có phân bố dữ liệu đặc trưng riêng và tránh trộn lẫn đặc điểm giữa các loại lưu lượng.

Tham số	Giá trị	Ý nghĩa
epochs	300	Số vòng huấn luyện CTGAN cho mỗi nhãn
batch_size	500	Số mẫu trong mỗi batch huấn luyện
discrete_columns	['traffic_type']	Chỉ định cột nhãn là biến rời rạc
random_seed	42	Đảm bảo khả năng tái lập

Giải thích các tham số:

– Số epoch được lựa chọn ở mức 300 nhằm:
+ Đảm bảo mô hình GAN hội tụ
+ Học đầy đủ phân bố xác suất của từng đặc trưng
+ Tránh underfitting đối với các nhãn có ít dữ liệu gốc
– Batch size được sử dụng để:
+ Ổn định quá trình huấn luyện GAN
+ Giảm dao động gradient giữa generator và discriminator
+ Phù hợp với kích thước dữ liệu của từng nhãn
– discrete_columns giúp CTGAN:
+ Nhận biết traffic_type là biến phân loại
+ Không sinh ra giá trị liên tục cho nhãn
+ Giữ nguyên tính rời rạc và tính hợp lệ của nhãn

Kết quả các bộ dữ liệu synthetic tạo ra

Để kiểm tra hiện tượng overfitting, 3 bộ synthetic được sinh:

+ 500,000 mẫu
+ 800,000 mẫu
+ 1,100,000 mẫu

Sau đó, cả Machine Learning và Deep Learning đều được train & test trên các bộ này để đánh giá chất lượng.

2.5 Mô hình Deep Learning

Tham số	Giá trị	Mục đích
Epochs	30	Giới hạn số vòng huấn luyện
Early Stopping	theo dõi val_loss	Tránh overfitting
Patience (ES)	8–10	Cho phép mô hình cải thiện
ReduceLROnPlateau	factor = 0.5	Giảm learning rate khi bão hòa
Batch size (Binary)	64	Phù hợp dữ liệu vừa
Batch size (Multi-class)	128	Ổn định gradient khi dữ liệu lớn

Validation split	0.15 – 0.2	Đánh giá trong huấn luyện
Output layer (Binary)	1 neuron, Sigmoid	Phân loại VPN / Non-VPN
Output layer (Multi-class)	*N* neuron, Softmax	Phân loại 14 loại lưu lượng
Loss function	Binary / Categorical Cross-Entropy	Phù hợp bài toán phân loại
Optimizer	Adam	Hội tụ nhanh, ổn định
Input shape	(23, 1)	Nhận vector đặc trưng lưu lượng mạng

2.5.1 MLP

Kiến trúc gồm 3 tầng:

+ Dense (hidden): 2
+ Dense (output): 1

Tham số	Giá trị	Mục đích
Dense layer 1	256 neuron, ReLU	Học quan hệ phi tuyến giữa các đặc trưng
Dropout	0.4	Giảm overfitting
Dense layer 2	128 neuron, ReLU	Tiếp tục trích xuất đặc trưng ở mức trừu tượng cao hơn
Dropout	0.3	Giảm overfitting ở tầng sâu

2.5.2 CNN

Kiến trúc gồm 4 tầng:

+ Conv1D: 2
+ Dense (hidden): 1
+ Dense (output): 1

Tham số	Giá trị	Mục đích
Conv1D layer 1	64 filters, kernel=3, ReLU	Trích xuất đặc trưng cục bộ
MaxPooling1D	Pool size = 2	Giảm kích thước không gian đặc trưng, giữ lại thông tin quan trọng
Conv1D layer 2	128 filters, kernel=3, ReLU	Học đặc trưng phức tạp hơn
Dropout	0.4	Giảm overfitting
Flatten	_	Chuyển tensor 3D sang vector 1D
Dense layer	128 neuron, ReLU	Kết hợp các đặc trưng đã trích xuất.

2.5.3 LSTM

Kiến trúc gồm 4 tầng:

+ LSTM: 2
+ Dense (hidden): 1
+ Dense (output): 1

Tham số	Giá trị	Mục đích
LSTM layer 1	128 units, return_sequences=True	Trích xuất chuỗi đặc trưng ở mức cao, giữ lại toàn bộ chuỗi đầu ra
Dropout	0.3	Giảm overfitting trong quá trình học chuỗi
LSTM layer 2	64 units	Tổng hợp thông tin chuỗi thành biểu diễn ngắn gọn
Dense layer	32 neuron, ReLU	Kết hợp thông tin trước khi phân loại

CHƯƠNG 3: KẾT QUẢ VÀ ĐÁNH GIÁ

3.1 Quá trình huấn luyện

Các thí mô hình được thực hiện trên cùng một tập dữ liệu gốc và cùng môi trường huấn luyện nhằm đảm bảo tính công bằng trong so sánh giữa các mô hình. Dataset ban đầu được xây dựng dựa trên ISCXVPN2016, sau đó được mở rộng bằng công cụ CTGAN để giải quyết bài toán mất cân bằng lớp trong phân loại đa lớp.

Các mô hình được đánh giá bao gồm:

- Machine Learning truyền thống (làm cơ sở): KNN, Decision Tree (C4.5)
- Deep Learning: MLP, CNN, LSTM

Hai bài toán được xem xét:

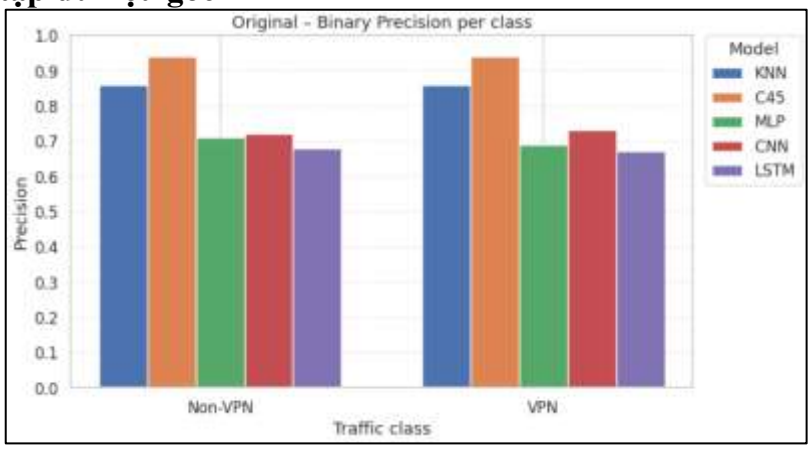
- Phân loại nhị phân: VPN vs Non-VPN
- Phân loại đa lớp: 14 loại traffic (BROWSING, VOIP, …, VPN-STREAMING)

Các tiêu chí đánh giá chính gồm:

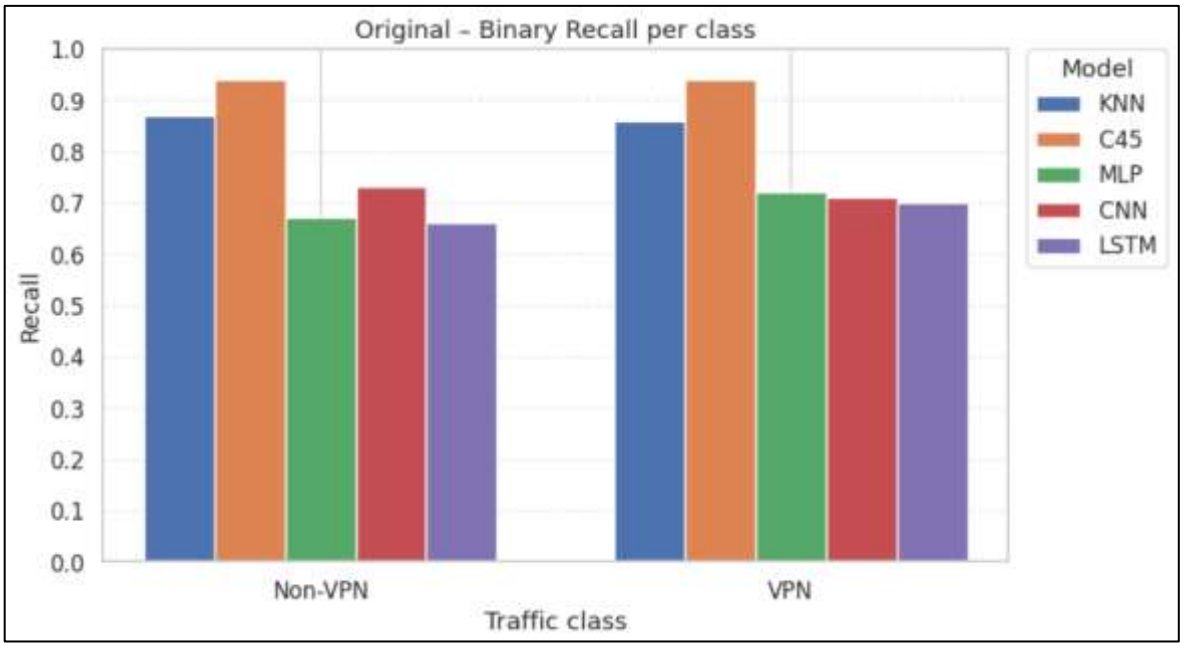
- Precision
- Recall
- F1-score
- Train time
- Test time

3.2 Kết quả các mô hình

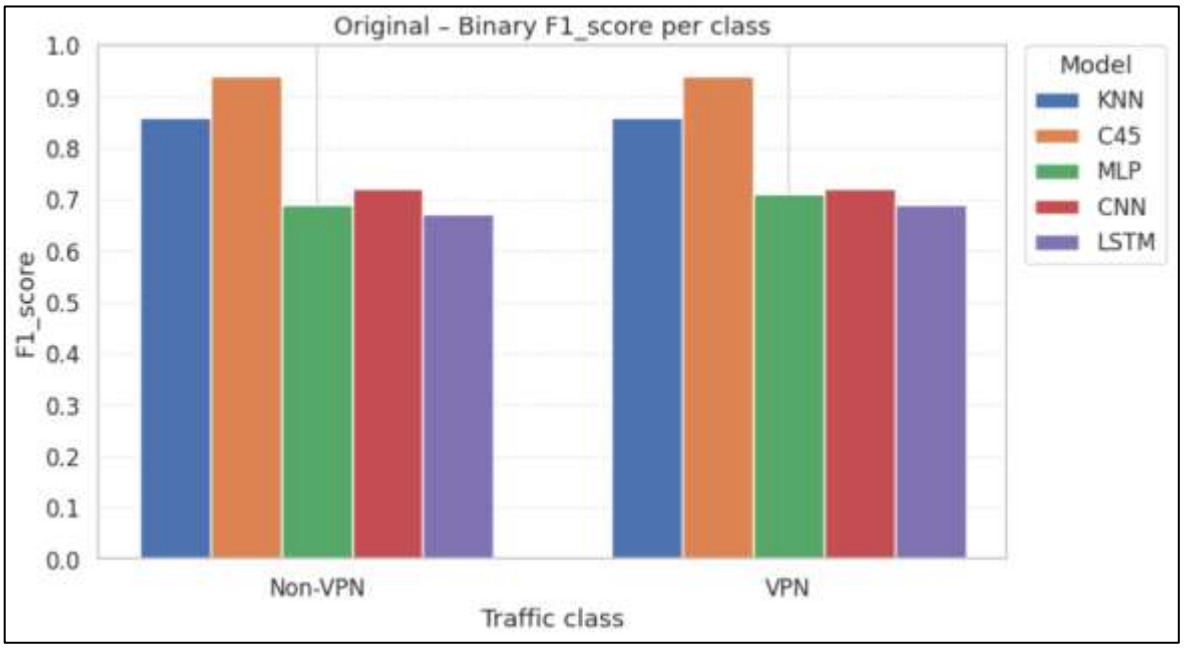
3.2.1 Trên tập dữ liệu gốc



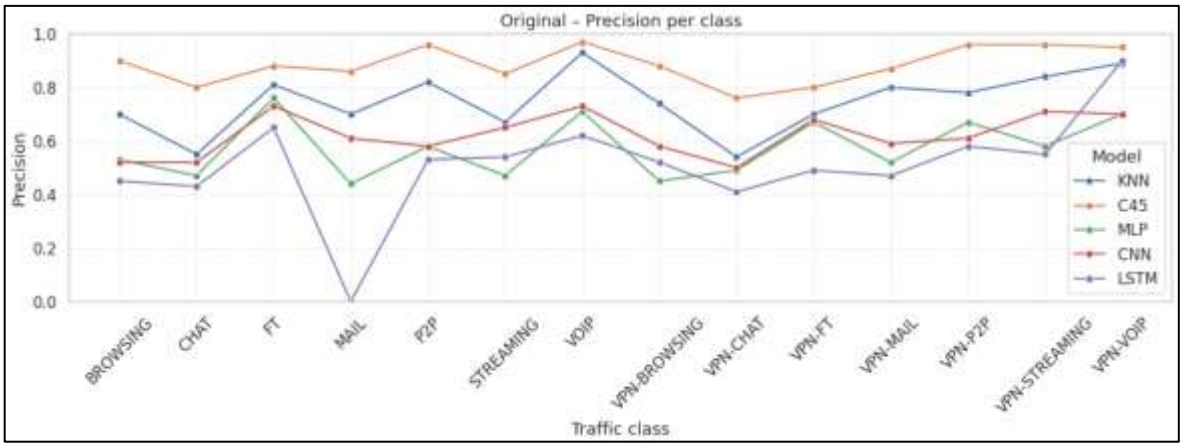
Hình 3.1 Precison nhị phân



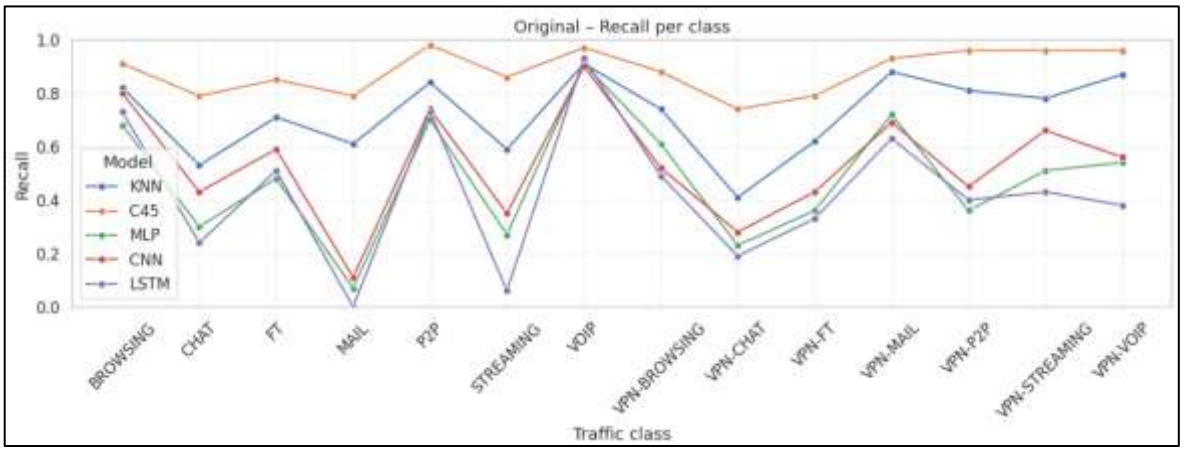
Hình 3.2 Recall nhị phân



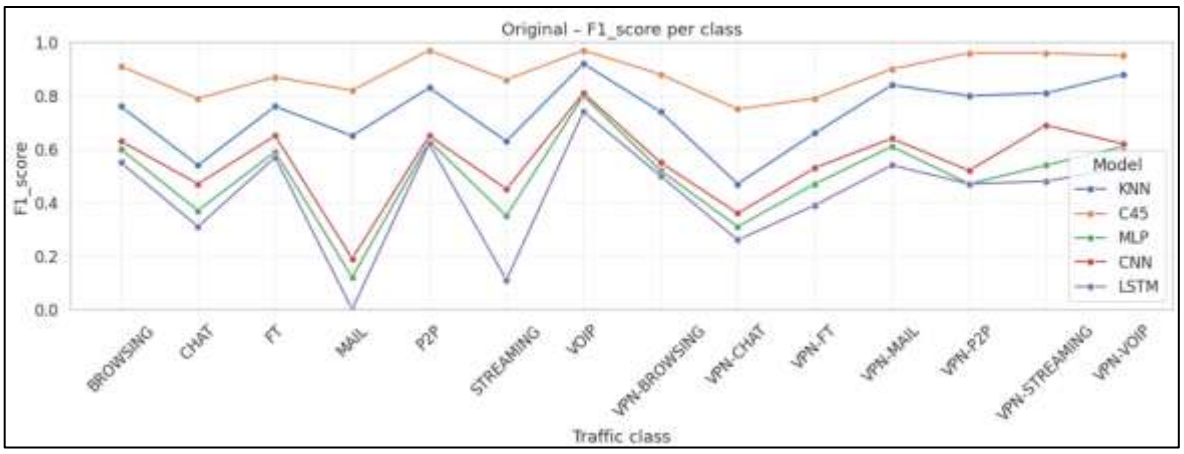
Hình 3.3 F1-score nhị phân



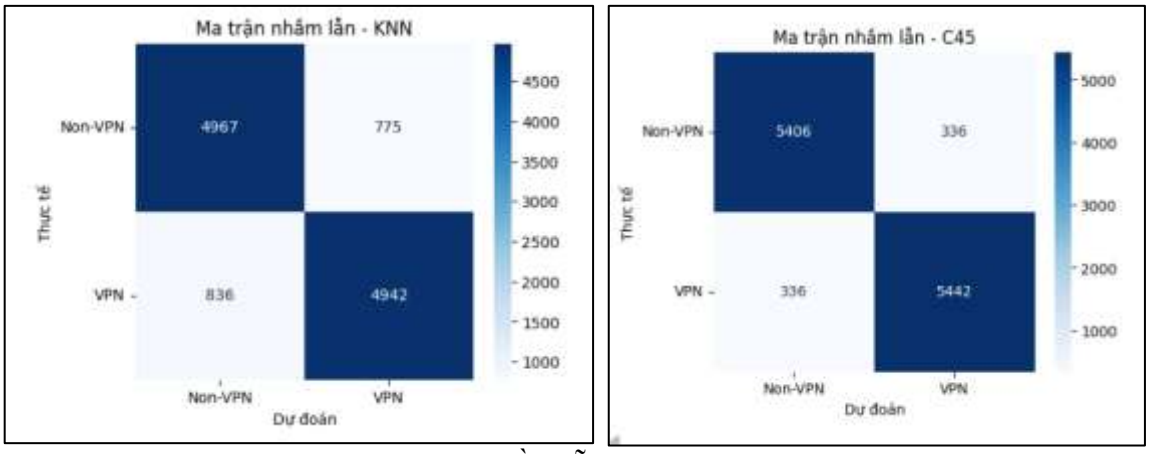
Hình 3.4 Precision đa lớp



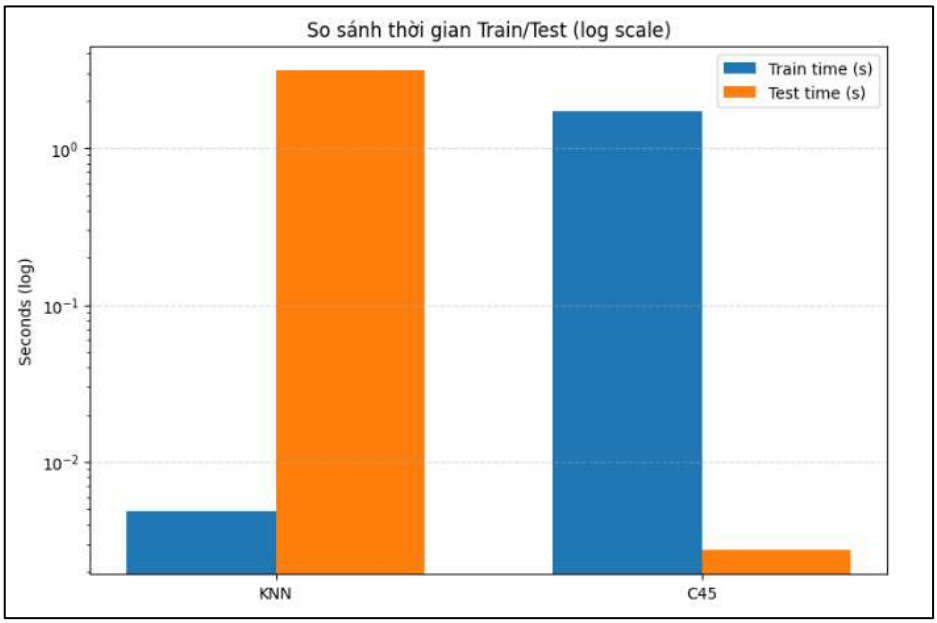
Hình 3.1 Recall đa lớp



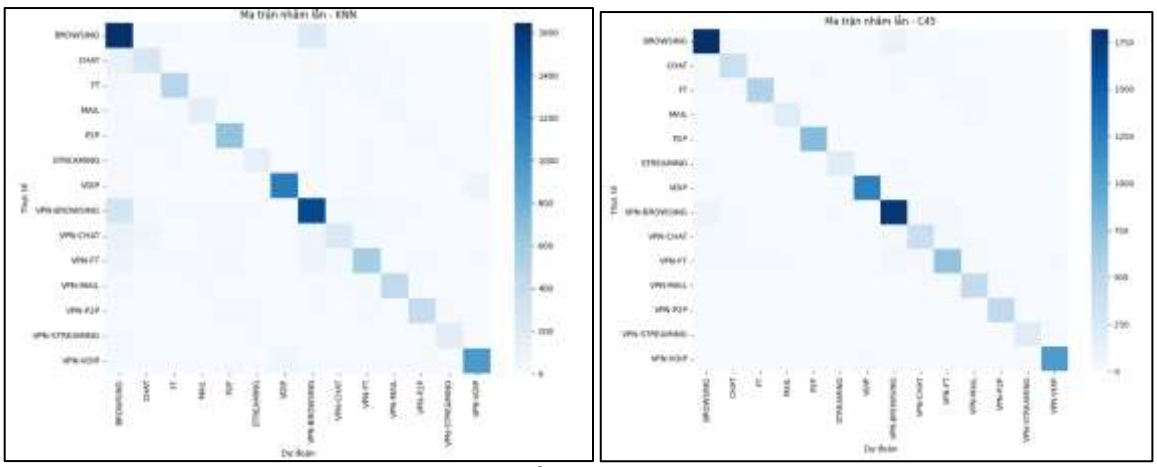
Hình 3.2 F1-score đa lớp



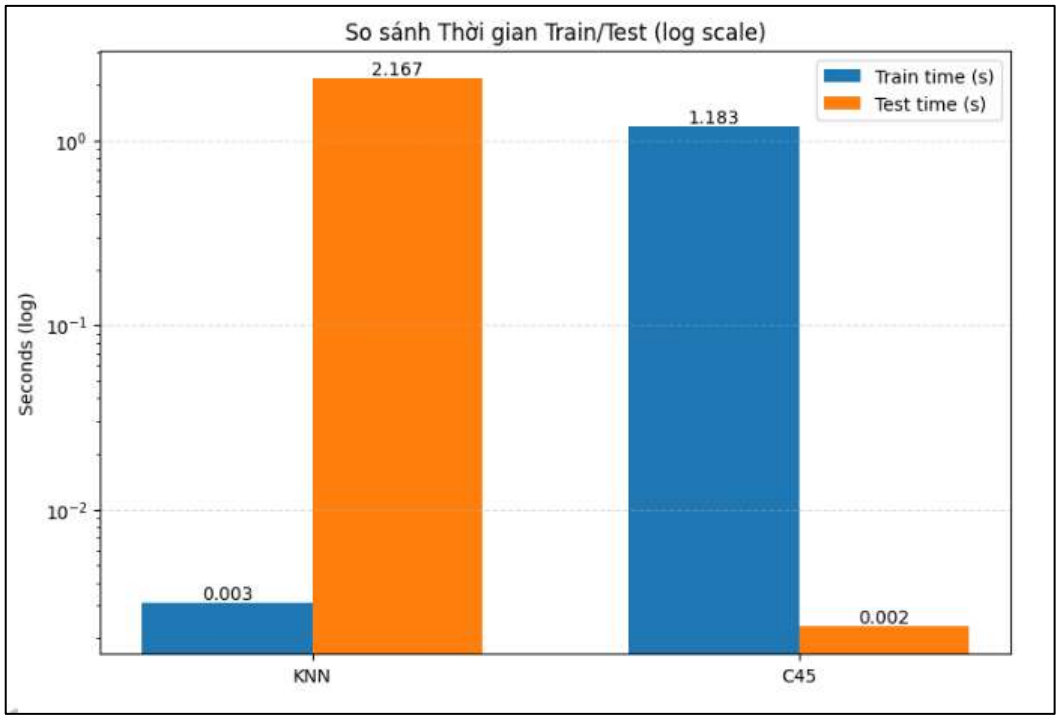
Hình 3.3 Ma trận nhầm lẫn KNN & C45 nhị phân



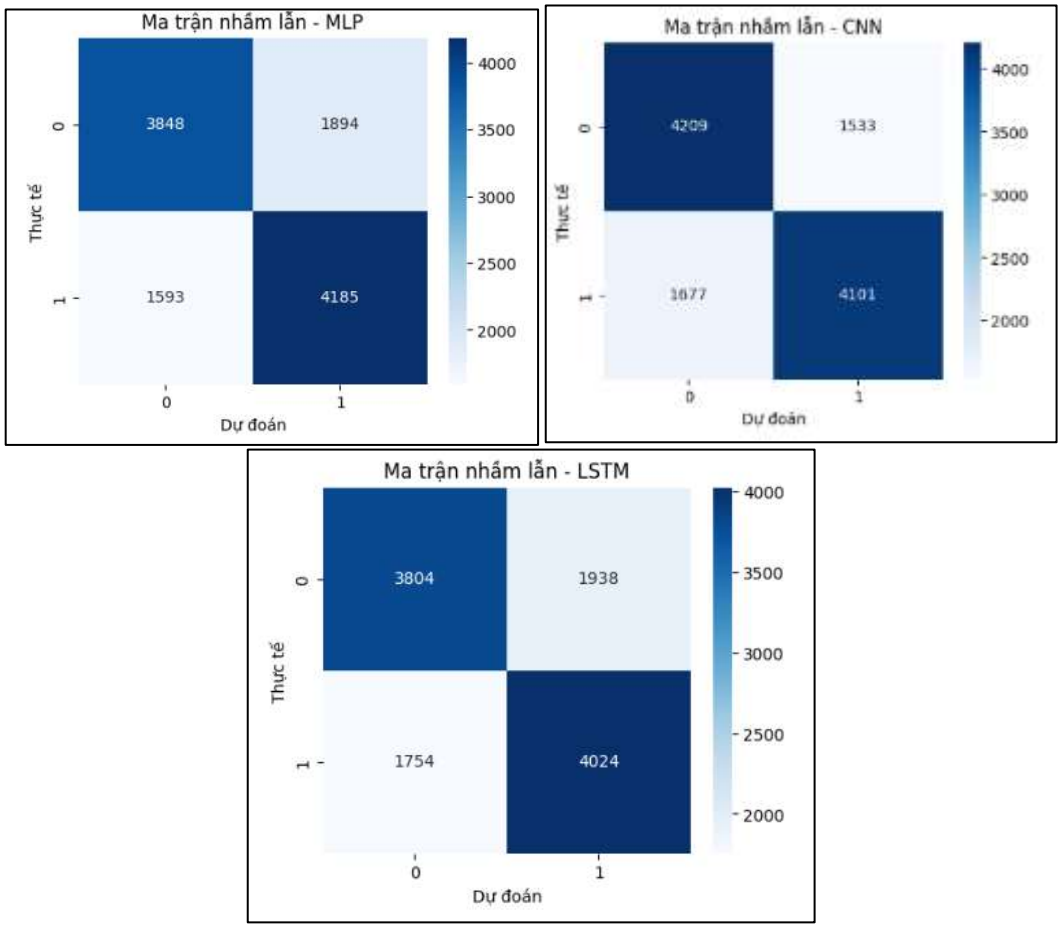
Hình 3.4 Train/Test KNN & C45 nhị phân



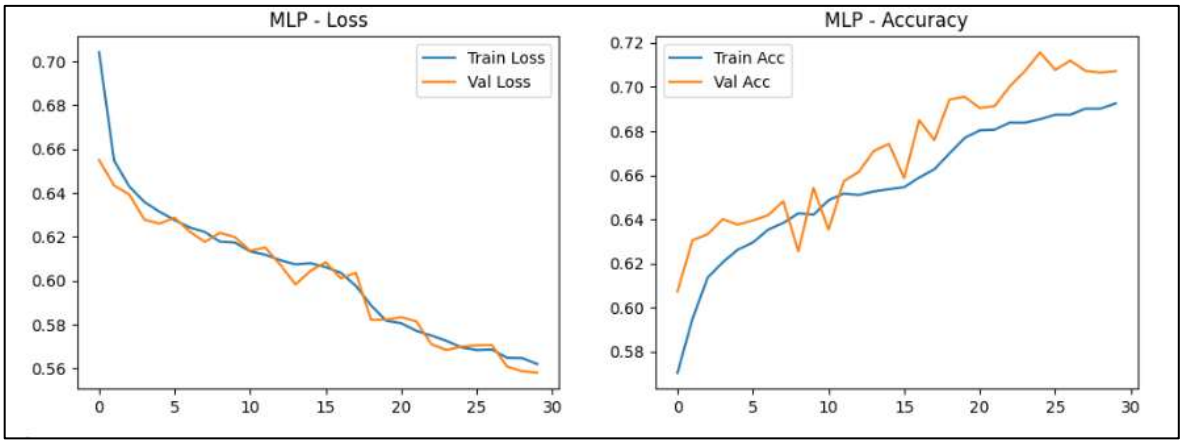
Hình 3.5 Ma trận nhầm lẫn KNN & C45 đa lớp



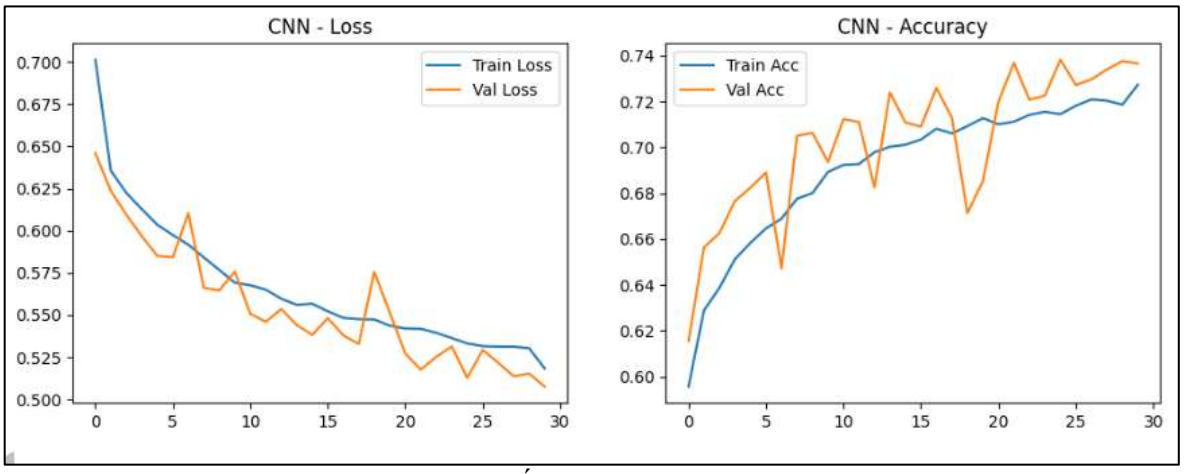
Hình 3.6 Train/Test time KNN & C45 đa lớp



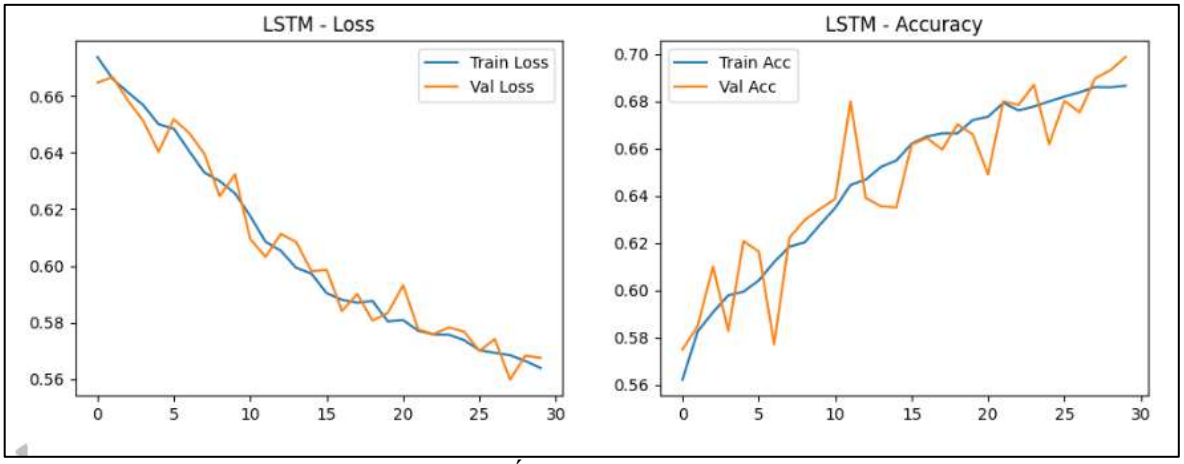
Hình 3.7 Ma trận nhầm lẫn MLP, CNN & LSTM đa lớp



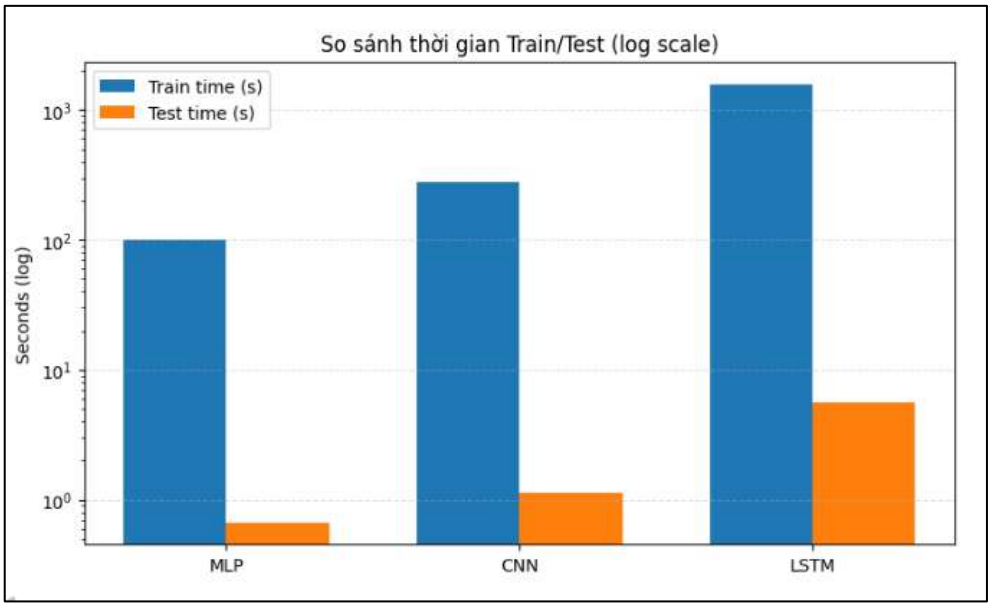
Hình 3.12 Quá trình huấn luyện mô hình MLP nhị phân



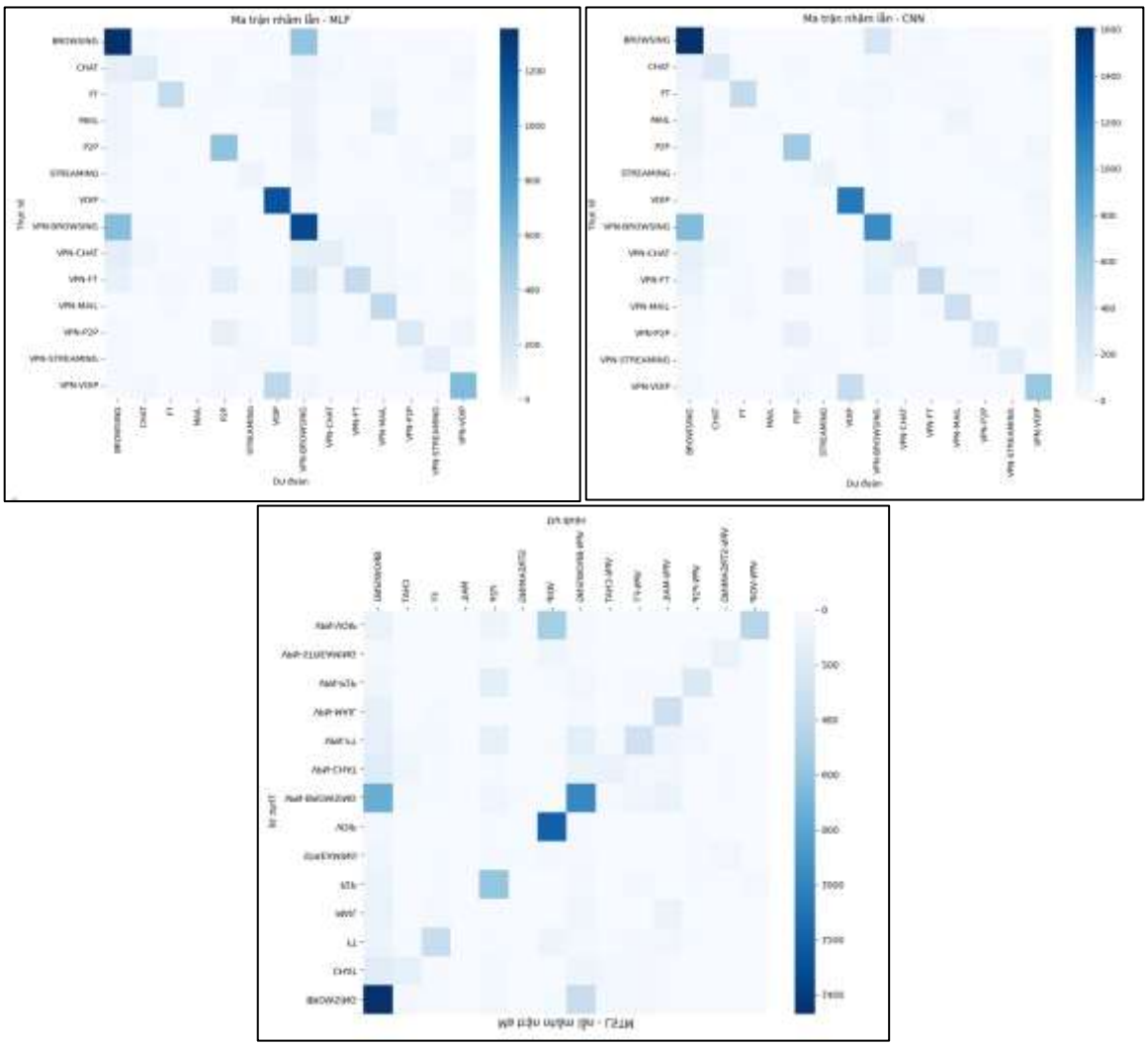
Hình 3.8 Quá trình huấn luyện mô hình CNN nhị phân



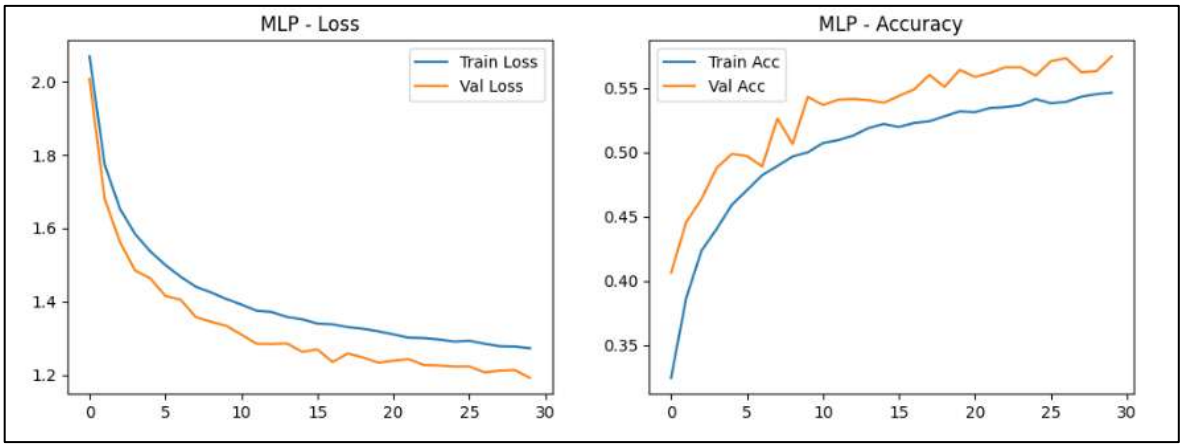
Hình 3.9 Quá trình huấn luyện mô hình LSTM nhị phân



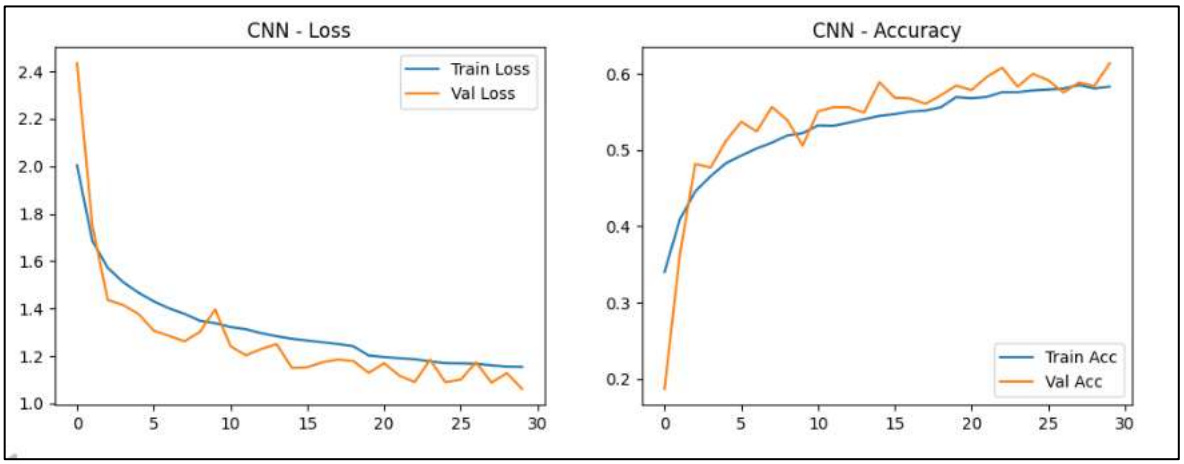
Hình 3.10 Train/Test time MLP, CNN & LSTM nhị phân



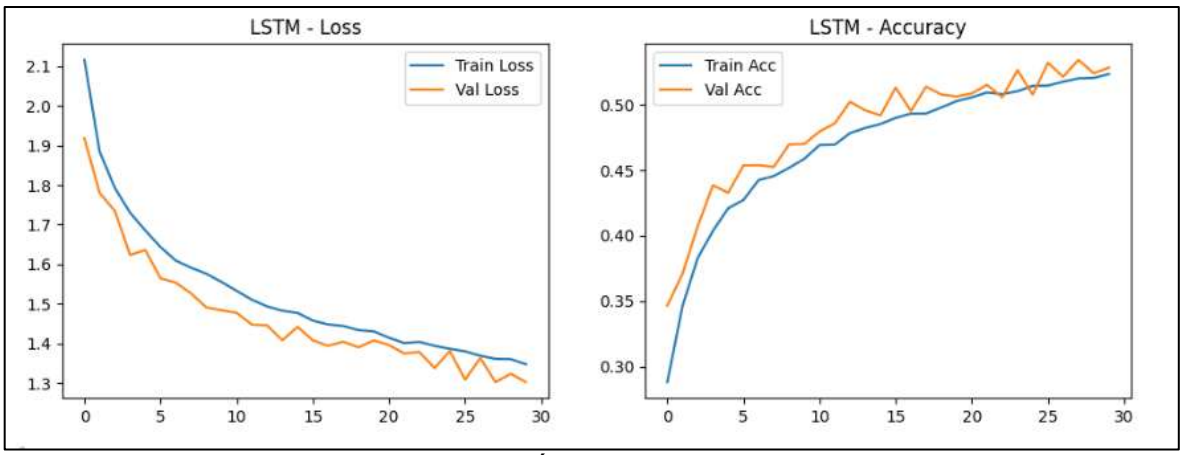
Hình 3.11 Ma trận nhầm lẫn MLP, CNN & LSTM đa lớp



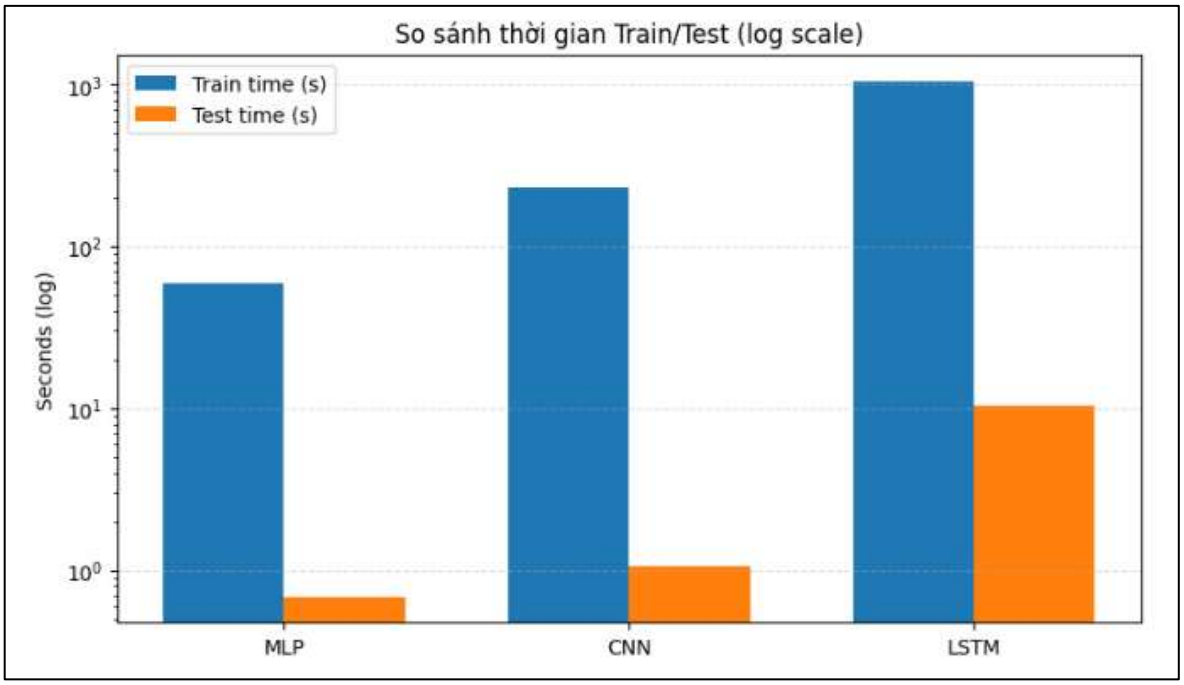
Hình 3.12 Quá trình huấn luyện mô hình MLP đa lớp



Hình 3.13 Quá trình huấn luyện mô hình CNN đa lớp



Hình 3.14 Quá trình huấn luyện mô hình LSTM đa lớp



Hình 3.15 Train/Test time MLP, CNN & LSTM đa lớp

Đánh giá

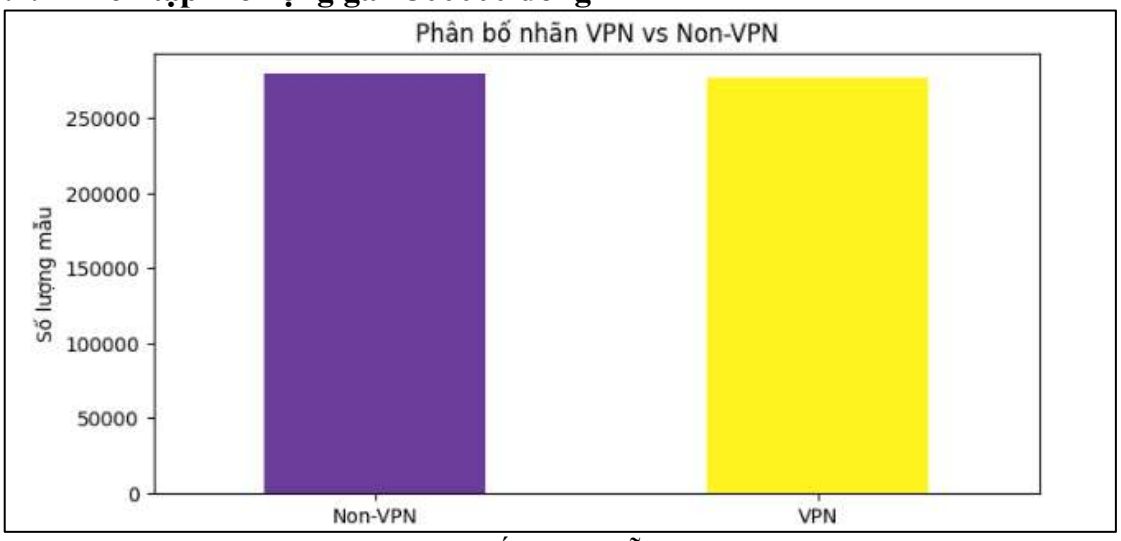
– Phân loại nhị phân:

+ Các mô hình phân loại khá tốt, đặc biệt là C45, nhưng các mô hình deep learning cho kết quả có phần kém hơn, tỷ lệ nhầm lẫn còn cao có thể do dữ liệu chưa đủ lớn để mô hình phát huy.
+ Precision - Recall - F1-score: KNN và C45 cho kết quả gần như bằng nhau, còn MLP, LSTM precision còn thấp, cho thấy các đặc trưng đủ mạnh để phân loại (C45) nhưng chưa đủ để DL học biểu diễn sâu.
+ Train/Test time: các mô hình deep learning mặc dù mất nhiều thời gian hơn nhưng kết quả lại không tương xứng.

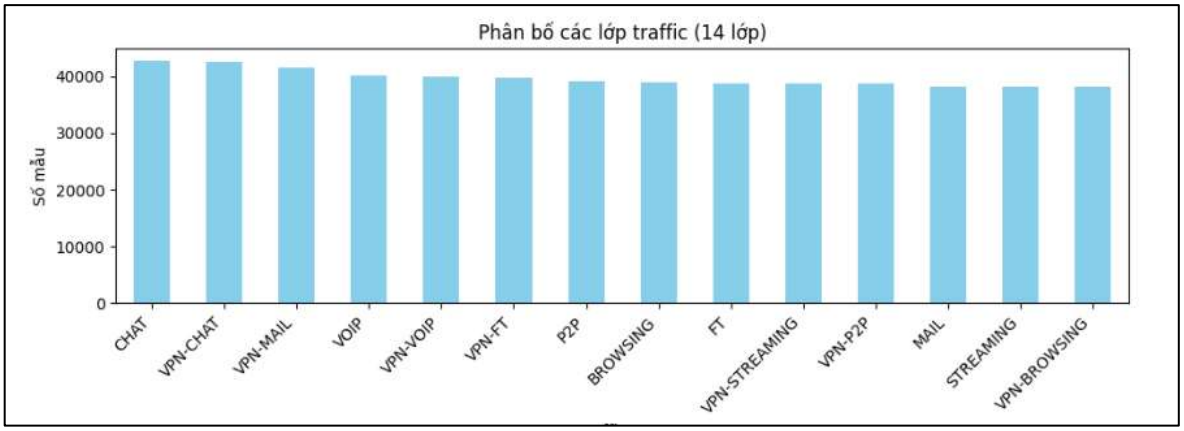
– Phân loại đa lớp:

+ Các nhãn VOIP/ VPN-VOIP, P2P/VPN-P2P, FT/VPN-FT dễ phân loại hơn do các đặc trưng rõ ràng, còn CHAT/MAIL, VPN-CHAT/VPN-MAIL khó phân loại dễ nhầm lẫn với nhau.
+ C45 cho F1-score cao và ổn định, ma trận nhầm lẫn có đường chéo rõ ràng, các mô hình DL giữa các kết quả biến động mạnh đặc biệt là LSTM.
+ Quá trình huấn luyện của các mô hình DL: loss MLP và CNN giảm khá nhanh giai đoạn đầu sau đó giảm chậm và ổn định, LSTM giảm chậm hơn, cần nhiều epoch hơn để ổn định, accuracy ở cả 3 đều tăng nhưng CNN có train và validation gần nhau tổng quát hóa tốt hơn nhưng kết quả cả 3 vẫn thấp hơn ML.

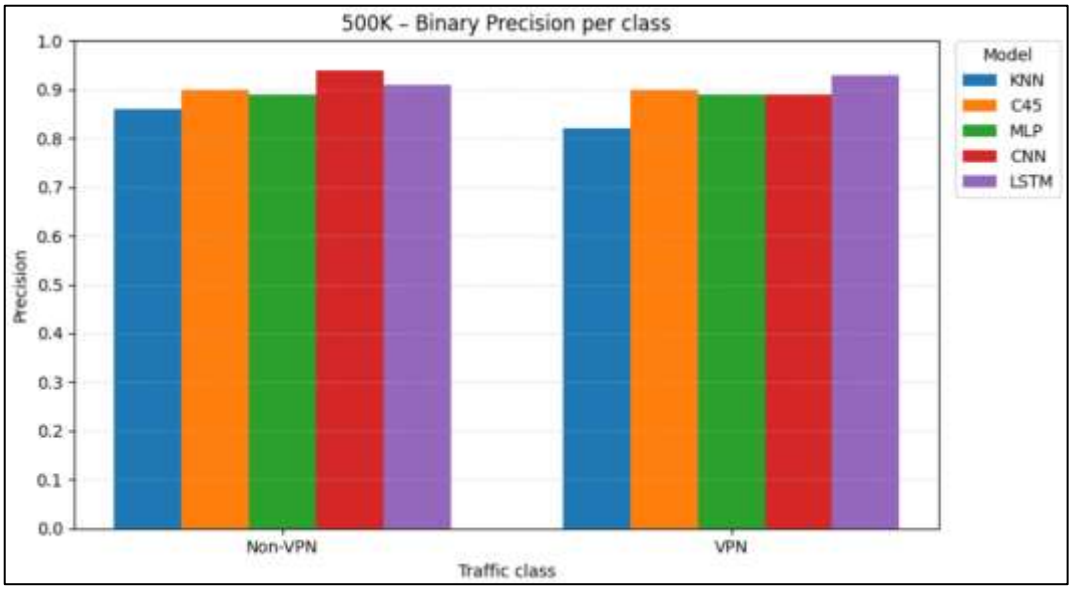
3.2.2 Trên tập mở rộng gần 500000 dòng



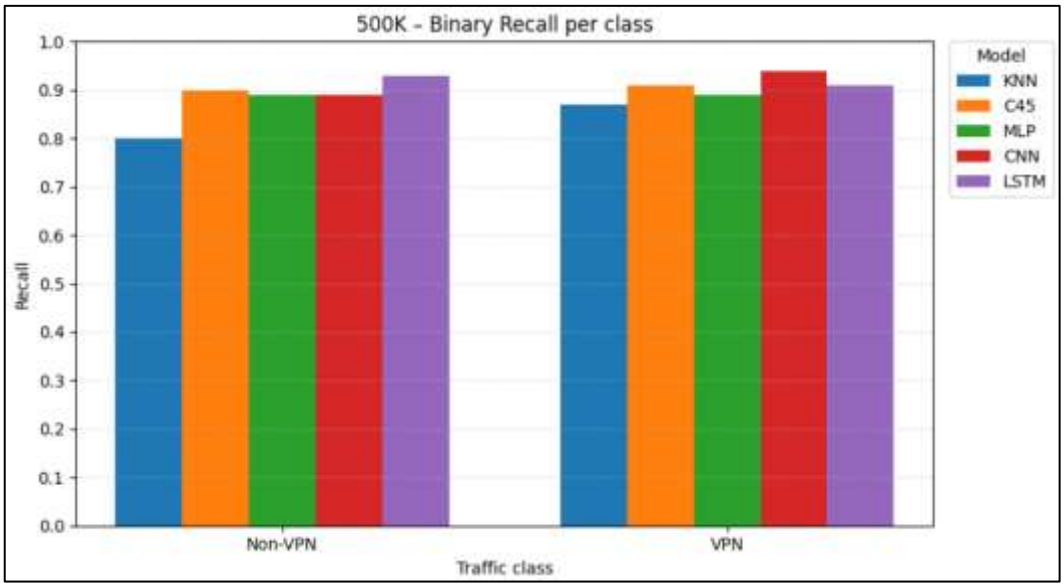
Hình 3.21 Số lượng mẫu nhị phân



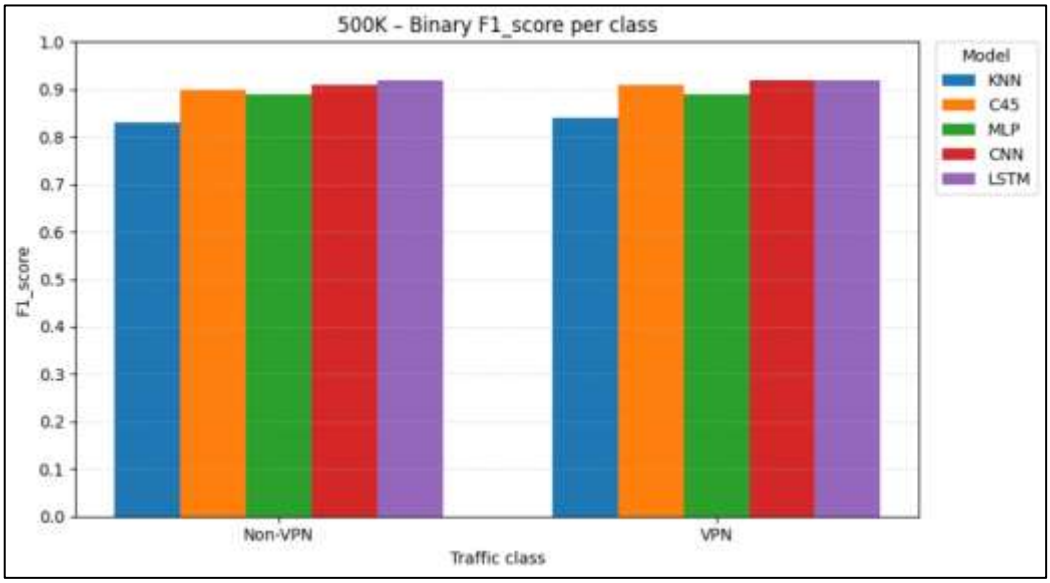
Hình 3.22 Số lượng mẫu đa lớp



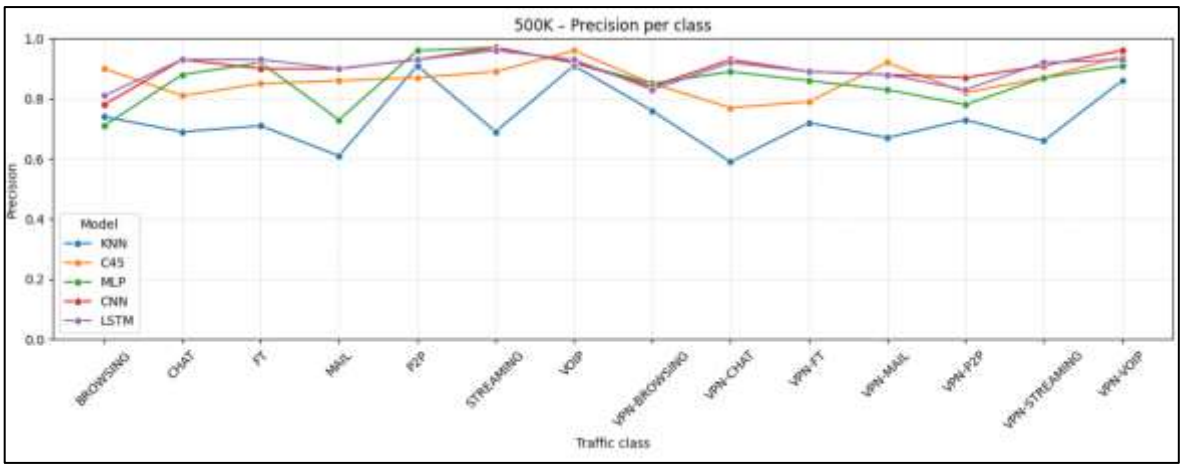
Hình 3.23 Precison nhị phân



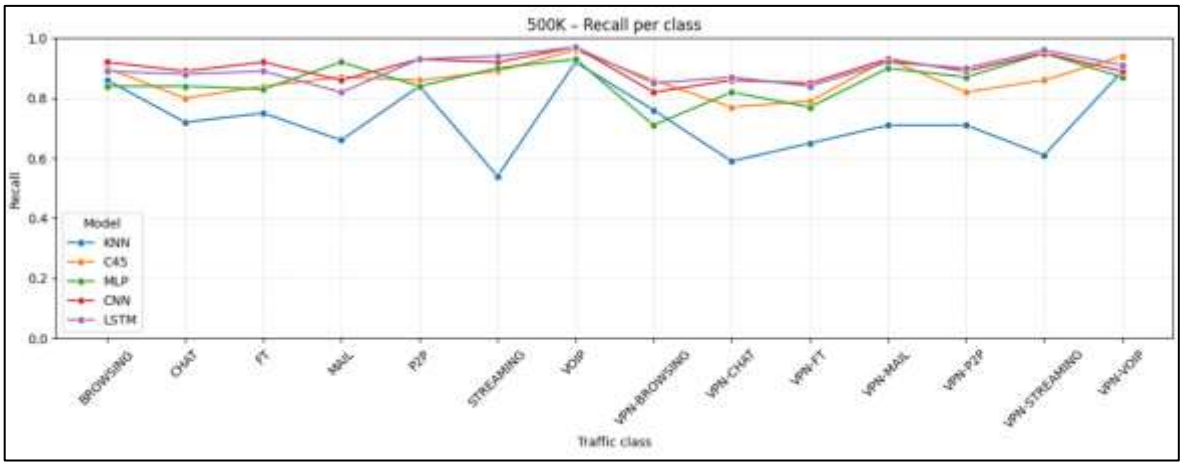
Hình 3.24 Recall nhị phân



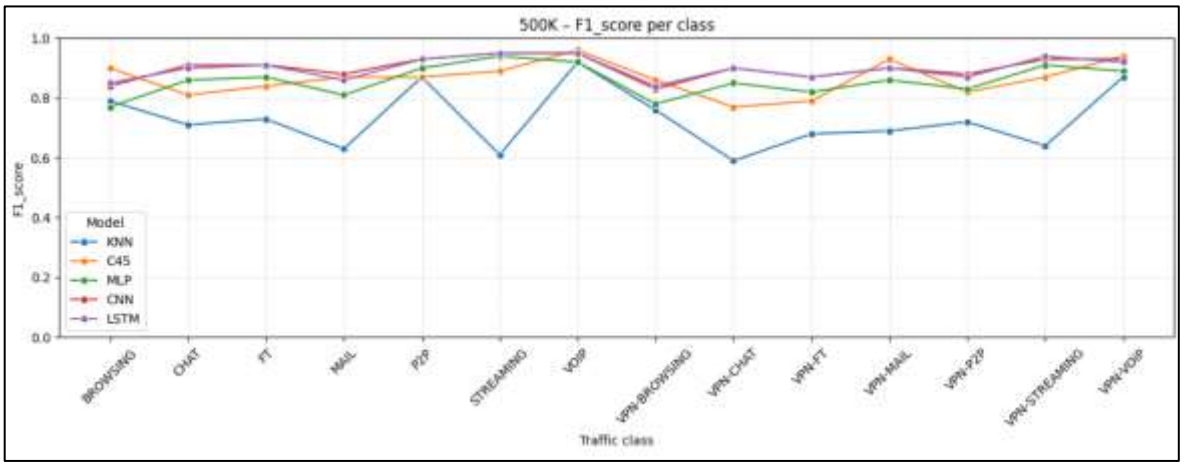
Hình 3.25 F1-score nhị phân



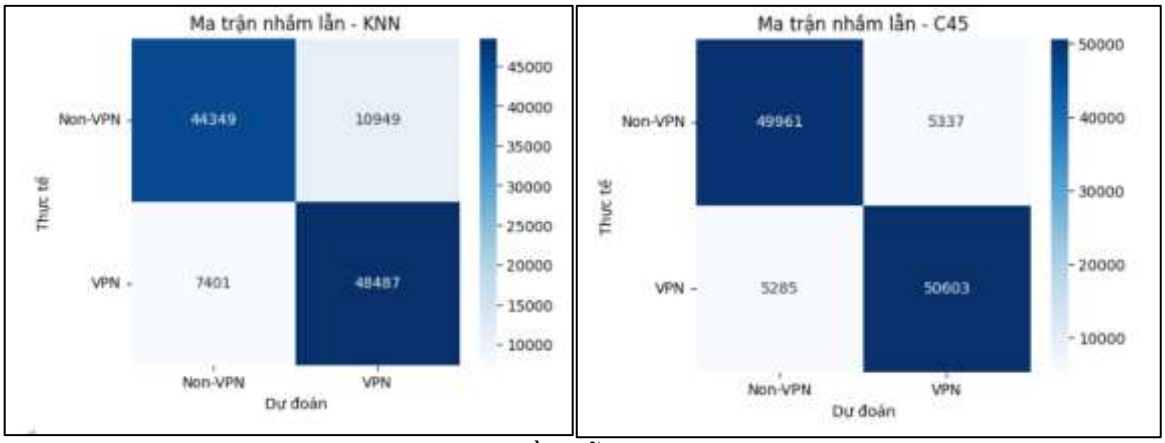
Hình 3.26 Precison đa lớp



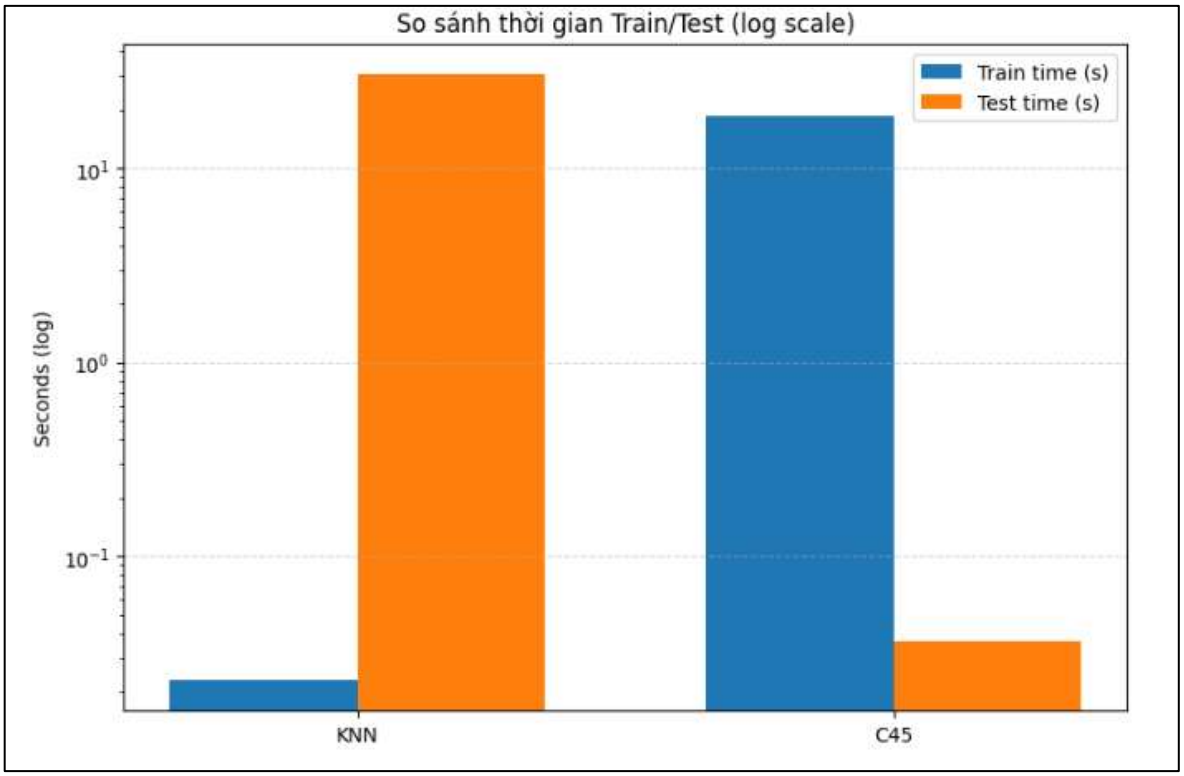
Hình 3.27 Recall đa lớp



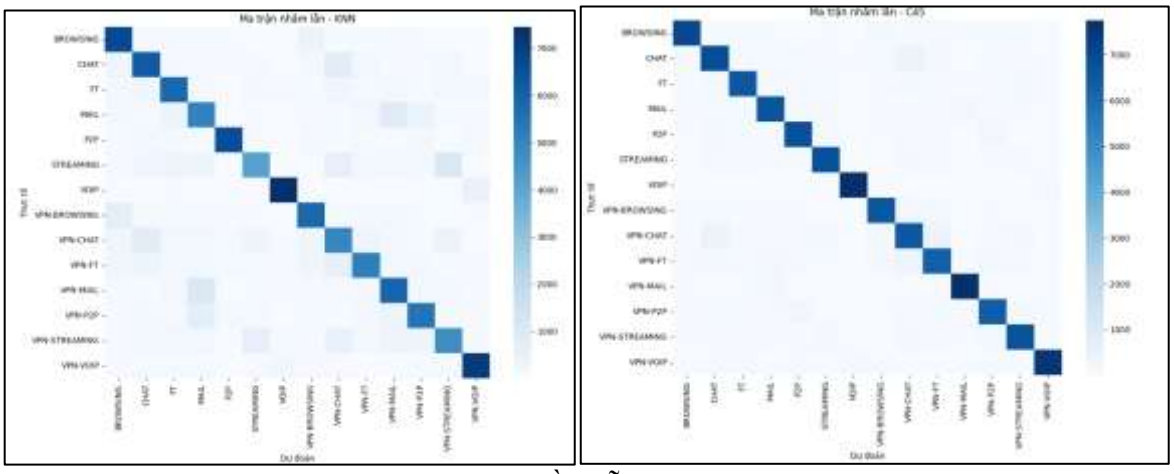
Hình 3.28 F1-score đa lớp



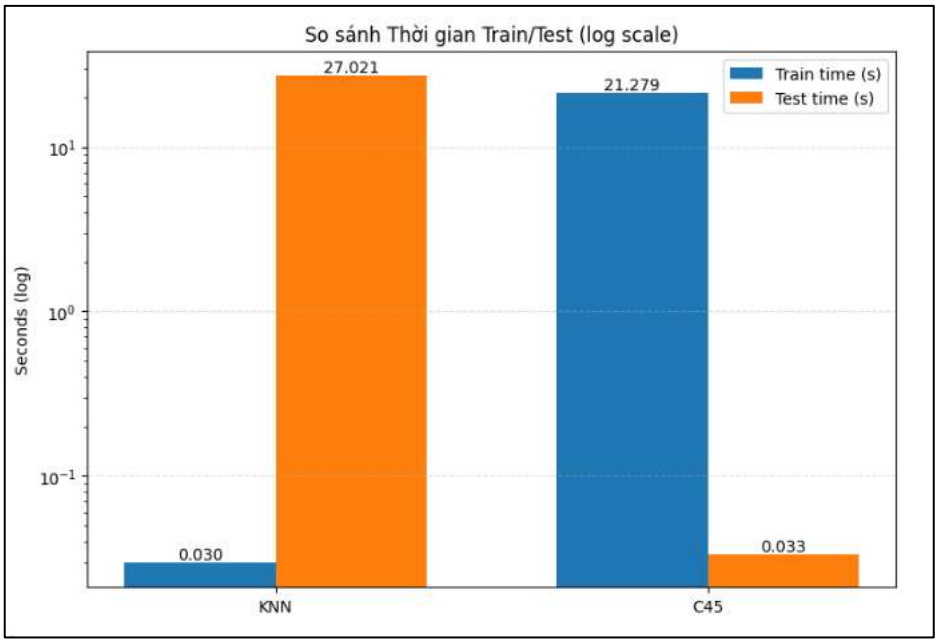
Hình 3.29 Ma trận nhầm lẫn KNN & C45 nhị phân



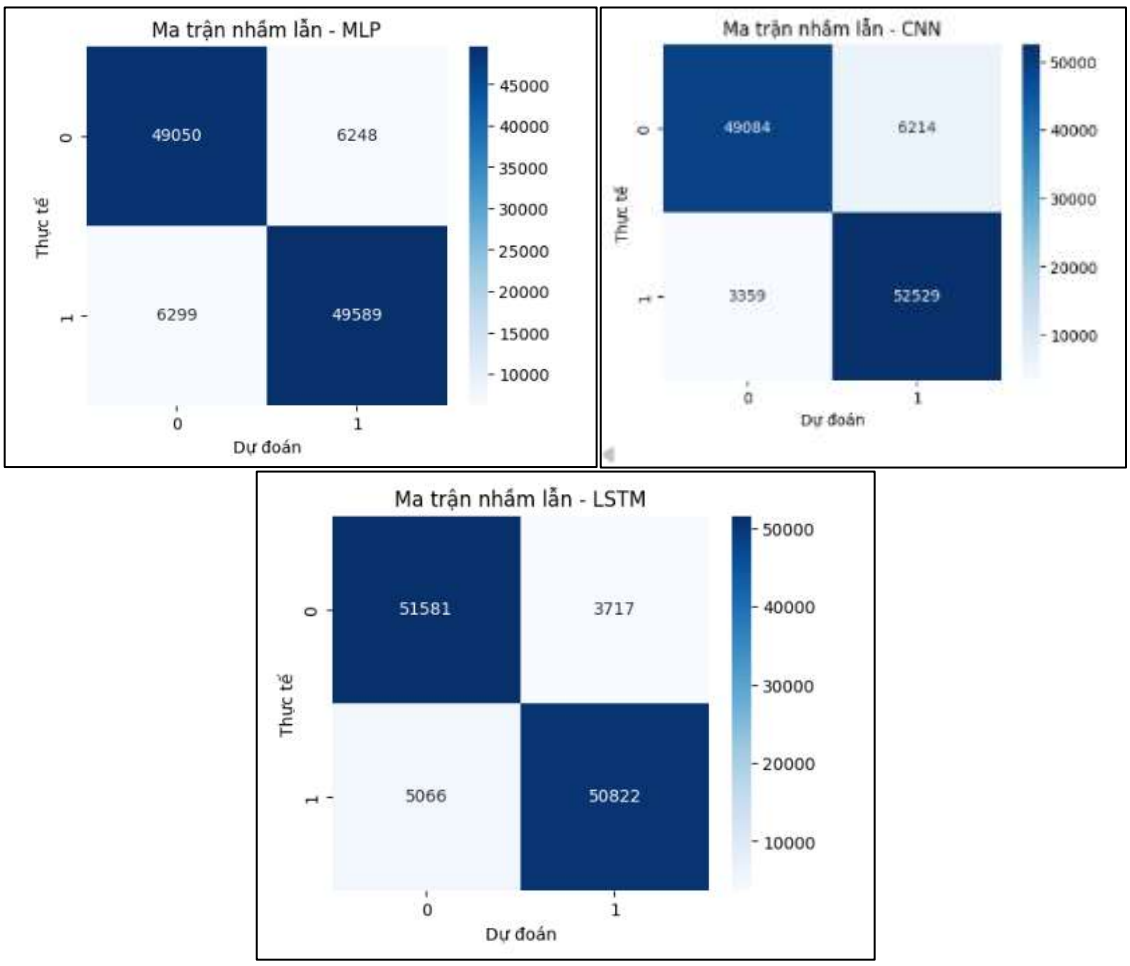
Hình 3.30 Train/Test time KNN & C45 nhị phân



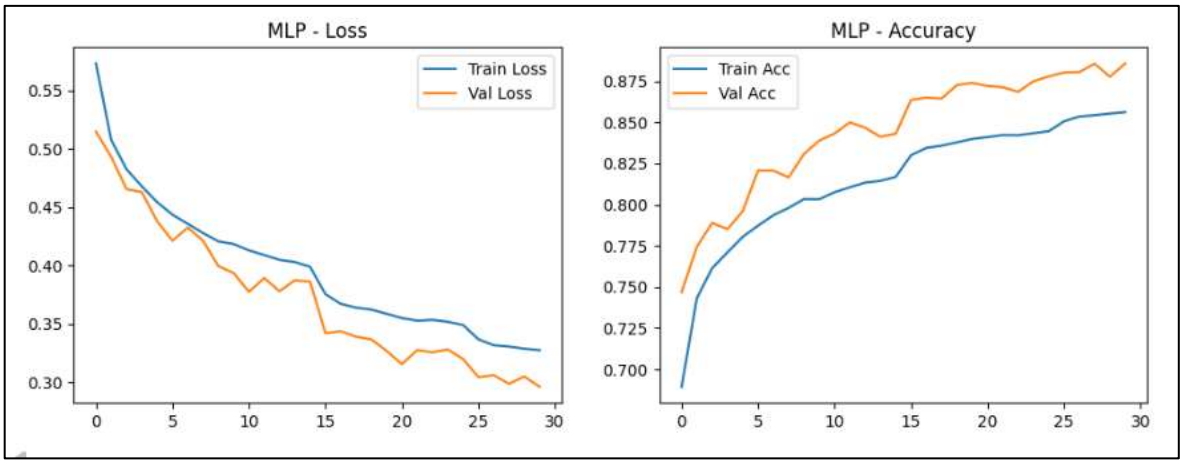
Hình 3.31 Ma trận nhầm lẫn KNN & C45 đa lớp



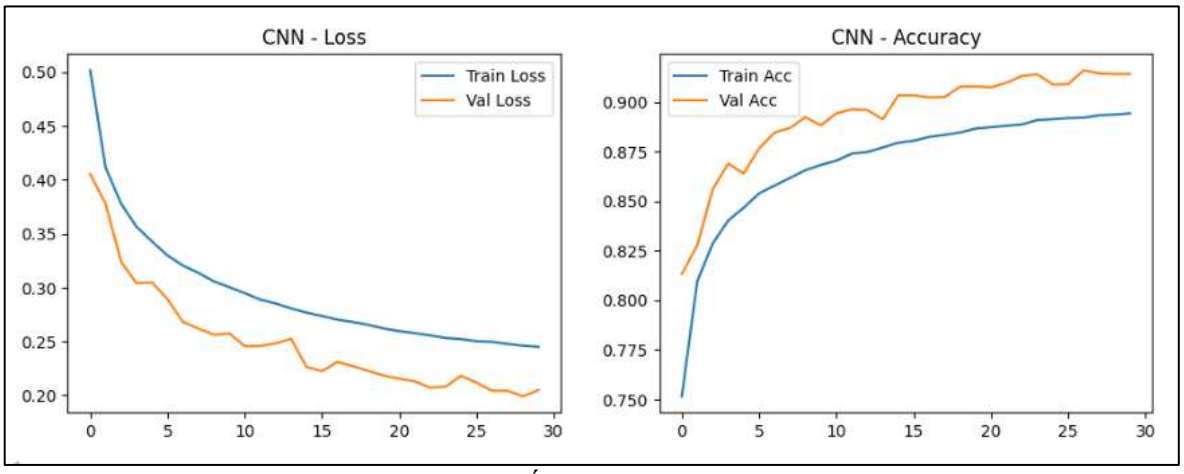
Hình 3.32 Train/Test time KNN & C45 đa lớp



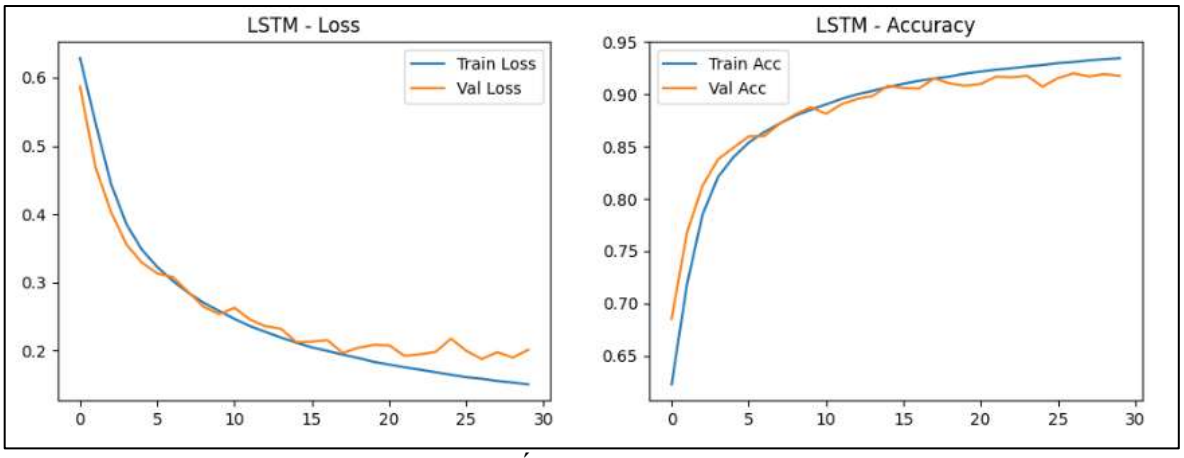
Hình 3.33 Ma trận nhầm lẫn MLP, CNN, LSTM nhị phân



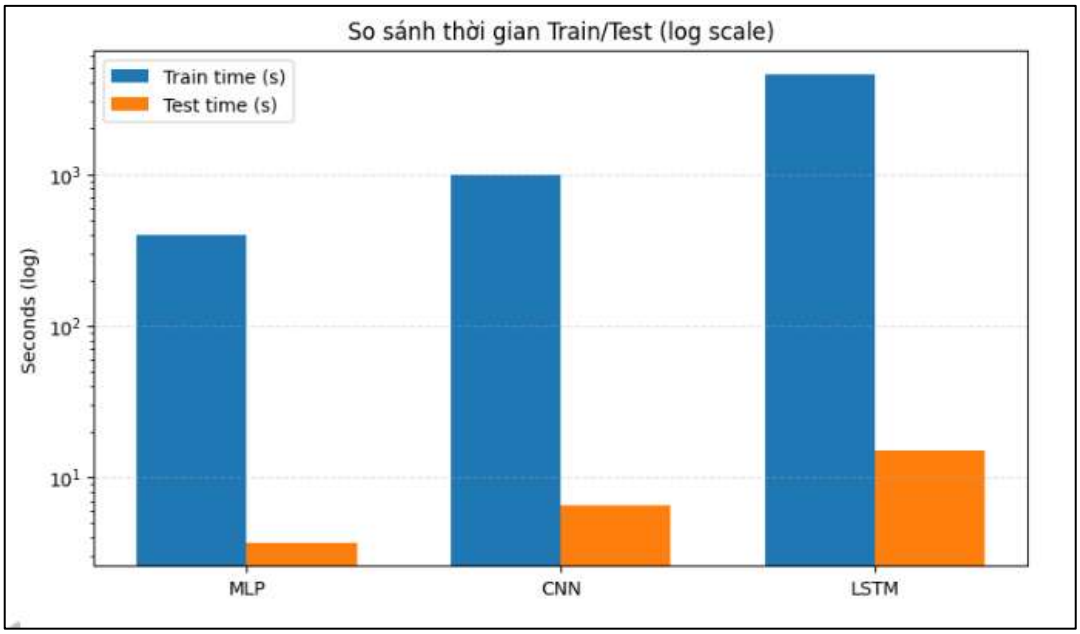
Hình 3.34 Quá trình huấn luyện mô hình MLP nhị phân



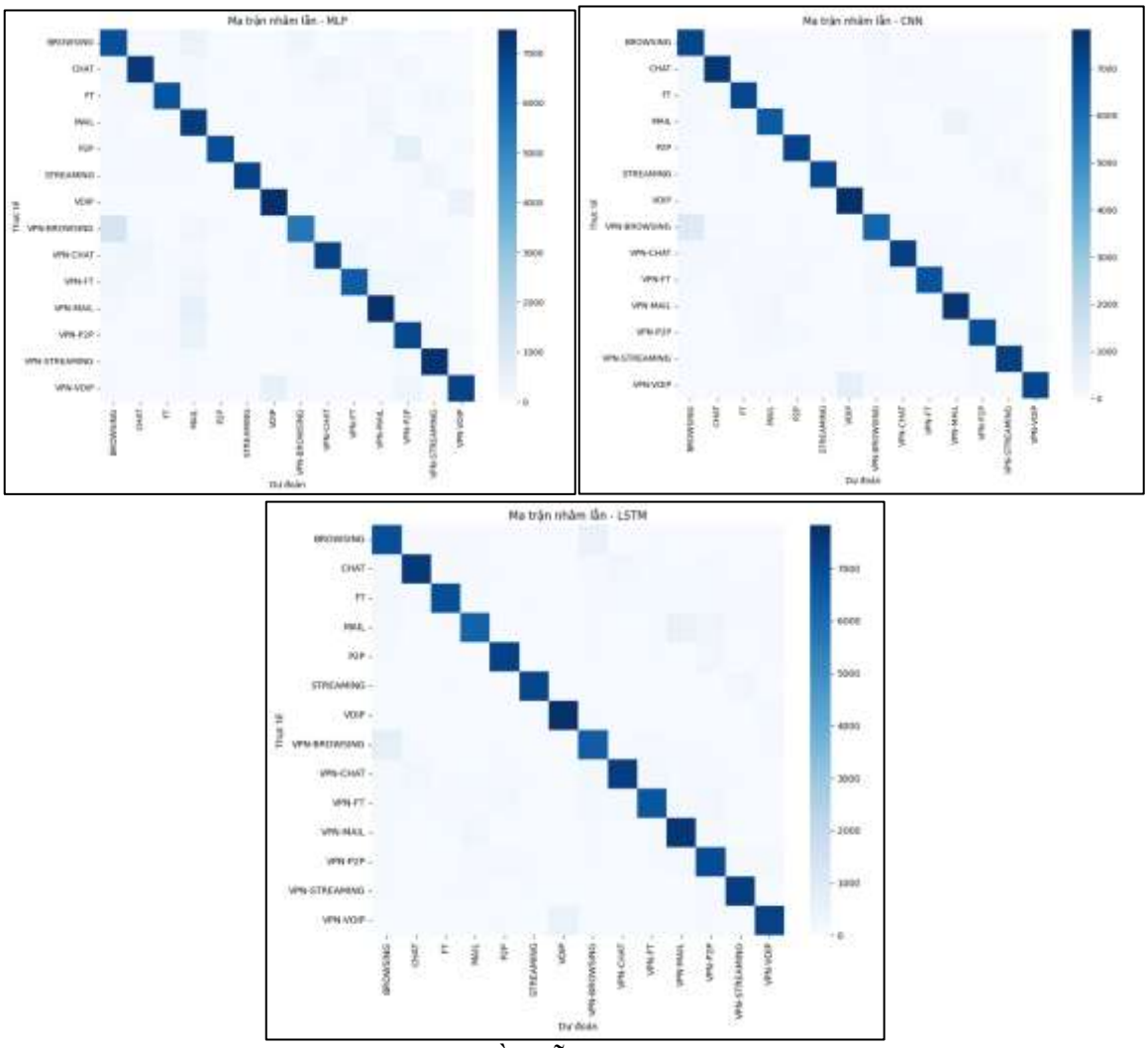
Hình 3.35 Quá trình huấn luyện mô hình CNN nhị phân



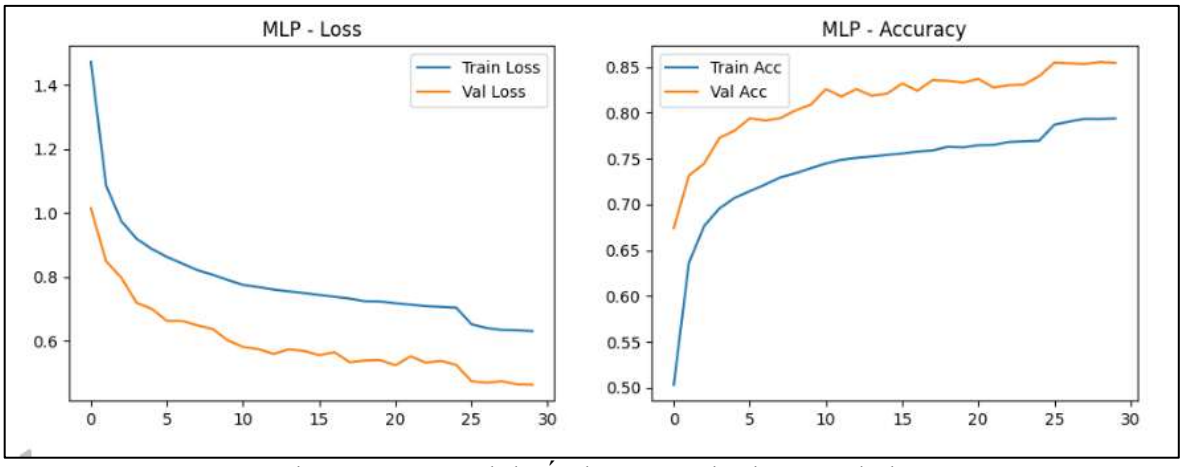
Hình 3.36 Quá trình huấn luyện mô hình LSTM nhị phân



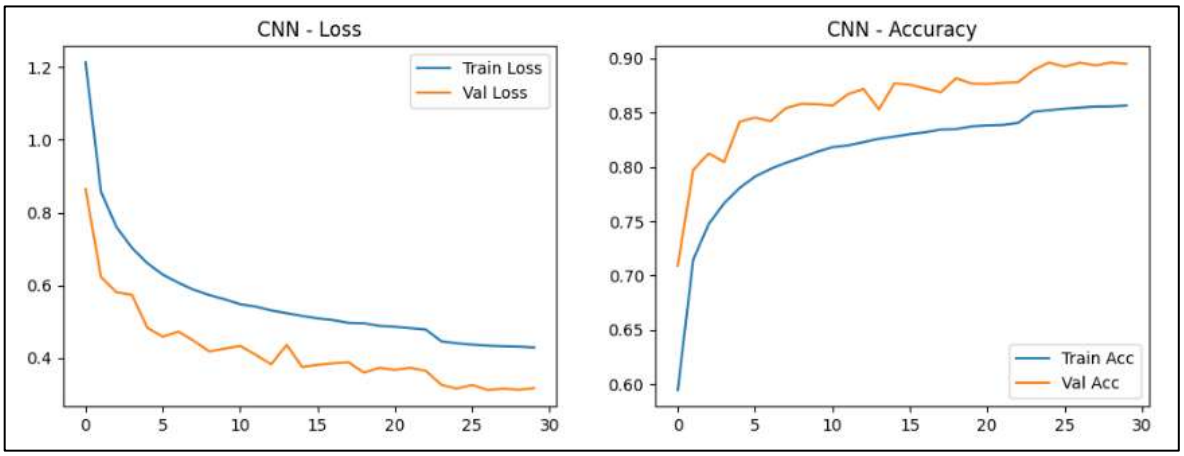
Hình 3.37 Train/Test time MLP, CNN, LSTM nhị phân



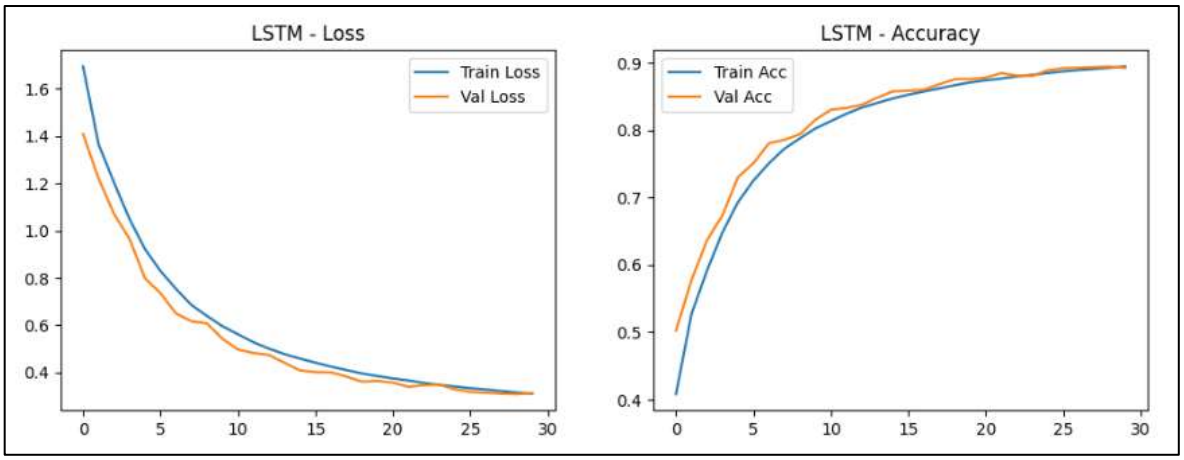
Hình 3.38 Ma trận nhầm lẫn MLP, CNN, LSTM đa lớp



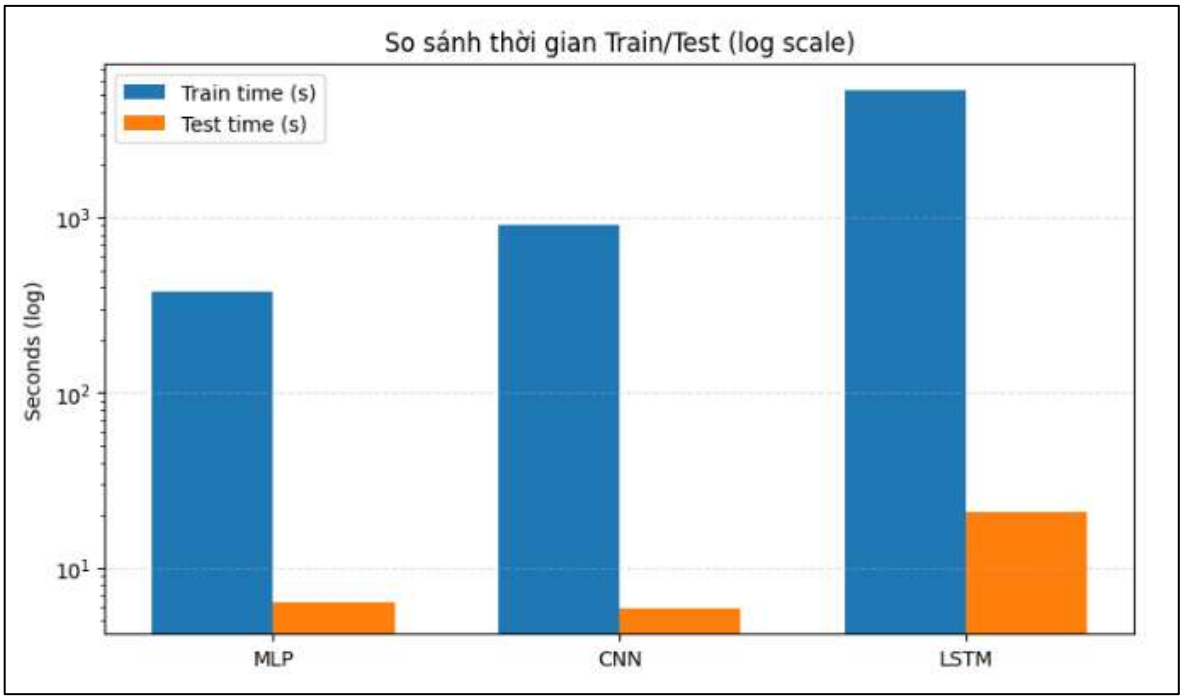
Hình 3.39 Quá trình huấn luyện mô hình MLP đa lớp



Hình 3.40 Quá trình huấn luyện mô hình CNN đa lớp



Hình 3.41 Quá trình huấn luyện mô hình LSTM đa lớp



Hình 3.42 Train/Test time MLP, CNN, LSTM đa lớp

Đánh giá

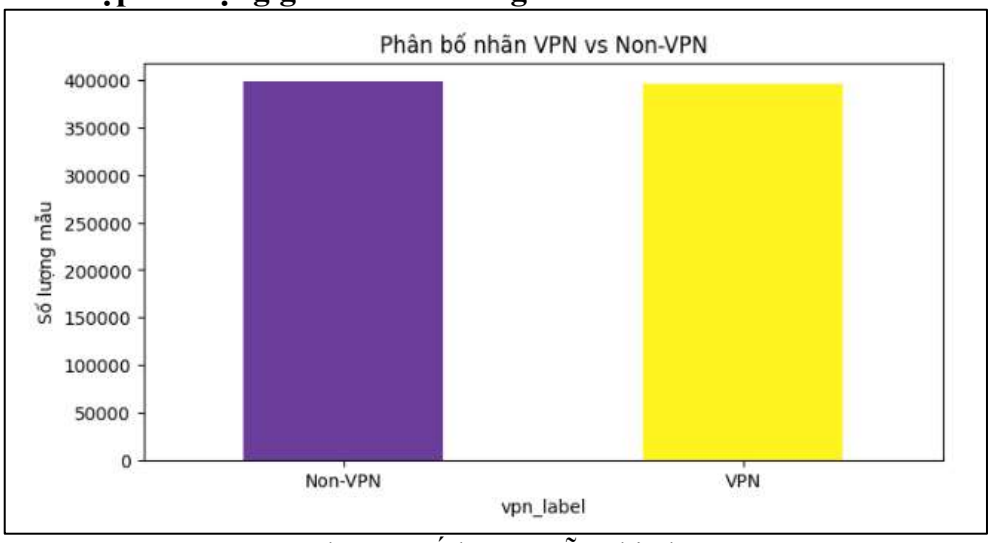
– Phân loại nhị phân:

+ Precision – Recall – F1-score:
  + C45 có Pre cao nhất cho cả hai lớp, CNN và LSTM có Pre VPN cao hơn Non-VPN, MLP ổn định nhưng thấp hơn CNN, LSTM.
  + Rec của CNN, LSTM trên VPN rất cao, C45 cân bằng giữa hai lớp, MLP tương đối cân bằng nhưng thấp hơn CNN, LSTM.
  + Các mô hình DL đã cải thiện rõ rệt.
+ Đối với ma trận nhầm lẫn CNN phân biệt VPN rất tốt, LSTM cân bằng tốt, ít chênh lệch giữa hai lớp, C45 phân biệt rõ ranh giới.
+ Quá trình huấn luyện: MLP, CNN Train và Val loss giảm đều, không cách xa, overfit không lớn, LSTM giảm và hội tụ rất nhanh, ở CNN, MLP Val acc lớn hơn Train acc cho thấy dữ liệu đủ đa dạng, LSTM Val acc luôn bám sát Train acc, dao động rất nhỏ đây là biểu đồ đẹp nhất trong ba mô hình DL
+ Train/Test time của DL cao hơn ML nhiều nhưng lại cho kết quả khá tốt.

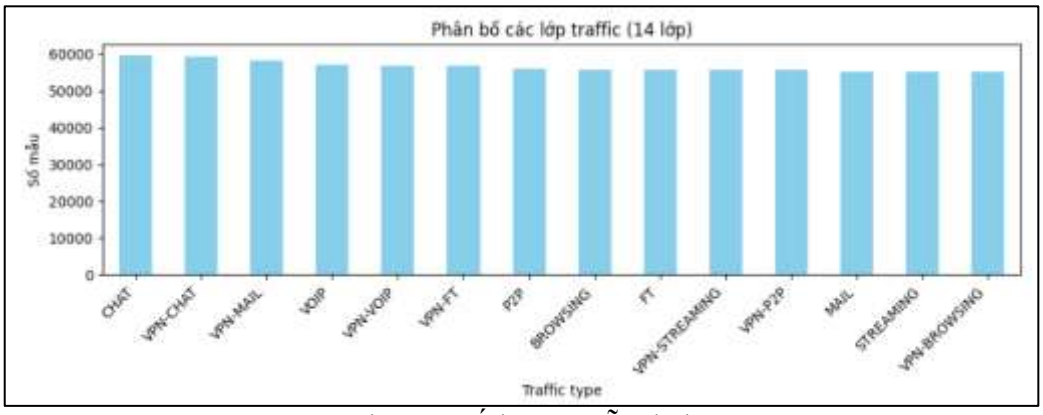
– Phân loại đa lớp:

+ Precision – Recall – F1-score: C45 cao và đồng đều hầu hết trên các lớp, đường chéo ma trận nhầm lẫn rất đậm, CNN & LSTM tiệm cận hoặc ngang C45 ở nhiều lớp, đường chéo trong ma trận nhầm lẫn rõ, một số ít còn nhầm nhẹ ở lớp VPN cùng loại, MLP đã tốt hơn rất nhiều và hơn KNN nhưng thấp hơn CNN, LSTM ở các lớp phức tạp.
+ Ở các lớp khó phân loại như CHAT/MAIL và khi qua VPN thì CNN, LSTM lại phân loại tốt hơn ML.
+ Train và Val loss giảm đều, không tách xa nhau, không có dấu hiệu overfit hay underfit, CNN hội tụ nhanh, LSTM chậm hơn CNN, MLP chậm hơn LSTM.

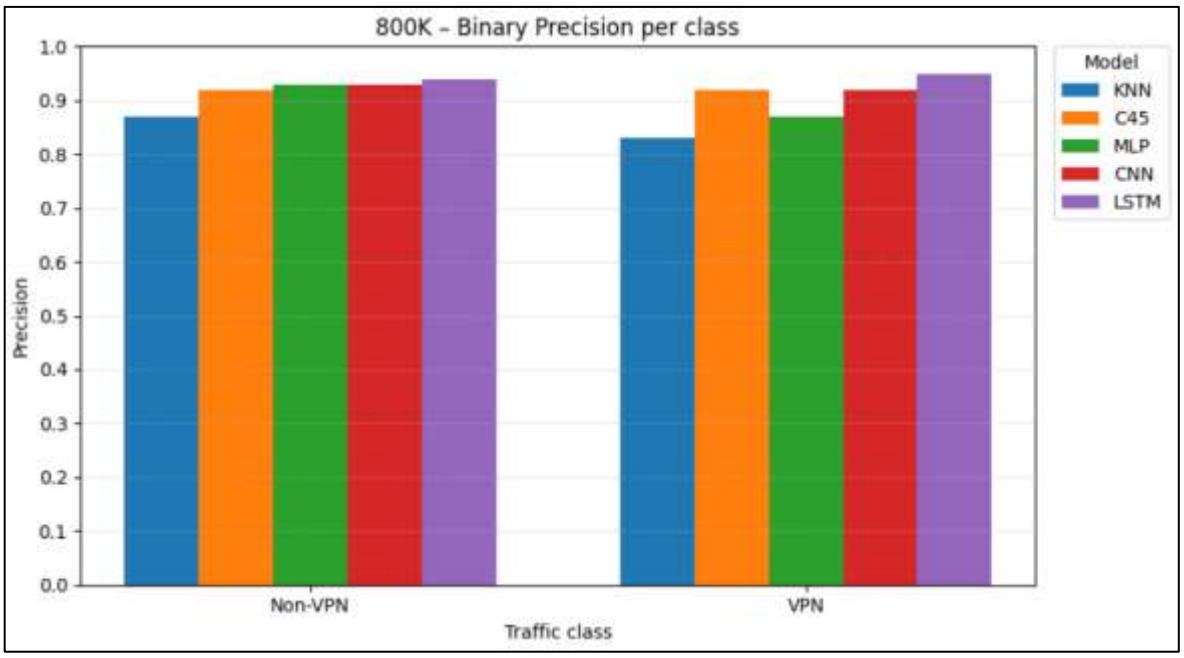
3.2.3 Trên tập mở rộng gần 800000 dòng



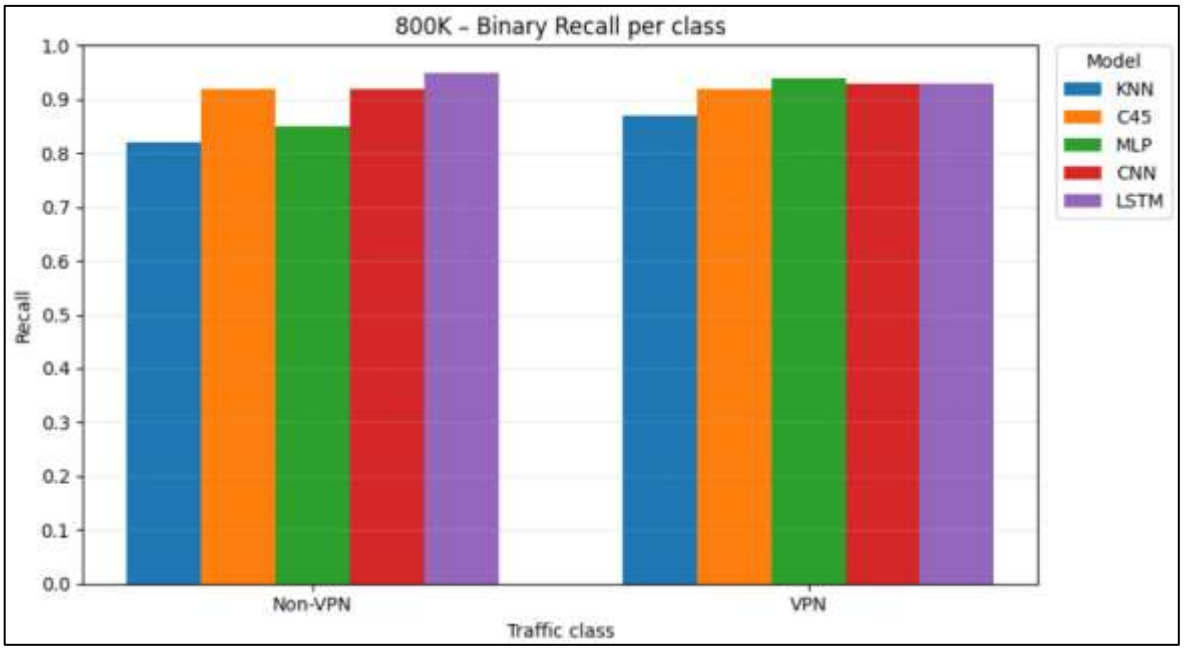
Hình 3.43 Số lượng mẫu nhị phân



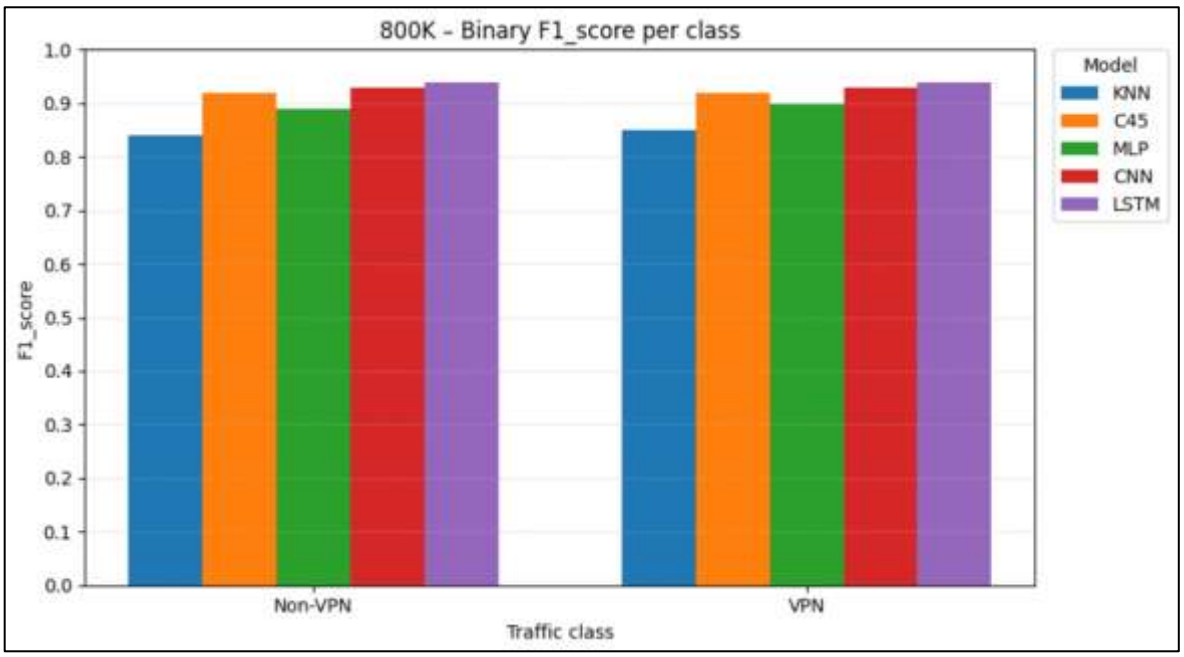
Hình 3.44 Số lượng mẫu đa lớp



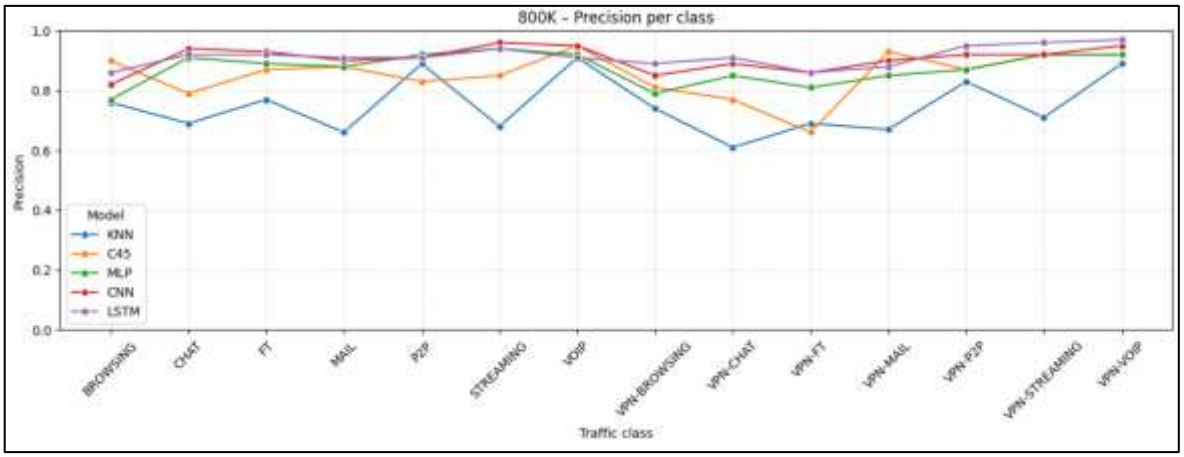
Hình 3.45 Precision nhị phân



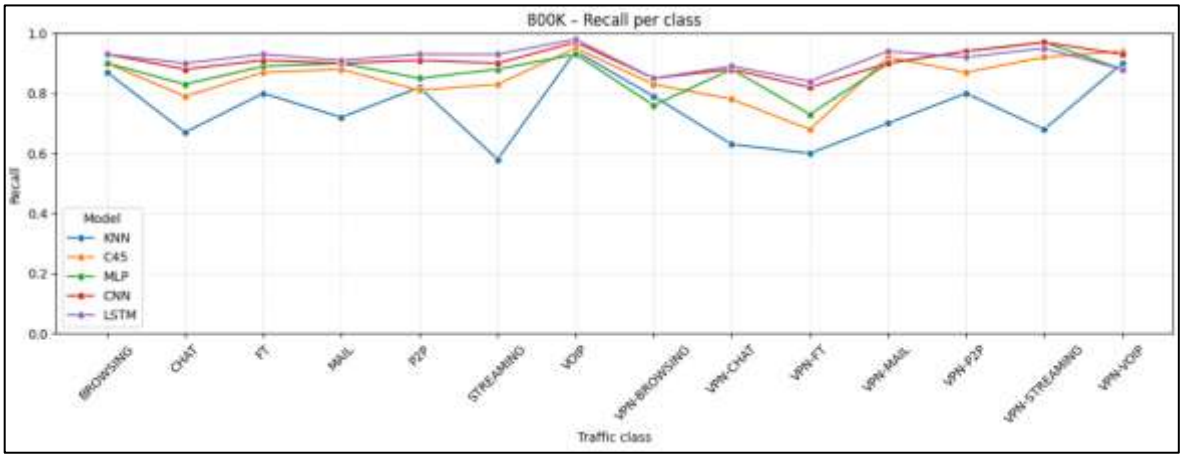
Hình 3.46 Recall nhị phân



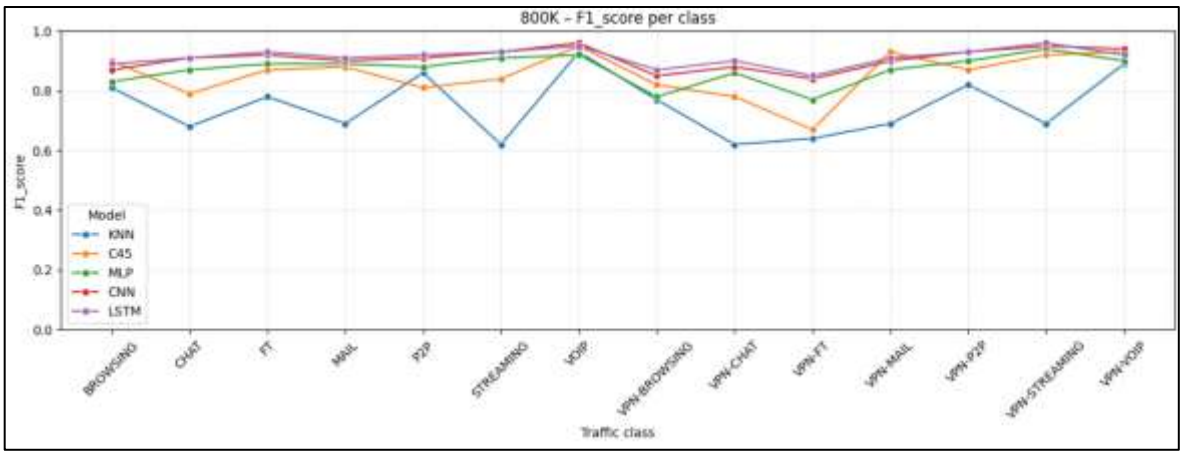
Hình 3.47 F1-score nhị phân



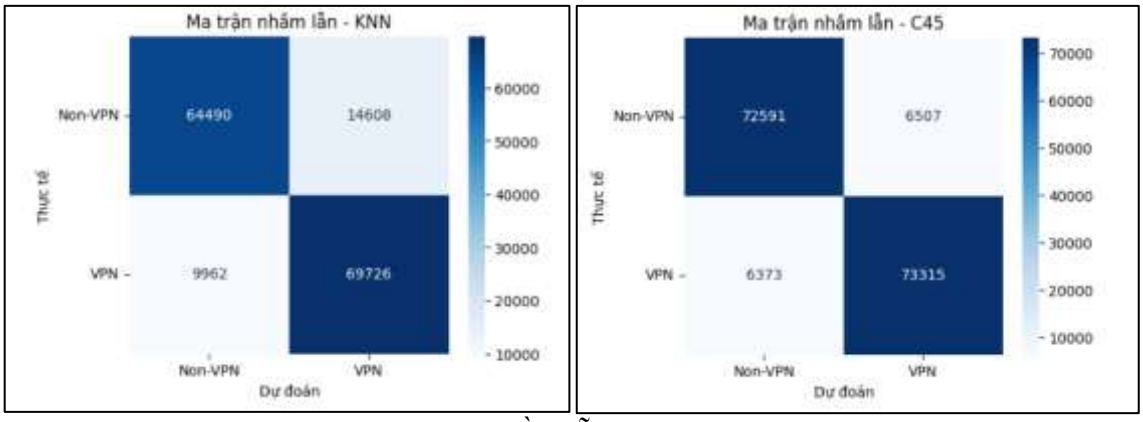
Hình 3.48 Precision đa lớp



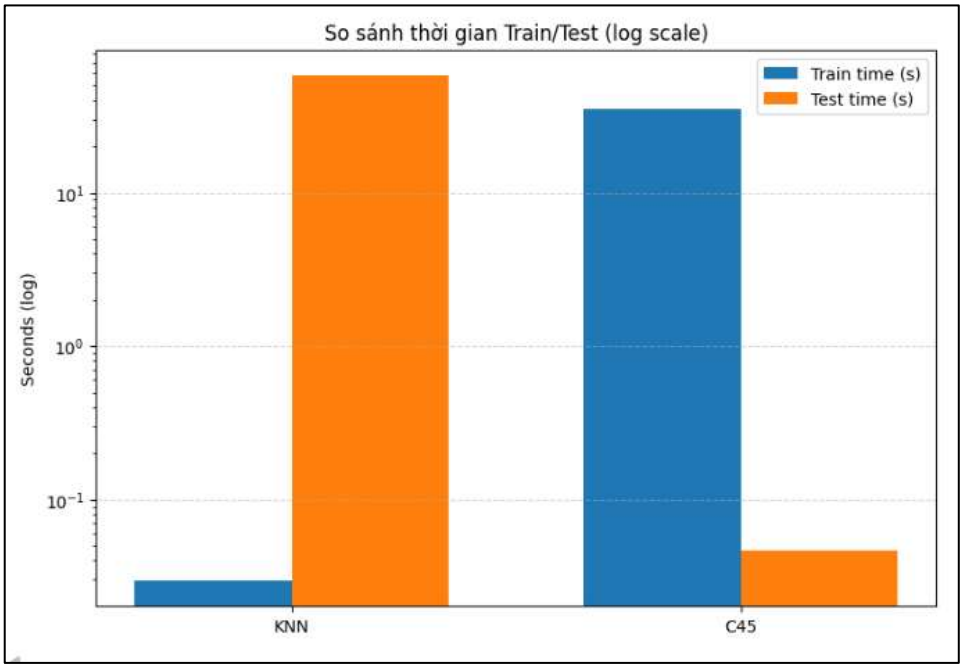
Hình 3.49 Recall đa lớp



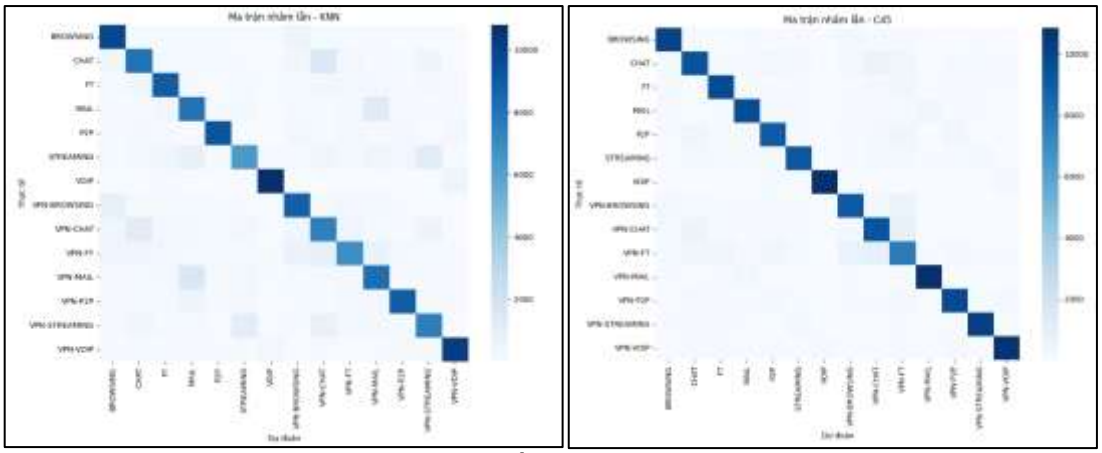
Hình 3.50 F1-score đa lớp



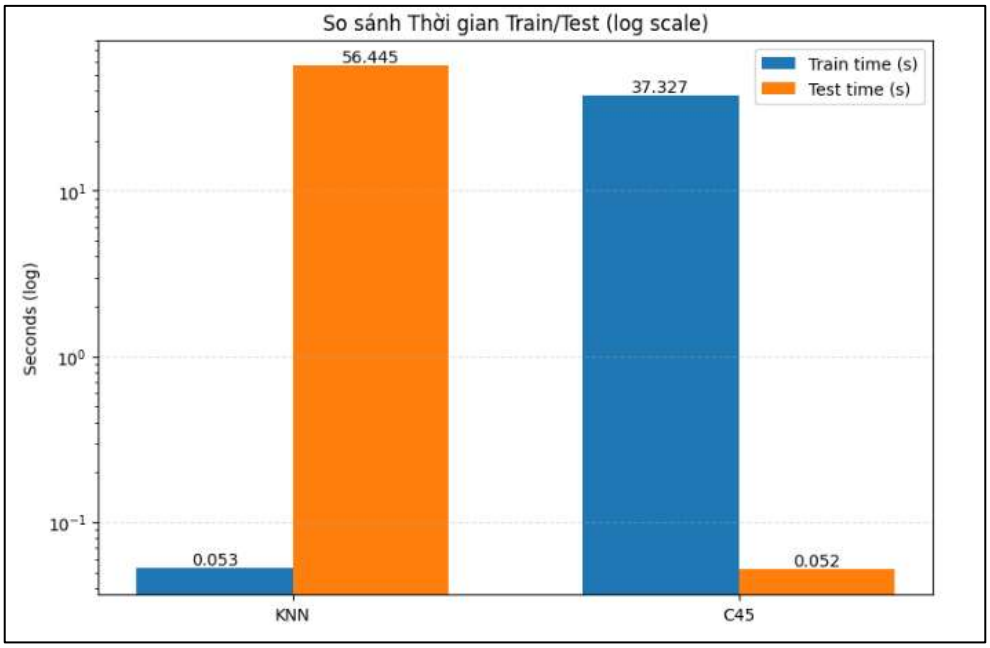
Hình 3.51 Ma trận nhầm lẫn KNN & C45 nhị phân



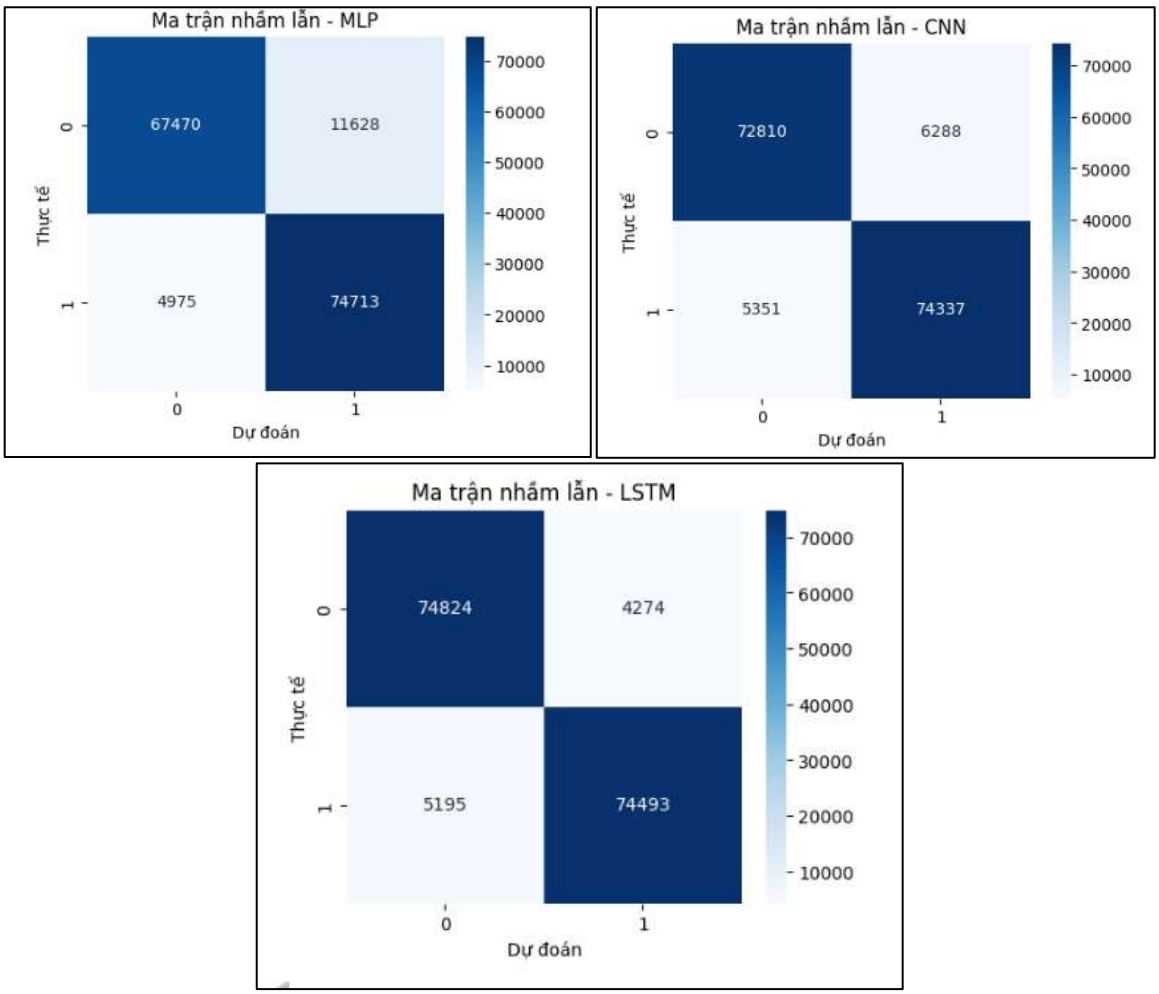
Hình 3.52 Train/Test time KNN & C45 nhị phân



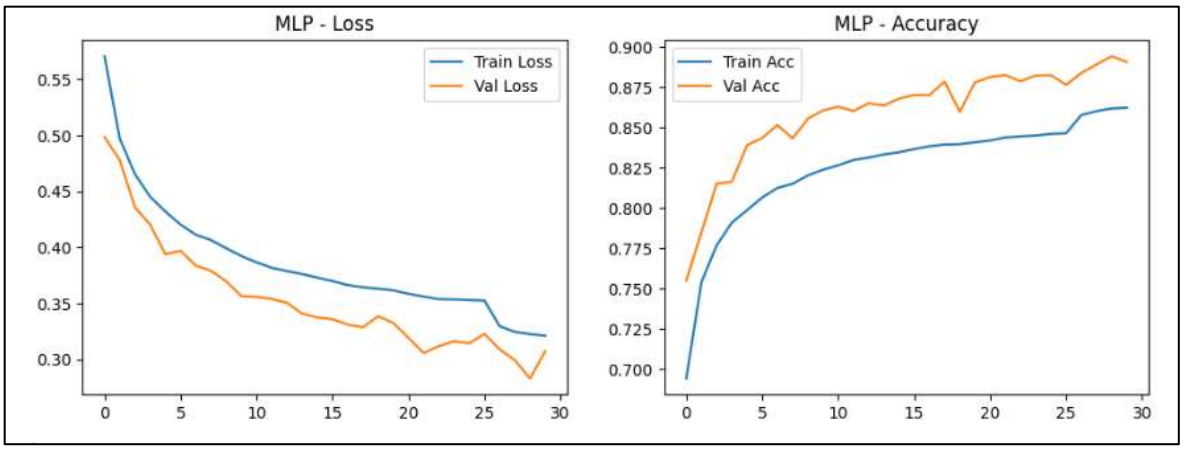
Hình 3.53 Ma trận nhầm lẫn KNN & C45 đa lớp



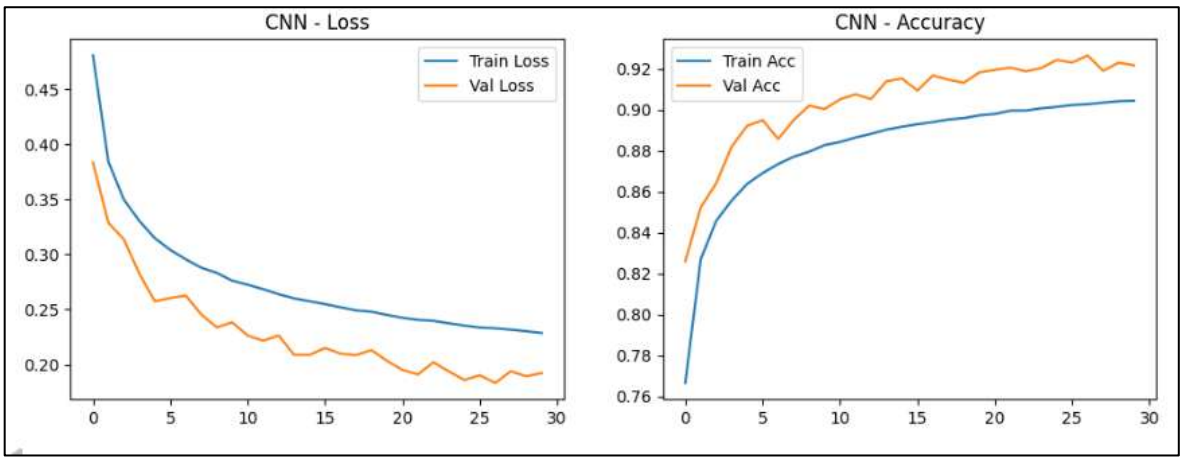
Hình 3.54 Train/Test time KNN & C45 đa lớp



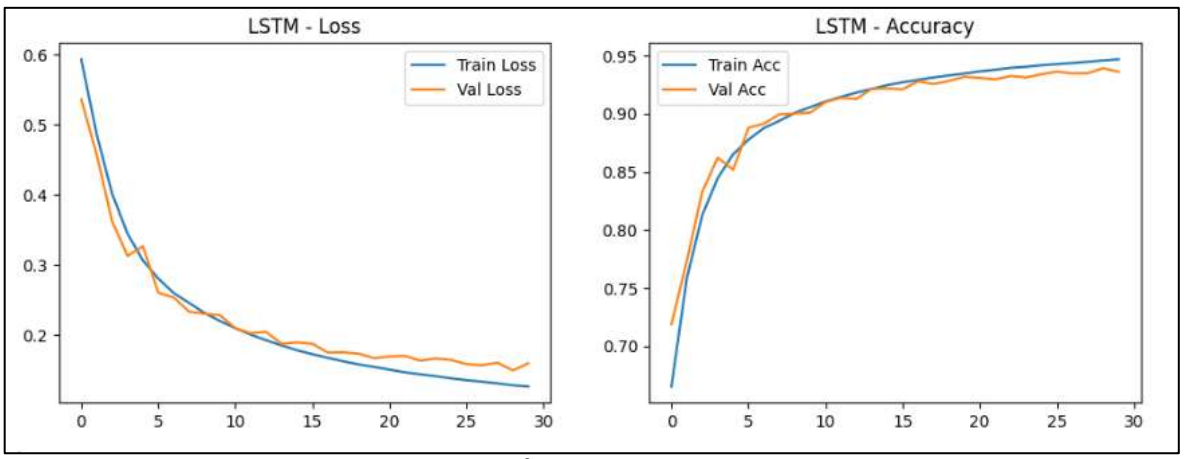
Hình 3.55 Ma trận nhầm lẫn MPL, CNN, LSTM nhị phân



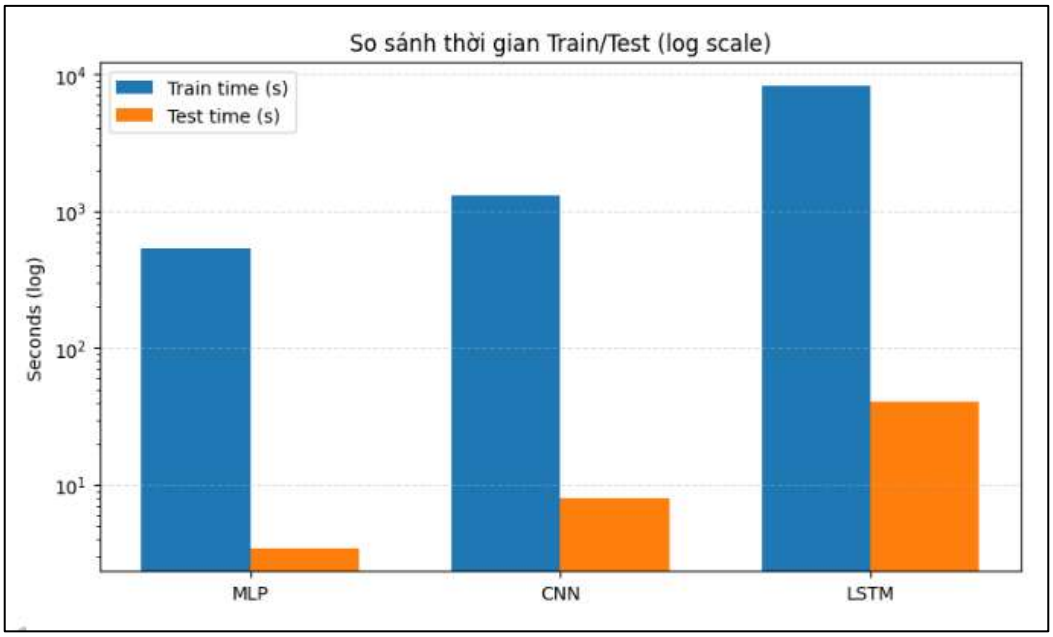
Hình 3.56 Quá trình huấn luyện mô hình MLP nhị phân



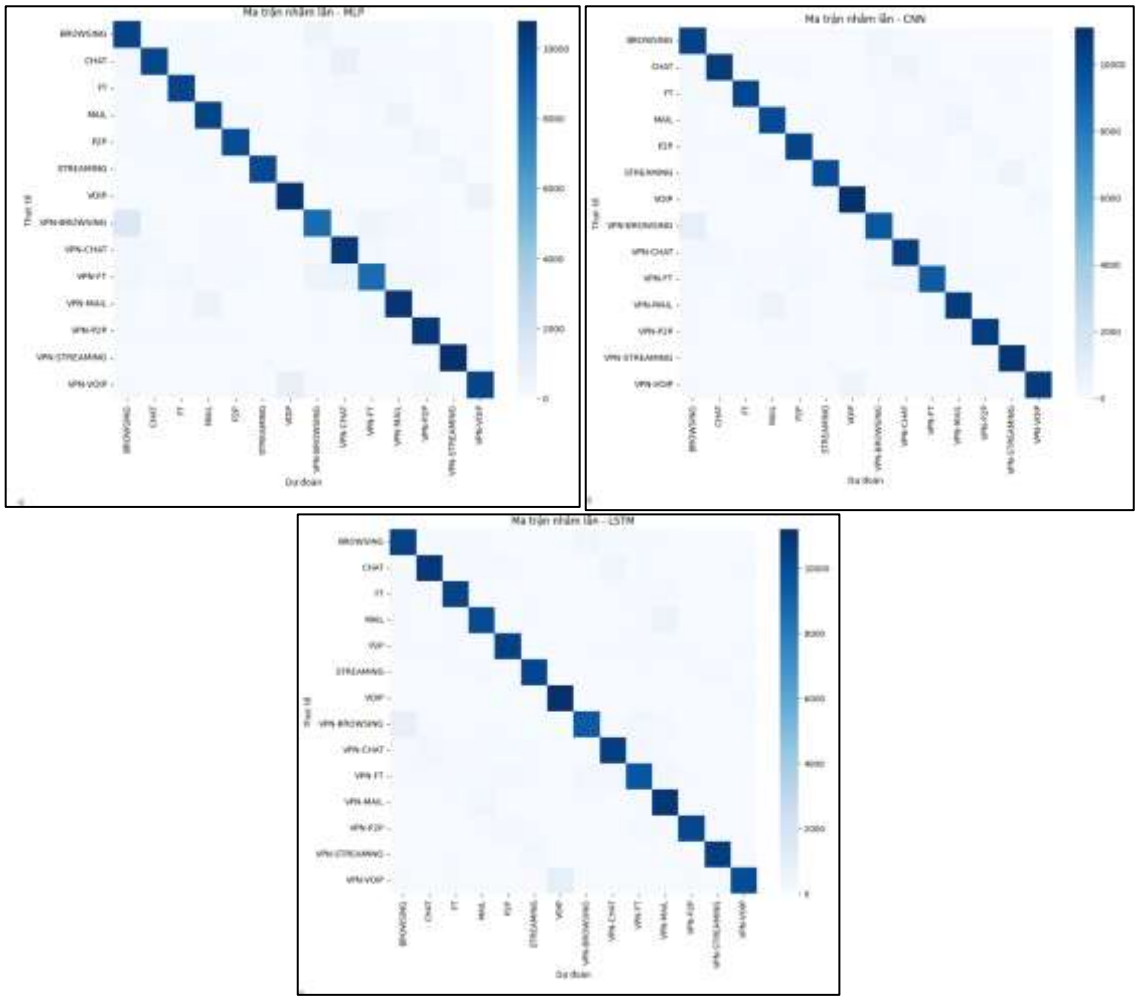
Hình 3.57 Quá trình huấn luyện mô hình CNN nhị phân



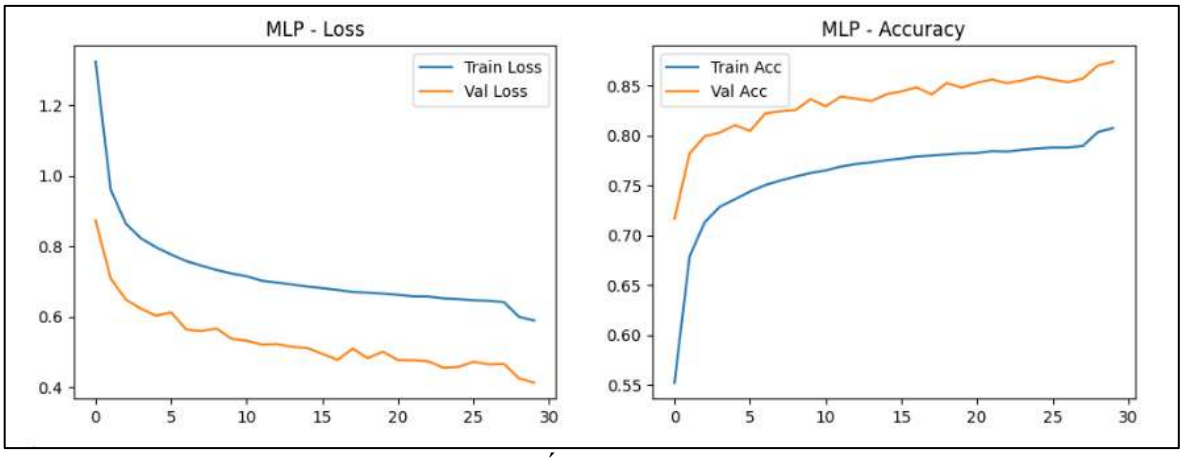
Hình 3.58 Quá trình huấn luyện mô hình LSTM nhị phân



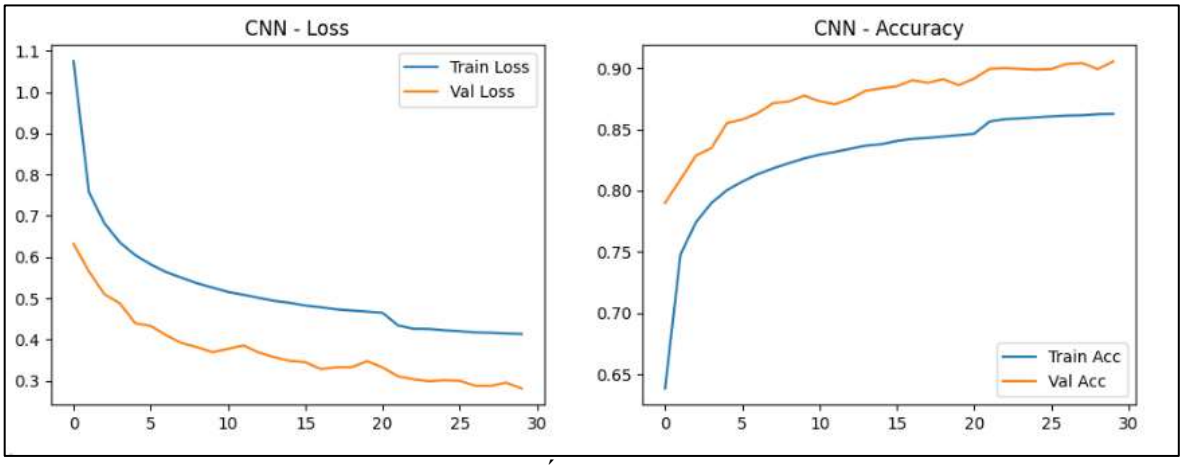
Hình 3.59 Train/Test time MLP, CNN, LSTM nhị phân



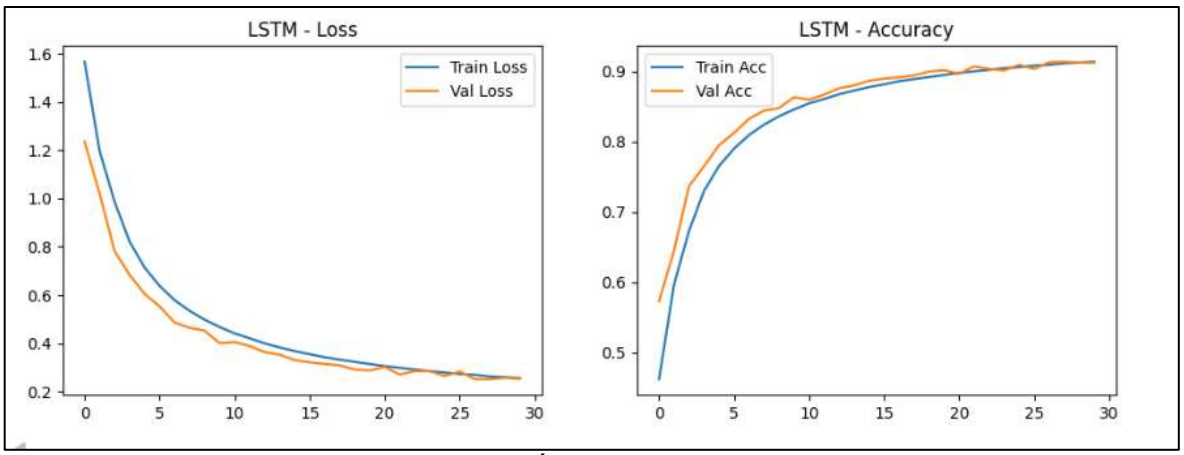
Hình 3.60 Ma trận nhầm lẫn MLP, CNN, LSTM đa lớp



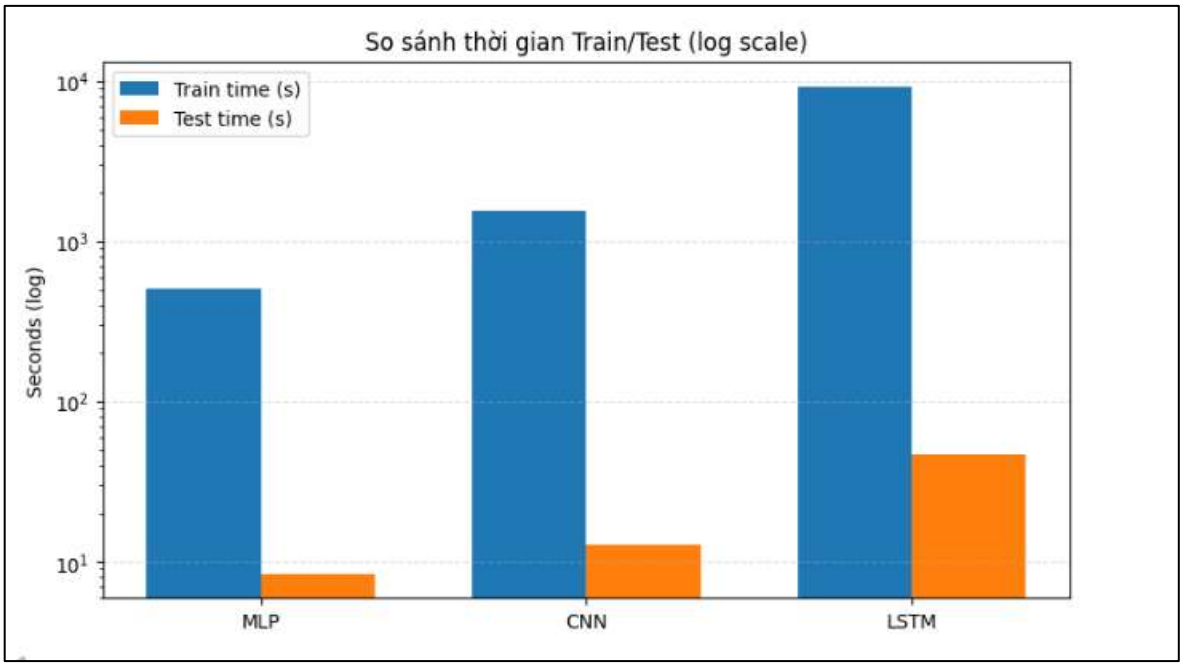
Hình 3.61 Quá trình huấn luyện mô hình MLP đa lớp



Hình 3.62 Quá trình huấn luyện mô hình CNN đa lớp



Hình 3.63 Quá trình huấn luyện mô hình LSTM đa lớp



Hình 3.64 Train/Test time MLP, CNN, LSTM đa lớp

Đánh giá

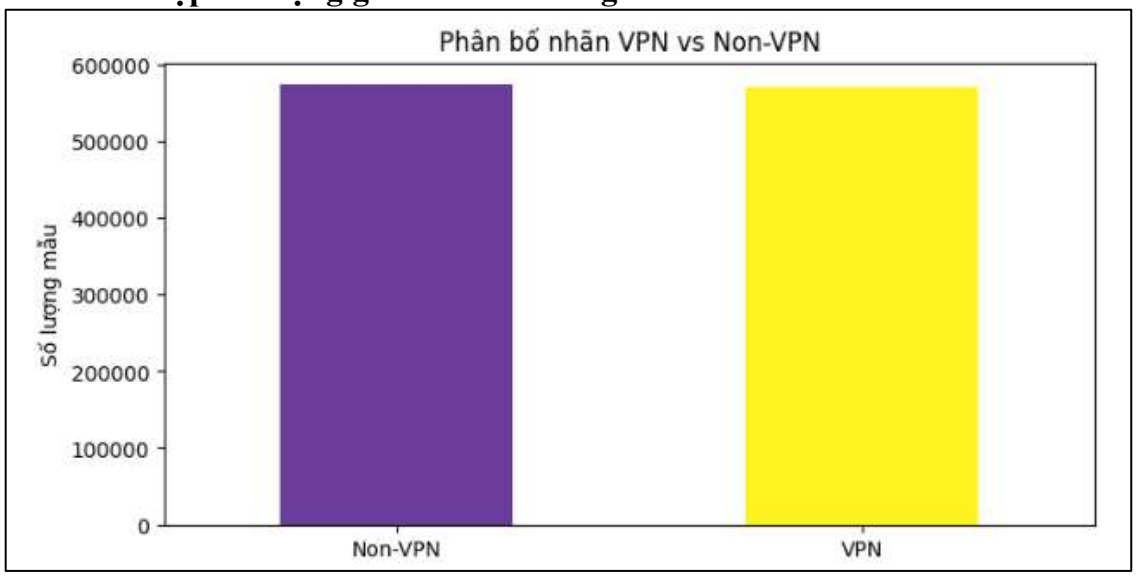
– Phân loại nhị phân:

+ Precision, Recall, F1-score: CNN, LSMT đạt cân bằng tốt giữa Prec và Rec, C45 có dấu hiệu khựng lại khi dữ liệu lớn.
+ Đường chéo trong ma trận nhầm lẫn CNN, LSTM rõ hơn C45, ít nhầm VPN thành Non-VPN hơn, MLP đã cải thiện hơn nhưng vẫn nhận nhầm nhiều VPN thành Non-VPN
+ Quá trình phân loại thì các mô hình DL loss giảm đều, Val acc lớn hơn Train acc, hội tụ tốt, không bị overfit.
+ Train/Test time: ML thực tế hơn DL

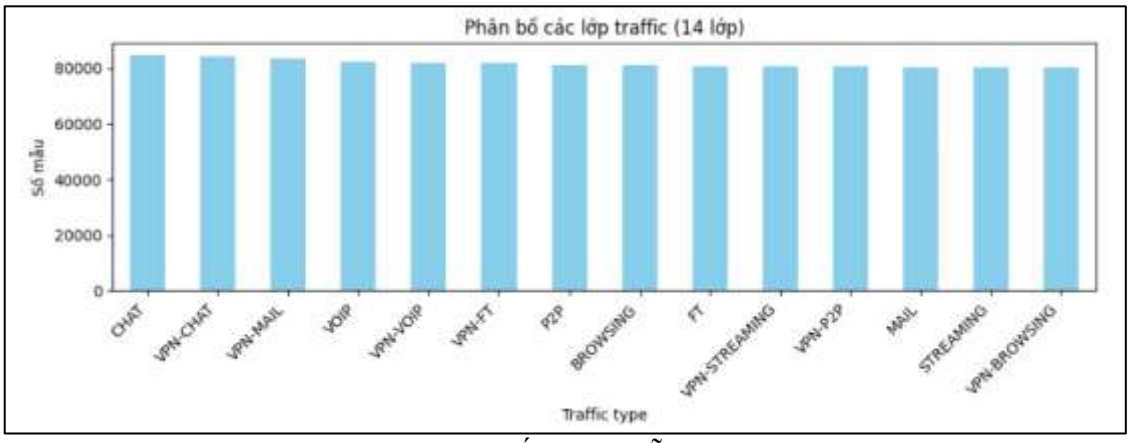
– Phân loại đa lớp:

+ C45 Prec và Rec ổn định, đường chéo ma trận nhầm lẫn rất rõ, ít nhầm lẫn nhưng vẫn kém các mô hình DL ở nhãn VPN-CHAT, VPN-FT.
+ MLP loss giảm đều, accuracy tăng đều, Prec, Rec và F1 ổn định, ít biến động hơn KNN, một số nhiễu ở VPN-BROWSING/VPN-CHAT, VPN-FT/VPN-MAIL.
+ CNN hội tụ nhanh hơn MLP, F1 cao và đặc biệt mạnh ở các nhãn STREAMING, VOIP, VPN-VOIP, trong ma trận nhầm lẫn rất ít nhiễu ngoài đường chéo.
+ LSTM loss giảm đều, Train gần bằng Val, accuracy tiến sát, Prec, Rec, F1 cao và ổn định nhất, nhất là CHAT, VPN- CHAT, STREAMING, VPN- STREAMING, ma trận nhầm lẫn rất tốt.

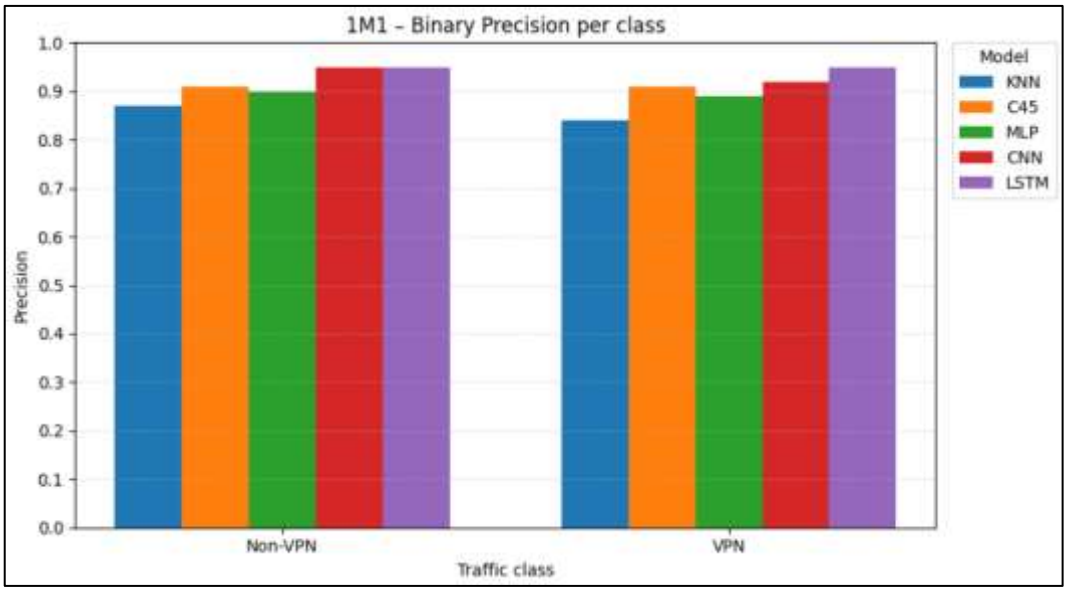
3.2.4 Trên tập mở rộng gần 1100000 dòng



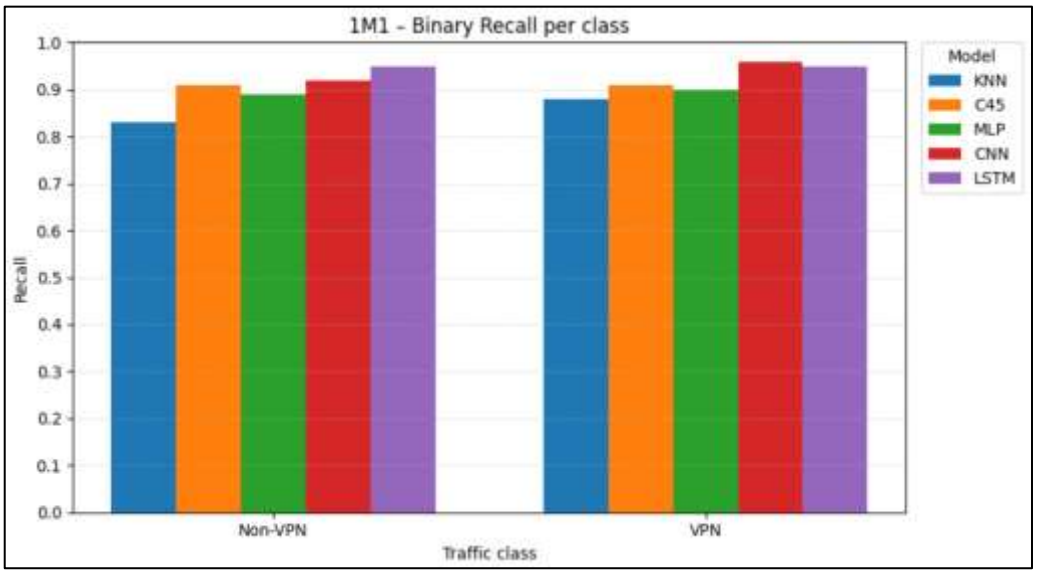
Hình 3.65 Số lượng mẫu nhị phân



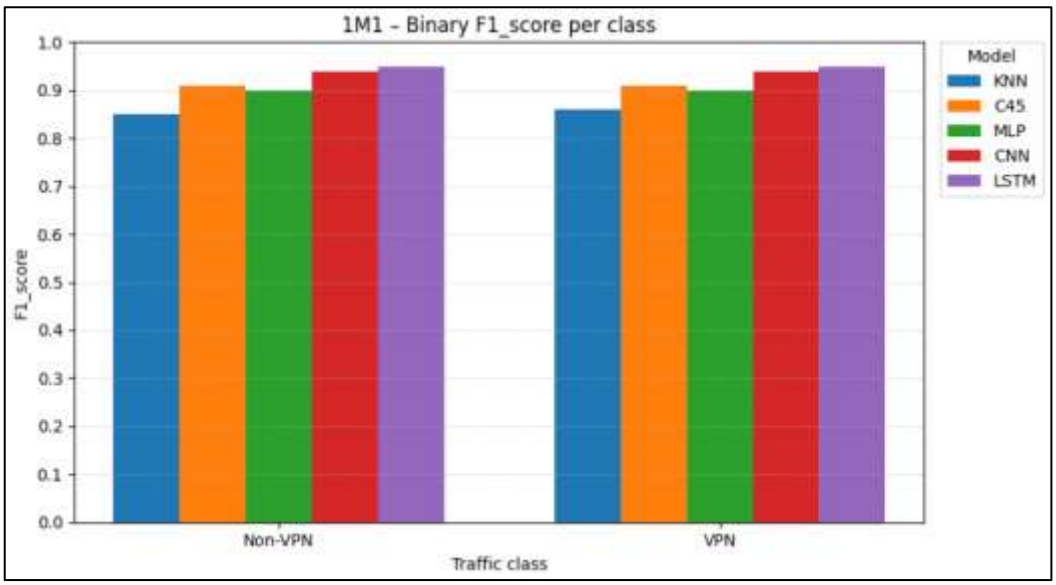
Hình 3.66 Số lượng mẫu đa lớp



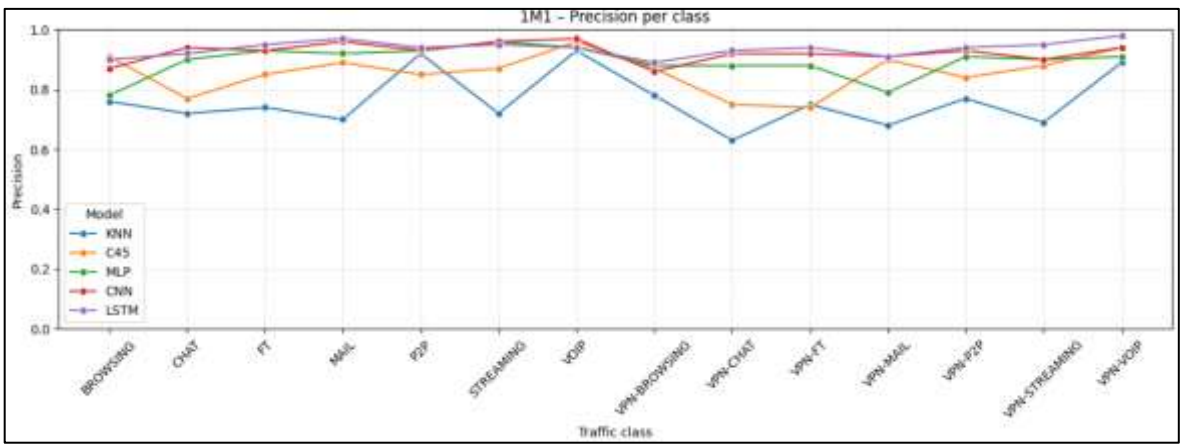
Hình 3.67 Precision nhị phân



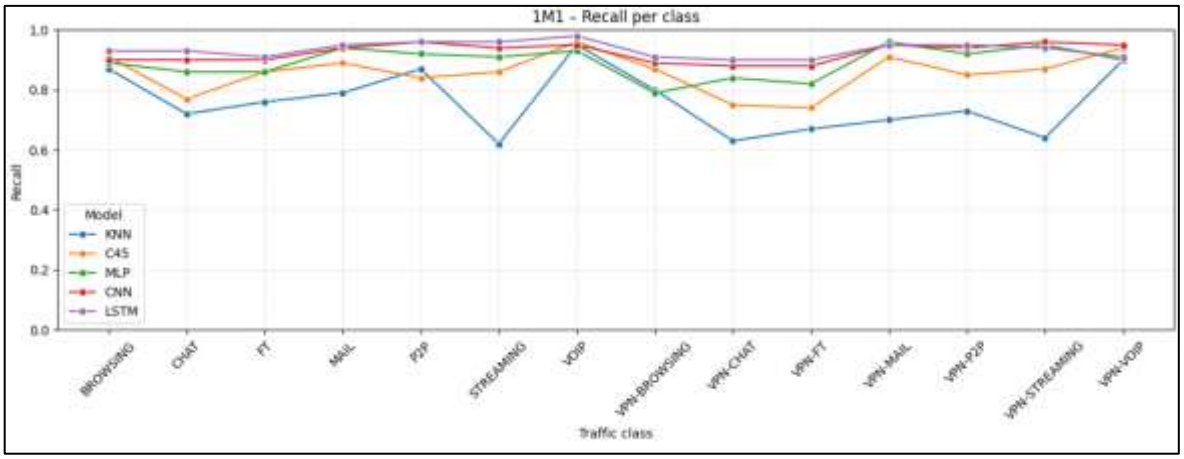
Hình 3.68 Recall nhị phân



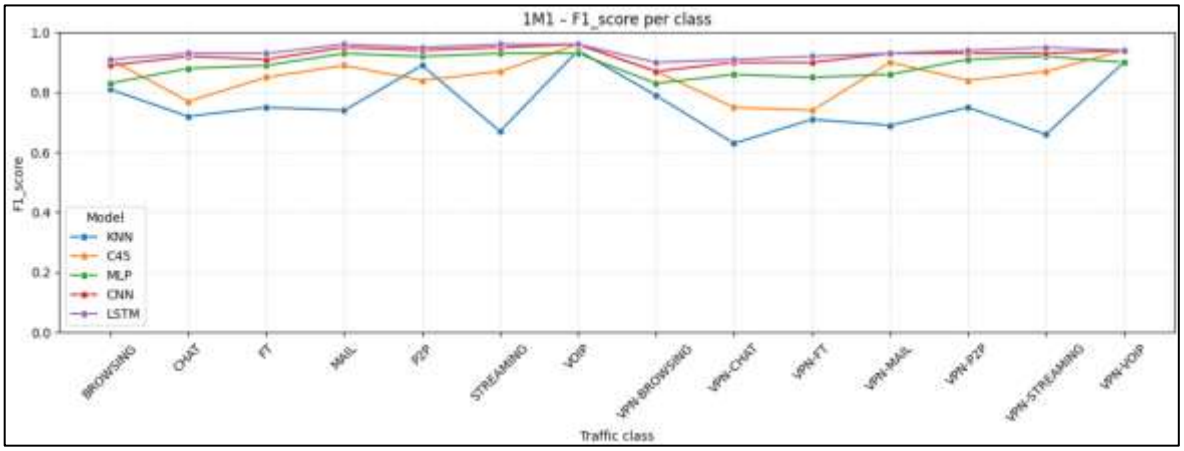
Hình 3.69 F1-score nhị phân



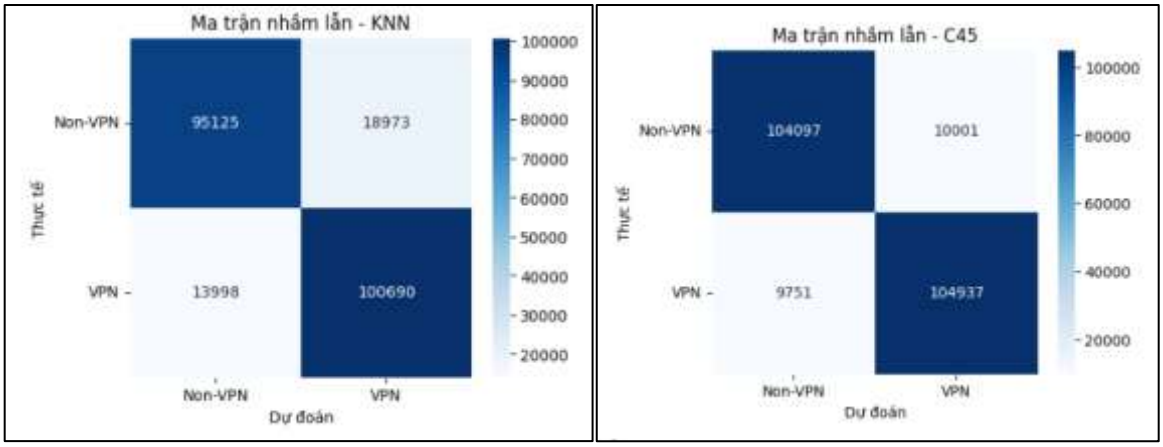
Hình 3.70 Precision đa lớp



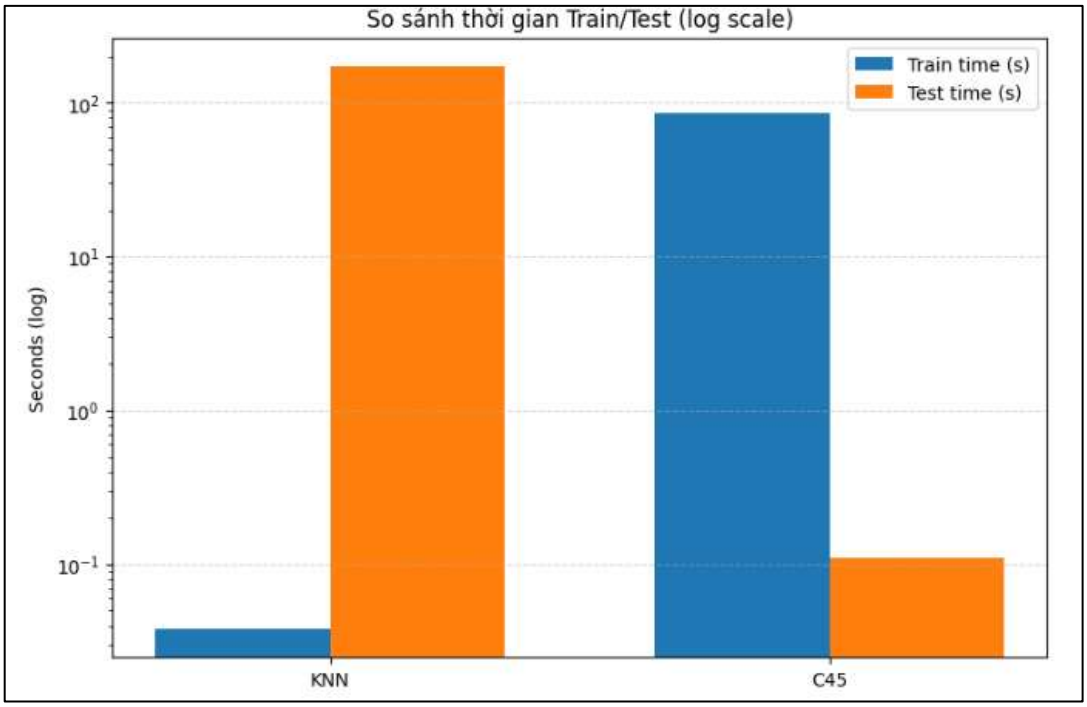
Hình 3.71 Recall đa lớp



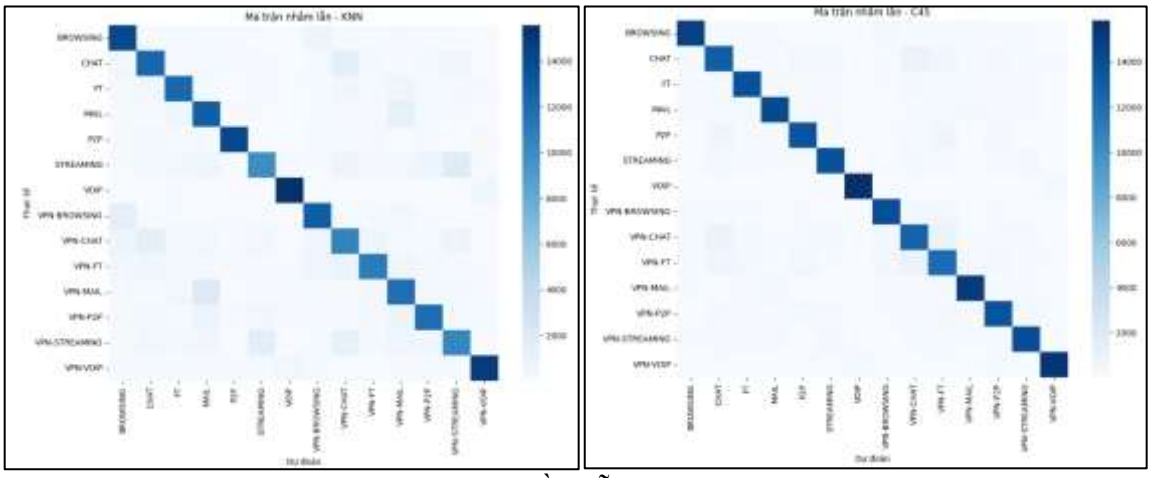
Hình 3.72 F1-score đa lớp



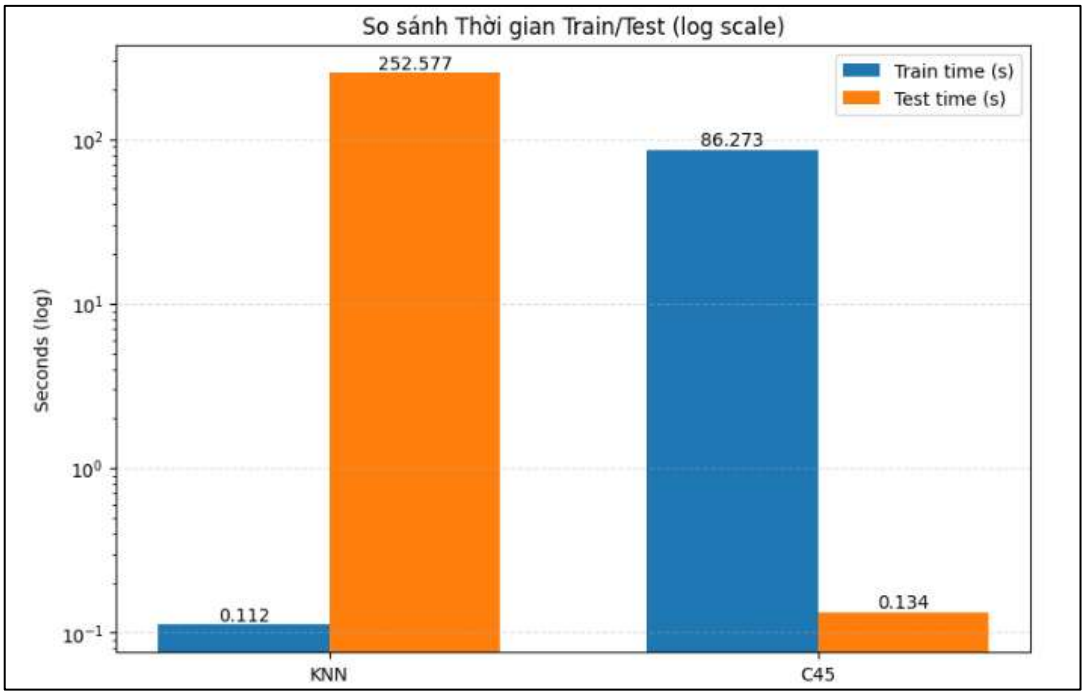
Hình 3.73 Ma trận nhầm lẫn KNN & C45 nhị phân



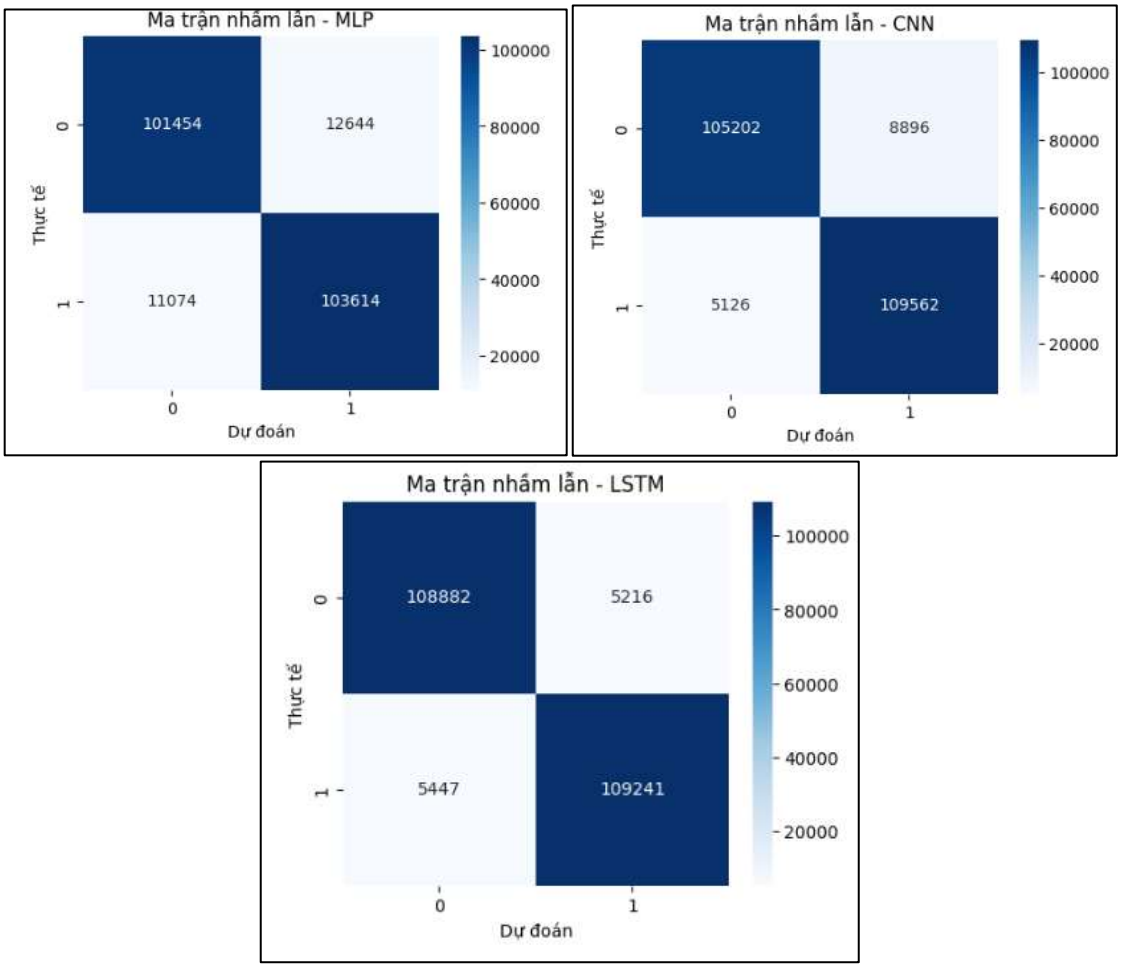
Hình 3.74 Train/Test time KNN & C45 nhị phân



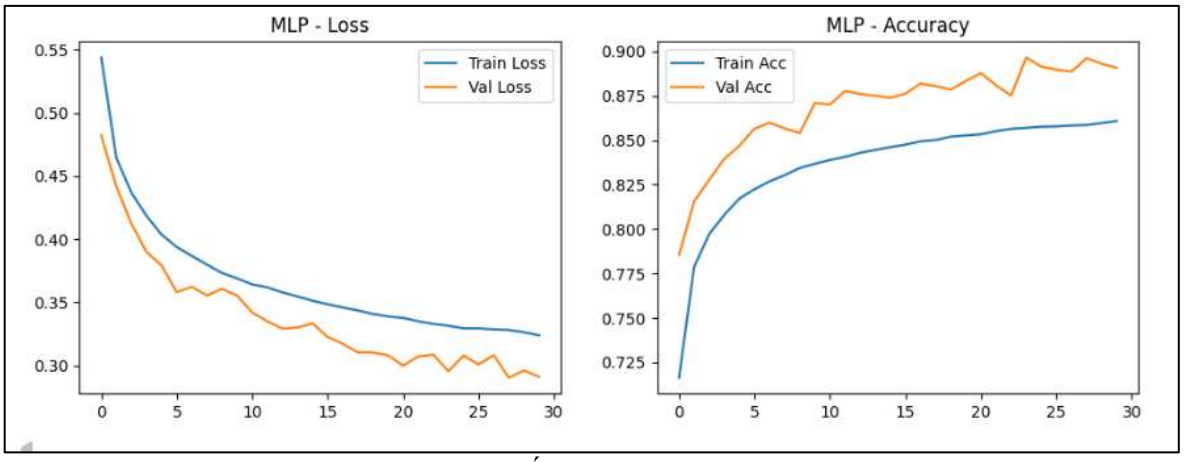
Hình 3.75 Ma trận nhầm lẫn KNN & C45 đa lớp



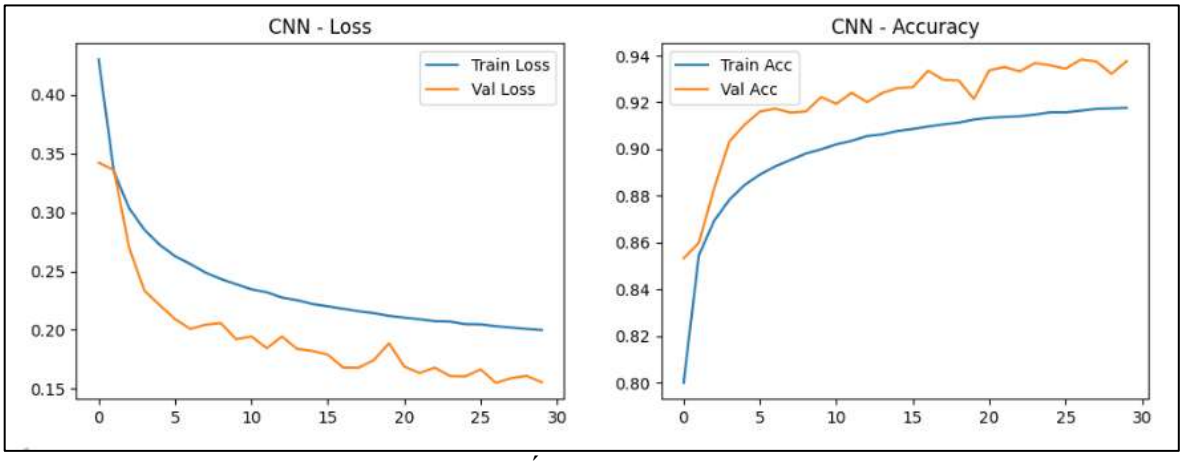
Hình 3.76 Train/Test time KNN & CNN đa lớp



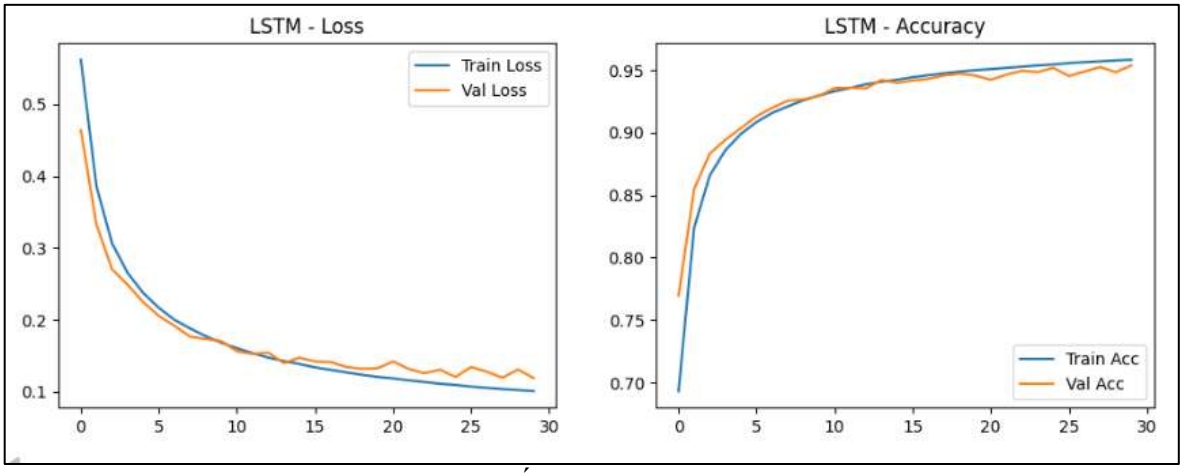
Hình 3.77 Ma trận nhầm lẫn MLP, CNN, LSTM nhị phân



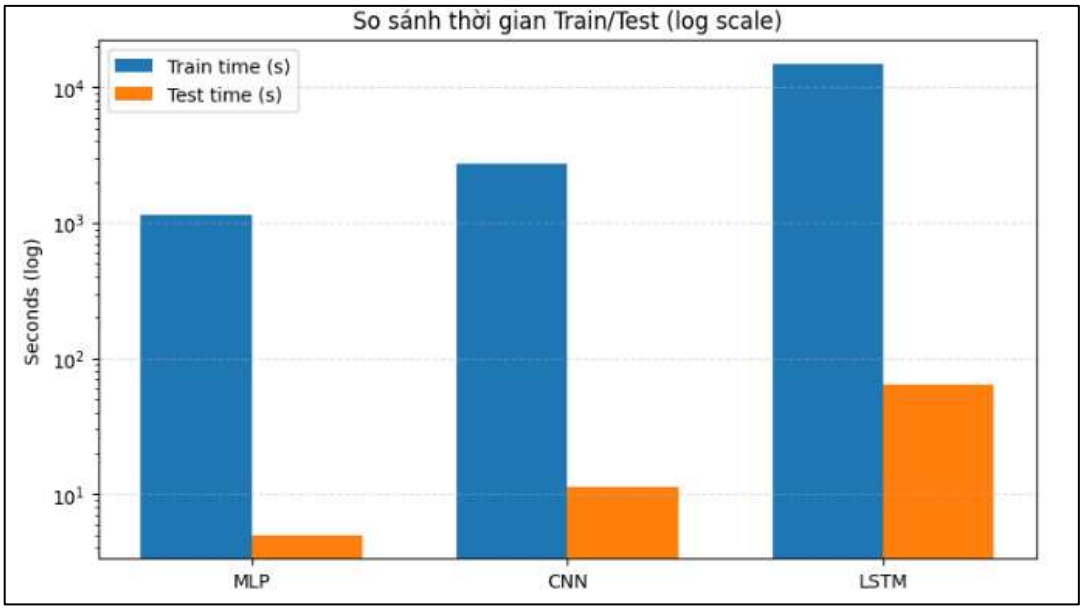
Hình 3.78 Quá trình huấn luyện mô hình MLP nhị phân



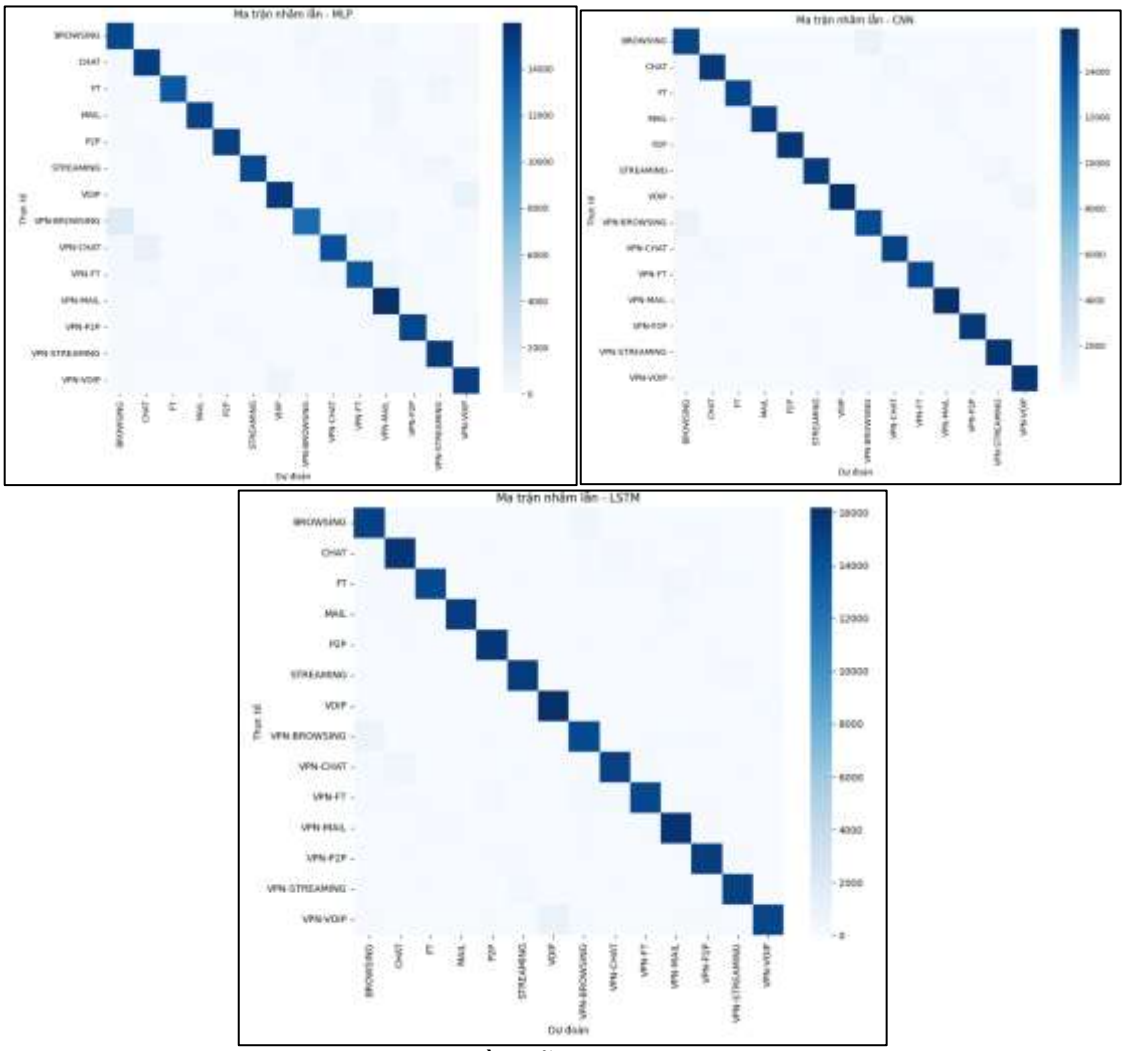
Hình 3.79 Quá trình huấn luyện mô hình CNN nhị phân



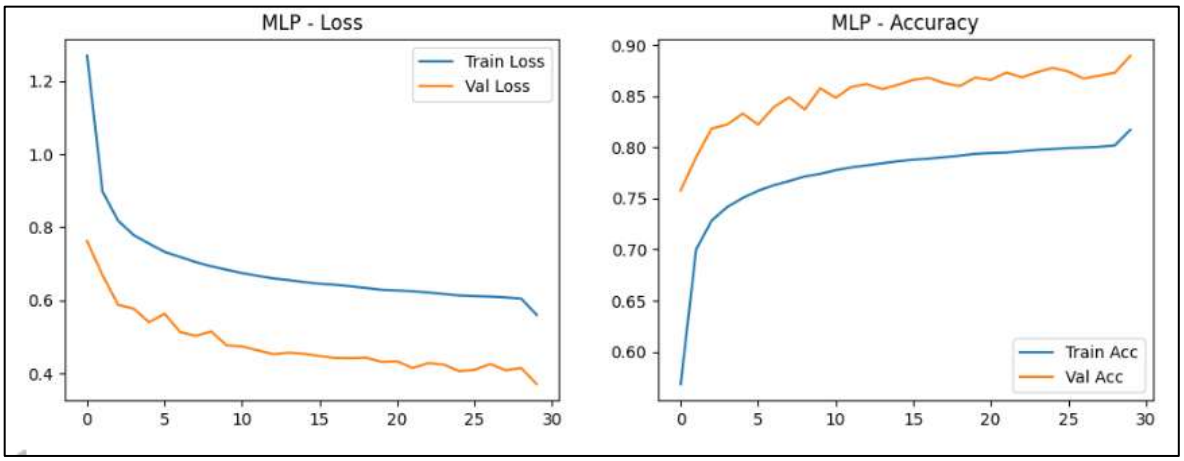
Hình 3.80 Quá trình huấn luyện mô hình LSTM nhị phân



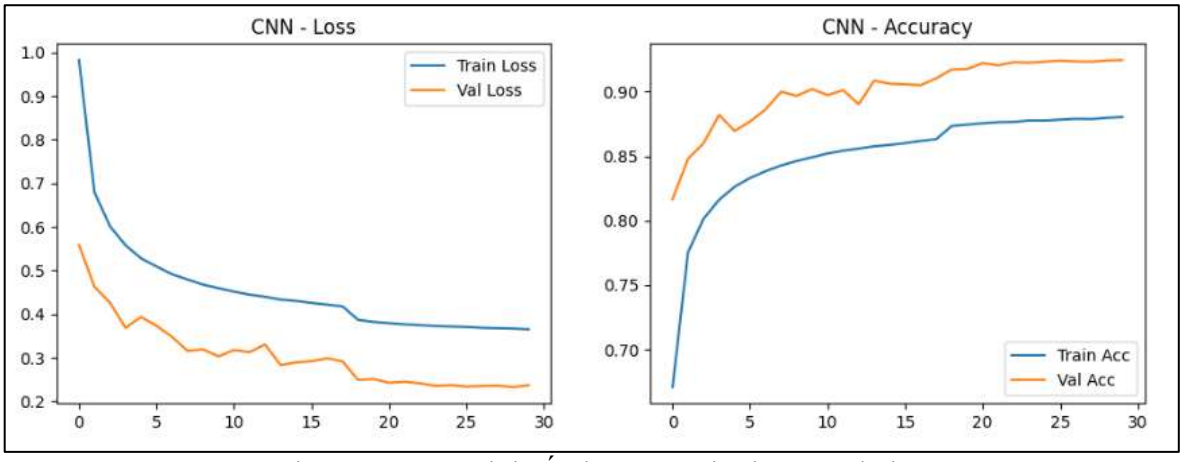
Hình 3.81 Train/Test time MLP, CNN, LSTM nhị phân



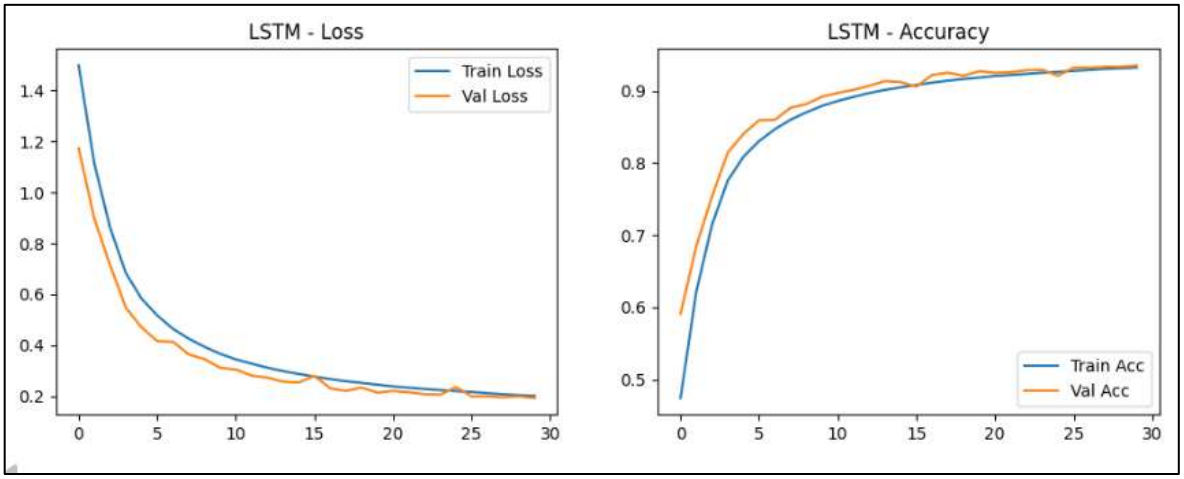
Hình 3.82 Ma trận nhầm lẫn MLP, CNN, LSTM đa lớp



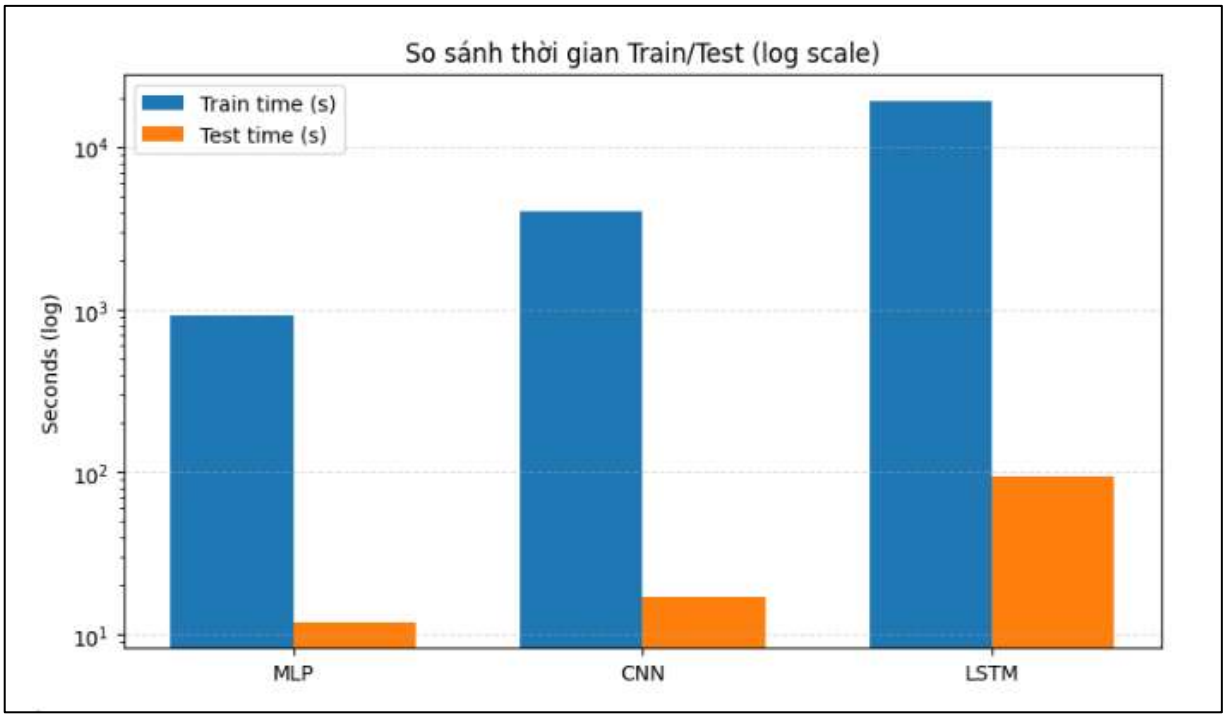
Hình 3.83 Quá trình huấn luyện mô hình MLP đa lớp



Hình 3.84 Quá trình huấn luyện mô hình CNN đa lớp



Hình 3.85 Quá trình huấn luyện mô hình LSTM đa lớp



Hình 3.86 Train/Test time MLP, CNN, LSTM đa lớp

Đánh giá

– Phân loại nhị phân:

\+ C45 nhận nhầm giảm đáng kể so với tập 500000 và 800000, kết quả vẫn kém hơn CNN và LSTM.

\+ KNN không còn phù hợp khi dữ liệu lớn.

\+ MLP kết quả khá tốt, số lượng nhầm Non-VPN thành VPN vẫn còn nhiều, loss giảm chậm hơn CNN/LSTM.

\+ CNN cho kết quả Prec, Rec, F1 tốt, chỉ sau LSTM, trong ma trận nhầm lẫn VPN bị đoán thành Non-VPN rất thấp, loss giảm đều, không bị overfit.

\+ LSTM là mô hình tốt nhất, ma trận nhầm lẫn ít lỗi nhất trong các mô hình, Train và Val hai đường gần trùng nhau, hội tụ sớm và ổn định.

– Phân loại đa lớp:

\+ Các nhãn CHAT/MAIL, VPN-CHAT/ VPN-MAIL rất khó phân biệt nhưng CNN và LSTM gần như không nhầm cho thấy VPN che nội dung nhưng không hoàn toàn che hành vi thống kê.

\+ Ma trận nhầm lẫn tốt, CNN và LSTM rất tốt.

\+ Cả ba mô hình DL loss giảm đều, accuracy tăng đều, LSTM ổn định và tốt nhất.

CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1 Kết luận:

4.1.1 Ảnh hưởng của kích thước tập dữ liệu:

Kết quả cho thấy khi kích thước tập dữ liệu tăng từ 500000 đến 1100000 lưu lượng, hiệu năng của tất cả các mô hình đều được cải thiện rõ rệt:

- Precision, Recall và F1-score tăng dần và ổn định hơn.
- Khoảng cách giữa hai lớp VPN và Non-VPN ngày càng thu hẹp.
- Một số cặp nhãn có lưu lượng khó phân biệt (CHAT/MAIL, VPN-CHAT/VPN-MAIL) được phân biệt tốt.
- Số lượng mẫu bị phân loại sai trong ma trận nhầm lẫn giảm đáng kể.

Điều này chứng tỏ rằng dữ liệu lớn đóng vai trò quan trọng trong việc học được các đặc trưng hành vi ổn định của lưu lượng VPN, vốn bị che giấu nội dung bởi mã hóa.

4.1.2 So sánh các mô hình

4.1.2.1 LSTM

- Đạt hiệu năng cao nhất trên tất cả các tập dữ liệu
- F1-score thường xuyên vượt 0.94-0.95 ở tập 1100000
- Ma trận nhầm lẫn cho thấy số lượng False Positive và False Negative thấp nhất
- Loss và accuracy hội tụ mượt, train và validation gần nhau

4.1.2.2 CNN

- Hiệu năng chỉ kém LSTM một mức nhỏ
- Hội tụ nhanh, ổn định
- Khả năng tổng quát hóa tốt
- Thời gian huấn luyện thấp hơn LSTM

4.1.2.3 C4.5

- Đạt kết quả bất ngờ cao
- F1-score đạt ~0.90-0.92 ở tập lớn
- Thời gian huấn luyện và suy luận rất nhanh

4.1.2.4 MLP

- Hiệu năng khá, ổn định
- Không bị overfitting
- Tuy nhiên kém CNN và LSTM trong việc giảm nhầm lẫn

4.1.2.5 KNN

- Hiệu năng thấp nhất
- Nhạy với dữ liệu lớn và phân bố phức tạp
- Không phù hợp cho bài toán quy mô lớn

Bài toán phát hiện VPN và phân loại chính xác từng loại ứng dụng có thể được giải quyết hiệu quả khi sử dụng dữ liệu đủ lớn. Trong đó, các mô hình học sâu, đặc biệt là LSTM và CNN, cho kết quả vượt trội. Tuy nhiên, các mô hình học máy truyền thống như C4.5 vẫn có giá trị thực tiễn nhờ chi phí tính toán thấp.

Nghiên cứu đã xây dựng và đánh giá thành công hệ thống phân loại lưu lượng mạng VPN/Non-VPN và phân loại ứng dụng dựa trên các mô hình học máy và học sâu. Kết quả cho thấy các mô hình học sâu, đặc biệt là LSTM, đạt hiệu năng cao và ổn định khi được huấn luyện trên dữ liệu lớn.

4.2 Hướng phát triển

Mặc dù nghiên cứu đã đạt được những kết quả khả quan trong bài toán phân loại lưu lượng mạng được mã hóa, vẫn còn nhiều hướng phát triển tiềm năng nhằm nâng cao hơn nữa độ chính xác, khả năng tổng quát hóa và tính ứng dụng thực tế của hệ thống.

Một trong những hạn chế chính của nghiên cứu là tập dữ liệu vẫn được thu thập trong các kịch bản và môi trường tương đối cố định. Trong tương lai, có thể thu thập dữ liệu trong môi trường mạng thực tế với nhiều điều kiện khác nhau (mạng doanh nghiệp, mạng di động, mạng IoT). Bổ sung thêm các loại VPN mới (WireGuard, Shadowsocks, obfsproxy) và các giao thức mã hóa hiện đại. Xem xét sự mất cân bằng dữ liệu theo lớp, đặc biệt đối với các ứng dụng ít phổ biến, nhằm tăng tính thực tế của mô hình.

Trong nghiên cứu hiện tại, các mô hình chủ yếu dựa trên các đặc trưng thống kê truyền thống của luồng lưu lượng. Hướng phát triển tiếp theo là khai thác các đặc trưng theo chuỗi thời gian chi tiết hơn như khoảng cách thời gian giữa các gói, phân bố kích thước gói theo cửa sổ trượt. Ứng dụng các mô hình representation learning để tự động học đặc trưng, thay vì thiết kế thủ công. Kết hợp đặc trưng đa mức nhằm nắm bắt đồng thời hành vi ngắn hạn và dài hạn của lưu lượng.

Kết quả thực nghiệm cho thấy CNN và LSTM đều có những ưu điểm riêng. Do đó, các hướng nghiên cứu tiếp theo có thể kết hợp CNN-LSTM nhằm khai thác đồng thời đặc trưng không gian và đặc trưng theo chuỗi thời gian. Áp dụng các mô hình Attention hoặc Transformer để tập trung vào các gói tin hoặc khoảng thời gian quan trọng trong luồng lưu lượng. Nghiên cứu Graph Neural Networks (GNN) để mô hình hóa mối quan hệ giữa các luồng hoặc các phiên giao tiếp.

Các kỹ thuật che giấu lưu lượng ngày càng tinh vi, đặc biệt trong môi trường VPN và traffic obfuscation. Do đó cần nghiên cứu khả năng chống lại các kỹ thuật làm nhiễu lưu lượng. Đánh giá độ bền vững của mô hình trước các kịch bản bị tấn công. Phát triển các mô hình có khả năng thích nghi với sự thay đổi hành vi lưu lượng theo thời gian.

---HẾT---

TÀI LIỆU THAM KHẢO:

[1] Dataset: VPN-nonVPN dataset (ISCXVPN2016).
[2] Gerard Drapper Gil, Arash Habibi Lashkari, Mohammad Mamun, Ali A. Ghorbani, "Characterization of Encrypted and VPN Traffic Using Time-Related Features", In Proceedings of the 2nd International Conference on Information Systems Security and Privacy(ICISSP 2016) , pages 407-414, Rome, Italy.
[3] Deep Learning for Network Traffic Monitoring and Analysis (NTMA): A Survey.
[4] Zero Trust VPN (ZT-VPN): A Systematic Literature Review and Cybersecurity Framework for Hybrid and Remote Work.