TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN HUỲNH ĐĂNG MINH - 52200277**

**NGUYỄN CÔNG TOÀN - 52200271**

**PHÁT HIỆN VÀ NGĂN CHẶN HÀNH VI BẤT THƯỜNG TRONG HỆ THỐNG CNTT BẰNG IDS/IPS KẾT HỢP MACHINE LEARNING**

**CHUYÊN ĐỀ NGHIÊN CỨU 1**

**MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

**DỮ LIỆU**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN HUỲNH ĐĂNG MINH - 52200277**

**NGUYỄN CÔNG TOÀN - 52200271**

**PHÁT HIỆN VÀ NGĂN CHẶN HÀNH VI BẤT THƯỜNG TRONG HỆ THỐNG CNTT BẰNG IDS/IPS KẾT HỢP MACHINE LEARNING**

**CHUYÊN ĐỀ NGHIÊN CỨU 1**

**MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

**DỮ LIỆU**

Người hướng dẫn

**TS. Trần Chí Thiện**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn quý giảng viên đã luôn tận tâm giảng dạy, hướng dẫn và truyền đạt kiến thức một cách chi tiết, giúp chúng em hiểu sâu hơn về chuyên ngành. Sự tận tụy của thầy, cùng với sự hỗ trợ và tạo điều kiện của nhà trường, đã giúp chúng em có cơ hội hoàn thành dự án này, góp phần gắn kết giữa học tập – thực hành – nghiên cứu ứng dụng trong thực tế nghề nghiệp. Bên cạnh đó chúng em cũng chân thành cảm ơn phía nhà trường luôn tạo cơ hội lắng nghe sinh viên, giúp đỡ sinh viên khi sinh viên cần điều đó quá tuyệt vời. Chúng em xin chân thành cảm ơn ạ!

*TP. Hồ Chí Minh, ngày 12 tháng 12 năm 2025.*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Minh*

*Nguyễn Huỳnh Đăng Minh*

*Toan*

*Nguyễn Công Toàn*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Trần Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 12 tháng 12 năm 2025.*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

*Minh*

*Nguyễn Huỳnh Đăng Minh*

*Toan*

*Nguyễn Công Toàn*

**PHÁT HIỆN VÀ NGĂN CHẶN HÀNH VI BẤT**

**THƯỜNG TRONG HỆ THỐNG CNTT BẰNG IDS/IPS KẾT**

**HỢP MACHINE LEARNING**

**TÓM TẮT**

Đề tài nghiên cứu kỹ thuật phát hiện bất thường (Anomaly-based Detection) trong hệ thống an ninh mạng, với mục tiêu nâng cao khả năng phát hiện và ngăn chặn tấn công thông qua IDS/IPS kết hợp Machine Learning. Nghiên cứu tập trung so sánh hai hướng tiếp cận chính là phát hiện dựa trên chữ ký (signature-based) và phát hiện dựa trên bất thường, phân tích các mô hình như threshold-based, profile-based, cùng các phương pháp statistical, knowledge-based và data-mining.

Phần thực nghiệm triển khai mô hình giám sát kết hợp host-based (Wazuh) và network-based (Suricata + pfSense), thử nghiệm với các kịch bản tấn công như Nmap scanning, DoS, phishing và tấn công tầng ứng dụng để đánh giá hiệu quả phát hiện – ngăn chặn. Kết quả được sử dụng để điều chỉnh rules hoặc áp dụng học máy nhằm tự động sinh và tối ưu rules mới.

Đề tài đề xuất chu trình cải tiến tự động gồm: phát hiện → phân tích → tinh chỉnh rules/ML → thử nghiệm lại, trong đó kết hợp unsupervised và supervised learning cùng các nguồn dữ liệu như synthetic data và threat intelligence feeds để nâng cao khả năng thích ứng và phát hiện các mối đe dọa mới (zero-day).

**DETECTING AND PREVENTING ABNORMAL ACTIVITIES IN IT SYSTEMS USING IDS/IPS COMBINED WITH MACHINE LEARNING**

**ABSTRACT**

The research focuses on anomaly-based detection techniques in cybersecurity systems, aiming to enhance the ability to detect and prevent attacks through IDS/IPS integration with Machine Learning. The study compares two main approaches: signature-based detection and anomaly-based detection, while analyzing models such as threshold-based, profile-based, and methods including statistical, knowledge-based, and data-mining techniques.

The experimental implementation deploys a hybrid monitoring model combining host-based (Wazuh) and network-based (Suricata + pfSense) systems. Various attack scenarios—including Nmap scanning, DoS, phishing, and application-layer attacks—are executed to evaluate detection and prevention effectiveness. The results are then used to refine rules or apply machine learning to automatically generate and optimize new detection rules.

The research proposes an automated improvement cycle consisting of: detection → analysis → rule/ML optimization → retesting. The system integrates both unsupervised and supervised learning, utilizing datasets from synthetic sources and threat intelligence feeds to improve adaptability and enhance detection of emerging threats, including zero-day attacks.

**MỤC LỤC**

[DANH MỤC HÌNH VẼ 8](#_Toc218355630)

[DANH MỤC BẢNG BIỂU 9](#_Toc218355631)

[DANH MỤC CÁC CHỮ VIẾT TẮT 10](#_Toc218355632)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc218355633)

[1.1 Lý do chọn đề tài 1](#_Toc218355634)

[1.2 Mục tiêu thực hiện đề tài 2](#_Toc218355635)

[1.3 Tổng quan về đề tài 3](#_Toc218355636)

[1.4 Lý do lựa chọn kiến trúc Hybrid NIDS/HIDS 5](#_Toc218355637)

[1.5 Lý do lựa chọn Suricata cho NIDS/IPS 5](#_Toc218355638)

[1.6 Lý do lựa chọn pfSense làm thiết bị biên 5](#_Toc218355639)

[1.7 Lý do lựa chọn CICFlowMeter để trích xuất đặc trưng 6](#_Toc218355640)

[1.8 Lý do lựa chọn Wazuh cho HIDS/SIEM 6](#_Toc218355641)

[1.9 Lý do lựa chọn Isolation Forest cho phát hiện bất thường 6](#_Toc218355642)

[1.10 Lý do lựa chọn Autoencoder và VAE 7](#_Toc218355643)

[1.11 Lý do lựa chọn XGBoost cho phân loại tấn công đã biết 7](#_Toc218355644)

[1.12 Lý do lựa chọn CIC-IDS-2017 làm tập dữ liệu huấn luyện 7](#_Toc218355645)

[1.13 Lý do lựa chọn cơ chế phản ứng tại pfSense 8](#_Toc218355646)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 9](#_Toc218355647)

[2.1 Tổng quan an ninh mạng hiện này: 9](#_Toc218355648)

[2.2 Hệ thống phát hiện và ngăn ngừa xâm nhập (IDS/IPS) 11](#_Toc218355649)

[2.3 So sánh IDS (Intrusion Detection System) và IPS (Intrusion Prevention System): 13](#_Toc218355650)

[2.4 Tổng quan về hệ thống phát hiện và ngăn chặn xâm nhập (IDS/IPS) 14](#_Toc218355651)

[2.4.1 Giới thiệu về hệ thống phát hiện xâm nhập (Intrusion Detection System – IDS) 14](#_Toc218355652)

[2.4.2 Phân loại hệ thống phát hiện xâm nhập: 15](#_Toc218355653)

[2.4.3 Nguyên lý vận hành của hệ thống phát hiện xâm nhập: 16](#_Toc218355654)

[2.4.4 So sánh signature-based và anomaly-based detection: 17](#_Toc218355655)

[2.4.5 Giới thiệu về hệ thống ngăn ngừa xâm nhập (Intrusion Prevention Systems - IPS): 17](#_Toc218355656)

[2.4.6 Phân loại hệ thống ngăn ngừa xâm nhập: 18](#_Toc218355657)

[2.4.7 Các ưu điểm và hạn chế của hệ thống ngăn ngừa xâm nhập - IPS: 18](#_Toc218355658)

[2.5 Kỹ thuật phát hiện dựa trên bất thường (anomaly-based intrusion detection): 19](#_Toc218355659)

[2.5.1 Phân loại hệ thống ngăn ngừa xâm nhập: 19](#_Toc218355660)

[2.5.2 Những ưu điểm và nhược điểm của phát hiện dựa trên bất thường: 19](#_Toc218355661)

[2.5.3 Mô hình của một IDS dạng Anomaly-Based Intrusion Detection: 20](#_Toc218355662)

[2.5.4 Tổng quan cơ bản về Threshold (ngưỡng) và Profile (hồ sơ): 20](#_Toc218355663)

[2.6 Các kỹ thuật xác định threshold và profiles: 22](#_Toc218355664)

[2.6.1 Kỹ thuật dựa trên thống kê (Statistical-based techniques): 22](#_Toc218355665)

[2.6.2 Kỹ thuật dựa trên nhận thức (Cognitive-based or knowledge-based techniques): 22](#_Toc218355666)

[2.6.3 Kỹ thuật dựa trên khai thác dữ liệu (Data mining-based techniques): 23](#_Toc218355667)

[2.7 Các vấn đề và thách thức đối với phát hiện bất thường trong hệ thống ids/ips: 23](#_Toc218355668)

[2.8 Machine learning là gì và ứng dụng machine learning trong phát hiện và ngăn chặn xâm nhập (IDS/IPS): 24](#_Toc218355669)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT 26](#_Toc218355670)

[3.1 Kiến trúc hệ thống 26](#_Toc218355671)

[3.2 Quy trình huấn luyện Quy trình bao gồm các bước chính: 26](#_Toc218355672)

[3.3 Pipeline hệ thống: 26](#_Toc218355673)

[3.4 Các agents và server được sử dụng 27](#_Toc218355674)

[3.5 Quy trình huấn luyện Quy trình bao gồm các bước chính: 28](#_Toc218355675)

[3.5.1 Cấu hình máy pfsense 2.7.2 28](#_Toc218355676)

[3.5.2 Cài đặt và cấu hình suricata trên pfsense 2.7.2 29](#_Toc218355677)

[3.5.3 Cài đặt WAZUH bằng QUICKSTART 30](#_Toc218355678)

[3.5.4 Cài đặt wazuh agent trên pfsense 2.7.2 32](#_Toc218355679)

[3.6 Pfsense Firewall GUI và CLI: 34](#_Toc218355680)

[3.7 Pfsense package Suricata: 35](#_Toc218355681)

[3.8 Wazuh server: 36](#_Toc218355682)

[3.9 Kali: 36](#_Toc218355683)

[CHƯƠNG 4. THỰC NGHIỆM 37](#_Toc218355684)

[4.1 Dữ liệu thực nghiệm 37](#_Toc218355685)

[4.2 Cài đặt thực nghiệm 37](#_Toc218355686)

[4.2.1 Thu thập dữ liệu 37](#_Toc218355687)

[CHƯƠNG 5. KẾT LUẬN 38](#_Toc218355688)

[5.1 Kết luận 38](#_Toc218355689)

[5.2 Hướng phát triển 38](#_Toc218355690)

[TÀI LIỆU THAM KHẢO 39](#_Toc218355691)

# DANH MỤC HÌNH VẼ

[Hình 1. Mô hình phận loại của IDS bao gồm: NIDS (Network Intrusion Detection System) và HIDS (Host-based Intrusion Detection System). 15](#_Toc218235583)

[Hình 2. Mô hình cơ bản của một IDS dạng Anomallly-Based Detection. 19](#_Toc218235584)

[Hình 4. Các agents và server được sử dụng 28](#_Toc218235585)

[Hình 5. Pfsense Firewall GUI và CLI 35](#_Toc218235586)

[Hình 6. Pfsense package Suricata 35](#_Toc218235587)

[Hình 7. Wazuh server 36](#_Toc218235588)

[Hình 8. Kali 36](#_Toc218235589)

[Hình 9. Dữ liệu thực nghiệm 37](#_Toc218235590)

# DANH MỤC BẢNG BIỂU

[Bảng 1: So sánh IDS (Intrusion Detection System) và IPS (Intrusion Prevention System) 14](#_Toc218235730)

[Bảng 2. So sánh signature-based và anomaly-based detection 16](#_Toc218235731)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| ML | Machine Learning – Học máy |
| CIC-IDS-2017 | Bộ dữ liệu tấn công mạng của Viện CIC |
| PCAP | Packet Capture – Tập tin bắt gói tin |
| WAN | Wide Area Network – Mạng diện rộng |
| LAN | Local Area Network – Mạng cục bộ |

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Trong tình hình tấn công trên không gian mạng ngày càng được gia tăng một cách nhanh chóng về quy mô, tốc độ và mức độ tinh vi, thì các tổ chức cần những giải pháp, cơ chế phát hiện linh hoạt và phản ứng nhanh, có khả năng đáng tin cậy cao. Những phương pháp truyền thống cơ bản dựa trên chữ ký (signature-based) có tính ưu thế về độ chính xác với các mẫu tấn công đã biết, nhưng lại gặp một số hạn chế rõ rệt trước các mối đe dọa mới (zero-day) và các biến thể liên tục thay đổi. Việc duy trì và cập nhật hệ thống chữ ký cũng tiêu tốn nhiều thời gian và nguồn lực, khiến cho khả năng thích ứng chưa theo kịp thực tế, thời đại.

Phát hiện dựa trên bất thường (anomaly-based) mở ra hướng tiếp cận bổ sung quan trọng, cho phép hệ thống có thể học được “hành vi bình thường” của lưu lượng mạng để nhận diện những dấu hiệu lạ chưa từng có chữ ký. Các mô hình unsupervised như Isolation Forest hay Variational Autoencoder (VAE) thường xuyên chứng minh được khả năng có thể bắt được các mẫu bất thường – nền tảng để phát hiện ra các zero-day. Khi sử dụng kết hợp thêm học có giám sát (supervised) như XGBoost được huấn luyện trên dữ liệu nhãn (ví dụ CIC-IDS-2017), hệ thống có thể vừa được phân loại chính xác các cuộc tấn công đã biết, vừa được mở rộng được phạm vi phát hiện đối với các dấu hiệu mới.

Từ đó, đề tài hướng tới kiến trúc hybrid kết hợp IDS/IPS với máy học ba lớp: signature-based, supervised và anomaly-based. Kiến trúc này tận dụng điểm mạnh của từng lớp – signature cho độ chính xác đối với những mẫu đã biết; supervised được dùng cho khả năng có thể phân loại đa dạng hơn; anomaly sử dụng cho vùng phủ zero-day – đồng thời giảm thiểu điểm yếu khi vận hành các hệ thống đơn lẻ, độc lập. Việc đặt IDS ở chế độ giám sát để xây dựng nền tảng quan sát rộng, rồi nâng các phát hiện có độ tin cậy cao thành rule IPS để chặn chủ động, giúp cân bằng giữa an toàn vận hành và hiệu quả ngăn chặn.

Một lý do quan trọng khác là khả năng cải tiến liên tục. Chu trình phát hiện → phân tích → tinh chỉnh rule/ML → thử nghiệm lại biến hệ thống thành một vòng lặp học hỏi theo thời gian. Dựa trên log, thống kê và dữ liệu mới (bao gồm cả synthetic và dữ liệu tự thu), mô hình có thể được tái huấn luyện, ngưỡng được điều chỉnh, và các rule được cập nhật nhằm giảm false positive/false negative. Khi kết hợp tương quan HIDS/NIDS (Wazuh trên host và Suricata trên mạng), hệ thống còn nâng độ tin cậy cảnh báo và nhận diện chuỗi tấn công tốt hơn.

Về tính thực tiễn, đề tài sử dụng các thành phần mã nguồn mở và phổ biến (Suricata, pfSense, Wazuh, CICFlowMeter/NFStream, Python), giúp dễ triển khai, tối ưu chi phí và linh hoạt mở rộng. Dữ liệu huấn luyện từ CIC-IDS-2017 cùng dữ liệu tự thu/labeled tạo nền tảng vững chắc cho cả supervised lẫn unsupervised. Khả năng tích hợp thêm threat intelligence feeds và cơ chế sinh/tối ưu rule tự động từ kết quả ML cho phép hệ thống thích ứng tốt hơn với các mối đe dọa mới.

Cuối cùng, đề tài mang giá trị cả về mặt khoa học lẫn ứng dụng: vừa đề xuất kiến trúc hybrid có tính hệ thống, vừa trình diễn quy trình tự động hóa để nâng cao khả năng phát hiện–ngăn chặn. Mục tiêu là xây dựng một hệ thống phát hiện xâm nhập hiện đại, vừa bền vững, vừa linh hoạt, sẵn sàng thích ứng trước tốc độ thay đổi của mối đe dọa trên không gian mạng.

## Mục tiêu thực hiện đề tài

Đề tài hướng tới nâng cao khả năng phát hiện và ngăn chặn tấn công trong hệ thống CNTT bằng cách kết hợp IDS/IPS với kỹ thuật phát hiện bất thường và Machine Learning. Mục tiêu là bao phủ cả các tấn công đã biết lẫn các mối đe dọa mới (zero-day), đồng thời giảm cảnh báo sai và cải thiện khả năng thích ứng hệ thống.

Về lý thuyết, đề tài trình bày tổng quan an ninh mạng, chuỗi tấn công và đặc trưng hành vi trên mạng/host. Tiếp theo, làm rõ khái niệm, vị trí triển khai và nguyên lý hoạt động của IDS/IPS, cùng các rủi ro vận hành như false positives/false negatives và lựa chọn fail-open/fail-close.

Đề tài so sánh hai hướng tiếp cận chính: phát hiện dựa trên chữ ký (signature-based) và phát hiện dựa trên bất thường (anomaly-based). Trong anomaly-based, nghiên cứu các mô hình dựa trên ngưỡng (threshold) và hồ sơ (profile), cùng các phương pháp statistical, knowledge-based và data-mining, để nêu ưu – nhược điểm và thách thức thực tế như alert fatigue, concept drift, chi phí xử lý log.

Về phương diện Machine Learning, đề tài lựa chọn kết hợp unsupervised (Isolation Forest, VAE) cho phát hiện bất thường và supervised (XGBoost) cho phân loại tấn công đã biết. Tập đặc trưng được trích xuất từ flow (CICFlowMeter), tối ưu cho hiệu quả mô hình và khả năng vận hành thực tế.

Trong phần thực nghiệm, đề tài triển khai môi trường giám sát hybrid: network-based (Suricata trên pfSense) kết hợp host-based (Wazuh). Các kịch bản thử nghiệm bao gồm quét cổng bằng Nmap, DoS/Slowloris, tấn công tầng ứng dụng (SQLi/XSS), và social engineering như phishing hay lừa người dùng chạy script/malware.

Hệ thống vừa cảnh báo (alert, log, webhook), vừa hỗ trợ ngăn chặn chủ động (block IP ở biên, Active Response trên Wazuh). Nếu một kịch bản chưa được ngăn chặn hiệu quả, nhóm sẽ điều chỉnh ngưỡng/rule hoặc áp dụng ML để sinh rule mới, sau đó thử nghiệm lại nhằm cải thiện chỉ số.

Đề tài cũng đề xuất cơ chế khởi tạo khi thiếu dữ liệu thực: sử dụng synthetic data, threat intelligence feeds và transfer learning để bootstrap. Các nguồn này giúp hệ thống sớm đạt ngưỡng phát hiện hữu ích và tiếp tục cải thiện khi có log thực tế.

Cuối cùng, đề tài đưa ra chu trình cải tiến liên tục: phát hiện → phân tích → tinh chỉnh rules/ML → thử nghiệm lại. Chu trình này giúp hệ thống học hỏi theo thời gian, giảm cảnh báo sai, tăng độ bao phủ tấn công, và duy trì cân bằng giữa an toàn vận hành (IDS-first) và hiệu quả ngăn chặn (IPS cho phát hiện có độ tin cậy cao).

## Tổng quan về đề tài

Đề tài tập trung vào phát hiện bất thường (anomaly-based detection) trong hệ thống an ninh mạng, được triển khai theo kiến trúc hybrid kết hợp IDS/IPS và Machine Learning. Mục tiêu là nâng cao năng lực phát hiện–ngăn chặn với độ bao phủ tốt cả tấn công đã biết lẫn zero-day, đồng thời giảm cảnh báo sai và tăng khả năng thích ứng của hệ thống trước thay đổi hành vi lưu lượng.

Về mặt lý thuyết, đề tài hệ thống hóa khái niệm, phân loại và nguyên lý hoạt động của IDS/IPS, vị trí triển khai (inline vs out-of-band), đặc điểm từng tiếp cận (signature-based vs anomaly-based), cùng rủi ro vận hành (false positives/false negatives, alert fatigue, concept drift). Phần chuyên sâu phân tích các mô hình anomaly theo hai hướng chính: threshold-based (điều chỉnh ngưỡng theo thống kê) và profile-based (học hồ sơ hành vi bình thường), đồng thời đặt chúng trong khung phương pháp phổ biến như statistical (phân tích phân phối/điểm bất thường), knowledge-based (rule/tri thức chuyên gia) và data-mining (học máy trên đặc trưng flow/host). Đề tài lý giải lựa chọn mô hình ML phù hợp: unsupervised (Isolation Forest, VAE) để phát hiện bất thường và supervised (XGBoost) để phân loại tấn công đã biết, gắn với tập đặc trưng trích xuất từ luồng mạng (flow features).

Về thực hành, đề tài xây dựng môi trường giám sát hybrid gồm network-based (Suricata trên pfSense) và host-based (Wazuh). Lưu lượng được thu thập và chuyển thành flow (ví dụ qua CICFlowMeter) để chạy qua pipeline 3 lớp: signature (Suricata), supervised (XGBoost) và anomaly (IForest/VAE). Các kịch bản thử nghiệm bao gồm: Nmap port scan, tấn công DoS/Hulk/Slowloris, tấn công tầng ứng dụng (SQLi/XSS) và tình huống social engineering như phishing/lừa người dùng chạy malware/script. Hệ thống vừa cảnh báo (log/alert/webhook/Wazuh) vừa hỗ trợ ngăn chặn (block IP trên pfSense, Active Response của Wazuh) cho các phát hiện có độ tin cậy cao.

Đề tài cũng xác lập quy trình cải tiến liên tục: phát hiện → phân tích → tinh chỉnh ngưỡng/rules/mô hình → thử nghiệm lại. Khi một kịch bản chưa được ngăn chặn hiệu quả, nhóm điều chỉnh rule hoặc áp dụng ML (tái huấn luyện, tối ưu ngưỡng anomaly, bổ sung đặc trưng) để nâng cao chỉ số. Trường hợp thiếu dữ liệu thực, đề tài khuyến nghị dùng synthetic data, threat intelligence feeds và transfer learning để bootstrap, sau đó hấp thụ dữ liệu mới vào vòng lặp cải tiến.

Tổng quan, đề tài vừa có cơ sở lý thuyết vững (khung IDS/IPS, so sánh signature–anomaly, mô hình và thách thức vận hành) vừa có triển khai thực nghiệm rõ ràng (môi trường hybrid, pipeline ML, kịch bản tấn công, chỉ số đánh giá và quy trình tối ưu). Kết quả mong đợi là một hệ thống giám sát–ngăn chặn có thể vận hành thực tế, giảm cảnh báo sai, tăng độ bao phủ và sẵn sàng thích ứng trước các mối đe dọa mới.

## Lý do lựa chọn kiến trúc Hybrid NIDS/HIDS

Hệ thống được thiết kế theo mô hình Hybrid NIDS/HIDS, kết hợp giữa phương pháp phát hiện dựa trên chữ ký (signature-based) và phát hiện bất thường bằng học máy (ML-based anomaly detection). Lý do lựa chọn kiến trúc này là vì các hệ thống IDS/IPS truyền thống chỉ có hiệu quả cao với các dạng tấn công đã biết, trong khi các tấn công mới (zero-day) hoặc tấn công biến thể thường không có chữ ký sẵn. Việc kết hợp thêm các mô hình học máy không giám sát và giám sát giúp hệ thống vừa duy trì độ chính xác cao với các tấn công phổ biến, vừa mở rộng khả năng phát hiện hành vi bất thường chưa từng xuất hiện trước đó. Ngoài ra, việc tích hợp HIDS/SIEM cho phép tương quan sự kiện từ nhiều nguồn và triển khai phản ứng tự động, nâng cao khả năng phòng thủ tổng thể của hệ thống.

## Lý do lựa chọn Suricata cho NIDS/IPS

Suricata được lựa chọn làm thành phần NIDS/IPS chính vì đây là một hệ thống IDS/IPS mã nguồn mở có hiệu năng cao, hỗ trợ xử lý đa luồng và có hệ sinh thái chữ ký (rules) phong phú. Suricata cho phép phát hiện hiệu quả các dạng tấn công đã biết thông qua signature, đồng thời xuất dữ liệu cảnh báo dưới định dạng EVE JSON, một định dạng chuẩn, dễ dàng tích hợp vào pipeline phân tích bằng Python và hệ thống SIEM như Wazuh. So với Snort, Suricata thuận tiện hơn trong việc xử lý dữ liệu JSON và mở rộng pipeline phân tích. Trong khi đó, Zeek mạnh về phân tích hành vi và tầng ứng dụng nhưng không phù hợp để đóng vai trò IPS signature truyền thống, do đó không đáp ứng đầy đủ yêu cầu chặn và phát hiện theo chữ ký trong hệ thống này.

## Lý do lựa chọn pfSense làm thiết bị biên

pfSense được lựa chọn làm thiết bị biên của hệ thống vì đây là nền tảng firewall/router mã nguồn mở phổ biến, ổn định và dễ quản trị trong môi trường thực tế. pfSense hỗ trợ sẵn các công cụ như tcpdump để thu thập lưu lượng mạng và easyrule để thực hiện chặn IP nhanh chóng, đồng thời cho phép điều khiển từ xa thông qua SSH/SCP. Việc chặn tấn công được thực hiện tại pfSense giúp giảm rủi ro gián đoạn dịch vụ so với triển khai IPS inline trực tiếp trên luồng dữ liệu. So với việc sử dụng iptables hoặc nftables thuần trên Linux, pfSense giúp giảm đáng kể công sức triển khai và vận hành, đặc biệt trong các môi trường mạng đã sẵn có firewall chuyên dụng.

## Lý do lựa chọn CICFlowMeter để trích xuất đặc trưng

CICFlowMeter được sử dụng để trích xuất đặc trưng mạng từ file PCAP vì công cụ này cung cấp bộ đặc trưng flow chuẩn, tương thích trực tiếp với các tập dữ liệu CIC-IDS. Việc sử dụng CICFlowMeter giúp bảo đảm tính nhất quán giữa dữ liệu huấn luyện và dữ liệu vận hành thực tế, giảm thiểu sai lệch đặc trưng (feature mismatch). So với các công cụ khác như NFStream hoặc log của Zeek, CICFlowMeter có ưu điểm là đầu ra CSV rõ ràng, dễ tích hợp vào pipeline học máy hiện có mà không cần xây dựng lại toàn bộ hệ thống ánh xạ đặc trưng.

## Lý do lựa chọn Wazuh cho HIDS/SIEM

Wazuh được lựa chọn làm nền tảng HIDS/SIEM vì đây là một giải pháp mã nguồn mở tích hợp tốt giữa giám sát máy chủ (HIDS), quản lý log và tương quan sự kiện (SIEM). Wazuh cho phép xây dựng decoder và rule tùy biến, giúp chuẩn hóa các sự kiện từ Suricata và mô hình học máy thành một định dạng thống nhất. Ngoài ra, Wazuh hỗ trợ cơ chế Active Response, cho phép tự động phản ứng khi phát hiện tấn công, phù hợp với kiến trúc hybrid của hệ thống. So với các giải pháp thương mại như Splunk hoặc Elastic Security, Wazuh có chi phí thấp hơn và phù hợp với mục tiêu nghiên cứu và triển khai thử nghiệm.

## Lý do lựa chọn Isolation Forest cho phát hiện bất thường

Isolation Forest được lựa chọn làm mô hình phát hiện bất thường chính vì thuật toán này hoạt động hiệu quả trên dữ liệu dạng bảng với số lượng đặc trưng lớn và không yêu cầu dữ liệu gán nhãn. Thuật toán có khả năng huấn luyện nhanh, dễ mở rộng và ít tham số cần tinh chỉnh so với các phương pháp unsupervised khác. Trong bối cảnh dữ liệu mạng thường có nhiễu và phân bố phức tạp, Isolation Forest cho phép phát hiện các mẫu lưu lượng bất thường một cách tổng quát và ổn định, phù hợp để triển khai trong môi trường vận hành thực tế.

## Lý do lựa chọn Autoencoder và VAE

Autoencoder được sử dụng để học biểu diễn phi tuyến của dữ liệu mạng và phát hiện bất thường dựa trên lỗi tái tạo, trong khi Variational Autoencoder (VAE) được lựa chọn để học phân phối xác suất của dữ liệu lành. VAE cho phép phát hiện các mẫu lưu lượng lệch phân phối, đặc biệt phù hợp với các kịch bản tấn công zero-day. So với các phương pháp dựa trên GAN, Autoencoder và VAE ổn định hơn, dễ huấn luyện và phù hợp hơn với dữ liệu tabular của đặc trưng mạng, do đó được ưu tiên trong hệ thống này.

## Lý do lựa chọn XGBoost cho phân loại tấn công đã biết

XGBoost được lựa chọn cho bài toán phân loại tấn công đã biết vì đây là thuật toán học có giám sát có hiệu suất cao trên dữ liệu dạng bảng. Thuật toán này xử lý tốt dữ liệu mất cân bằng, có khả năng tổng quát hóa tốt và được chứng minh hiệu quả trong nhiều nghiên cứu IDS. So với Random Forest, XGBoost thường đạt độ chính xác cao hơn, trong khi vẫn đảm bảo tính ổn định và khả năng triển khai thực tế. Việc sử dụng XGBoost giúp hệ thống phân loại chính xác các loại tấn công đã có nhãn trong tập dữ liệu CIC-IDS.

## Lý do lựa chọn CIC-IDS-2017 làm tập dữ liệu huấn luyện

CIC-IDS-2017 được lựa chọn làm tập dữ liệu chính vì đây là một trong những bộ dữ liệu IDS phổ biến nhất trong nghiên cứu, có đầy đủ nhãn tấn công và bộ đặc trưng tương thích trực tiếp với CICFlowMeter. Việc sử dụng tập dữ liệu này giúp hệ thống dễ dàng so sánh với các nghiên cứu trước và giảm công sức tiền xử lý dữ liệu. Mặc dù các tập dữ liệu mới hơn như CSE-CIC-IDS2018 có thể bổ sung thêm kịch bản, CIC-IDS-2017 vẫn đáp ứng tốt yêu cầu xây dựng baseline và huấn luyện mô hình ban đầu.

## Lý do lựa chọn cơ chế phản ứng tại pfSense

Cơ chế phản ứng của hệ thống được thiết kế để chặn IP tại firewall biên pfSense thay vì chặn inline trực tiếp trong Suricata. Cách tiếp cận này giúp giảm rủi ro gián đoạn dịch vụ và dễ kiểm soát hơn trong môi trường thực tế. Việc kết hợp cảnh báo từ Suricata, kết quả suy luận của mô hình học máy và tương quan sự kiện trong Wazuh cho phép đưa ra quyết định chặn có kiểm soát, tránh chặn nhầm và đảm bảo tính ổn định của hệ thống.

# CƠ SỞ LÝ THUYẾT

## Tổng quan an ninh mạng hiện này:

Trong vài năm gần đây, bức tranh an toàn mạng tiếp tục phức tạp và khó lường. Tội phạm mạng chuyển dịch mạnh sang mô hình “dịch vụ hoá” (Ransomware ‑as‑a‑Service, Phishing‑as‑a‑Service), giúp tăng tốc độ, quy mô và mức độ tinh vi của các chiến dịch tấn công. Kịch bản ngày càng đa dạng: từ tấn công chuỗi cung ứng, lợi dụng cập nhật phần mềm, tới khai thác lỗ hổng zero‑day và kỹ thuật “living‑off‑the‑land” (lợi dụng công cụ hợp pháp có sẵn trong hệ thống). Những thay đổi này làm mờ ranh giới giữa hoạt động thông thường và hành vi độc hại, đặt ra thách thức lớn cho việc phát hiện sớm.

Ransomware vẫn là mối đe doạ nổi bật, thường đi kèm đánh cắp dữ liệu (double/triple extortion) và gây gián đoạn nghiêm trọng cho vận hành. Song song, lừa đảo thư điện tử doanh nghiệp (BEC), giả mạo danh tính, và lừa đảo xã hội (phishing, spear‑phishing) tiếp tục gia tăng vì chi phí thấp và hiệu quả cao. Xu hướng sử dụng AI bởi kẻ tấn công để tạo nội dung giả (deepfake, email/phishing “viết tay” tự động hoá) khiến việc phân biệt thông tin thật – giả khó hơn, đặc biệt ở khâu xác thực con người.

Môi trường đám mây và SaaS bùng nổ đem lại lợi ích vận hành, nhưng cũng mở rộng bề mặt tấn công. Sai cấu hình (misconfiguration), thiếu kiểm soát quyền truy cập theo nguyên tắc tối thiểu (least privilege), và quản lý khoá/secret chưa chặt chẽ là nguyên nhân phổ biến dẫn đến rò rỉ dữ liệu. Trong khi đó, hạ tầng IoT/OT (thiết bị công nghiệp, cảm biến, camera, PLC) thường thiếu cơ chế vá lỗi và bảo vệ tiêu chuẩn, khiến việc giám sát và phản ứng trở nên khó khăn, đặc biệt ở môi trường sản xuất.

Về kỹ thuật, mã hoá (TLS, E2E) giúp bảo vệ riêng tư nhưng cũng che khuất nội dung lưu lượng, khiến phát hiện dựa trên nội dung thuần tuý trở nên kém hiệu quả. Tấn công theo danh tính (đánh cắp tài khoản, nhảy phiên, MFA fatigue, credential stuffing) gia tăng làm nổi bật yêu cầu bảo vệ khối IAM: quản trị vòng đời tài khoản, MFA đúng cách, phát hiện đăng nhập bất thường theo ngữ cảnh. Lỗ hổng chuỗi cung ứng phần mềm (phụ thuộc thư viện, CI/CD, package registry) và thiếu kiểm soát SBOM cũng trở thành điểm yếu lớn.

Ở góc độ vận hành, khối lượng log và cảnh báo tăng nhanh dẫn đến “mệt mỏi cảnh báo” (alert fatigue). Đội ngũ an ninh phải xử lý dữ liệu đa nguồn (endpoint, network, cloud, ứng dụng), nhiều định dạng, với yêu cầu thời gian thực; trong khi nguồn lực hạn chế, đặc biệt với doanh nghiệp vừa và nhỏ. Hiện tượng concept drift (mô hình hành vi bình thường thay đổi theo thời gian) làm giảm hiệu năng phát hiện nếu không có cơ chế cập nhật mô hình/luật định kỳ.

Các tổ chức dần chuyển sang mô hình Zero Trust (không tin cậy mặc định), tăng cường EDR/XDR, SIEM/SOAR và threat intelligence để hợp nhất tín hiệu, tự động hoá quy trình phản ứng (playbook) và giảm thời gian xử lý sự cố. Phát hiện dựa trên bất thường (anomaly‑based) kết hợp Machine Learning được áp dụng rộng hơn để bổ sung cho chữ ký truyền thống, nhằm phát hiện các hành vi mới hoặc biến thể tinh vi. Trọng tâm là dung hợp dữ liệu đa nguồn (host + network + cloud), xây dựng baseline theo ngữ cảnh (theo hệ thống, dịch vụ, khung thời gian), và điều chỉnh ngưỡng dựa trên KPI thực tế (precision/recall, FPR/TPR, độ trễ).

Về quản trị, tuân thủ và pháp lý ngày càng khắt khe: yêu cầu minh bạch về rò rỉ dữ liệu, chuẩn mực bảo vệ thông tin cá nhân, và nghĩa vụ thông báo sự cố. Điều này thúc đẩy tổ chức xây dựng năng lực quan sát (observability), lưu vết (audit) và quy trình ứng phó bài bản. Kèm theo là kỷ luật ở các “nền tảng” cơ bản: quản lý bản vá, sao lưu/khôi phục, phân đoạn mạng (segmentation), nguyên tắc tối thiểu (least privilege), quản lý khoá/secret, và đào tạo nhận thức an ninh cho người dùng.

Tình hình an toàn mạng hiện nay là sự kết hợp giữa bề mặt tấn công mở rộng, kỹ thuật ngày càng tinh vi, và áp lực vận hành lớn. Chiến lược khả thi là phòng thủ nhiều lớp, kết hợp chữ ký với anomaly‑based và ML, tiêu chuẩn hoá dữ liệu log, tự động hoá phản ứng (SOAR), và liên tục đo lường – cải tiến theo vòng lặp: phát hiện → phân tích → tinh chỉnh → thử nghiệm lại. Cách tiếp cận này giúp tăng khả năng phát hiện sớm, giảm cảnh báo sai, rút ngắn thời gian phản ứng và nâng cao tính bền vững của hệ thống an ninh mạng.

## Hệ thống phát hiện và ngăn ngừa xâm nhập (IDS/IPS)

Hệ thống phát hiện và ngăn ngừa xâm nhập (IDS/IPS) là lớp phòng thủ trọng yếu trong an ninh mạng, nhằm phát hiện và phản ứng trước các hành vi bất thường hoặc tấn công vào tài nguyên hệ thống. IDS (Intrusion Detection System) tập trung vào phát hiện và cảnh báo, vận hành theo chế độ thụ động (passive); IPS (Intrusion Prevention System) mở rộng thêm khả năng can thiệp theo thời gian thực (inline), tự động chặn hoặc sửa đổi lưu lượng để ngăn chặn tấn công. Trong thực tế, nhiều nền tảng hỗ trợ cả hai vai trò, tuỳ cách triển khai (giám sát thụ động hoặc chèn vào đường truyền).

Về phân loại theo vị trí, có hai nhánh chính: NIDS (Network-based IDS/IPS), đặt cảm biến trên đường truyền mạng để giám sát lưu lượng (ví dụ Suricata trên pfSense); và HIDS (Host-based IDS), giám sát hành vi và log trên máy chủ/điểm cuối (ví dụ Wazuh agent). NIDS quan sát được bức tranh lưu lượng tổng thể, nhận diện tấn công lớp mạng/ứng dụng; HIDS đi sâu vào tiến trình, file, cấu hình và nhật ký hệ điều hành, phát hiện hành vi bất thường tại máy chủ. Sự kết hợp NIDS và HIDS giúp bao phủ nhiều lớp phòng thủ, tăng độ tin cậy khi đối chiếu tín hiệu từ hai nguồn.

Về phương pháp phát hiện, có hai hướng chủ đạo: phát hiện dựa trên chữ ký (signature-based) và phát hiện dựa trên bất thường (anomaly-based). Signature-based dùng tập quy tắc/chuỗi mẫu đã biết để so khớp, hiệu quả với mối đe doạ phổ biến, dễ giải thích và kiểm soát; tuy nhiên phụ thuộc mạnh vào cập nhật chữ ký, dễ bỏ sót zero‑day hoặc biến thể tinh vi. Anomaly-based xây dựng baseline hành vi bình thường và tìm lệch chuẩn; ưu thế ở khả năng phát hiện kịch bản mới, nhưng đòi hỏi đặc trưng tốt, ngưỡng phù hợp và cơ chế giảm cảnh báo sai. Trong anomaly-based, các mô hình phổ biến gồm ngưỡng (threshold) và hồ sơ (profile), cùng các hướng statistical, knowledge-based, data‑mining; gần đây, Machine Learning (ML) kết hợp unsupervised (Isolation Forest, Autoencoder) và supervised (XGBoost) được áp dụng để nâng cao chất lượng phát hiện và giảm phụ thuộc vào rule tĩnh.

Về kiến trúc dữ liệu và triển khai, NIDS như Suricata có thể xuất log theo định dạng EVE JSON chuẩn hoá, thuận tiện cho pipeline tiền xử lý và học máy.

Trên nền pfSense, Suricata quan sát lưu lượng, tạo sự kiện (chữ ký và chỉ số hành vi) đẩy về hệ thống phân tích. HIDS như Wazuh gom và chuẩn hoá log hệ thống/ứng dụng, cung cấp bối cảnh host để dung hợp với tín hiệu mạng. Dữ liệu PCAP/luồng có thể chuyển thành đặc trưng nhất quán (khoảng 84 feature) bằng CICFlowMeter/NFStream, đưa vào mô hình ML để phân loại/đánh giá mức độ nghiêm trọng. Thiết kế tách lớp “thu thập → đặc trưng → ML → hành động” giúp thay thế/ mở rộng từng khối (mô hình, nguồn log, cơ chế hành động) mà không cần tái kiến trúc toàn hệ thống.

Về cơ chế phản ứng, IPS hoặc IDS có thể ánh xạ mức độ nghiêm trọng sang hành động tương ứng: ghi log, phát cảnh báo, kích hoạt webhook, hoặc chặn IP/tạm thời vô hiệu hoá kết nối. Thực hành vận hành an toàn thường kèm whitelist (bảo vệ tài nguyên tin cậy), rate limiting (giới hạn tần suất hành động), auto‑unblock (tự bỏ chặn sau thời lượng) và chế độ dry‑run/log‑only (chỉ ghi nhận, không áp dụng chặn) để giảm nguy cơ gián đoạn dịch vụ. Các chính sách này giúp cân bằng giữa độ nhạy phát hiện và rủi ro chặn nhầm trong môi trường sản xuất.

Dù quan trọng, IDS/IPS đối mặt nhiều thách thức: cảnh báo sai (false positives) gây “mệt mỏi cảnh báo” (alert fatigue); tấn công mã hoá (TLS) làm giảm khả năng quan sát nội dung; vị trí triển khai và hiệu năng (đa luồng, băng thông cao) ảnh hưởng chất lượng giám sát; và hiện tượng concept drift khiến baseline lỗi thời nếu không được cập nhật. Khắc phục cần kết hợp quản trị rule/ chữ ký, tinh chỉnh ngưỡng theo ngữ cảnh, chọn đặc trưng phù hợp, tái huấn luyện mô hình định

kỳ, cùng các biện pháp vận hành như quay vòng log, dọn dẹp dữ liệu, và giám sát sức khoẻ pipeline.

Đánh giá IDS/IPS nên dựa trên KPI định lượng và có thể điều chỉnh: tỷ lệ phát hiện (TPR/Recall), tỷ lệ cảnh báo sai (FPR), độ chính xác (Precision/F1), độ trễ từ phát hiện đến hành động, và khả năng duy trì chu kỳ phân tích ổn định theo thời gian. Trong mô hình lai signature–anomaly kết hợp ML, quy trình cải tiến lặp “phát hiện → phân tích → tinh chỉnh rules/ML → thử nghiệm lại” là trọng tâm để nâng cao hiệu năng, giảm FP, và tăng độ tin cậy vận hành.

Tổng quan, IDS/IPS là thành phần nòng cốt trong hệ thống phòng thủ, và cách tiếp cận lai giữa signature‑based, anomaly‑based và ML phù hợp với bối cảnh mối đe doạ hiện đại: vừa bảo toàn khả năng phát hiện mối đe doạ đã biết, vừa mở rộng phạm vi nhận diện hành vi mới; đồng thời tự động hoá phản ứng để giảm thời gian xử lý sự cố, nâng cao hiệu quả và tính bền vững của vận hành an ninh mạng.

## So sánh IDS (Intrusion Detection System) và IPS (Intrusion Prevention System):

|  |  |  |
| --- | --- | --- |
| Tiêu chí | IDS | IPS |
| Là gì? | Hệ thống phát hiện tấn công | Hệ thống phát hiện và chặn tấn công |
| Mục đích | Phát hiện vấn đề và báo cho quản trị | Phát hiện và chặn ngay |
| Cách đặt trong mạng | Đặt ngoài luồng mạng | Đặt trực tiếp trên luồng mạng |
| Ảnh hưởng đến mạng | Không ảnh hưởng lưu lượng | Có thể ảnh hưởng nếu cấu hình sai |
| Khi phát hiện tấn công | Ghi lại và gửi cảnh báo | Chặn gói tin, ngắt kết nối, khóa IP |
| Mức độ rủi ro | Thấp, an toàn khi vận hành | Cao hơn vì có thể chặn nhầm |
| Độ trễ xử lý | Thấp | Cao hơn IDS |
| Yêu cầu phần cứng | Không cao | Cao (xử lý toàn bộ dữ liệu) |
| Cách phát hiện | Dựa vào mẫu + hành vi bất thường | Dựa vào mẫu/ngưỡng rất chắc chắn |
| Xử lý nhầm lẫn | Chấp nhận báo nhầm | Hạn chế tối đa chặn nhầm |
| Vai trò chính | Giám sát, phân tích sự cố | Bảo vệ chủ động hệ thống |
| Dùng khi nào? | Theo dõi, điều tra, học tập | Bảo vệ hệ thống thật |
| Quản lý luật | Dễ thử nghiệm | Phải kiểm tra rất kỹ |
| Kết hợp AI/ML | Phát hiện bất thường | Chỉ chặn khi rất chắc chắn |
| Công cụ tiêu biểu | Suricata (IDS), Zeek, Wazuh | Suricata inline, Snort, Firewall thế hệ mới |

Bảng 1: So sánh IDS (Intrusion Detection System) và IPS (Intrusion Prevention System)

## Tổng quan về hệ thống phát hiện và ngăn chặn xâm nhập (IDS/IPS)

Trong bối cảnh an ninh mạng hiện nay, các hệ thống IDS/IPS được xem là thành phần quan trọng trong việc xây dựng kiến trúc phòng vệ mạng. Việc tích hợp nhiều cơ chế phát hiện, bao gồm phát hiện theo chữ ký, phát hiện hành vi bất thường và các kỹ thuật học máy, giúp hệ thống vừa đảm bảo khả năng nhận diện chính xác các mối đe doạ đã được xác định trước, vừa mở rộng khả năng phát hiện các hình thức tấn công mới hoặc chưa từng xuất hiện. Bên cạnh đó, việc áp dụng cơ chế phản ứng tự động góp phần rút ngắn thời gian xử lý sự cố, nâng cao hiệu quả bảo vệ và tăng cường tính ổn định cũng như khả năng vận hành bền vững của hệ thống an ninh mạng.

### Giới thiệu về hệ thống phát hiện xâm nhập (Intrusion Detection System – IDS)

Hệ thống phát hiện xâm nhập (Intrusion Detection System – IDS) là một thành phần khá quan trọng trong các giải pháp bảo mật mạng hiện nay, nó được triển khai dưới dạng phần mềm hoặc thiết bị chuyên dụng nhằm có thể giám sát và phân tích được các hoạt động của hệ thống. Thông qua việc thu thập và xử lý lưu lượng dữ liệu đa dạng cũng như các sự kiện phát sinh thêm, IDS có khả năng phát hiện được những hành vi bất thường hoặc truy cập trái phép. Khi xuất hiện dấu hiệu xâm nhập mới, hệ thống sẽ đưa ra cảnh báo tương ứng để hỗ trợ người quản trị có thể kịp thời nhận biết và triển khai nhanh chóng các biện pháp xử lý nhằm tăng bảo vệ hệ thống.



Hình 1. Mô hình phân loại IDS với hai nhóm chính: NIDS (Network-based IDS) và HIDS (Host-based IDS).

### Phân loại hệ thống phát hiện xâm nhập:

Dựa trên phạm vi triển khai và giám sát, hệ thống phát hiện xâm nhập được phân thành hai loại cơ bản là hệ thống phát hiện xâm nhập mạng (NIDS) và hệ thống phát hiện xâm nhập dựa trên máy chủ (HIDS).

**Hệ thống phát hiện xâm nhập mạng (Network Intrusion Detection System – NIDS):** NIDS là hệ thống được triển khai tại các điểm trọng yếu trong mạng mục đích nhằm thu thập và phân tích lưu lượng dữ liệu truyền qua các thiết bị mạng. Thông qua việc xử lý các gói tin và luồng dữ liệu, NIDS có khả năng nhận diện các dấu hiệu bất thường hoặc nguy cơ tấn công tiềm ẩn mà không làm can thiệp hay thay đổi nội dung của các gói tin trong quá trình truyền dẫn đến và đi.

**Hệ thống phát hiện xâm nhập dựa trên máy chủ (Host-based Intrusion Detection System – HIDS):** HIDS được cài đặt trực tiếp trên các máy chủ hoặc máy trạm cần bảo vệ, cho phép tăng khả năng giám sát chi tiết hoạt động diễn ra bên trong của từng thiết bị. Hệ thống này sẽ theo dõi lưu lượng vào ra, các tiến trình dẫn truyền và sự kiện hệ thống nhằm phát hiện và cảnh báo kịp thời các hành vi truy cập trái phép hoặc bất thường đáng ngờ.

Trong các hệ thống an ninh hiện đại hiện tại, NIDS và HIDS thường được kết hợp nhằm thu thập dữ liệu từ nhiều nguồn khác nhau. Việc phối hợp này giúp nâng cao tối ưu độ chính xác cao và hiệu quả trong các quá trình phát hiện xâm nhập, đồng thời có thể giảm thiểu được khả năng bỏ sót các mối đe doạ trên toàn bộ hệ thống mạng.

### Nguyên lý vận hành của hệ thống phát hiện xâm nhập:

Trong quá trình hoạt động của hệ thống, hệ thống phát hiện xâm nhập tiến hành thu thập dữ liệu từ lưu lượng mạng hoặc các sự kiện hệ thống, sau đó thực hiện phân tích nhằm đánh giá mức độ an toàn của các hoạt động này. Dữ liệu thu thập được sẽ được đối chiếu với các mô hình hoặc tiêu chí đã được xây dựng trước để phát hiện các dấu hiệu bất thường hoặc hành vi có khả năng gây xâm nhập. Khi phát hiện nguy cơ tiềm ẩn, hệ thống sẽ tạo ra cảnh báo để thông báo đến cho người quản trị, từ đó hỗ trợ việc đưa ra các biện pháp xử lý nhanh chóng phù hợp nhằm hạn chế rủi ro và thiệt hại cho hệ thống.

Về cơ bản, các hệ thống IDS hiện nay thường áp dụng hai cơ chế phát hiện chính là phát hiện xâm nhập dựa trên chữ ký và phát hiện xâm nhập dựa trên hành vi bất thường.

**Phát hiện xâm nhập dựa trên chữ ký (Signature-based Detection):** Phát hiện dựa trên chữ ký là cơ chế trong đó hệ thống luôn so sánh lưu lượng mạng hoặc nhật ký hệ thống với các mẫu tấn công đã được xác định và lưu trữ trước đó. Các mẫu chữ ký này thường được xây dựng dựa trên đặc trưng của từng loại tấn công cụ thể, chẳng hạn như chuỗi lệnh độc hại, mẫu byte đặc trưng hoặc những hành vi đã được ghi nhận. Nhờ sử dụng cơ chế đối chiếu này, hệ thống có thể nhanh chóng phát hiện ra và cảnh báo các mối đe doạ đã biết với độ chính xác cao.

**Phát hiện xâm nhập dựa trên hành vi bất thường (Anomaly-based Detection):** Phát hiện dựa trên hành vi bất thường tập trung vào việc xác định các hoạt động sai lệch so với trạng thái hoạt động bình thường của hệ thống. Cơ chế này thường sử dụng các kỹ thuật học máy nhằm xây dựng mô hình hành vi chuẩn, còn gọi là đường cơ sở từ dữ liệu mạng hợp lệ. Các hành vi mới sẽ được so sánh với mô hình này để phát hiện những sai lệch có thể biểu hiện cho các cuộc tấn công chưa từng có được ghi nhận được ghi nhận. Tuy nhiên, do đặc tính linh hoạt của lưu lượng mạng, phương pháp này có thể phát sinh nhiều cảnh báo giả trong trường hợp các hoạt động hợp lệ nhưng chưa từng xuất hiện trước đó bị nhận diện nhầm là hành vi độc hại.

### So sánh signature-based và anomaly-based detection:

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Phát hiện dựa trên chữ ký (Signature-based) | Phát hiện dựa trên bất thường (Anomaly-based) |
| Nguyên lý | So sánh lưu lượng hoặc sự kiện với tập các mẫu tấn công đã được định nghĩa trước | Đánh giá sự sai lệch của hành vi hiện tại so với mô hình các hành vi thông thường |
| Điểm mạnh | Thời gian để phát hiện khá nhanh, kết quả ổn định, ít phát sinh cảnh báo không chính xác | Tăng khả năng nhận diện được các hình thức tấn công mới hoặc chưa từng có xuất hiện |
| Hạn chế | Không phát hiện zero-day | Dễ cảnh báo nhầm (false positive) |
| Ứng dụng | IDS truyền thống, antivirus | IDS hiện đại, Machine Learning |

Bảng 2. So sánh phát hiện dựa trên chữ ký (Signature-based) và phát hiện dựa trên bất thường (Anomaly-based)

### Giới thiệu về hệ thống ngăn ngừa xâm nhập (Intrusion Prevention System – IPS)

Hệ thống ngăn ngừa xâm nhập (Intrusion Prevention System – IPS) là một giải pháp an ninh mạng được triển khai nhằm giám sát lưu lượng dữ liệu và chủ động ngăn chặn các hành vi tấn công hoặc truy cập trái phép vào hệ thống. Khác với hệ thống phát hiện xâm nhập (IDS) chỉ việc thực hiện chức năng giám sát và cảnh báo, thì IPS còn có thêm khả năng can thiệp trực tiếp được vào những luồng dữ liệu để mà hạn chế hoặc loại bỏ ngay mối đe dọa khi chúng mới xuất hiện.

Trong quá trình vận hành của hệ thống, IPS sẽ liên tục thu thập và phân tích các gói tin đi qua mạng, đối chiếu với các tập các quy tắc, chữ ký tấn công hoặc mô hình hành vi đã được thiết lập trước đó. Khi phát hiện ra các dấu hiệu khả nghi bất thường hoặc hành vi nguy hại xấu, hệ thống nó sẽ tự động thực hiện loạt các biện pháp phòng vệ phù hợp như chặn đứng gói tin, ngắt phiên kết nối, cô lập nguồn tấn công hoặc cập nhật chính sách lọc trên tường lửa nhằm ngăn chặn sự xâm nhập lan rộng hơn trong hệ thống. Đồng thời thì hệ thống IPS cũng sẽ ghi nhận và lưu trữ thông tin chi tiết về sự kiện để tiếp tục phục vụ công tác giám sát, phân tích và đánh giá an ninh.

Về bản chất thì hệ thống IPS này có thể được xem là sự phát triển nâng cao của IDS, khi kết hợp được cả khả năng phát hiện với cơ chế phản ứng tự động. Các quy tắc và phương pháp nhận diện được sử dụng trong IPS thường tương tự như IDS, tuy nhiên được cấu hình ở mức độ nghiêm ngặt hơn để cho phép hệ thống tự động thực thi được các hành động ngăn ngừa, thay vì chỉ dừng lại ở việc tạo cảnh báo cho người quản trị biết.

### Phân loại trong hệ thống ngăn ngừa xâm nhập:

Đối với việc triển khai và đối tượng cần bảo vệ, hệ thống ngăn ngừa xâm nhập (IPS) thường được chia thành hai nhóm chính, bao gồm hệ thống ngăn ngừa xâm nhập chỉ ở mức mạng và hệ thống ngăn ngừa xâm nhập chỉ ở mức máy chủ.

**Hệ thống ngăn ngừa xâm nhập dựa trên mạng (Network-based Intrusion Prevention System – NIPS)** là hệ thống được triển khai trực tiếp trên hạ tầng mạng nhằm mục đích giám sát và kiểm soát cũng như rà soát toàn bộ lưu lượng dữ liệu ra vào của hệ thống. NIPS thường được đặt tại các vị trí trọng yếu như trước hoặc phía sau của tường lửa. Khi mà nó được triển khai phía trước tường lửa, NIPS có khả năng bảo vệ toàn bộ mạng nội bộ, bao gồm các vùng như vùng mạng trung tâm (DMZ) và các thiết bị an ninh trung gian. Trong trường hợp triển khai phía sau của tường lửa, NIPS thì tập trung vào việc phát hiện và ngăn chặn các mối đe dọa phát sinh từ bên trong hệ thống, chẳng hạn như các thiết bị đầu cuối hoặc các thiết bị kết nối từ xa thông qua VPN có khả năng bị khai thác lỗ hổng bảo mật.

**Hệ thống ngăn ngừa xâm nhập dựa trên máy chủ (Host-based Intrusion Prevention System – HIPS** là hệ thống được cài đặt trực tiếp trên từng máy chủ hoặc máy trạm để thực hiện liên tục giám sát và bảo vệ ở mức với các thiết bị. HIPS sẽ theo dõi các hoạt động của hệ điều hành, tiến trình đang chạy và các thay đổi liên quan đến tệp tin hoặc cấu hình hệ thống. Khi mà nó phát hiện hành vi bất thường hoặc dấu hiệu tấn công khả nghi mới thì hệ thống có thể chủ động ngăn chặn mã độc tấn công và sẽ hạn chế quyền truy cập và gửi cảnh báo cho người quản trị. Nhờ khả năng hoạt động độc lập trên từng host, HIPS đặc biệt này phù hợp để bảo vệ các máy chủ quan trọng hoặc các hệ thống lưu trữ dữ liệu nhạy cảm.

### Các ưu điểm và hạn chế của hệ thống ngăn ngừa xâm nhập - IPS:

**Ưu điểm:**

* Cung cấp nên các giải pháp bảo vệ toàn diện hơn đối với tài nguyên hệ thống.
* Ngăn chặn kịp thời các tấn công đã biết hoặc chưa được biết.

**Hạn chế:**

* Có thể gây ra tình trạng phát hiện nhầm (faulse positives), có thể không cho phép các truy cập hợp lệ tới hệ thống.

## Kỹ thuật phát hiện xâm nhập dựa trên bất thường (anomaly-based intrusion detection):

Phát hiện dựa trên bất thường (Anomaly-based Intrusion Detection) là một kỹ thuật an ninh mạng tập trung vào việc sẽ nhận diện được các hành vi hoặc mẫu dữ liệu khác biệt so với trạng thái hoạt động bình thường hằng ngày trên hệ thống. Thay vì tìm kiếm các dấu hiệu tấn công đã được xác định trước, phương pháp này xây dựng một mô hình đại diện cho hành vi hợp lệ, sau đó sẽ đánh giá mức độ sai lệch của các quan sát mới so với mô hình đó.

Trong bối cảnh an ninh mạng như hiện tại thì các trường hợp bất thường có thể xuất hiện dưới dạng lưu lượng truy cập không điển hình, tần suất kết nối bất thường hoặc các hành vi truy cập không tuân theo quy luật thông thường. Mặc dù những sai lệch này không phải lúc nào cũng là tấn công, nhưng trong nhiều trường hợp, chúng có thể phản ánh các mối đe dọa nghiêm trọng như là xâm nhập trái phép, khai thác lỗ hổng hoặc là gian lận hệ thống.

Về mặt kỹ thuật, phát hiện bất thường thường được triển khai dựa trên việc phương pháp phân tích dữ liệu và học máy nhằm trích xuất đặc trưng, xây dựng nên đường cơ sở của hành vi bình thường và xác định được các điểm dữ liệu có mức độ lệch đáng kể. Cách tiếp cận này cho phép hệ thống phát hiện được các hình thức tấn công mới hoặc chưa từng được ghi nhận, qua đó nâng cao năng lực cũng như khả năng tự bảo vệ trước các mối đe dọa chưa có mẫu nhận diện sẵn.

### Tầm quan trọng của việc phương pháp phát hiện dựa trên bất thường:

Phát hiện bất thường đóng một vai trò đặc biệt quan trọng đối với các ngành như tài chính, bán lẻ và an ninh mạng nhưng tuy nhiên mọi doanh nghiệp cũng cần cân nhắc giải pháp phát hiện bất thường. Giải pháp này sẽ cung cấp một phương tiện tự động phát hiện các giá trị ngoại lai gây hại và bảo vệ dữ liệu của bạn một cách an toàn.

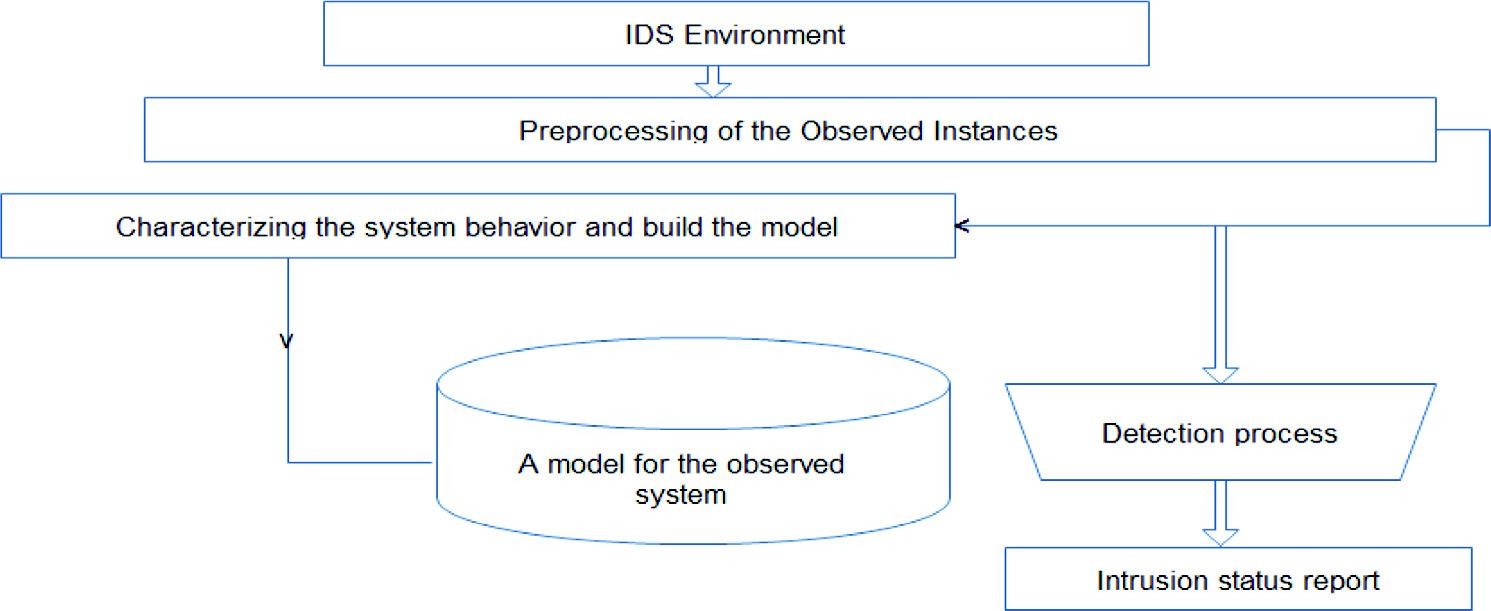
Phát hiện bất thường xác định hoạt động khả nghi khác với khuôn mẫu hành vi bình thường đã thiết lập của bạn. Một giải pháp bảo vệ hệ thống của bạn theo thời gian thực khỏi những phiên bản có thể dẫn đến thiệt hại đáng kể về tài chính, vi phạm dữ liệu hoặc các sự kiện có hại khác.

### Những ưu điểm và nhược điểm của phát hiện dựa trên bất thường:

Phương pháp phát hiện xâm nhập dựa trên bất thường mang lại nhiều lợi thế trong bối cảnh các mối đe dọa an ninh mạng ngày càng đa dạng và khó lường. Trước hết là do không phụ thuộc vào các mẫu tấn công được định nghĩa sẵn, kỹ thuật này có khả năng nhận diện được những hình thức tấn công mới hoặc chưa từng xuất hiện trước đó. Điều này giúp hệ thống mở rộng phạm vi phát hiện, đặc biệt hiệu quả đối với các mối đe dọa zero-day hoặc là với các hành vi xâm nhập tinh vi chưa có chữ ký để có thể nhận dạng cụ thể. Bên cạnh đó thì các kết quả thu được từ quá trình phát hiện bất thường còn có thể được khai thác để bổ sung dữ liệu huấn luyện, hỗ trợ xây dựng hoặc cập nhật cơ sở dữ liệu chữ ký cho các phương pháp phát hiện truyền thống.

Tuy nhiên thì việc phát hiện dựa trên bất thường cũng tồn tại một số hạn chế nhất định. Do hệ thống phải đánh giá hành vi dựa trên mức độ sai lệch so với mô hình bình thường nên các hoạt động hợp lệ nhưng ít phổ biến hoặc mang tính đột xuất của người dùng và hệ thống có thể bị nhận diện nhầm là những hành vi mang tính tấn công, dẫn đến xảy ra hiện tượng cảnh báo giả, gây thông báo nhầm. Ngoài ra, việc xây dựng và duy trì mô hình hành vi bình thường thường có độ phức tạp cao, đòi hỏi phải thu thập và cập nhật liên tục các bản ghi sự kiện để phản ánh chính xác trạng thái hoạt động thực tế của hệ thống.

### Mô hình của một IDS dạng Anomaly-Based Intrusion Detection:

****

Hình 2. Mô hình cơ bản của một IDS dạng Anomallly-Based Detection.

Về cơ bản thì hệ thống được tổ chức thành ba giai đoạn chính, bao gồm tiền xử lý dữ liệu, quan sát và phân tích hành vi, và thực hiện chức năng phát hiện.

**Giai đoạn tiền xử lý dữ liệu (Formation of Attributes)** là bước đầu tiên trong mô hình. Ở giai đoạn này thì các dữ liệu thu thập từ môi trường IDS như lưu lượng mạng, nhật ký hệ thống hoặc là các sự kiện hoạt động được làm sạch, sàng lọc cẩn thận và trích xuất các thuộc tính cần thiết. Quá trình này giúp loại bỏ được nhiều dữ liệu dư thừa, chuẩn hóa định dạng mới và tạo tập đặc trưng phù hợp cho các bước phân tích tiếp theo.

**Giai đoạn quan sát và xây dựng mô hình (Observation Stage)** thì tập trung vào việc phân tích hành vi của hệ thống dựa trên dữ liệu đã được tiền xử lý. Thông qua các cơ chế tự động hoặc các mô hình phân tích, hệ thống tiến hành học và mô tả đặc trưng của hành vi bình thường. Kết quả của giai đoạn này là một mô hình đại diện cho trạng thái hoạt động hợp lệ của hệ thống, được lưu trữ để làm cơ sở so sánh.

**Giai đoạn phát hiện và xử lý chức năng (Functional Stage)** là bước thực hiện đánh giá và ra quyết định. Dữ liệu quan sát mới được so sánh với mô hình đã xây dựng nhằm xác định mức độ sai lệch. Khi phát hiện các hành vi có độ lệch vượt quá ngưỡng cho phép, hệ thống sẽ đánh dấu đó là bất thường và sinh ra cảnh báo xâm nhập. Kết quả phát hiện được tổng hợp thành báo cáo trạng thái xâm nhập, hỗ trợ quản trị viên trong việc giám sát và xử lý sự cố an ninh mạng.

### Khái quát về khái niệm Threshold (ngưỡng) và Profile (hồ sơ hành vi):

Trong các hệ thống phát hiện xâm nhập dựa trên bất thường, threshold (ngưỡng) và profile (hồ sơ hành vi) là hai khái niệm nền tảng được sử dụng nhiều để có thể phân biệt giữa trạng thái hoạt động bình thường và các dấu hiệu bất thường xảy ra với hệ thống.

**Threshold (ngưỡng)** được hiểu là một giá trị giới hạn được xác định trước nhằm làm mốc so sánh cho các tham số giám sát. Khi mà giá trị đo được của một tham số, chẳng hạn như lưu lượng mạng, số lượng kết nối hoặc mức sử dụng tài nguyên, vượt quá hoặc thấp hơn so với ngưỡng cho phép, hệ thống sẽ coi đó là dấu hiệu bất thường và sẽ tự kích hoạt cơ chế cảnh báo. Việc thiết lập ngưỡng có thể được thực hiện thủ công dựa trên kinh nghiệm vận hành của quản trị viên, hoặc được xác định tự động thông qua các phương pháp thống kê và thuật toán học máy dựa trên dữ liệu lịch sử.

**Profile (hồ sơ hành vi)** là mô hình đại diện cho hành vi bình thường của người dùng hoặc hệ thống trong điều kiện vận hành ổn định. Hồ sơ này được xây dựng bằng cách phân tích dữ liệu lịch sử hoặc là các quan sát trong thời gian dài nhằm có thể mô tả được chính xác các đặc trưng của hoạt động điển hình. Khi một hành vi mới có sự sai lệch đáng kể so với hồ sơ đã được thiết lập trước đó, hệ thống sẽ xem đó là biểu hiện bất thường và sinh ra cảnh báo ngay. Cách tiếp cận này đặc biệt phù hợp với các kỹ thuật phát hiện dựa trên bất thường, nơi mà các hành vi được theo dõi liên tục và đánh giá theo sự thay đổi theo thời gian.

Sự khác biệt cốt lõi giữa hai khái niệm này thì nằm ở cách thức ra quyết định: threshold thì dựa trên các giá trị giới hạn cố định, trong khi đó thì profile dựa trên mô hình hóa hành vi bình thường và mức độ sai lệch so với mô hình đó.

Ví dụ minh họa cho phương pháp threshold:

Giả sử như hệ thống giám sát nhiệt độ của một phòng máy chủ ghi nhận mức nhiệt ổn định hẳng ngày chỉ quanh quẩn 22°C là trong điều kiện vận hành bình thường. Nhưng khi nó xảy ra sự cố về hệ thống làm mát khiến nhiệt độ có thể tăng lên 40°C, nếu như ngưỡng cảnh báo được thiết lập trước đó chỉ ở mức 25°C, hệ thống sẽ nhanh chóng phát hiện sự vượt ngưỡng này và nhanh chóng gửi cảnh báo đến quản trị viên để kịp thời có các biện pháp nhanh chóng xử lý và khắc phục tối ưu nhất thiệt hại gây ra.

Ví dụ minh họa cho phương pháp profile:

Xét một hệ thống phát hiện gian lận theo dõi hành vi chi tiêu của một người dùng. Dữ liệu lịch sử hằng ngày cho thấy người này thường thực hiện các giao dịch với giá trị vừa phải và tập trung vào các loại chi tiêu quen thuộc. Tuy nhiên nhưng khi có xuất hiện các giao dịch có giá trị lớn hơn bất thường và phát sinh tại những địa điểm hoặc lĩnh vực chưa từng xuất hiện trước đó của người dùng này thì các hành vi như vậy sẽ bị đánh giá là lệch khỏi hồ sơ chi tiêu thông thường. Do đó, hệ thống sẽ ngay lập tức gắn cờ cảnh báo nhằm kiểm tra khả năng tài khoản đã bị xâm phạm.

## Các kỹ thuật xác định ngưỡng (threshold) và hồ sơ hành vi (profiles):

### Kỹ thuật dựa trên thống kê (Statistical-based techniques):

**Khái niệm:** Kỹ thuật dựa trên thống kê sử dụng các đại lượng thống kê cơ bản như là giá trị trung bình, phương sai và độ lệch chuẩn để từ đó có thể xây dựng nên các mô hình mô tả hành vi hoạt động bình thường của hệ thống.

**Nguyên lý hoạt động:** Trong giai đoạn huấn luyện ban đầu, hệ thống tiến hành thu thập dữ liệu khi môi trường được xem là an toàn và không có tấn công. Từ tập dữ liệu này, các tham số thống kê được tính toán cho từng đặc trưng giám sát, chẳng hạn như lưu lượng gói tin, tần suất truy cập hoặc thời gian phản hồi. Khi hệ thống đi vào vận hành, các hành vi mới phát sinh sẽ được so sánh với mô hình thống kê đã xây dựng. Nếu giá trị quan sát vượt ra ngoài phạm vi cho phép (ví dụ lớn hơn hoặc nhỏ hơn giá trị trung bình cộng/trừ một số lần độ lệch chuẩn xác định trước) thì hệ thống sẽ phát hiện xâm nhập và kích hoạt cảnh báo.

**Ưu điểm:**

* Không cần biết về signature của cuộc tấn công (có thể phát hiện được zeroday attack).
* Việc ko cần phụ thuộc vào signature nên update không cần thường xuyên.
* Các cuộc tấn công xảy ra trong thời gian dài có thể xác định chính xác như dos.

**Nhược điểm:**

* Độ chính xác của phương pháp phụ thuộc lớn vào việc xây dựng mô hình phân phối thống kê phù hợp với dữ liệu thực tế.
* Quá trình thu thập dữ liệu và hình thành hồ sơ hành vi bình thường đòi hỏi thời gian đủ dài để đảm bảo tính ổn định và độ tin cậy của mô hình.

### Kỹ thuật dựa trên nhận thức (Cognitive-based or knowledge-based techniques):

**Khái niệm:** Kỹ thuật dựa trên nhận thức khai thác tri thức thu được từ các cuộc tấn công đã xảy ra cũng như những lỗ hổng bảo mật đã được phát hiện trước đó. Các phương pháp trong nhóm này thường được chia thành hai hướng chính: phân tích sự chuyển đổi trạng thái của hệ thống và phân tích dựa trên chữ ký tấn công.

**Nguyên lý hoạt động:** Hệ thống tiến hành thu thập, phân tích và tổng hợp thông tin liên quan đến các kịch bản tấn công, điểm yếu của các bảo mật và hành vi xâm nhập đã được ghi nhận trước. Trên cơ sở đó, các quy tắc, mô hình hoặc tập chữ ký được xây dựng nhằm mô tả đặc trưng của từng dạng tấn công. Trong quá trình vận hành, hệ thống giám sát liên tục so sánh các hành vi quan sát được với tập tri thức đã thiết lập; khi phát hiện sự tương đồng hoặc trùng khớp, cơ chế cảnh báo sẽ được kích hoạt.

**Ưu điểm:**

* Độ chính xác cao và tỷ lệ cảnh báo sai thấp nhờ vào các mẫu tấn công đã được xác minh, kiểm chứng từ trước.
* Hỗ trợ nhà quản trị và chuyên gia an ninh dễ dàng phân tích nguyên nhân, đánh giá mức độ rủi ro và triển khai biện pháp phòng vệ phù hợp.

**Nhược điểm:**

* Quá trình xây dựng và cập nhật cơ sở tri thức đòi hỏi nhiều thời gian và công sức nghiên cứu.
* Khả năng phát hiện bị hạn chế đối với các hình thức tấn công mới hoặc chưa từng được ghi nhận, do phụ thuộc vào tập tri thức sẵn có.

### Kỹ thuật dựa trên khai thác dữ liệu (Data mining-based techniques):

**Khái niệm:** Kỹ thuật khai thác dữ liệu tập trung vào việc phân tích các tập dữ liệu có quy mô lớn nhằm phát hiện các khuôn mẫu tiềm ẩn, xu hướng và mối quan hệ giữa các thuộc tính dữ liệu. Thông qua các phương pháp phân loại, phân cụm và học mẫu, kỹ thuật này hỗ trợ quá trình ra quyết định và giải quyết các bài toán phức tạp dựa trên dữ liệu.

**Các bước thực hiện:**

* Làm sạch dữ liệu: làm sạch tạp hoặc bất thường.
* Tích hợp dữ liệu: kết hợp nhiều nguồn.
* Chuyển đổi dữ liệu: để phân tích và tổng hợp.
* Khai phá dữ liệu: trích xuất từ hiện có.
* Đánh giá mẫu: thỏa yêu cầu.
* Trình bày thông tin: dạng cây, bảng, biểu đồ và ma trận.

**Ưu điểm:**

* Có khả năng xử lý hiệu quả các tập dữ liệu lớn và phức tạp.
* Thời gian xử lý nhanh trong giai đoạn vận hành nhờ các mô hình đã được huấn luyện trước.
* Có thể áp dụng các phương pháp học không giám sát để tự động phát hiện cấu trúc và hành vi tiềm ẩn trong dữ liệu.

**Nhược điểm:**

* Do có nhiều kỹ thuật dựa trên phân cụm hoặc phân loại vì vậy các mô hình này không được thiết kế chuyên biệt cho nhiệm vụ phát hiện bất thường.
* Yêu cầu tài nguyên lưu trữ lớn và có thể làm chậm quá trình phân tích khi kích thước dữ liệu tăng cao.

## Các vấn đề và thách thức đối với phát hiện bất thường trong hệ thống ids/ips:

Các hệ thống IDS/IPS trong thực tế triển khai vẫn phải đối mặt với nhiều vấn đề và thách thức bao gồm như:

**Cảnh báo sai (false alarm):** Việc giảm tỷ lệ cảnh báo sai trong khi vẫn đảm bảo là không bỏ sót các hành vi tấn công thực sự là một nhiệm vụ rất khó khăn. Nếu ngưỡng phát hiện được đặt quá chặt thì hệ thống có thể bỏ lọt xâm nhập nhưng ngược lại, nếu như quá lỏng sẽ dẫn đến nhiều cảnh báo không cần thiết, cảnh báo giả báo sai.

**Xây dựng dữ liệu và mô hình ban đầu:** Việc thu thập dữ liệu huấn luyện đại diện cho đầy đủ trạng thái hoạt động bình thường của hệ thống cũng như phát triển một phương pháp phát hiện mang tính tổng quát đều là một thách thức, vấn đề lớn do trong môi trường mạng luôn có sự thay đổi.

**Cập nhật mẫu tấn công**: Các mẫu tấn công mới cần được bổ sung thường xuyên vào cơ sở dữ liệu của hệ thống. Tuy nhiên thì việc cập nhật này phải luôn được xem xét, đảm bảo không làm suy giảm hiệu suất và khả năng phản hồi của hệ thống.

**Độ phức tạp thuật toán:** Nhiều thuật toán phát hiện bất thường có độ phức tạp cao, đặc biệt trong giai đoạn tiền xử lý dữ liệu và huấn luyện mô hình, gây khó khăn cho việc triển khai trong các hệ thống thời gian thực.

**Đa dạng hình thức tấn công:** Mỗi loại tấn công có đặc điểm và hành vi riêng biệt, do đó rất khó để có thể xây dựng một phương pháp phát hiện duy nhất tổng quát phù hợp cho tất cả các kịch bản tấn công.

## Khái niệm Machine Learning và ứng dụng trong phát hiện, ngăn chặn xâm nhập (IDS/IPS):

Machine Learning (ML – học máy) là một nhánh quan trọng của trí tuệ nhân tạo (Artificial Intelligence – AI), nó tập trung vào việc xây dựng các mô hình cho phép hệ thống máy tính học hỏi từ dữ liệu và cải thiện hiệu quả hoạt động theo thời gian mà không cần lập trình các quy tắc cố định. Thay vì phụ thuộc hoàn toàn vào các luật do con người thiết kế sẵn, học máy khai thác dữ liệu lịch sử để nhận diện các mẫu, mối liên hệ và hành vi đặc trưng, từ đó có thể đưa ra dự đoán hoặc quyết định đối với dữ liệu mới.

Về cơ bản thì Machine Learning được chia thành ba nhóm chính:

**Học có giám sát (Supervised Learning):** Mô hình được huấn luyện trên tập dữ liệu đã được gán nhãn, xác minh từ trước, trong đó đầu ra mong muốn đã được xác định trước. Phương pháp này thường được sử dụng cho các bài toán phân loại hoặc dự đoán, chẳng hạn như phân biệt lưu lượng mạng hợp lệ và lưu lượng tấn công.

**Học không giám sát (Unsupervised Learning):** Mô hình làm việc với dữ liệu chưa được gán nhãn và tự động tìm ra cấu trúc hoặc quy luật tiềm ẩn trong dữ liệu. Cách tiếp cận này đặc biệt phù hợp cho việc phát hiện bất thường, nơi mà các hành vi tấn công chưa được xác định rõ từ trước.

**Học tăng cường (Reinforcement Learning):** Hệ thống học thông qua quá trình tương tác liên tục với môi trường, dựa trên cơ chế thưởng và phạt để điều chỉnh hành vi nhằm đạt được mục tiêu hiệu quả tối ưu theo thời gian.

Trong lĩnh vực an ninh mạng, đặc biệt là các hệ thống phát hiện và ngăn chặn xâm nhập (IDS/IPS), Machine Learning đóng vai trò quan trọng trong việc tự động hóa quá trình nhận dạng và phân tích hành vi tấn công. Thay vì chỉ dựa vào mẫu tấn công có sẵn như phương pháp signature-based, các mô hình học máy có thể học và phát hiện các hành vi bất thường chưa từng xuất hiện (zero-day attacks), giúp nâng cao độ chính xác và khả năng thích ứng của hệ thống.

Một hướng nghiên cứu phổ biến là kết hợp học có giám sát và học không giám sát để cải thiện hiệu quả phát hiện. Trong đó, unsupervised learning được sử dụng để phát hiện các mẫu dữ liệu lạ hoặc bất thường, còn supervised learning được huấn luyện để phân loại, gán nhãn và đánh giá mức độ rủi ro của từng sự kiện. Cách tiếp cận này cho phép hệ thống tự động sinh và ưu tiên tập luật phát hiện (rules) dựa trên các đặc trưng mạng như địa chỉ IP, mẫu payload, giao thức và cổng truy cập (port patterns).

Nhờ khả năng tự học và thích nghi thì hệ thống IDS/IPS ứng dụng Machine Learning không chỉ dừng lại ở việc phát hiện và cảnh báo mà còn có thể tự điều chỉnh để nâng cao mức hiệu quả theo thời gian. Khi được tích hợp với các kỹ thuật xử lý dữ liệu lớn (Big Data), học máy giúp góp phần làm giảm đi đáng kể tỷ lệ cảnh báo giả, đồng thời nâng cao khả năng phát hiện các hình thức tấn công phức tạp trong môi trường mạng hiện đại hiện nay. Đây được xem là xu hướng phát triển tất yếu trong việc xây dựng các hệ thống an ninh mạng thông minh, đáp ứng yêu cầu bảo mật ngày càng cao trong kỷ nguyên chuyển đổi số.

# MÔ HÌNH ĐỀ XUẤT

## Kiến trúc hệ thống

Hệ thống được triển khai trên hạ tầng ảo hóa VMware với mô hình mạng như sau:

• pfSense Firewall/Router: Quản lý traffic WAN/LAN, chạy Suricata IDS/IPS.

• Wazuh All-in-One Server: Thu thập log, phân tích và hiển thị Dashboard.

• Machine Learning Environment (Ubuntu): Chạy các script Python để train model và detect anomaly.

Sơ đồ luồng dữ liệu:

1. Thu thập: Log từ Wazuh Indexer được export qua API OpenSearch.

2. Tiền xử lý (Preprocessing): Làm sạch dữ liệu, mã hóa (Label Encoding) các trường text như event\_desc, agent. Feature Engineering tạo ra 71 đặc trưng từ dữ liệu thô.

3. Phát hiện (Detection): Sử dụng mô hình Ensemble (kết hợp Isolation Forest, LOF, SVM) để chấm điểm bất thường.

4. Phản ứng (Response): Nếu phát hiện tấn công nghiêm trọng, hệ thống gọi module pfsense\_integration.py để block IP thông qua SSH.

## Quy trình huấn luyện Quy trình bao gồm các bước chính:

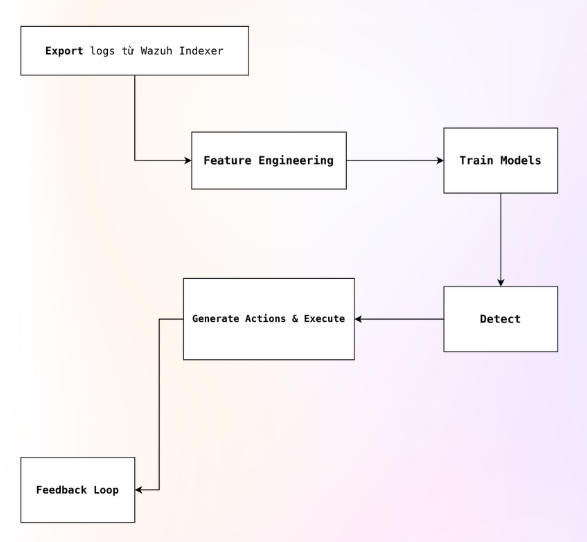
• Bước 1: Export log từ Elasticsearch/OpenSearch.

• Bước 2: Train model Isolation Forest với tỉ lệ contamination=0.1.

• Bước 3: Train model Classification để phân loại kiểu tấn công (Attack Type).

• Bước 4: Lưu model (.pkl) và Feature Selector.

## Pipeline hệ thống:



Hình 3. Pipeline hệ thống

Pipeline hệ thống gồm 6 bước chính:

* Trích xuất logs từ Wazuh Indexer bằng query
* Feature Engineering: Tạo 71 features từ raw data
* Train Models Anomaly detection (Isolation-Forest/Ensemble) + Classification
* Detect: Phát hiện và phân loại anomalies
* Tự động generate actions (block IP/Port, notify) và thực thi actions trên pfSense firewall
* Feedback Loop: Analyze → Tune → Retrain → Test (cải thiện liên tục)

## Các agents và server được sử dụng

A screenshot of a computer program

AI-generated content may be incorrect.

Hình 4. Các agents và server được sử dụng

## Quy trình huấn luyện Quy trình bao gồm các bước chính:

### Cấu hình máy pfsense 2.7.2

- Cấu hình máy ảo cho pfsense

* Tùy chọn cho NETWORK ADAPTER: NAT (192.168.180.129/24)
* Tùy chọn cho NETWORK ADAPTER 02: mạng nội bộ (172.16.158.100/24)

- Kiểm tra thiết lập cho WAN

* Chọn INTERFACE và chọn WAN
* WAN thì mình lấy DHCP từ VMWARE
* Tắt: “Block private networks and loopback addresses” và “Block bogon networks”

- Thêm rule cho WAN

* FIREWALL, chọn RULES và chọn WAN
* Chọn ADD để thêm một rule mới
* Chọn PASS, chọn ANY (Protocol)

- Có thể kiểm tra bằng ping từ KALI đến IP WAN của pfsense (ping phải thành công)

- KIỂM TRA thiết lập cho LAN

* Có thể để mặc định

- Cấu hình DNS Forwarders

* Chọn SERVICES, chọn DNS RESOLVER và tắt dịch vụ này
* Chọn SERVICES, chọn DNS FORWARDER và bật dịch vụ này

### Cài đặt và cấu hình suricata trên pfsense 2.7.2

- Cài đặt dịch vụ suricata trên pfsense

* SYSTEM và chọn PACKAGE MANAGER
* Chọn AVAILABLE PACKAGES và nhập suricata (chọn INSTALL để cài đặt)
* Sau khi hoàn tất thì sẽ có mục suricata trong SERVICES

- Cấu hình suricata cho WAN (interface)

* Chọn INTERFACE và chọn ADD
* Mục INTERFACE: chọn WAN
* Mục LOGGING: để mặc định
* Chọn mục EVE OUTPUT và tùy chọn

- Cấu hình hardware checksum offloading

* Chọn SYSTEM, chọn ADVANCED và chọn NETWORKING
* Tích Hardware Checksum Offloading
* Chọn SAVE và pfsense sẽ khởi động lại

- Cấu hình ETOPEN (dùng ruleset)

* Chọn GLOBAL SETTINGS và chọn ETOPEN
* Chọn UPDATES và chọn update để cập nhập bộ rule của etopen
* Chọn vào INTERFACE và chọn EDIT
* Chọn WAN CATEGORIES và chọn SELECT ALL (chọn SAVE để áp dụng thiết lập)
* Chọn RESTART để khởi động lại INTERFACE

- Kiểm tra kết quả

* Ping thử từ máy KALI vào IP WAN
* Phát hiện PING trong ALERTS
* Chọn LOGS VIEW và xem eve.json

### Cài đặt WAZUH bằng QUICKSTART

- Cấu hình ip tĩnh cho máy wazuh

* # /etc/netplan/00-installer-config.yaml
* network:
* ethernets:
* ens33:
* dhcp4: false
* addresses: [172.16.158.150/24]
* routes:
* - to: default
* via: 172.16.158.100
* nameservers:
* addresses:
* - 172.16.158.100
* - 8.8.8.8
* version: 2
* # áp dụng thiết lập: sudo netplan apply
* # kiểm tra ping thành công đến google.com

- Cấu hình hostname

# dùng lệnh hostnamectl: sudo hostnamectl set-hostname wazuh-server

- Cấu hình sử dụng tài khoản root

# gán hay tạo mật khẩu: sudo passwd

# dùng root: su –

- Khởi động lại máy ảo

# dùng lệnh: sudo reboot

- Cập nhập hệ thống bằng apt

# dùng lệnh apt: sudo apt update && sudo apt full-upgrade

- Link doc của wazuh cho quickstart:

https://documentation.wazuh.com/current/quickstart.html

- Cài đặt wazuh bằng script

# dùng lệnh : curl -sO https://packages.wazuh.com/4.14/wazuh-install.sh && sudo bash ./wazuh-install.sh -a

### Cài đặt wazuh agent trên pfsense 2.7.2

- Dùng ssh cho pfsense 2.7.2

• Chọn SYSTEM và chọn ADVANCED

• Bật Enable Secure Shell và chọn SAVE

• Truy cập từ máy MANAGEMENT

- Cài đặt nano để tùy chỉnh tập tin

# cập nhập repo : pkg update

# cài đặt nano : pkg install nano

- Thiết lập FREEBSD để dùng cho repo

# /usr/local/etc/pkg/repos/pfSense.conf

nano /usr/local/etc/pkg/repos/pfSense.conf

FreeBSD: {enabled: yes}

# /usr/local/etc/pkg/repos/FreeBSD.conf

nano /usr/local/etc/pkg/repos/FreeBSD.conf

FreeBSD: { enabled: yes }

# cập nhập lại : pkg update

- Cài đặt WAZUH AGENT cho pfsense

# tìm kiếm gói wazuh-agent : pkg search wazuh

# cài đặt bằng pkg : pkg install wazuh-agent-4.12.2

- Cấu hình WAZUH-AGENT trên pfsense

# copy file cấu hình sẵn vào wazuh : cp /etc/localtime /var/ossec/etc/

# tùy chỉnh : /var/ossec/etc/ossec.conf

• cd /var/ossec/etc

• nano /var/ossec/etc/ossec.conf

• # thay IP bằng IP của WAZUH-SERVER

• <address>172.16.158.150<address>

• # thay PROTOCOL bằng TCP

• <protocol>tcp</protocol>

- Khởi chạy dịch vụ wazuh-agent

# bật tính năng autorun cho dịch vụ

• sysrc wazuh\_agent\_enable="YES"

# tạo symbolick để chạy script

• ln -s /usr/local/etc/rc.d/wazuh-agent /usr/local/etc/rc.d/wazuh-agent.sh

# khởi chạy dịch vụ

• service wazuh-agent start

- Cấu hình giám sát eve.json

# tìm vị trí của tập tin eve.json

• Truy cập logs view và chọn eve.json

# tùy chỉnh : /var/ossec/etc/ossec.conf

• nano /var/ossec/etc/ossec.conf

# thêm localfile cho eve.json

• <localfile>

• <log\_format>json</log\_format>

• <location>/var/log/suricata/suricata\_em039435/eve.json</location>

• </localfile>

# khởi động lại dịch vụ wazuh-agent

• service wazuh-agent restart

## Pfsense Firewall GUI và CLI:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 5. Pfsense Firewall GUI và CLI

## Pfsense package Suricata:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 6. Pfsense package Suricata

## Wazuh server:

A screenshot of a computer

AI-generated content may be incorrect.

Hình 7. Wazuh server

## Kali:

A screenshot of a computer

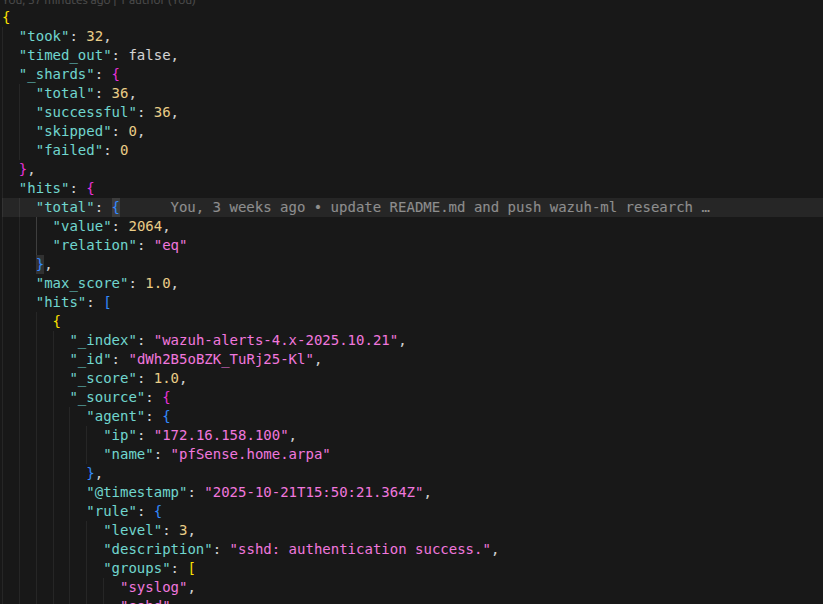
AI-generated content may be incorrect.

Hình 8. Kali

# THỰC NGHIỆM

## Dữ liệu thực nghiệm

Dữ liệu thực nghiệm Dữ liệu bao gồm 2064 bản ghi từ wazuh và pfsense tổng hợp lại và trích xuất qua opensearch

Hình 9. Dữ liệu thực nghiệm

## Cài đặt thực nghiệm

### Thu thập dữ liệu

# KẾT LUẬN

## Kết luận

Hệ thống Wazuh ML đã hoàn thiện chu trình từ thu thập log (Elastic/Wazuh Indexer) → xử lý đặc trưng → huấn luyện song song mô hình phát hiện bất thường (ensemble Isolation Forest + LOF + One-Class SVM) và hai mô hình phân loại (attack type, event category).

Pipeline tuning tự động tìm contamination/voting tối ưu, lưu lịch sử (tuning\_history.json) và đóng gói mô hình để tái sử dụng cho bước detect/classify.

Kết quả thực nghiệm cho thấy mô hình phát hiện được 3.69% log bất thường với phân tích chi tiết theo thời gian, agent và loại sự kiện; classifier đạt accuracy ~0.99

## Hướng phát triển

Gia cố mô hình: thử giảm độ phức tạp RandomForest (giới hạn max\_depth, tăng min\_samples\_leaf), giảm số feature RFE xuống 20–25 và bổ sung regularization cho SVM/LOF để mô hình nhạy hơn với mẫu mới.

Tự động hóa quy trình: thêm subcommand không tương tác (ví dụ python main.py train-all --auto) để dễ tích hợp CI/CD hoặc cron; mở rộng feedback loop thành pipeline retrain định kỳ dựa trên tuning\_history.

Giám sát & cảnh báo: dựng dashboard trực quan cho các thống kê bất thường (theo giờ, agent, severity), đồng thời tích hợp webhook gửi cảnh báo khi ensemble phát hiện sự kiện unanimous mức cao.

Áp dụng LLM vào phân tích log anomalies được trả về

Phát triển thêm đề xuất và điều chỉnh rule trên pfsense bằng ssh qua kết quả giám sát anomaly

# TÀI LIỆU THAM KHẢO

**Tiếng Việt**

* Hệ điều hành Windows – Cài đặt, sử dụng và xử lý sự cố cơ bản, NXB Lao động – Xã hội, 2019.
* Microsoft Việt Nam, Hướng dẫn sử dụng Windows 10 và Windows 11, truy cập tại: https://support.microsoft.com/vi-vn
* Công ty Cổ phần MISA, Tài liệu hướng dẫn sử dụng phần mềm kế toán MISA, truy cập tại: https://help.misa.vn
* Công ty Fast Software, Hướng dẫn sử dụng phần mềm quản lý bán hàng Fast Sales, truy cập tại: https://fast.com.vn
* Google Việt Nam, Hướng dẫn sử dụng Google Drive, truy cập tại: https://support.google.com/drive
* Microsoft, Hướng dẫn sử dụng OneDrive, truy cập tại: https://support.microsoft.com/onedrive

**Tiếng Anh**

* 1. Documentation of Suricata IDS/IPS, The Open Information Security Foundation (OISF), available at: https://suricata.io.
* 2. Documentation of Wazuh Security Platform, Wazuh Inc., available at: https://documentation.wazuh.com.
* 3. Documentation of pfSense Firewall, Netgate, available at: https://docs.netgate.com/pfsense.
* 4. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. IEEE Symposium on Computational Intelligence for Security and Defense Applications.
* 5. Pfahringer, B., & Bifet, A. (2018). Handling Concept Drift in Streaming Data with Adaptive Learning. Machine Learning Journal, 107(11), 1985–2020.
* 6. Scarfone, K., & Mell, P. (2007). Guide to Intrusion Detection and Prevention Systems (IDPS). National Institute of Standards and Technology (NIST).