Đánh giá hiệu suất của các mô hình học máy dùng nhận dạng xâm nhập mạng

Nguyễn Văn Thành*, Nguyễn Thị Phong Dung**
Khoa Công nghệ thông tin, Trường Đại học Nguyễn Tất Thành, Thành phố Hồ Chí Minh
*thanhnv@ntt.edu.vn, **ntpdung@ntt.edu.vn

Tóm tắt

An ninh mạng ngày càng trở nên quan trọng đối với các hệ thống thông tin. Việc phát triển các hệ thống tự động có khả năng phát hiện và phân loại các hoạt động nguy hại là vô cùng cần thiết. Nghiên cứu này tập trung vào việc thực nghiệm và đánh giá hiệu suất của các mô hình học máy dùng để phát hiện các cuộc tấn công mạng, nhằm nâng cao hiệu quả của hệ thống nhận dạng xâm nhập (IDS). Các mô hình được huấn luyện và đánh giá trên một bộ dữ liệu chuẩn về xâm nhập mạng với mục tiêu phân loại lưu lượng mạng bình thường và lưu lượng tấn công mạng. Kết quả thực nghiệm cho thấy mỗi mô hình có những ưu điểm và hạn chế riêng trong việc phát hiện xâm nhập. Đặc biệt, mô hình Rừng ngẫu nhiên (Random Forest) cho thấy hiệu suất vượt trội trong việc đạt được độ chính xác cao và khả năng xử lý tốt với các đặc trưng phức tạp của dữ liệu mạng. Nghiên cứu này cung cấp một cái nhìn tổng quan về tiềm năng của các kỹ thuật học máy trong việc xây dựng các hệ thống IDS tự động và hiệu quả, đóng góp vào việc nâng cao khả năng phòng thủ cho các hệ thống mạng hiện đại. Các kết quả thu được có thể được sử dụng làm cơ sở để phát triển và triển khai các hệ thống nhận dạng xâm nhập thực tế.

Nhận 17/06/2025
Được duyệt 25/07/2025
Công bố 30/10/2025

Từ khóa
an ninh mạng, nhận dạng xâm nhập, học máy, rừng ngẫu nhiên, hồi quy logistic.

1. Giới thiệu

Sự phát triển nhanh chóng của công nghệ thông tin và mạng máy tính đã mang lại nhiều tiện ích cho xã hội hiện đại, đồng thời cũng kéo theo sự gia tăng đáng kể của các mối đe dọa an ninh mạng. Các cuộc tấn công mạng ngày càng tinh vi và thường xuyên hơn, gây tổn thất lớn về dữ liệu, tài chính, uy tín cho các cá nhân cũng như doanh nghiệp. Trong bối cảnh đó, việc phát triển các hệ thống nhận dạng xâm nhập mạng (Intrusion Detection Systems – IDS) hiệu quả trở thành một yêu cầu cấp thiết trong chiến lược đảm bảo an toàn thông tin mạng

Các IDS truyền thống dựa trên dấu hiệu nhận dạng (signature-based) có hiệu quả trong việc phát hiện các tấn công đã biết, nhưng gặp nhiều hạn chế khi phải đối mặt với các hình thức tấn công mới, chưa được biết trước [1]. Để khắc phục điểm yếu này, các kỹ thuật học máy (machine learning – ML) đã được vận dụng vào lĩnh vực an ninh mạng nhằm xây dựng các hệ thống IDS thông minh, có khả năng học và thích nghi từ dữ liệu, đồng thời phát hiện những hành vi bất thường tiềm ẩn trong lưu lượng mạng (traffic) ra vào hệ thống.

Đã có nhiều nghiên cứu áp dụng học máy trong nhận dạng xâm nhập mạng. Tuy nhiên, câu hỏi về việc lựa chọn mô hình học máy nào là tối ưu vẫn chưa có câu trả lời thống nhất. Mỗi mô hình học máy có những ưu và nhược điểm riêng, ảnh hưởng đến hiệu suất phân loại, nhận dạng, khả năng tổng quát hóa, chi phí tính toán và thời gian xử lý. Do đó, việc đánh giá và so sánh hiệu suất của các mô hình học máy trên cùng một bộ dữ liệu chuẩn là cần thiết, nhằm đưa ra những gợi ý phù hợp cho việc triển khai IDS trong thực tế.

Trong nghiên cứu này, chúng tôi sử dụng một tập dữ liệu nhận dạng xâm nhập đã được gán nhãn để tiến hành huấn luyện và đánh giá bằng nhiều mô hình học máy, bao gồm: Gaussian Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Gradient Boosting Classifier [3] và ANN Multi-layer Perceptron [4]. So sánh và đánh giá hiệu suất của các mô hình được dựa trên nhiều chỉ số như: độ chính xác, độ nhạy, thời gian huấn luyện và kiểm thử cùng với các chỉ số F1-score, AUC-ROC. Kết quả nghiên cứu nhằm cung cấp góc nhìn toàn diện về ưu điểm, nhược điểm của từng mô hình, từ đó đề xuất lựa chọn phù hợp cho việc xây dựng hệ thống IDS hiệu quả và khả thi.

2. Dữ liệu và tiền xử lý dữ liệu

2.1. Mô tả tập dữ liệu nhận dạng xâm nhập mạng

Nghiên cứu sử dụng một tập dữ liệu nhận dạng xâm nhập mạng được công bố công khai, bao gồm gần năm trăm nghìn mẫu ghi lại các thông tin của lưu lượng mạng. Tập dữ liệu này được thu thập trong thời điểm hệ

https://doi.org/10.55401/fbyam689

thống mạng đang bị nhiều hình thức tấn công cùng lúc. Mỗi mẫu trong tập dữ liệu đại diện cho một phiên kết nối với các đặc trưng kỹ thuật chi tiết và được gán nhãn "bình thường" (normal) cho những phiên kết nối của người dùng thông thường và nhãn "bất thường" (abnormal) cho những phiên kết nối tấn công mạng.

Tập dữ liệu bao gồm 42 đặc trưng, chia thành các nhóm chính:

- Nhóm *Thông tin kết nối*: gồm các đặc trưng biểu thị trạng thái và loại hình kết nối mạng như: thời gian kết nối, loại giao thức (*protocol_type*), dịch vụ được truy cập (*service*), cờ điều khiển (*flag*).
- Nhóm *Đặc trưng hành vi*: gồm các đặc trưng biểu thị loại yêu cầu xử lý, rà quét mạng, khai thác yếu điểm của giao thức mạng, tấn công xâm nhập như: số lần truy cập đến mục tiêu (*dst_host_count, srv_count*), tỷ lệ kết nối lỗi (*serror_rate, rerror_rate*).
- Nhóm *Thống kê định lượng*: gồm các đặc trưng biểu thị chỉ số thống kê như: số lần đăng nhập thất bại (*lgerror*), số đoạn bị phân mảnh (*frag_seg*), số byte truyền (*sbytes*), số byte nhận (*rbytes*)…

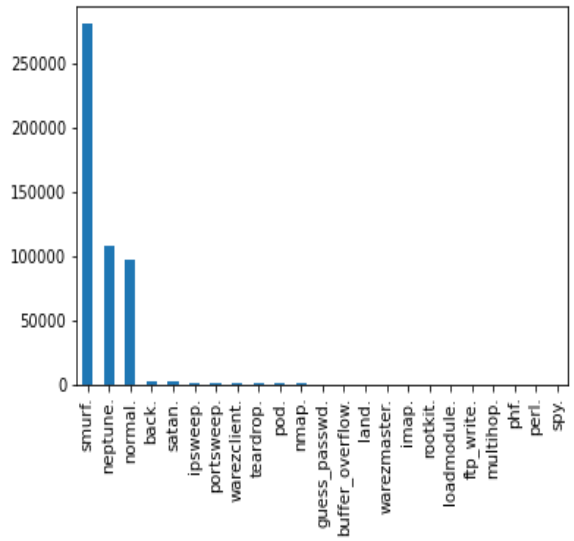
Các mẫu dữ liệu gán nhãn "*bất thường*" khi chúng có những đặc trưng được nhận dạng và phân loại các dạng tấn công mạng điển hình như:

- *DoS* (*Denial-of-Service*): tấn công từ chối dịch vụ.
- *Probe*: Quét mạng và thu thập thông tin.
- *R2L* (*Remote-to-Local*): kẻ tấn công từ bên ngoài cố gắng xâm nhập trái phép hệ thống.
- *U2R* (*User-to-Root*): kẻ tấn công đã có quyền truy cập và cố gắng leo thang đặc quyền.

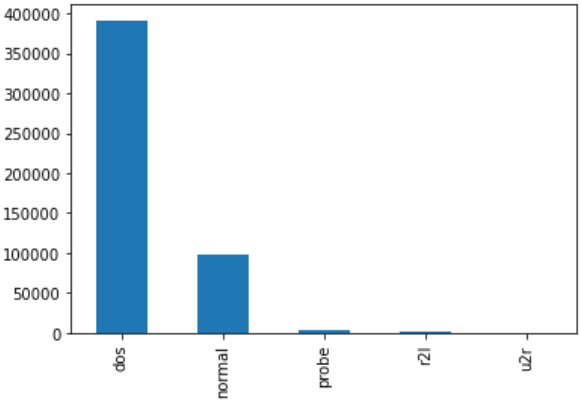
Phân tích sơ bộ cho thấy một số giá trị đặc trưng trong tập dữ liệu có phân bố mất cân bằng mạnh, ảnh hưởng đến quá trình huấn luyện và phân loại. Hai đặc trưng sau đây thể hiện rõ sự mất cân bằng đó:

- Loại ứng dụng nguồn tại phía người dùng: các phiên kết nối từ loại ứng dụng tấn công xâm nhập như: *smurt, neptune*… chiếm tỷ trọng rất cao so với những ứng dụng bình thường (Hình 1).
- Hình thức tấn công mạng: có sự chênh lệch lớn giữa các phiên kết nối với mục đích tấn công mạng so với phiên bình thường, trong đó phiên tấn công DoS chiếm tỷ lệ lớn nhất (Hình 2).

Để huấn luyện và đánh giá chính xác hơn, tập dữ liệu này cần được tiền xử lý và lựa chọn chỉ số đánh giá phù hợp.



Hình 1. Thống kê các loại ứng dụng từ người dùng.



Hình 2. Thống kê các hình tấn công mạng

2.2. Tiền xử lý dữ liệu

Nhằm đảm bảo độ chính xác và tính khả dụng cho các mô hình học máy, tập dữ liệu nhận dạng xâm nhập mạng được tiền xử lý [5] qua các bước sau:

2.2.1. Mã hóa các đặc trưng phân loại

Ba cột phân loại gồm *protocol_type, service* và *flag* có dạng chuỗi văn bản, cần được chuyển đổi sang dạng số bằng kỹ thuật Mã hóa nhãn (*Label Encoding*). Việc xử lý này giúp mô hình học máy xử lý các thuộc tính phân loại như dạng số, đồng thời giữ nguyên được thông tin định danh.

Một số giá trị được chúng tôi Mã hóa nhãn:

SF′: 0, ′S0′: 1, ′REJ′: 2, ′RSTR′: 3, ′RSTO′: 4, ′SH′: 5, ′S1′: 6,′S2′: 7,′RSTOS0′: 8,′S3′: 9,′OTH′: 10

2.2.2. Chuẩn hóa dữ liệu

Những đặc trưng định lượng có đơn vị và thang giá trị khác nhau được chuẩn hóa về cùng miền giá trị (0-1) bằng phép biến đổi *Min-Max Scaling*. Điều này giúp tăng tốc độ hội tụ và tránh hiện tượng mô hình bị chi phối bởi đặc trưng có giá trị lớn.

Phép biến đổi *Min-Max Scaling* được tính theo công thức (1)

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Trong đó:

- x' là giá trị đã được chuẩn hóa.
- x là giá trị gốc của đặc trưng.
- x_{min} là giá trị nhỏ nhất của đặc trưng trong tập dữ liệu.
- x_{max} là giá trị lớn nhất của đặc trưng trong tập dữ liệu

2.2.3. Kiểm tra và xử lý mất cân bằng dữ liệu

Do sự chênh lệch mạnh giữa các lớp (ví dụ: tấn công U2R và R2L rất hiếm), nghiên cứu cân nhắc sử dụng các kỹ thuật như SMOTE (*Synthetic Minority Over-sampling Technique*) [6] để tăng số lượng mẫu cho lớp thiểu số (minority class). Cùng với đó là sử dụng kỹ thuật Under-sampling để giảm số lượng mẫu của lớp đa số (majority class). Quá trình xử lý này nhằm giảm mất cân bằng của các lớp mà không làm sai lệch độ chính xác trong quá trình huấn luyện mô hình.

2.2.4. Kiểm tra tương quan và giảm chiều dữ liệu

Phân tích tương quan dữ liệu (*data correlation analysis*) được thực hiện để xác định các cặp đặc trưng có tương quan quá cao (lớn hơn |0,9|). Những đặc trưng xảy ra hiện tượng đa cộng tuyến (*multicollinearity*) rõ rệt thì một trong các đặc trưng đó sẽ được loại bỏ để tránh dư thừa thông tin và cải thiện hiệu quả mô hình.

Hình 3. Biểu đồ tương quan các đặc trưng trong tập dữ liệu

Dựa vào biểu đồ thể hiện tương quan các đặc trưng trong tập dữ liệu (Hình 3), chúng tôi loại bỏ các đặc trưng xảy ra hiện tượng đa cộng tuyến (màu nhạt nhất) và các đặc trưng có tương quan rất thấp (màu đậm nhất) với biến mục tiêu. Có 09 đặc trưng loại bỏ gồm:

srv_serror_rate, srv_rerror_rate,
num_root.dst_host_srv_serror_rate,
dst_host_rerror_rate, dst_host_same_srv_rate,
dst_host_serror_rate, dst_host_srv_rerror_rate.

Tập dữ liệu giảm còn 33 đặc trưng.

3. Chọn mô hình huấn luyện và Chỉ số đánh giá, so sánh

3.1. Chọn các mô hình học máy

Sau khi hoàn tất bước tiền xử lý, tập dữ liệu được chia thành 2 phần: tập huấn luyện (70 %) và tập kiểm thử (30 %), đảm bảo các lớp được phân bố đồng đều. Trong một số thử nghiệm, dữ liệu còn được chia bằng cách giữ tỉ lệ tương ứng với từng loại tấn công để tránh mất cân bằng.

Sau khi hoàn tất bước tiền xử lý, tập dữ liệu được chia thành 2 phần: tập huấn luyện (70 %) và tập kiểm thử (30 %), đảm bảo các lớp được phân bố đồng đều. Trong một số thử nghiệm, dữ liệu còn được chia bằng cách giữ tỉ lệ tương ứng với từng loại tấn công để tránh mất cân bằng. Các mô hình học máy được chọn trong nghiên cứu này bao gồm:

1) *Gaussian Naive Bayes* (GNB): mô hình đơn giản với giả định phân phối chuẩn, phù hợp với bài toán nhanh và tuyến tính.
2) *Decision Tree* (DT): mô hình phân tách dữ liệu dựa trên thuộc tính, dễ giải thích nhưng dễ bị quá khớp.
3) *Random Forest* (RF): mô hình tập hợp nhiều cây quyết định nhằm tăng độ chính xác và tính khái quát.
4) *Logistic Regression* (LR): mô hình tuyến tính dùng cho phân loại nhị phân.
5) *Support Vector Machine* (SVM): mô hình xác định siêu phẳng (*hyperplane*) nhằm tối ưu phân loại hai lớp dữ liệu.
6) *Gradient Boosting Classifier* (GBC): mô hình tăng cường tuần tự giúp tối ưu hoá hiệu suất phân loại.
7) *Multi-layer Perceptron* (MLP): mô hình mạng nơ-ron nhân tạo nhiều tầng giúp học các đặc trưng phức tạp và là nền tảng cho học máy sâu (*deep-learning*) [7].

Toàn bộ quá trình huấn luyện dựa trên 07 mô hình học máy trên được thực hiện trên cùng một môi trường tính toán để đảm bảo tính đồng nhất về điều kiện thực nghiệm. Hầu hết các mô hình được thiết lập với tham số mặc định trong thư viện *Scikit-learn*. Riêng đối với mô hình MLP, chúng tôi xây dựng một tầng ẩn (*hidden layer*) với số lượng neuron được điều chỉnh thử nghiệm. MLP được huấn luyện bằng thư viện *Keras* với thuật toán tối ưu *Adam* và hàm mất mát *Binary Cross-entropy*.

3.2. Các chỉ số đánh giá mô hình

Để đo lường hiệu quả của các mô hình học máy được triển khai trong nhận dạng xâm nhập mạng, nghiên cứu sử dụng một hệ thống các chỉ số đánh giá đa chiều, bao gồm *độ chính xác phân loại, hiệu suất tính toán* và *mức*

độ tổng quát hóa mô hình. Các chỉ số đánh giá này không chỉ giúp xác định khả năng phân loại giữa các phiên kết nối bình thường và bất thường, mà còn đánh giá mức độ phù hợp của từng mô hình trong môi trường thực tế yêu cầu tính ổn định, nhanh và chính xác.

Hiệu suất mô hình được đánh giá bằng các chỉ số sau:

- Thời gian huấn luyện (*Training Time*): là khoảng thời gian cần thiết để mô hình học từ tập dữ liệu huấn luyện. Đánh giá khả năng mở rộng và tính thực tiễn khi triển khai.
- Thời gian kiểm thử (*Testing Time*): là khoảng thời gian cần để mô hình phân loại toàn bộ tập kiểm thử, phản ánh tốc độ dự đoán và tính khả dụng trong môi trường thời gian thực.
- Độ chính xác (*Accuracy*): là thước đo phản ánh khả năng nhận diện chính xác lớp "bình thường" và "bất thường" là chỉ số đo mức độ chính xác của mô hình. Nghiên cứu này đánh giá Độ chính xác cho 2 quá trình: Độ chính xác trong huấn luyện (Training Accuracy) và Độ chính xác trong kiểm thử (*Testing Accuracy*).
- Điểm F1 (*F1-score*): chỉ số đánh giá độ tin cậy của chỉ số *Accuracy* khi huấn luyện mô hình trên tập dữ liệu mất cân bằng.
- AUC-ROC (*Area Under Curve - Receiver Operating Characteristic*): Biểu diễn khả năng của mô hình trong việc phân biệt hai lớp thông qua đường cong ROC. Diện tích dưới đường cong càng lớn, khả năng phân biệt càng tốt.

Chỉ số Độ chính xác (*Accuracy*) tính bằng tỷ lệ phần trăm của số mẫu được mô hình dự đoán "*bất thường*" trên tổng số mẫu "*bất thường*" trong tập dữ liệu (công thức 2)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (2)$$

Trong đó:

- TP (*True Positives*): số mẫu dự đoán "bất thường" là chính xác.
- TN (*True Negatives*): số mẫu dự đoán là "bất thường" nhưng thực tế là "bình thường".
- FP (*False Positives*): số mẫu "bất thường" mô hình dự đoán sai.
- FN (*False Negatives*): số mẫu "bình thường" mô hình dự đoán sai.

Khi huấn luyện trên dữ liệu mất cân bằng, độ tin cậy của chỉ số *Accuracy* sẽ giảm đi do các mô hình có khuynh hướng gia tăng số lượng mẫu dự đoán "bất thường", kể cả những mẫu "bình thường". Chỉ số F1-score đánh giá điểm tin cậy của chỉ số *Accuracy* bằng cách tính trung bình điều hòa giữa độ chính xác (*precision*) và độ nhạy (*recall*).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Trong đó:

- *Precision*: tỉ lệ giữa các mẫu dự đoán "bất thường" là chính xác (TP) trên tổng số mẫu "bất thường" trong tập dữ liệu (TP + FP).
- *Recall*: tỉ lệ giữa các mẫu dự đoán "bất thường" là chính xác (TP) trên tổng số mẫu mà mô hình dự đoán là "bất thường" (TP + FN).

3.3. Các tiêu chí so sánh

Các mô hình được so sánh dựa trên 4 tiêu chí:

1) So sánh *Độ chính xác tuyệt đối*: kiểm tra khả năng tổng quát hóa bằng cách đối chiếu kết quả huấn luyện và kiểm thử.
2) So sánh *Tính ổn định mô hình*: đánh giá độ chênh lệch giữa accuracy huấn luyện và kiểm thử để nhận biết hiện tượng quá khớp.
3) So sánh *Hiệu suất tính toán*: so sánh thời gian huấn luyện và kiểm thử để cân nhắc tính khả thi triển khai.
4) So sánh *Hiệu năng trên dữ liệu mất cân bằng*: F1-score và AUC được dùng để đảm bảo mô hình không thiên lệch theo lớp chiếm ưu thế.

Tất cả các đánh giá được tính toán với sự hỗ trợ của thư viện **Scikit-learn**, **Keras** (đối với ANN), và được tổng hợp định lượng trong bảng kết quả để dễ so sánh.

4. Kết quả huấn luyện và kiểm thử

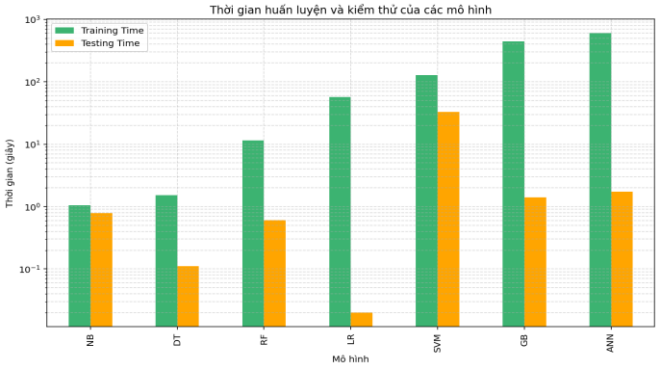
Kết quả quá trình huấn luyện và kiểm thử trên 07 mô hình học máy, được trình bày dưới dạng bảng và biểu đồ dưới đây:

Bảng 1. Tổng hợp kết quả thực nghiệm các mô hình sau quá trình huấn luyện và kiểm thử:

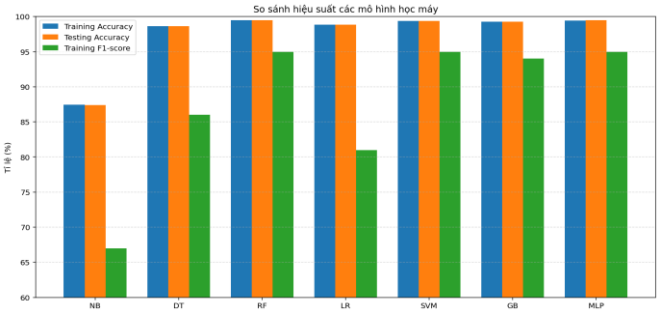
Mô hình	Thời gian huấn luyện (giây)	Thời gian kiểm thử (giây)	Độ chính xác huấn luyện (%)	Độ chính xác kiểm thử (%)	Điểm F1 (F1-score)
Gaussian Naive Bayes (GNB)	1,05	0,79	87,45	87,40	0,67
Decision Tree (DT)	1,51	0,11	98,66	98,65	0,86
Random Forest (RF)	11,45	0,61	99,51	99,47	0,95
Logistic Regression (LR)	56,67	0,02	98,85	98,85	0,81
Support Vector Machine (SVM)	126,96	32,73	99,38	99,38	0,95
Gradient Boosting (GB)	446,69	1,41	99,27	99,27	0,94
Multi-layer Perceptron (MLP)	605,75	1,73	99,41	99,49	0,95

Dữ liệu kết quả thực nghiệm được lập thành 2 biểu đồ nhằm trực quan cho đánh giá và thảo luận.

- Một biểu đồ so sánh hiệu suất thời gian huấn luyện (cột 1) và kiểm thử (cột 2) giữa các mô hình (Hình 4).
- Một biểu đồ so sánh độ chính xác khi huấn luyện (cột 1), độ chính xác kiểm thử (cột 2) và điểm F1 (cột 3) của các mô hình (Hình 5).



Hình 4. So sánh hiệu suất thời gian giữa các mô hình học máy.



Hình 5. So sánh độ chính xác và điểm F1 các mô hình học máy.

Từ kết quả thực nghiệm trên, chúng tôi đưa ra những nhận xét sơ bộ như sau:

- Các mô hình phức tạp như *Random Forest* (RF), *Gradient Boosting* (GB) và *Multi-layer Perceptron* (MLP) đạt độ chính xác và *F1-score* cao vượt trội, vượt ngưỡng 99 % về độ chính xác kiểm thử và trên 94 % về F1-score. Tuy nhiên, chúng đòi hỏi thời gian huấn luyện dài hơn đáng kể, đặc biệt là MLP với thời gian huấn luyện lên đến 605,75 giây.
- Ngược lại, các mô hình đơn giản như *Naive Bayes* (NB) hay *Decision Tree* (DT) có thời gian huấn luyện và kiểm thử ngắn, nhưng hiệu quả nhận dạng thấp hơn đáng kể, đặc biệt với NB chỉ đạt 87,40 % độ chính xác và 67 % F1-score trong kiểm thử.

5. Thảo luận kết quả

Kết quả thực nghiệm cho thấy sự đánh đổi giữa độ chính xác và chi phí tính toán khi lựa chọn mô hình học máy cho bài toán phát hiện xâm nhập mạng. Chúng tôi đưa ra một số nhận xét sau:

- Về hiệu suất mô hình: các mô hình có độ chính xác cao như RF, GB và MLP thể hiện khả năng học đặc trưng phức tạp và tổng quát hóa tốt trên tập kiểm thử. Trong đó, MLP đạt độ chính xác kiểm thử cao nhất (99,49 %) và F1-score đạt 95,0 %, cho thấy khả năng nhận diện tốt các mẫu tấn công và bình thường.
- Về chi phí tính toán: các mô hình có hiệu suất cao sẽ đi kèm với chi phí huấn luyện đáng kể. MLP và GB có thời gian huấn luyện rất lớn (trên 400 giây), điều này cần cân nhắc nếu triển khai trong môi trường thời gian thực hoặc tài nguyên tính toán hạn chế.
- Về sự ổn định của độ chính xác: các mô hình DT, LR, SVM, GB có chênh lệch độ chính xác giữa tập huấn luyện và kiểm thử gần như bằng 0, cho thấy tính ổn định của các mô hình này rất cao. Trong khi đó, MLP lại có chênh lệch độ chính xác lớn nhất (0,08 %) cho thấy mô hình này tiềm ẩn nguy cơ overfitting. Cần tinh chỉnh mô hình tối ưu hơn.
- Về sự ổn định của tính tin cậy: Điểm *F1-score* đánh giá mức tin cậy của độ chính xác. Chênh lệch giữa *F1-score* với *Accuracy* càng lớn thì mô hình càng bỏ sót các mẫu tấn công nhiều hơn. Các mô hình như RF, SVM, GB, MLP có chênh lệch giữa *Accuracy* và *F1-score* khá nhỏ, khả năng nhận dạng mẫu tấn công là ổn định và đáng tin cậy. Ngược lại, NB và LR có độ chênh khá cao (17-20) %, chứng tỏ hiệu suất nhận diện tấn công kém đáng kể, ảnh hưởng nghiêm trọng khi triển khai thực tế.
- Sự đánh đổi giữa tính đơn giản và hiệu quả: DT và LR mang lại hiệu suất tốt ở mức cơ bản (trên 98,6 % độ chính xác), đồng thời có thời gian huấn luyện ngắn, phù hợp cho các hệ thống nhẹ hoặc có yêu cầu thời gian đáp ứng nhanh.
- Mô hình NB thể hiện hiệu suất thấp hơn đáng kể, cả về độ chính xác và *F1-score*, cho thấy giả định độc lập giữa các thuộc tính đầu vào là không phù hợp với đặc trưng phức tạp của dữ liệu xâm nhập mạng.

Tổng kết lại, lựa chọn mô hình triển khai thực tế cần dựa trên mục tiêu cụ thể của hệ thống:

- Nếu ưu tiên độ chính xác tối đa và có năng lực tính toán mạnh thì chọn các mô hình như *Gradient Boosting*, *Multi-layer Perceptron* là phù hợp.
- Nếu cần độ chính xác tối đa nhưng năng lực tính toán không cao thì nên chọn mô hình *RF*, *SVM*.
- Với các ứng dụng thời gian thực yêu cầu nhẹ về tài nguyên, *DT* hoặc *LR* là những lựa chọn hiệu quả.
- Mô hình *GNB* không phù hợp cho nhận dạng xâm nhập.

6. Kết luận, hạn chế và hướng phát triển

Nghiên cứu này đã tiến hành đánh giá hiệu suất của bảy mô hình học máy trên bài toán nhận dạng xâm nhập mạng. Kết quả thực nghiệm cho thấy các mô hình *ANN*, *RF* và *GB* có khả năng phân loại rất tốt với độ chính xác trên 99,7 %, trong khi *GNB* có hiệu suất thấp hơn rõ rệt.

Sự chênh lệch về thời gian xử lý và độ chính xác giữa các mô hình phản ánh rằng không có lựa chọn tối ưu tuyệt đối, mà cần cân nhắc tùy theo mục tiêu ứng dụng cụ thể.

Hạn chế của nghiên cứu:

- Về phạm vi dữ liệu: nghiên cứu sử dụng tập dữ liệu chỉ bao gồm các hình thức tấn công mạng truyền thống như DoS, Probe, R2L, U2R… Thiếu dữ liệu của một số kiểu tấn công mạng mới phát sinh ngày nay như: tấn công APT (Advanced Persistent Threats), khai thác lỗ hổng bảo mật (vulnerability exploit).
- Về thiết lập mô hình: ngoại trừ mô hình ANN-MLP, các mô hình còn lại trong nghiên cứu được thiết lập tham số mặc định, chưa có thực hiện tối ưu siêu tham số (*hyperparameter tuning*) để tìm cấu hình tốt nhất. Các mô hình học sâu tiên tiến hơn như LSTM, Autoencoder cũng không được sử dụng trong nghiên cứu này do đòi hỏi cao về tài nguyên tính toán, tính khả dụng thấp với các hệ thống nhận dạng và ngăn chặn xâm nhập (IDS/IPS) nhỏ lẻ.

Hướng phát triển tiếp theo của nghiên cứu có thể bao gồm:

- Mở rộng sang các mô hình học sâu (deep learning) phức tạp hơn, như LSTM hoặc Autoencoder, để khai thác cấu trúc tuần tự trong lưu lượng mạng.
- Áp dụng các kỹ thuật phát hiện bất thường (anomaly detection) trên tập dữ liệu không gán nhãn (unsupervised learning).
- Tối ưu hóa mô hình bằng học bán giám sát để tận dụng phần lớn dữ liệu chưa được phân loại trong thực tế.
- Tích hợp mô hình vào hệ thống IDS thời gian thực và kiểm nghiệm trên dữ liệu mạng thực tế.
- Tiến hành tối ưu hóa siêu tham số (hyperparameter tuning) cho các mô hình để đạt được hiệu suất tối ưu hơn.
- Đánh giá khả năng mở rộng và chi phí tính toán khi triển khai các mô hình ở quy mô lớn.

Lời cảm ơn

Chúng tôi xin cảm ơn Trường Đại học Nguyễn Tất Thành, Thành phố Hồ Chí Minh đã hỗ trợ cho nghiên cứu này.

Tài liệu tham khảo

1. Tavallaee M., Bagheri E., Lu W., Ghorbani A. (2009). A detailed analysis of the KDD CUP 99 data set. *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications* (pp. 1-6).
2. Ahmed M., Mahmood A. N., Hu J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19-31.
3. Scikit-learn developers. (2025). *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org
4. Keras Developer guides (2025), *Training & evaluation with the built-in methods*. https://keras.io/guides/
5. DataCamp. (2025). *Data Preprocessing: A Complete Guide with Python Examples*. https://www.datacamp.com/blog/data-preprocessing
6. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
7. Goodfellow I., Bengio Y., Courville A. (2016). *Deep Learning*. MIT Press.

Evaluating the performance of machine learning models for network intrusion detection

Nguyen Van Thanh*, Nguyen Thi Phong Dung**
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City.
*thanhnv@ntt.edu.vn, **ntpdung@ntt.edu.vn

Abstract Cybersecurity is becoming increasingly important for information systems. The development of automated systems capable of detecting and classifying malicious activities is essential. This study focuses on testing and evaluating the performance of machine learning models used to detect cyber-attacks, improving the effectiveness of intrusion detection systems (IDS). The models are trained and evaluated on a standard network intrusion dataset, with the goal of classifying normal and attack network traffic. The experimental results demonstrate that each model has its own advantages and limitations in intrusion detection. In particular, the Random Forest overperforms other

models by achieving high accuracy and handling complex network data features well. This study provides an overview of the potential of machine learning techniques in building automated and efficient IDS systems, contributing to improving the defense capabilities of modern network systems. The obtained results can be used as a basis for developing and implementing practical intrusion recognition systems.
Keywords cybersecurity, intrusion detection, machine learning, random forest, logistic regression