HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LÊ HOÀNG BẢO

PHÂN LOẠI LƯU LƯỢNG MẠNG INTERNET DÙNG MACHINE LEARNING

Chuyên ngành: HỆ THỐNG THÔNG TIN
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ
(*Theo định hướng ứng dụng*)

TP.HỒ CHÍ MINH - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TIẾN SĨ.NGUYỄN HỒNG SƠN……

(Ghi rõ học hàm, học vị)

Phản biện 1: ……

Phản biện 2: ……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Phân loại lưu lượng mạng Internet là một trong những đề tài được quan tâm hàng đầu trong cộng đồng nghiên cứu và phát triển mạng Internet. Rất nhiều phương án khác nhau được đã đề xuất nhằm quản lý vấn đề bảo mật cũng như đảm bảo chất lượng sử dụng dịch vụ (Quality of Service – QoS). Tuy nhiên, một số phương pháp phân loại truyền thống, đã không còn phù hợp do những hạn chế trong việc xử lý đặc trưng mới trong lưu lượng mạng Internet (phân bố cổng động, tạo đường hầm, mã hóa, …). Trong vài năm trở lại đây, việc áp dụng các phương pháp học máy (Machine Learning – ML) cho phân loại lưu lượng mạng Internet đã đạt được những kết quả đáng chú ý. Với khả năng xử lý nhiều thông tin phức tạp từ nhiều đặc trưng khác nhau, các mô hình học máy có thể phân loại các dữ liệu đầu vào với độ chính xác khá cao. Đây cũng là lý do tôi chọn đề tài "Phân loại lưu lượng mạng Internet bằng phương pháp học máy".

Trong Đề tài này sử dụng phương pháp nghiên cứu lý thuyết kết hợp với xây dựng ứng dụng mô phỏng:

Nghiên cứu và phân tích những khái niệm cơ bản trong lĩnh vực học máy, khảo sát những mô hình học máy từng được đề xuất và áp dụng trong lĩnh vực phân loại mạng Internet trong cộng đồng nghiên cứu.

So sánh các phương pháp và mô hình phân loại kể trên, đề xuất một mô hình học máy phù hợp với mục tiêu tổng quan có độ chính xác cao.

Kiểm tra, đánh giá những kết quả của mô hình đề xuất ví dụ như chỉ số Accuracy, Precision, F1 value…,

Ngoài phần mở đầu, mục lục, kết luận và kiến nghị, danh mục hình vẽ, danh mục bảng biểu, tài liệu tham khảo, phụ lục, phần chính của luận văn gồm 4 chương như sau:

Chương 1: Nghiên cứu tổng quan các phương pháp học máy, các phướng pháp cơ bản, các pháp tiên xử lý dữ liệu.

Chương 2: Tổng quan về học máy, nêu lên các phương pháp học máy và các bài toán cơ bản về học máy.

Chương 3: Phát triển mô hình dữ trên tập dữ liệu đã được thông qua, xây dựng tập dữ liệu,mô hình phân loại lưu lượng,tiền xử lý dữ liệu,

Chương 4: Đánh giá kết quả thực hiện dựa trên các mô hình đã nêu K – Lân cận (KNN – K-Nearest Neighbors),Mạng Neuron nhân tạo (ANN – Artificial Neural Networks),Rừng ngẫu nhiên (RF - Random Forest)

Chương 1: NGHIÊN CỨU TỔNG QUAN

1.1 Nhu cầu phân tích lưu lượng mạng Internet

Trong lĩnh vực phân loại lưu lượng Internet, những phương pháp truyền thống có một số hạn chế nhất định. Đầu tiên, đánh dấu gói (packet marking) được đề xuất để phân biệt lưu lượng dựa trên lớp QoS của nó. Một số ví dụ về các trường được sử dụng để đánh dấu gói là Loại dịch vụ (Type of Sevice - ToS), Điểm mã dịch vụ phân biệt (Differentiated Services Code Point - DSCP) và Thông báo tắc nghẽn rõ ràng (Explicit Congestion Notification - ECN).

Ngoài ra, có hai phương pháp phân loại truyền thống được ứng dụng rộng rãi, bao gồm phương pháp phân loại dựa trên cổng (Port – based) và phương pháp phân loại dựa trên tải trọng (Payload – based).

Phân loại dựa trên cổng (Port-based technique): Kỹ thuật phân loại dựa trên cổng là kỹ thuật phổ biến và thông dụng nhất để phân loại lưu lượng mạng Internet. Trong kỹ thuật này, mỗi một gói dữ liệu (packet) trong lưu lượng mạng IP đều mang số cổng (số cổng nguồn và số cổng đích) do tổ chức IANA (Internet Assigned Number Authority – Tổ chức cấp phát số hiệu Internet) ấn định. Ví dụ: các ứng dụng Email sử dụng số cổng 25 (SMTP) để gửi email và cổng 110 (POP3) được sử dụng để nhận email, các ứng dụng web sử dụng số cổng 80.

Phân loại dựa trên tải trọng (Payload-based technique): thường được biết đến dưới cái tên phương pháp kiểm tra gói chuyên sâu (Deep Packet Inspection - DPI). Trong kỹ thuật này, nội dung của gói dữ liệu được kiểm tra dựa trên đặc trưng của các ứng dụng mạng trong lưu lượng Internet. Kỹ thuật này đặc biệt được đề xuất cho các ứng dụng Peer-to-Peer (P2P), hoặc cho những ứng dụng tương đương có sử dụng số cổng động nhằm xác định lưu lượng mạng Internet.

Một số giao thức học máy cơ bản

Cây quyết định: Các phương pháp dựa trên cây quyết định (Decision Tree - DT) đã từng được đề xuất nhằm phân loại lưu lượng mạng Internet. Phương pháp phân loại cây quyết định DT là một phương pháp dựa trên quy tắc. Nó chủ yếu bao gồm việc trả lời một chuỗi các câu hỏi ở các nút câu hỏi (non-leaf node) để đưa về phía các nút lá (leaf node), mà tại đó mỗi node lá đại diện cho một nhãn dán được dự đoán. Tuy nhiên, phương pháp này lại có xu hướng tạo ra hiện tượng quá khớp (overfitting).

Bộ phân loại Naïve Bayes (NB): NB là một phương pháp học máy khác đã được sử dụng để phân loại lưu lượng mạng Internet. NB là một phương pháp xác suất dựa vào định lý Bayes mà tại đó, NB là phương thức đơn giản nhất trong họ phương pháp Bayes.

K lân cận (KNN): KNN là một phương pháp huấn luyện học máy phi tham số. Là một thuật toán được phân vào loại mô hình huấn luyện lười học (lazy learning), KNN không bao gồm giai đoạn huấn luyện. Vì vậy, thời gian phân loại phụ thuộc vào kích thước dữ liệu. Ở giai đoạn phân loại, thuật toán tiến hành phân loại dữ liệu dựa trên việc đo khoảng cách giữa mẫu thử nghiệm với tất cả các mẫu được gắn nhãn. Mẫu thử nghiệm sẽ được gán cho lớp có K – lân cận gần nhất của nó.

Mạng Nơ-ron (NN): là một phương pháp huấn luyện học máy đã được thiết kế lấy nguồn gốc từ hệ thống thần kinh của con người. Một bài nghiên cứu áp dụng mô hình mạng nơ-ron học sâu trong phân loại lưu lượng Internet vào ứng dụng phát hiện xâm nhập mạng đã cho thấy, các mô hình mạng nơ-ron cũng có thể tạo ra những kết quả đáng chú ý.

Những mô hình trên đã liệt kê ra những khả năng thích hợp ứng dụng học máy vào phân loại lưu lượng mạng Internet. Trong phạm vi đề tài luận văn sẽ chỉ tập trung vào một số mô hình phù hợp dựa trên kỹ thuật học có giám sát như NB, K-NN hay ANN, CNN.

Chương 2: TỔNG QUAN VỀ HỌC MÁY

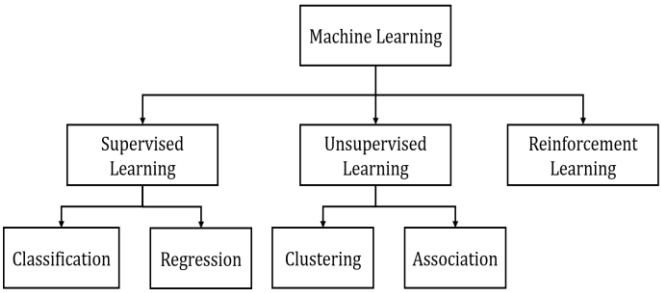
2.1 Giới thiệu

Học máy (machine learning hay ML) là một tập hợp con của trí tuệ nhân tạo và được coi là một trong những lĩnh vực trong khoa học máy tính với khả năng tự học dựa trên dữ liệu đầu vào mà không cần phải có sự lập trình cụ thể.

Đầu vào và đầu ra của quá trình học máy

Các phương pháp học trong quá trình học máy:

Thông thường, các bài toán học máy sẽ được phân loại như hình dưới đây:



Hình 2.1: Phân loại thuật toán trong máy học

Học có giám sát (supervised learning) là việc xây dựng mô hình học dự đoán các mẫu dữ liệu mới thành các nhãn đã cho trước dựa trên các mẫu dữ liệu huấn luyện.

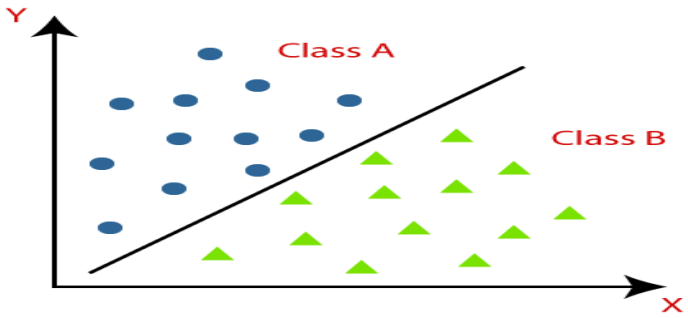
Học không giám sát (unsupervised learning) là thuật toán học máy trích xuất được những thông tin quan trọng dựa trên mối liên hệ của các điểm dữ liệu. Nói cách khác, điểm dữ liệu

trong phương pháp học này sẽ không được gán nhãn và không có đầu ra tương ứng. Học không giám sát được áp dụng trong các bài toàn phân cụm hay giảm chiều dữ liệu.

Học củng cố (reinforcement learning) là lĩnh vực liên quan đến việc dạy cho máy (agent) thực hiện tốt một nhiệm vụ (task) bằng cách tương tác với môi trường (environment) thông qua hành động (action) và nhận được phần thưởng (reward). Các bài toán học củng cố giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance).

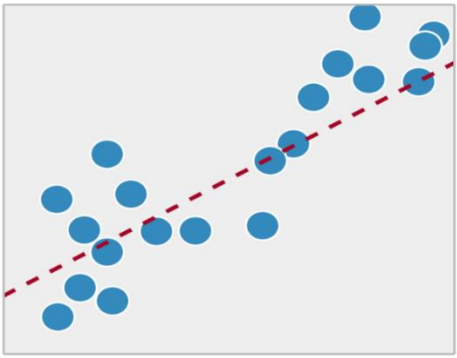
2.2 Các loại bài toán cơ bản trong học máy

Phân loại (classification) là một trong những bài toán được quan tâm và nghiên cứu nhiều nhất trong học máy. Đối với các bài toán này, nhiệm vụ được yêu cầu xác định nhãn của một điểm dữ liệu trong số các nhãn khác nhau. Các cặp dữ liệu sẽ được ký hiệu là (x, y) tương đương với (dữ liệu, nhãn). Số nhãn trong tập dữ liệu được ký hiệu là C, khi đó, việc xây dựng mô hình thật chất là việc tìm một hàm số f ánh xạ một điểm dữ liệu x vào một phần tử y.



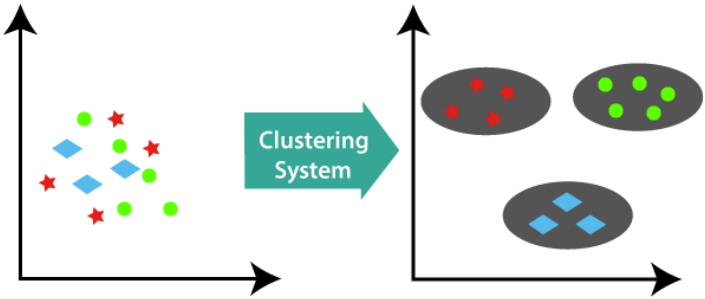
Hình 2.2: Hình minh hoạ cho bài toán phân loại

Hồi quy (regression) bao gồm một tập hợp các phương pháp học máy cho phép chúng ta dự đoán một biến kết quả liên tục y dựa trên giá trị của một hoặc nhiều đặt trưng của điểm dữ liệu x. Hồi quy được dùng để ước tính mối quan hệ giữa mục tiêu (target) và biến (variable) độc lập. Điều này giúp cho việc dự đoán các biến liên tục và được ứng dụng vào các trường hợp thực tế như phân tích giá cả thị trường, khuynh hướng doanh số bán hàng, v.v .



Hình 2.3: Hình minh hoạ cho bài toán hồi quy

Phân cụm (clustering) là thực hiện nhóm các đối tượng có các đặc điểm giống nhau thành các cụm mà không có sự tách động trước nào. Bài toán phân cụm sẽ chia toàn bộ dữ liệu thành các cụm nhỏ dựa trên sự tương quan giữa các dữ liệu trong mỗi cụm.



Hình 2.4: Hình minh hoạ cho bài toán phân cụm

Nhìn chung, bài toán phân loại là một trong những bài toán được quan tâm và nghiên cứu nhiều trong mảng học máy :K lân cận (K-nearest neighbor hay KNN), Bộ phân loại Naive Bayes (Naive Bayes Classifier hay NBC) ,Mạng neuron nhân tạo (Artificial Neural Network hay ANN)

Chương 3: PHÁT TRIỂN MÔ HÌNH

3.1. Tập dữ liệu

Trong nghiên cứu này, tập dữ liệu VPN-nonVPN (ISCXVPN2016) sẽ được sử dụng cho quá trình huấn luyện và quá trình kiểm tra. Tập dữ liệu VPN-nonVPN (ISCXVPN2016) được đề xuất bởi [21]. Dữ liệu dược thu thập từ Đại học New Brunswick ở Canada và được tạo ra bằng việc tạo 2 tài khoản người dùng để sử dụng các dịch vụ như Skype, Facebook, v.v. Bảng 3.1 thể hiện các loại lưu lượng và ứng dụng khác nhau trong tập dữ liệu.

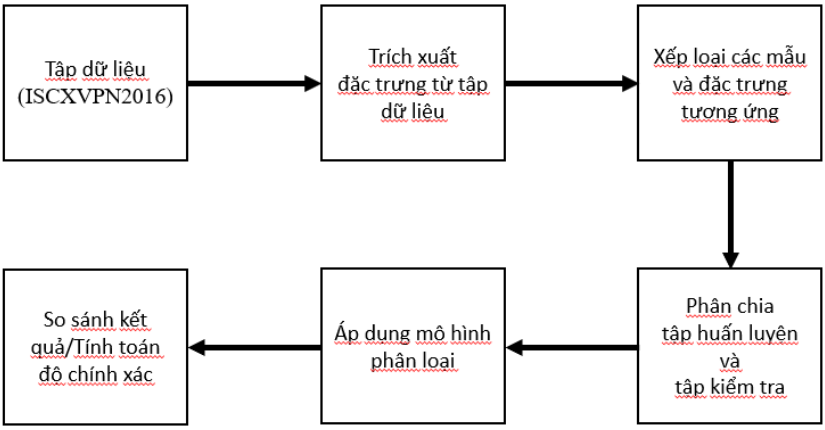
Bảng 3.1: Tóm tắt của tập dữ liệu [21]

Loại lưu lượng	Chi tiết
WebBrowsing	Firefox và Ch
Email	SMPTS, POP3S và IMAPS
Chat	ICQ, AIM, Skype, Facebook và Hangouts
Streaming	Vimeo và Youtube
File Transfer	Skype, FTPS và SFTP
VoIP	Facebook, Skype và Hangouts voice calls (trong 1 giờ)
P2p	uTorrent và Bittorrent

Ngoài ra, trong tập dữ liệu này, các loại lưu lượng đã thu tập còn được chia thành 2 loại: được mã hoá bởi VPN và không được mã hoá bởi VPN. Như vậy, với mỗi trường hợp trên, ta sẽ có 7 loại lưu lượng được thu thập.

Phân chia tập huấn luyện và tập kiểm tra một cách ngẫu nhiên nhằm đảm bảo kết quả dự đoán của mô hình phân loại mang đầy đủ tính khách quan.

Sơ đồ khối tổng quan của mô hình phân loại lưu lượng Internet được biểu diễn bằng hình 3.1.



Hình 3.1: Sơ đồ khối mô hình phân loại lưu lượng Internet

Xây dựng mô hình

Tiền xử lý dữ liệu

Chuẩn hóa Tối đa – Tối thiểu: Dựa trên bản chất của tập dữ liệu vốn chứa rất nhiều các đặc trưng khác nhau với tổng cộng 23 đặc trưng cụ thể, mỗi đặc trưng lại chứa các giá trị trong phạm vi khác nhau, đề tài quyết định chuẩn hóa các trường dữ liệu khác nhau về phạm vi giá trị chuẩn [0,1]. Trong ứng dụng cụ thể cho quá trình tiền xử lý dữ liệu, biến *scaler* được dùng để thiết lập chuẩn hóa tối đa – tối thiểu *MinMaxScaler*. Biến *scaler* sau đó được dùng để áp dụng chung quá trình chuẩn hóa cho cả tập huấn luyện và tập kiểm tra.

```
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Quá trình chuẩn hóa tối đa – tối thiểu trong tập dữ liệu này được thực hiện nhằm làm bước chẩn bị cho quá trình huấn luyện cho các mô hình học máy ở các bước sau, đồng thời làm giảm độ phức tạp trong quá trình tính toán.

Mã hóa nhãn: Tương tự, nhằm hướng đến mục tiêu giảm tải khối lượng tính toán và phân loại các nhãn nhóm mạng Internet các nhãn phân loại được mã hóa dựa trên biến *encoder* trong quá trình áp dụng tiền xử lý dữ liệu. Biến *encoder* sẽ mã hóa nhãn dán các lớp lưu lượng mạng, hỗ trợ quá trình xử lý dữ liệu tại bước phân loại mẫu dữ liệu được nhanh chóng hơn, giảm tải thời gian tính toán lên mô hình.

```
# encode the class name as int 64 for training
encoder = LabelEncoder()
encoder.fit(Y)
Y = encoder.transform(Y)
```

Kết quả mã hóa tương ứng cho từng lớp lưu lượng mạng được miêu tả trong bảng 3.2.

Bảng 3.2: Mã hóa nhãn các lớp lưu lượng mạng Internet

Lưu lượng Internet	**Giá trị mã hóa tương ứng**
VoIP	0
Web Browsing	1
File transfer	2
P2P	3
Chat	4
Streaming	5
Email	6

Mẫu dữ liệu và đặc trưng tương ứng

Như đã đề cập sơ bộ tại mục 3.1, tập dữ liệu ISCXVPN2016 bao gồm các mẫu dữ liệu ghi nhận từ 7 lớp lưu lượng mạng Internet khác nhau. Mỗi một mẫu dữ liệu được miêu tả 8 loại đặc trưng chính, bao gồm *fiat, biat, flowiat, active, idle, fb_psec* và *fp_psec*. Trong đó các đặc trưng fiat, biat, flowiat, active và idle được miêu tả cụ thể hơn bằng 4 thông số như giá trị trung bình (mean), giá trị nhỏ nhất (min), giá trị lớn nhất

(max) và độ lệch chuẩn (standard deviation). Tổng hợp 23 đặc trưng của tập dữ liệu được tóm tắt trong bảng 3.3.

Bảng 3.3: Tổng hợp nhóm 23 đặc trưng của tập dữ liệu ISCXVPN2016

Đặc trưng	***Miêu tả***
duration	khoảng thời gian lưu lượng
fiat_mean	Giá trị trung bình thời gian giữa hai gói được gửi theo hướng đi
fiat_max	Giá trị lớn nhất thời gian giữa hai gói được gửi theo hướng đi
fiat_min	Giá trị nhỏ nhất thời gian giữa hai gói được gửi theo hướng đi
fiat_std	Giá trị độ lệch chuẩn thời gian giữa hai gói được gửi theo hướng đi
biat_mean	Giá trị trung bình thời gian giữa hai gói được gửi theo hướng ngược về
biat_max	Giá trị lớn nhất thời gian giữa hai gói được gửi theo hướng ngược về
biat_min	Giá trị nhỏ nhất thời gian giữa hai gói được gửi theo hướng ngược về
biat_std	Giá trị độ lệch chuẩn thời gian giữa hai gói được gửi theo hướng ngược về

flowiat_mean	Giá trị trung bình thời gian giữa hai gói được gửi theo một trong hai hướng
flowiat_max	Giá trị lớn nhất thời gian giữa hai gói được gửi theo một trong hai hướng
flowiat_min	Giá trị nhỏ nhất thời gian giữa hai gói được gửi theo một trong hai hướng
flowiat_std	Giá trị độ lệch chuẩn thời gian giữa hai gói được gửi theo một trong hai hướng
active_mean	Giá trị trung bình khoảng thời gian lưu lượng hoạt động trước khi ngừng hoạt động
active_max	Giá trị lớn nhất khoảng thời gian lưu lượng hoạt động trước khi ngừng hoạt động
active_min	Giá trị nhỏ nhất khoảng thời gian lưu lượng hoạt động trước khi ngừng hoạt động
active_std	Giá trị độ lệch chuẩn khoảng thời gian lưu lượng hoạt động trước khi ngừng hoạt động
idle_mean	Giá trị trung bình khoảng thời gian luồng không hoạt động trước khi bắt đầu hoạt động lại
idle_max	Giá trị lớn nhất khoảng thời gian luồng không hoạt động trước khi bắt đầu hoạt động lại

idle_min	Giá trị nhỏ nhất khoảng thời gian luồng không hoạt động trước khi bắt đầu hoạt động lại
idle_std	Giá trị độ lệch chuẩn khoảng thời gian luồng không hoạt động trước khi bắt đầu hoạt động lại
fb_psec	lưu lượng byte trên một giây
fp_psec	lưu lượng gói tin trên một giây

Mỗi mẫu dữ liệu trong tập dữ liệu ISCXVPN2016 đều sở hữu 23 loại đặc trưng miêu tả như trên. Tương tự, mỗi một lớp lưu lượng mạng bao gồm nhiều mẫu dữ liệu khác nhau và chúng được sử dụng làm dữ liệu đầu vào cho các mô hình huấn luyện, từ đó các mô hình sẽ phân tích các đặc trưng có trong mẫu dữ liệu làm căn cứ phân loại các lớp lưu lượng mạng khác nhau.

Dữ liệu đầu vào – phân chia tập huấn luyện và kiểm tra

Để đánh giá hiệu suất của mô hình học máy, thuật toán đánh giá cần sử dụng các mẫu dữ liệu không tham gia vào trong quá trình huấn luyện. Nếu không, việc đánh giá mô hình sẽ

không mang tính khách quan và dễ dẫn đến sai sót trong quá trình đánh giá. Phương pháp đơn giản nhất là chia toàn bộ tập dữ liệu thành hai tập dữ liệu bao gồm tập huấn luyện và tập kiểm tra. Sau đó, sử dụng một tập dữ liệu để huấn luyện mô hình học máy, và một để đánh giá khả năng phân loại của mô hình được chọn. Đây được gọi là phương pháp giữ lại (hold-out method). Hình 3.2 minh họa các mẫu dữ liệu đại diện có sẵn trong tập dữ liệu sau khi hoàn thành quá trình sắp xếp các mẫu dữ liệu với 23 đặc trưng được miêu tả như trên, tương ứng với từng lớp lưu lượng mạng được mã hóa trong khâu tiền xử lý dữ liệu.

	duration	min_fiat	min_biat	max_fiat	max_biat	mean_fiat	mean_biat	...	std_flowiat	min_active	mean_active	max_active	std_active	min_idle	mean_idle	max_idle	std_idle	class1
0	14993462.0	0.0	0.0	823486.0	854818.0	873.134288	287.914617	...	7050.781273	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'BROWSING'
1	14463281.0	0.0	0.0	742368.0	742339.0	1321.330258	312.290951	...	6982.036846	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'BROWSING'
2	14997099.0	1.0	0.0	537201.0	565232.0	1850.116210	344.855753	...	5560.260369	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'BROWSING'
3	14999986.0	2.0	0.0	954084.0	954062.0	1796.827863	382.537496	...	9375.105249	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'BROWSING'
4	14999090.0	2.0	0.0	1014690.0	1016593.0	1668.792029	394.588556	...	9205.461338	9578068.0	9578068.0	9578068.0	0.000000	1014624.0	1014624.0	1014624.0	0.000000	b'BROWSING'
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8960	14997258.0	69.0	203.0	21131.0	28475.0	19659.708661	19785.300792	...	4198.398294	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'VOIP'
8961	400609.0	400609.0	200451.0	400609.0	200451.0	400609.000000	200451.000000	...	115497.999657	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'VOIP'
8962	11257986.0	66.0	168.0	225830.0	63675.0	19960.968065	19753.841355	...	6398.053203	-1.0	0.0	-1.0	0.000000	-1.0	0.0	-1.0	0.000000	b'VOIP'
8963	4009751.0	154.0	331.0	2005165.0	2004379.0	801903.800000	801930.800000	...	810535.993475	2003874.0	2004518.0	2005162.0	910.752985	2003446.0	2003912.5	2004379.0	659.729869	b'VOIP'
8964	8272911.0	62.0	1.0	1030078.0	1029668.0	459606.166667	288368.000000	...	360925.668879	1002717.0	1213655.6	2003627.0	441785.036352	1001616.0	1012422.4	1029668.0	14330.336844	b'VOIP'

8965 rows × 24 columns

Hình 3.2: Một số giá trị đại diện từ những đặc trưng của mẫu trong tập dữ liệu

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2)
```

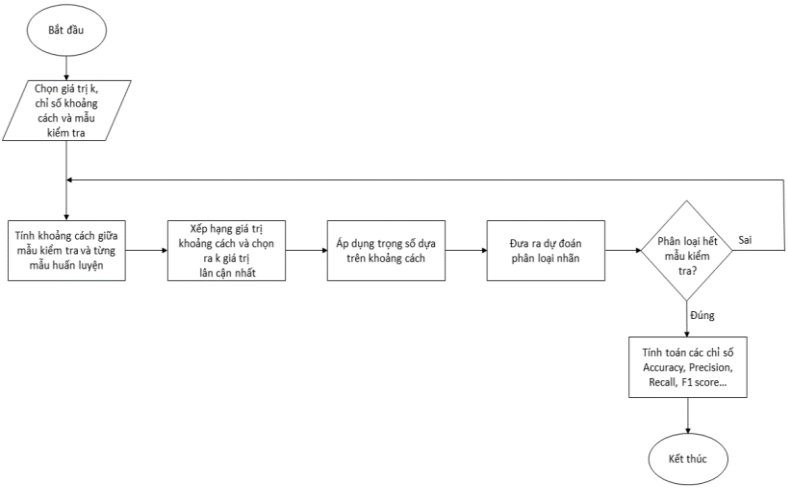
Trong quá trình áp dụng mô hình học máy, đề tài sử dụng tỷ lệ 8:2 nhằm phân chia tập dữ liệu huấn luyện và tập kiểm tra. Trong đó, *X_train* và *X_test* là tập hợp mẫu dữ liệu lần lượt có trong tập huấn luyện và tập kiểm tra. Kế tiếp, *y_train* và *y_test* là các nhãn dán tương ứng cho các mẫu dữ liệu trong 2 tập trên, trong đó *y_train* sẽ được sử dụng để huấn luyện mô hình, và sau đó *y_test* sẽ được dùng để so sánh kết quả phân loại, từ đó đánh giá hiệu suất phân loại của mô hình.

Tỷ lệ này được áp dụng sau khi tham khảo tỷ lệ phân chia tập dữ liệu từ các mô hình học máy khác nhau trong quá trình khảo sát trong mục 1.3. Ngoài ra, tỷ lệ này đảm bảo độ chênh lệch giữa tập huấn luyện và tập kiểm tra không quá lớn, đảm bảo tính khách quan trong quá trình phân loại và đánh giá mô hình huấn luyện.

3.1.1 KNN

Phương pháp KNN là một thuật toán phân loại bằng cách nhóm tất cả các mẫu thể hiện những đặc trưng tương tự của tập dữ liệu lại với nhau [22].

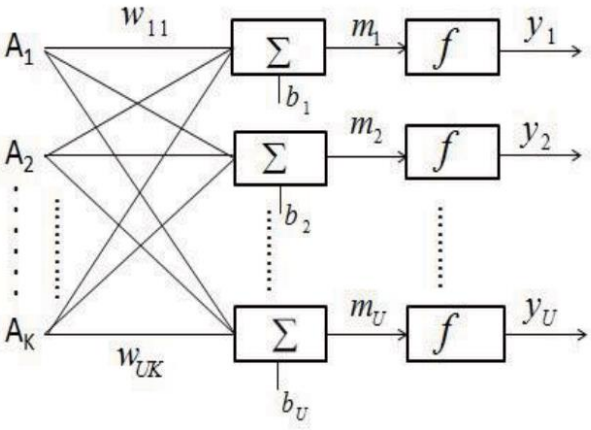
Sơ đồ khối: Quy trình phân loại lưu lượng Internet áp dụng cho mô hình KNN trọng số điểm lân cận được miêu tả bằng mô hình sơ đồ khối tại hình 3.4.



Hình 3.3: Sơ đồ khối mô hình phân loại KNN trọng số điểm lân cận

3.1.2 ANN

Nguyên tắc hoạt động: Trên thực tế, mạng máy tính nhân tạo có cấu trúc bao gồm sự liên kết của những giá trị dữ liệu đầu vào, đầu ra, và một số lượng lớn các tế bào neuron thần kinh đan xen. Mặc dù một mạng máy tính có thể được điều chỉnh để thích nghi với những tập dữ liệu đầu vào khác nhau, hoặc phụ thuộc vào mục đích của bài toán ứng dụng, cấu trúc của chúng thường bao gồm lớp dữ liệu đầu vào, lớp dữ liệu ẩn và lớp dữ liệu đầu ra.



Hình 3.4: Cấu trúc của một lớp dữ liệu ẩn

Trong hình 3.8, $\boldsymbol{A_i}$ $(1 \leq i \leq K)$ đại biểu các giá trị dữ liệu đầu vào, $\boldsymbol{w_{ij}}$ $(1 \leq i \leq K; 1 \leq j \leq U)$ là những giá trị trọng số, $\boldsymbol{b_j}$ $(1 \leq j \leq U)$ là giá trị độ lệch tương ứng với các nút mạng neuron, và $\boldsymbol{y_j}$ $(1 \leq j \leq U)$ là các giá trị đầu ra lớp dữ liệu ẩn này. Tổng trọng số của các giá trị đầu vào tại một neuron có thể được miêu tả theo công thức sau:

$$m_j = \sum_{i=1}^{K} w_{ij} A_i + b_j$$

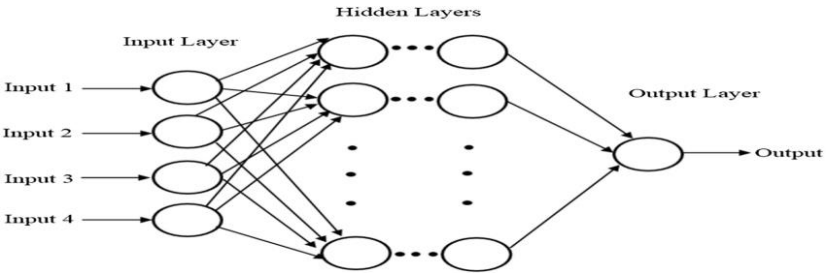
Như vậy từ công thức (3.7), với mỗi hàm kích hoạt $\boldsymbol{f}$ tùy theo nhu cầu của thuật toán, giá trị dữ liệu đầu ra của một nút mạng neuron được liên hệ trong công thức (3.10):

$$y_j = f(m_j)$$

Trong quá trình áp dụng mô hình ANN nhằm xây dựng mạng học sâu, các chỉ số *Dense* đại diện cho số lượng nút tại một lớp mạng ẩn, *activation* đại diện cho hàm kích hoạt được sử dụng tại lớp tương ứng, bao gồm hai giá trị là '*relu*' và 'softmax'. Mô hình huấn luyện sẽ được xây dựng dựa trên các chỉ số sau, từ đó áp dụng tập huấn luyện làm dữ liệu đầu và tiến hành huấn luyện. Sau đó, mô hính sẽ áp dụng nhóm dữ liệu kiểm tra, từ đó tiến hành phân loại và so sánh kết quả đầu ra nhằm đánh giá khả năng phân loại của mô hình.

```
#Define model
model = Sequential()
model.add(Dense(1024, input_shape=input_shape, activation = 'relu'))
model.add(Dense(512, activation = 'relu'))
model.add(Dense(256, activation = 'relu'))
model.add(Dense(256, activation = 'relu'))
model.add(Dense(128, activation = 'relu'))
model.add(Dense(64, activation = 'relu'))
model.add(Dense(num_classes, activation = 'softmax'))
```

Sơ đồ khối: Cấu trúc của mô hình huấn luyện ANN dung để phân loại lưu lượng Internet được miêu tả bằng mô hình sơ đồ tại hình 3.6.



Hình 3.5: Cấu trúc mô hình mạng ANN được áp dụng

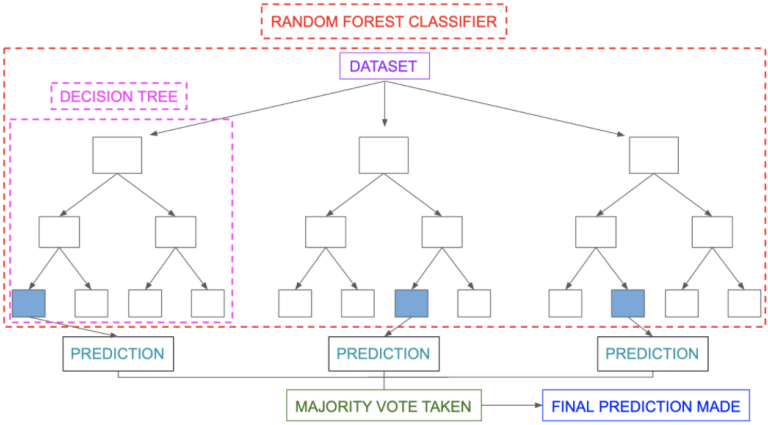
3.1.3 Rừng ngẫu nhiên (RF - Random Forest):

Nguyên tắc hoạt động của Rừng ngẫu nhiên: Sự khác biệt chính giữa mô hình huấn luyện cây quyết định và mô hình rừng ngẫu nhiên là việc thiết lập các nút gốc và các nút quyết định được thực hiện ngẫu nhiên trong những thuật toán sau. Mô hình rừng ngẫu nhiên sử dụng phương pháp đóng gói (bagging) để tạo ra các phép dự đoán cần thiết.

Quá trình thực hiện phương pháp đóng gói bao gồm việc sử dụng các tập dữ liệu khác nhau (dữ liệu huấn luyện) thay vì chỉ áp dụng một tập huấn luyện duy nhất. Tập dữ liệu huấn luyện bao gồm các mẫu dữ liệu và các đặc trưng tương ứng được sử dụng để đưa ra dự đoán. Các mô hình cây quyết định khác nhau tạo ra các kết quả phân loại nhãn đầu ra khác nhau, tùy thuộc vào dữ liệu huấn luyện được cung cấp tại đầu vào của thuật toán rừng ngẫu nhiên. Hình 3.14 miêu tả đơn giản quá trình đưa ra kết quả phân loại dựa trên ứng dụng mô hình huấn luyện rừng ngẫu nhiên.

Các kết quả đầu ra này sẽ được xếp hạng, và kết quả cao nhất sẽ được chọn làm kết quả phân loại đầu ra cuối cùng. Việc lựa chọn kết quả dự đoán cuối cùng tuân theo nguyên tắc đa số. Do đó, kết quả phân loại được lựa chọn bởi số lượng lớn các cây quyết định sẽ trở thành kết quả đầu ra cuối cùng của mô hình rừng ngẫu nhiên. Nói cách khác, việc tăng số lượng

cây quyết định trong việc thiết kế mô hình rừng ngẫu nhiên gia tăng độ chính xác trong quá trình phân loại kết quả dữ liệu đầu ra. Ngoài ra, mô hình rừng ngẫu nhiên cũng đạt hiệu quả cao hơn trong việc đưa ra dự đoán mà không cần quá trình tinh chỉnh siêu tham số. Kết quả phân loại cũng giải quyết được hiện tượng quá khớp hay gặp trong mô hình cây huấn luyện.



Hình 3.6: Sơ đồ khối quá trình phân loại nhãn của mô hình Rừng ngẫu nhiên

Trong mô hình thiết kế phạm vi đề tài này, sau quá trình thử nghiệm cho ra kết quả khả quan nhất đồng thời tham khảo từ các báo cáo nổi tiếng cùng lĩnh vực, số lượng mô hình cây quyết định tối đa được đặt trong mô hình rừng ngẫu nhiên, tương ứng với chỉ số *n_estimator*. Tương tự, số lượng tối đa nút có thể phân nhánh của một nhóm cây được khai báo bằng

biến *max_depth*. Đồng thời, thuật toán đánh giá chỉ tiêu phân nhánh trước khi đưa ra quyết định phân nhóm dữ liệu dựa trên giá trị Độ lợi thông tin ứng với *criterion* và *max_feature*. Từ đó, thuật toán *GridSearchCV* lần lượt áp dụng các thông số được liệt kê trong các biến trên và luân phiên áp dụng chúng nhằm tìm ra giá trị tối ưu nhất cho từng chỉ số, qua đó áp dụng vào mô hình Rừng Ngẫu nhiên để giải quyết bài toán phân loại.

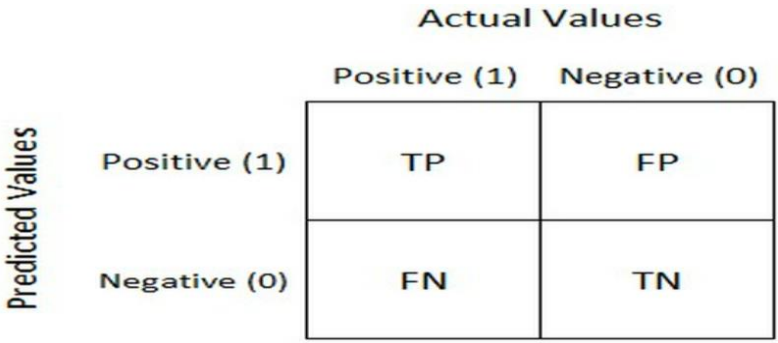
```
4  model = RandomForestClassifier()
5  param_grid = {'max_depth': [40,50,60],
6                'n_estimators': [90,100],
7                'max_features': ['auto','log2'],
8                'criterion': ['gini','entropy']}
9  GR = GridSearchCV(estimator = model, param_grid = param_grid,
   scoring = 'accuracy', cv = 6)
```

Chương 4: KẾT QUẢ THỰC NGHIỆM

4.1 Môi trường thực hiện

4.1.1 Ma trận hỗn loạn (Confusion Matrix)

Trong lĩnh vực máy học và cụ thể là trong các bài toán phân loại thống kê, ma trận hỗn loạn, hay còn được gọi là ma trận lỗi, là một ma trận thể hiện bố cục cụ thể cho phép người dùng hình dung được hiệu suất phân loại của một thuật toán. 4.1



Hình 4.1: Ma trận hỗn loạn

Dựa trên các thông số trên, những chỉ số đánh giá sau có thể được tính toán như Độ chính xác Accuracy, Precision, Recall, F1 score, v v…

4.1.2 Các chỉ số đánh giá

Độ chính xác (Accuracy): là tổng số lần mẫu kiểm tra được phân loại đúng lớp của mình, được minh họa bằng công thức sau:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (4.1)$$

Precision: Chỉ số đánh giá khi mẫu kiểm tra được phân loại thuộc về một lớp, số lần thực sự mà mẫu thuộc về lớp đó. Ví dụ trên tổng số lần mẫu kiểm tra được phân loại về lớp P2P, chỉ số Precision chỉ ra được tỷ lệ bao nhiêu mẫu thật sự thuộc về lớp đó.

$$Precision = \frac{TP}{TP + FP}, \quad (4.2)$$

Recall: thông số chỉ ra khi mẫu kiểm tra thực sự thuộc về một lớp, số lần mẫu kiểm tra được phân loại thuộc về lớp đó trong những lần dự đoán. Ví dụ như, khi mà một mẫu kiểm tra thuộc về lớp Web Browsing, chỉ số Recall tính toán tỷ lệ số lần hệ thống thường xuyên phân loại mẫu kiểm tra thuộc về lớp đó.

$$Recall = \frac{TP}{TP + FN}, \quad (4.3)$$

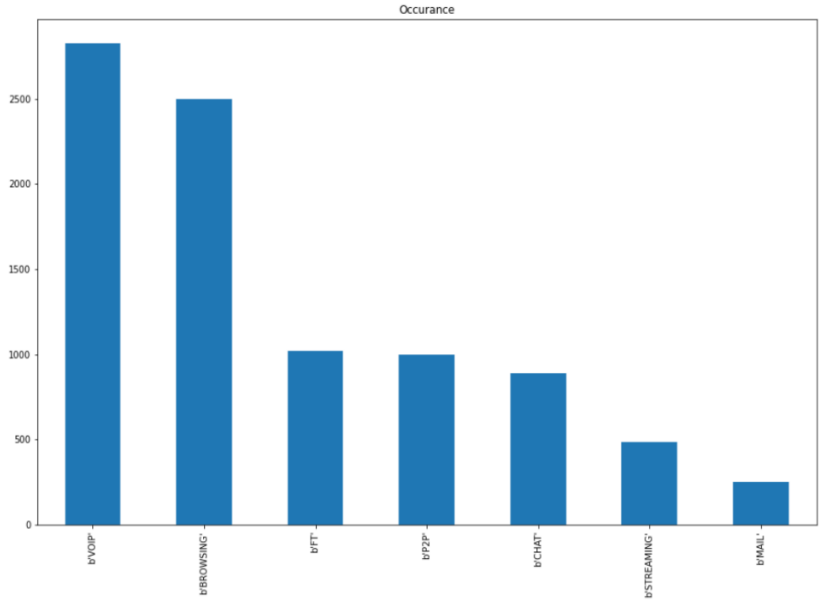
F1 score: Tổng hợp 2 chỉ số Precision và Recall, một mô hình có chỉ số F-score cao chỉ khi cả 2 chỉ số Precision và Recall đều cao. Một trong hai chỉ số này thấp đều sẽ kéo điểm F1-score xuống. Trường hợp xấu nhất khi 1 trong hai chỉ số Precison và

Recall bằng 0 sẽ kéo điểm F-score về 0. Trường hợp tốt nhất khi cả điểm chỉ số đều đạt giá trị bằng 1, khi đó điểm F-score sẽ là 1

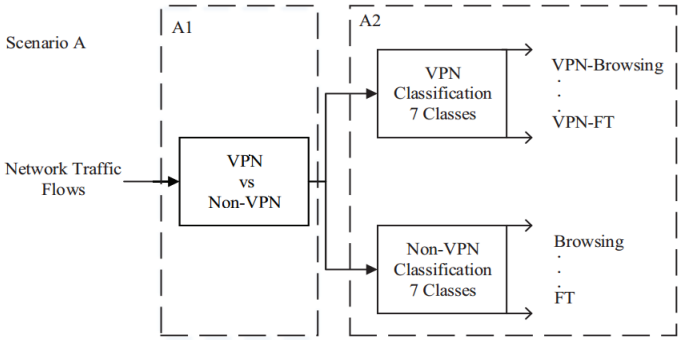
$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

4.1.3 Miêu tả các bối cảnh thí nghiệm

Dựa trên mật độ phân bổ mẫu dữ liệu không đồng đều của các lớp, những chỉ số được sử dụng để đánh giá hiệu suất của mô hình đề xuất sẽ là chỉ số Accuracy, Precision, Recall and F1 score.



Hình 4.2: Mật độ phân bố mẫu dữ liệu của các lớp

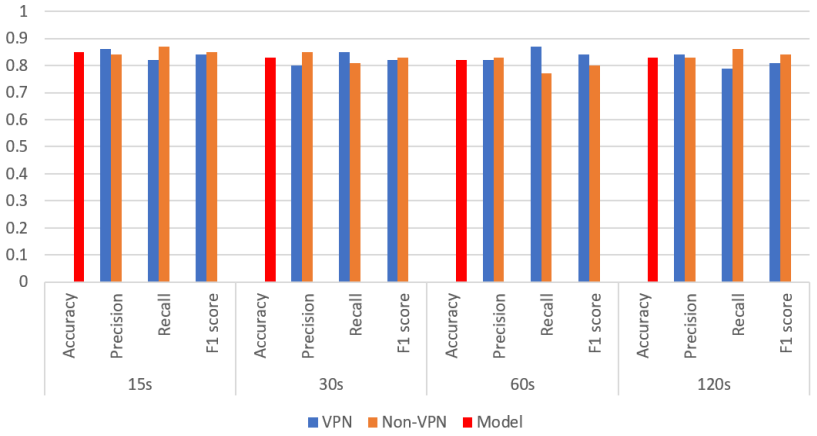


Hình 4.3: Mô hình minh họa 2 bối cảnh thí nghiệm A1 và A2

4.2 Kết quả thu được

4.2.1 Mô hình KNN

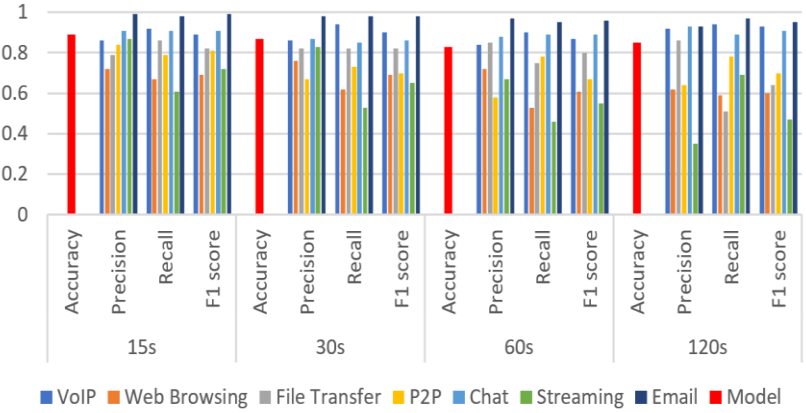
Bối cảnh A1: Trong bước phân loại này, mục tiêu của mô hình nhằm đến việc phân loại liệu các lớp dữ liệu đầu vào có được mã hóa VPN hoặc không, tương đương với một mô hình phân loại nhị phân. Tại bối cảnh này, tổng cộng có 4 nhóm dữ liệu dựa theo khung thời gian 15, 30, 60 và 120s. Trong hình 4.4, những giá trị Accuracy, Precision, Recall và F1 score, được báo cáo và biểu diễn dưới dạng biểu đồ cột.



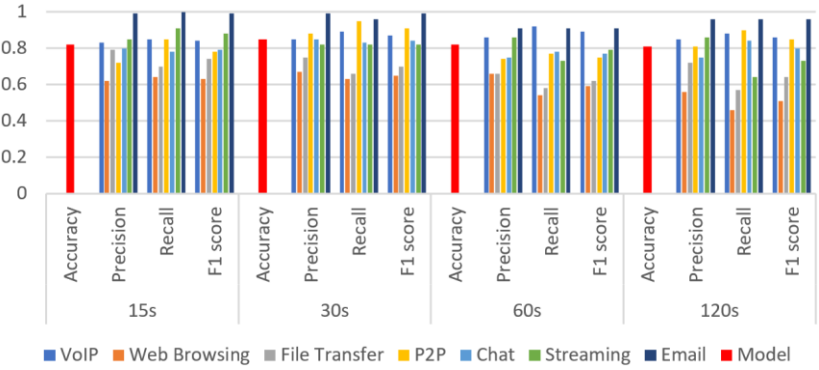
Hình 4.4: Biểu đồ chỉ số đánh giá kết quả phân loại bằng KNN - Bối cảnh A1

Bối cảnh A2: Trong bối cảnh A2, dữ liệu mạng Internet được mã hóa VPN và không mã hóa Non-VPN được phân loại thành 7 lớp lưu lượng mạng tương ứng. Biểu đồ so sánh chỉ số phân

loại cho từng loại dữ liệu mạng được cung cấp lần lượt trong hình 4.5 và 4.6



Hình 4.5: Biểu đồ chỉ số đánh giá không mã hóa Non-VPN bằng KNN - Bối cảnh A2



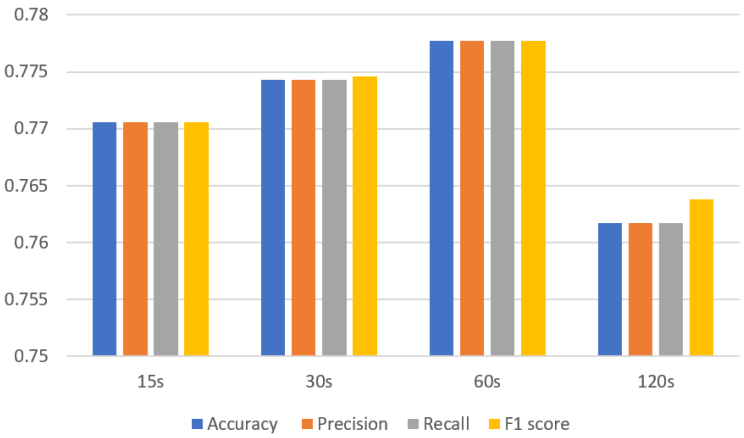
Hình 4.6: Biểu đồ chỉ số đánh giá mã hóa VPN bằng KNN - Bối cảnh A2

Trong nhóm các lớp dữ liệu lưu lượng mạng, lớp Email một lần nữa đạt được kết quả phân loại chính xác cao nhất. Kết quả Recall và Precision của các lớp VoIP, Chat, P2P và Streaming dao động khá nhiều qua các nhóm dữ liệu thời gian khác nhau, nhìn chung đạt kết qua cao hơn 80%..

4.2.2 Kết quả thu được – Mô hình ANN

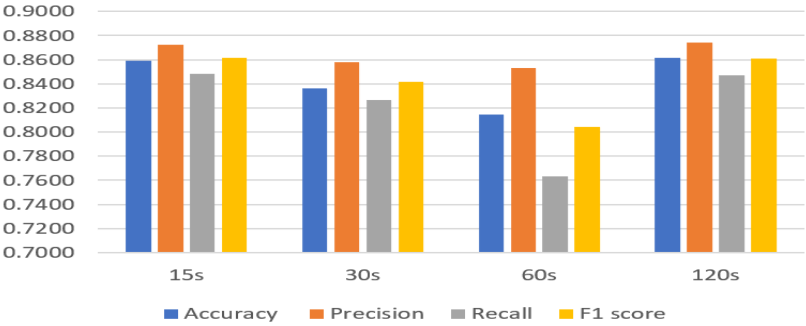
Tương tự với mô hình KNN, mô hình ANN cũng được đánh giá qua 2 bối cảnh A1 và A2 nhằm đánh giả khả năng phân loại lưu lượng Internet.

Bối cảnh A1: Tại bối cảnh phân loại dữ liệu không được mã hóa Non-VPN và có mã hóa VPN, tổng cộng có 4 nhóm dữ liệu dựa theo khung thời gian 15, 30, 60 và 120s.

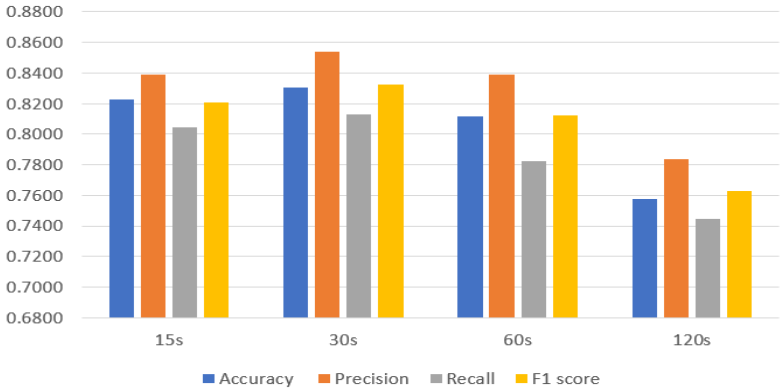


Hình 4.7: Chỉ số đánh giá kết quả phân loại bằng ANN - Bối cảnh A1

Bối cảnh A2: Trong bối cảnh A2, dữ liệu mạng Internet được mã hóa VPN và không mã hóa Non-VPN được phân loại thành 7 lớp lưu lượng mạng tương ứng. Hình 4.9 báo cáo kết quả phân loại 7 lớp lưu lượng không mã hóa VPN bằng mô hình huấn luyện ANN.



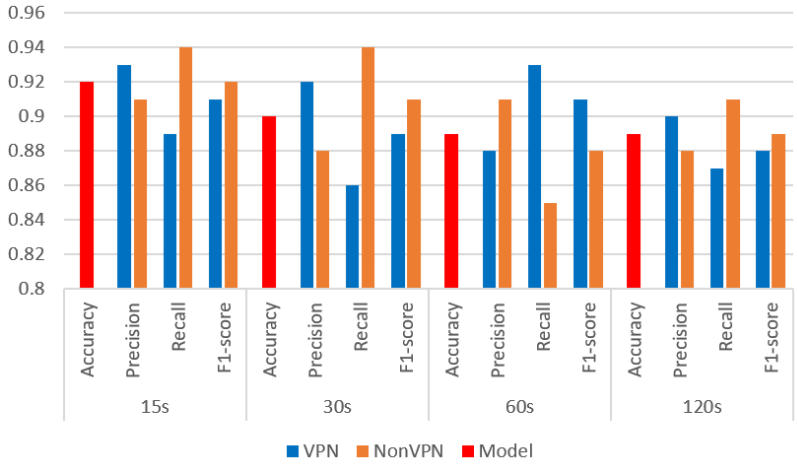
Hình 4.8: Chỉ số đánh giá kết quả phân loại bằng ANN – bối cảnh A2 Non-VPN



Hình 4.9: Chỉ số đánh giá kết quả phân loại bằng ANN – bối cảnh A2 mã hoá VPN

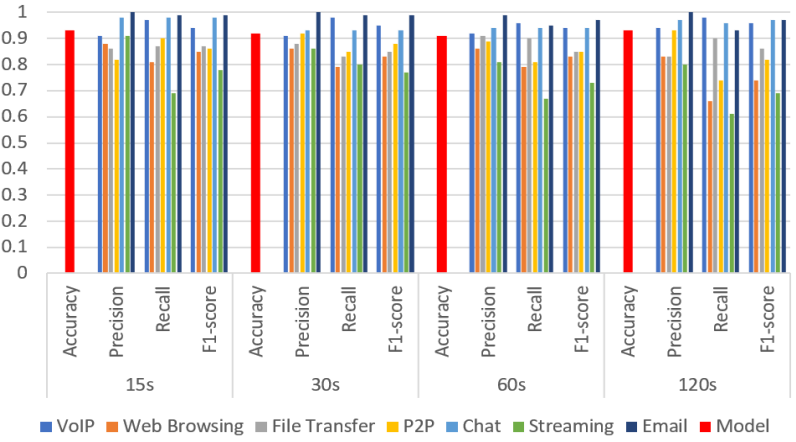
4.2.3 Kết quả thu được – Mô hình RF

Bối cảnh A1: Tại bối cảnh phân loại nhị phân cho tập dữ liệu gồm 2 nhóm dữ liệu không được mã hóa Non-VPN và có mã hóa VPN, tổng cộng có 4 nhóm dữ liệu dựa theo khung thời gian 15, 30, 60 và 120s.



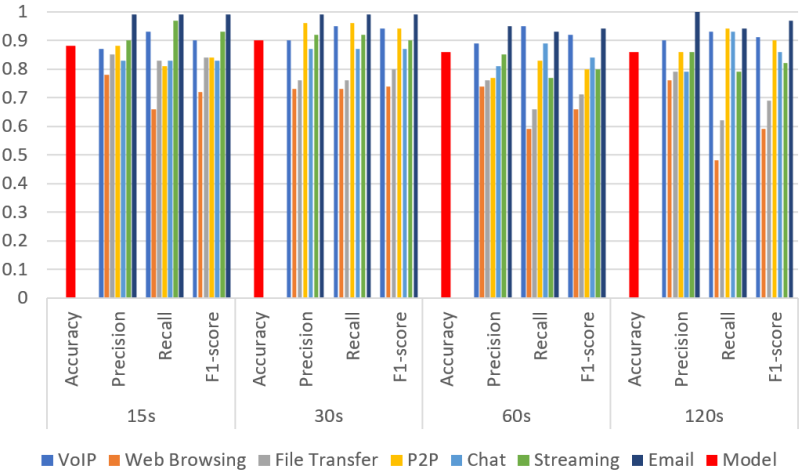
Hình 4.10: Biểu đồ chỉ số đánh giá kết quả phân loại bằng RF - Bối cảnh A1

Bối cảnh A2: Tương tự với 2 mô hình trên, trong bối cảnh này, mô hình Rừng ngẫu nhiên sẽ phân loại từng tập dữ liệu mạng mã hóa VPN và không mã hóa Non-VPN thành 7 lớp lưu lượng mạng khác nhau.



Hình 4.11: Biểu đồ chỉ số đánh giá không mã hóa Non-VPN bằng RF - Bối cảnh A2

.



Hình 4.12: Biểu đồ chỉ số đánh giá mã hóa VPN bằng RF - Bối cảnh A2

4.3 Kết quả tổng quan từ 3 mô hình

Bảng 4.1 đưa ra các chỉ số đánh giá kết quả phân loại tổng hợp của 3 mô hình huấn luyện KNN, ANN và RF. Trong 3 nhóm mô hình, giá trị cao nhất được đánh dấu đỏ.

Bảng 4.1: Kết quả tổng hợp phân loại mạng Internet của 3 mô hình KNN, ANN và RF – Bối cảnh A1

	Time	15s				30s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Scenario A1	KNN	84.86%	84.97%	86.68%	85.82%	82.58%	83.06%	84%	83.53%
	ANN	77.06%	77.06%	77.06%	77.06%	77.43%	77.43%	77.43%	77.46%
	RF	91.61%	90.60%	93.84%	92.19%	89.92%	88.01%	93.54$	90.70%
	Time	60s				120s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
	KNN	82.02%	82.14%	76.33%	79.13%	82.44%	80.60%	86.79%	83.58%
	ANN	77.77%	77.77%	77.77%	77.77%	76.17%	76.17%	76.17%	76.38%
	RF	89.45%	91.02%	84.99%	87.90%	88.87%	88.14%	90.77%	89.44%

Dựa vào bảng tổng kết trên, có thể thấy được trong bối cảnh A1, nhìn chung mô hình Rừng ngẫu nhiên RF cho kết quả phân loại cao nhất, tiếp đó là mô hình KNN và thấp nhất là ANN. Xuyên suốt tất cả chỉ số Accuracy, Precision, Recall và F1-score trong 4 nhóm dữ liệu thời gian, mô hình Rừng ngẫu nhiên luôn cho giá trị từ 88-93%, là một trong những kết quả cao nhất cho tập dữ liệu này. Tương tự, mô hình KNN cung cấp kết quả tương đối với giá trị vào khoảng 80-86%, với 2 giá trị Recall và F1-score 76.33% và 79.13%. Thấp nhất trong nhóm bối cảnh này là mô hình ANN khi phần lớn giá trị đánh giá nằm trong khoảng 77%.

Bảng 4.2: Kết quả tổng hợp phân loại mạng Internet của 3 mô hình KNN, ANN và RF – bối cảnh A2 - NonVPN

	Time	15s				30s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Scenario A2 Non-VPN	KNN	87.81%	71.02%	65.66%	68.24%	85.55%	73.86%	61.08%	66.86%
	ANN	85.91%	87.27%	84.83%	86.16%	83.62%	85.80%	82.66%	84.17%
	RF	93.46%	88.35%	81.48%	84.78%	92.10%	86.31%	79.23%	82.62%
	Time	60s				120s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
	KNN	84.77%	62.67%	60.99%	61.82%	92.76%	82.69%	66.15%	73.50%
	ANN	81.47%	85.32%	76.34%	80.44%	86.16%	87.45%	84.73%	86.11%
	RF	90.87%	86.38%	79.31%	82.70%	92.76%	82.69%	66.15%	73.50%

Ngược lại trong bối cảnh A2, các giá trị cao nhất được phân chia cho 2 mô hình ANN và RF. Trong tập dữ liệu Non-VPN, mô hình RF chiếm giá trị cao nhất đa phần trong các chỉ số Accuracy và Precision. Cụ thể hơn, chỉ số Accuracy qua các nhóm dữ liệu thời gian khác nhau đều lớn hơn 90%, và chỉ số Precision cũng đạt giá trị từ 86-88%. Còn lại trong mô hình ANN, nhóm chỉ số đạt giá trị cao nhất là Recall trong khoảng 82-84%, với ngoại lệ duy nhất trong nhóm dữ liệu 60s thuộc mô hình RF có chỉ số Recall trung bình là 79.31%. Tương tự, chỉ số F1-score đều mang giá trị cao trên 80% trên tất cả các nhóm dữ liệu thời gian ghi nhận thông tin.

Bảng 4.3: Kết quả tổng hợp phân loại mạng Internet của 3 mô hình KNN, ANN và RF – bối cảnh A2 - VPN

	Time	15s				30s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Scenario A2 VPN	KNN	83.05%	61.98%	62.67%	62.33%	83.28%	67.11%	59.53%	63.09%
	ANN	82.30%	83.88%	80.43%	82.09%	83.02%	85.38%	81.30%	83.24%
	RF	87.99%	77.63%	66.48%	71.62%	90.22%	76.15%	72.81%	74.44%
	Time	60s				120s			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
	KNN	82.75%	63.93%	54.93%	59.09%	88.29%	61.45%	68.92%	64.97%
	ANN	81.16%	83.92%	78.23%	81.23%	75.74%	78.38%	74.44%	76.29%
	RF	86.35%	73.95%	59.06%	65.67%	86.45%	76.06%	48.21%	59.02%

Như vậy, có thể kết luận mô hình Rừng ngẫu nhiên RF phù hợp nhất với mục tiêu phân loại Internet theo thứ tự bối cảnh A1 và A2. Kết quả phân loại của mô hình ANN cũng đồng thời cung cấp những kết quả phân loại có giá trị đáng kể. Từ đó, tùy thuộc vào mục đích phân loại lưu lượng của từng lớp sử dụng mạng khác nhau, mô hình ANN có thể được cân nhắc để áp dụng vào bài toán phân loại Internet tùy thuộc vào mục đích phân loại. Mô hình KNN tuy cung cấp một số giá trị cao, nhưng đánh giá tổng quan không phù hợp với tập dữ liệu mạng Internet được áp dụng trong phạm vi đánh giá của đề tài này.

KẾT LUẬN

1. Kết quả đạt được

Trong phạm vi nghiên cứu của đề tài này, các mô hình huấn luyện học máy khác nhau đã được đề xuất nhằm mục tiêu phân loại lưu lượng mạng Internet sử dụng tập dữ liệu mở ISCXVPN2016, bao gồm mô hình huấn luyện K – lân cận, Mạng Neuron nhân tạo và Rừng ngẫu nhiên. Trong đó, dựa trên các chỉ số đánh giá ghi nhận được cho mỗi mô hình huấn luyện cho 2 bối cảnh A1 và A2, Rừng ngẫu nhiên là mô hình đạt được kết quả phân loại cao và phù hợp nhất so với 2 mô hình huấn luyện còn lại. Nhìn chung, việc phân loại lưu lượng được mã hóa khó hơn nhiều so với việc phân loại lưu lượng không được mã hóa, chủ yếu vì các chuyên gia phân tích không thể thực hiện phương pháp phân tích gói chuyên sâu.

Điều này cũng được phản ánh rõ ràng trong kết quả phân loại lưu lượng Internet bối cảnh A2. Các chỉ số đánh giá kết quả phân chia đều cho các mô hình phân loại Rừng ngẫu nhiên và Mạng Neuron nhân tạo, thậm chí ở chỉ số phân loại F1-score kết quả của mô hình Rừng ngẫu nhiên tỏ ra yếu kém hơn khá nhiều so với mô hình mạng neuron. Tuy nhiên, khi cân nhắc toàn bộ bối cảnh thí nghiệm, xét từ kết quả bối cảnh A1 nối tiếp sau đó là bối cảnh A2, Rừng ngẫu nhiên có thể được xem là mô hình phù hợp nhất cho bài toán phân loại lưu lượng Internet trong

phạm vi đề tài này. Khi kết quả phân loại lưu lượng Internet được áp dụng với kết quả thành công cao, những chuyên gia an ninh mạng có thể xây dựng các chính sách bảo mật tiến hành tự động ngăn chặn khi phát hiện loại có sự truy cập của lưu lượng không mong muốn.

2 Phương hướng nghiên cứu

Với sự gia tăng tỷ lệ sử dụng internet, các ứng dụng mới và giao thức mới đang xuất hiện. Các ứng dụng mã hóa lưu lượng truy cập qua internet bằng các phương pháp và giao thức mã hóa để đảm bảo giao tiếp bí mật và an toàn. Trong các nghiên cứu trong tương lai, một tập dữ liệu mới có thể được thu thập để đánh giá các ứng dụng và giao thức mới, qua đó tiến hành thực hiện phân loại lưu lượng Internet. Khi các nghiên cứu trong lĩnh vực này tiếp tục được mở rộng với càng nhiều mô hình huấn luyện được thiết kế phù hợp hơn cho bài toán phân loại, phương hướng nghiên cứu tiếp theo là áp dụng các mô hình huấn luyện phức tạp hơn với khả năng tinh chỉnh siêu tham số. Từ đó, những mô hình phù hợp nhất cho bài toán này có thể được phát hiện và ứng dụng phù hợp với nhu cầu nghiên cứu và phát triển.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] B. M. Leiner, V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff, "A Brief History of the Internet," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 5, pp. 22–31, Oct. 2009.

[2] O. Salman, I. Elhajj, A. Chehab, and A. Kayssi, "IoT survey: An SDN and fog computing perspective," *Computer Networks*, vol. 143, 2018.

[3] L. Stewart, G. Armitage, P. Branch, and S. Zander, "An Architecture for Automated Network Control of QoS over Consumer Broadband Links," 2005, pp. 1–6.

[4] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 2451–2455.

[5] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Challenges in Network Application Identification," 2012.

[6] A. Razaghpanah, A. Akhavan Niaki, N. Vallina-Rodriguez, S. Sundaresan, J. Amann, and P. Gill, "Studying TLS Usage in Android Apps," 2017, pp. 350–362.

[7] B. Park, Y. Won, J. Chung, M. Kim, and J. W.-K. Hong, "Fine-grained traffic classification based on functional separation," *International Journal of Network Management*, vol. 23, no. 5, pp. 350–381, 2013.

[8] G. Aceto, A. Dainotti, W. Donato, and A. Pescapè, "PortLoad: Taking the Best of Two Worlds in Traffic Classification," 2010, pp. 1–5.

[9] D. Qin, J. Yang, J. Wang, and B. Zhang, "IP traffic classification based on machine learning," 2011.

[10] J. Dromard, P. Owezarski, V. Mozo, A. Ordozgoiti, and B. Vakaruk, “Delivrable Algorithms Description: Traffic pattern evolution and unsupervised network anomaly detectionONTIC D4.2,” 2016.

[11] N. Namdev, S. Agrawal, and S. Silkari, “Recent Advancement in Machine Learning Based Internet Traffic Classification,” *Procedia Computer Science*, vol. 60, pp. 784–791, 2015.

[12] k. claffy, “Internet traffic characterization,” UC San Diego, 1994.

[13] V. Paxson, “Empirically derived analytic models of wide-area TCP connections,” *IEEE/ACM Transactions on Networking*, vol. 2, no. 4, pp. 316–336, 1994.

[14] C. Dewes, A. Wichmann, and A. Feldmann, “An analysis of Internet chat systems,” *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 2003.

[15] Z. Yuan and C. Wang, “An improved network traffic classification algorithm based on Hadoop decision tree,” *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*. pp. 53–56, 2016.

[16] Y. Ma, Z. Qian, G. Shou, and Y. Hu, “Study of Information Network Traffic Identification Based on C4.5 Algorithm,” in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, 2008, pp. 1–5.

[17] M. Dixit, R. Sharma, S. Shaikh, and K. Muley, “Internet Traffic Detection using Naïve Bayes and K-Nearest Neighbors (KNN) algorithm,” in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1153–1157.

[18] L. Zhipeng, Z. Qin, K. Huang, X. Yang, and S. Ye, “Intrusion Detection Using Convolutional Neural

Networks for Representation Learning," 2017, pp. 858–866.

[19] I. Witten and I. H. Frank, *Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations*, vol. 31. Morgan Kaufmann Publishers, 2005.

[20] I. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Ed. Morgan Kaufmann Publishers, 2016.

[21] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. 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 (ICISSP2016)*, 2016, pp. 407–414.

[22] L. Jun, Z. Shunyi, L. Yanqing, and Z. Zailong, "Internet Traffic Classification Using Machine Learning," in *2007 Second International Conference on Communications and Networking in China*, 2007, pp. 239–243.

[23] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 665–671.

[24] W. Zhou, L. Dong, L. Bic, M. Zhou, and L. Chen, "Internet traffic classification using feed-forward neural network," in *2011 International Conference on Computational Problem-Solving (ICCP)*, 2011, pp. 641–646.