ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHAN THẾ DUY

TĂNG CƯỜNG KHẢ NĂNG PHÒNG CHỐNG TẤN CÔNG TRONG MẠNG SDN

Ngành: **Công Nghệ Thông Tin**
Mã số: **9.48.02.01**

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

TP. HỒ CHÍ MINH – Năm 2025

Công trình được hoàn thành tại:

PHÒNG THÍ NGHIỆM AN TOÀN THÔNG TIN,

TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN – ĐHQG TP. HCM.

Người hướng dẫn khoa học:

TS. Phạm Văn Hậu

PGS. TS. Lê Đình Duy

Phản biện 1: .....

Phản biện 2: .....

Phản biện 3: .....

Phản biện độc lập 1: .....

Phản biện độc lập 2: .....

Luận án sẽ/đã được bảo vệ trước hội đồng chấm luận án cấp Trường tại:

.....

.....

vào lúc… giờ … ngày … tháng … năm … .

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Thư viện ĐHQG-HCM
- Thư viện Trường Đại học Công nghệ Thông tin – Đại học Quốc gia Tp. Hồ Chí Minh.

MỤC LỤC

TÓM TẮT

Với kiến trúc mạng khả lập trình (Software-Defined Networking - SDN), việc điều phối an toàn thông tin được thực hiện thông qua các ứng dụng mạng và bộ điều khiển. Các ứng dụng mạng này có thể kiểm soát quyền truy cập, áp dụng kiểm soát bảo mật và theo dõi hoạt động mạng để phát hiện và ngăn chặn các mối đe dọa, tạo ra một môi trường mạng tin cậy và an toàn cho các hoạt động kinh doanh và giao tiếp trực tuyến. Tuy nhiên, để xây dựng các hệ thống phát hiện xâm nhập (IDS) hiệu quả trong môi trường này, cần phải vượt qua một số rào cản đến từ sự hiểu biết của chuyên gia nhiều hơn để thu thập đủ dấu hiệu tấn công phục vụ mục đích đào tạo các mô hình học máy phát hiện xâm nhập trong hệ thống. Việc thu thập này đòi hỏi tiêu tốn nhiều thời gian, băng thông truyền tải, lưu trữ dữ liệu, cũng như quyền riêng tư dữ liệu bị xâm phạm trong các cách tiếp cận học máy truyền thống. Học liên kết - Federated Learning (FL) sẽ giúp giải quyết các vấn đề của IDS học máy tập trung, như thu thập dữ liệu lớn, đảm bảo quyền riêng tư, và cải thiện khả năng nhận diện mối đe dọa.

Việc triển khai các IDS dựa trên FL mang lại nhiều lợi ích trong việc bảo vệ dữ liệu nhạy cảm, nhưng cũng gặp phải một số yếu điểm và rào cản. Đầu tiên, các tấn công xâm phạm quyền riêng tư (privacy attack) nhắm vào các hệ thống dựa trên FL có thể xảy ra trong quá trình truyền tải dữ liệu giữa các máy khách huấn luyện và máy chủ, cũng như trong quá trình lưu trữ và tính toán tại máy chủ huấn luyện. Những tấn công này có thể dẫn đến việc lộ thông tin nhạy cảm từ khai thác các bản cập nhật của mô hình từ máy khách huấn luyện hay máy chủ tổng hợp, làm giảm tính bảo mật và tin cậy của hệ thống, và gây ra hậu quả nghiêm trọng cho sự an toàn của toàn bộ mạng. Thứ hai, các mô hình IDS khi triển khai theo mô hình FL có thể bị đầu độc bởi các bên tham gia ác ý do có sự cộng tác từ nhiều bên tham gia mà không thể kiểm soát tính trung thực. Để ngăn chặn ảnh hưởng của tấn công đầu độc mô hình liên kết phát hiện xâm nhập, cần xác định và loại bỏ các bản cập nhật mô hình độc hại nhằm ngăn chặn tác động tiêu cực của các máy khách huấn luyện ác ý. Thứ ba, triển khai kiểm tra, đánh giá và tăng cường tính bền vững của các ứng dụng phát hiện xâm nhập cũng là một thách thức lớn. Điều này là cần thiết vì các cuộc tấn công mạng liên tục biến đổi để trốn tránh sự nhận diện từ hệ thống phòng thủ. Bằng cách đánh giá và nâng cao tính bền vững, các ứng dụng phát hiện xâm nhập sẽ có khả năng phát hiện và ngăn chặn các cuộc tấn công mới một cách hiệu quả. Cuối cùng, việc xác thực quá trình kết nối giữa các ứng dụng mạng tới bộ điều khiển còn thiếu cơ chế kiểm soát truy cập linh động và tin cậy. Để tăng cường tính an toàn cho việc giám sát và thu thập dữ liệu huấn luyện các ứng dụng IDS liên kết, cơ chế kiểm soát truy cập ứng dụng cần được phi tập trung hóa. Thay vì dựa vào một điểm tập trung để quản lý quyền truy cập, các quyết định về kiểm soát truy cập sẽ được

phân tán và thực hiện trên các thiết bị mạng cục bộ, giúp tăng cường khả năng mở rộng, linh hoạt và tin cậy cho hệ thống.

Từ những hiện trạng và động lực như trên, luận án này tập trung vào việc giải quyết các vấn đề trong triển khai hệ thống phát hiện xâm nhập (IDS) liên kết trong mạng khả lập trình (SDN) bằng cách sử dụng mô hình học liên kết (Federated Learning - FL). Đầu tiên, luận án sẽ tập trung giải quyết các tấn công xâm phạm quyền riêng tư (privacy attack) trong quá trình truyền tải và lưu trữ dữ liệu giữa máy khách huấn luyện và máy chủ. Thứ hai, luận án sẽ đề xuất các giải pháp ngăn chặn tấn công đầu độc mô hình (poisoning attack) bằng cách xác định và loại bỏ các bản cập nhật mô hình độc hại từ các máy khách ác ý. Thứ ba, luận án sẽ trình bày cơ chế kiểm tra, đánh giá và tăng cường tính bền vững của các ứng dụng phát hiện xâm nhập để đối phó với các cuộc tấn công mạng liên tục biến đổi. Cuối cùng, nó sẽ phát triển cơ chế kiểm soát truy cập linh động và tin cậy, phi tập trung hóa quá trình quản lý quyền truy cập để tăng cường giám sát và thu thập dữ liệu mạng, giúp hệ thống trở nên linh hoạt và tin cậy hơn.

CHƯƠNG 1. Giới thiệu tổng quan

1.1. Động lực nghiên cứu

Đứng ở góc nhìn của quản trị viên, lớp ứng dụng trong kiến trúc SDN là cầu nối để quản trị viên có thể thực hiện các thao tác giám sát, quản lý, điều phối các hoạt động và chính sách trong mạng. Do đó, để đảm bảo sự hoạt động hiệu quả của toàn hệ thống mạng, yêu cầu xây dựng tầng ứng dụng an toàn, bảo mật là một trong những ưu tiên quan trọng trong hệ thống. Trong số các ứng dụng đảm bảo an toàn thông tin cho hệ thống mạng, các ứng dụng phát hiện xâm nhập, săn tìm mối đe dọa được xem là lá chắn bảo vệ hệ thống mạng trước những hành vi phá hoại từ bên ngoài. Để xây dựng được hệ thống hệ thống phát hiện xâm nhập, tấn công mạng hiệu quả cho hệ thống mạng thì vấn đề thiếu hụt dữ liệu và quản lý lưu trữ dữ liệu phục vụ huấn luyện mô hình học máy là những thách thức cần giải quyết. Vấn đề thiếu hụt, khan hiếm dữ liệu thu thập được có thể bắt đầu từ những lo ngại về quyền riêng tư dữ liệu bị xâm phạm, tiết lộ một khi chủ sở hữu dữ liệu chia sẻ nó cho bên khai thác dữ liệu. Điều này làm giảm sự đa dạng của dữ liệu từ việc thu thập từ nhiều nguồn, do các bên không sẵn lòng chia sẻ với các tổ chức khác.

Trong tình huống này, học liên kết (Federated Learning - FL) được xem như một giải pháp thúc đẩy sự hợp tác chia sẻ thông tin, dữ liệu về các mối đe dọa từ chính nguồn dữ liệu nội bộ trong khi vẫn đảm bảo được sự toàn vẹn và riêng tư cho dữ liệu. Cụ thể, mô hình học liên kết không yêu cầu thu thập và lưu trữ, huấn luyện mô hình học máy tập trung. Nó hoạt động theo hình thức phân tán với nguyên tắc mô hình học máy được huấn luyện cục bộ tại các bên nắm giữ dữ liệu và tổng hợp mô hình huấn luyện từ các mô hình học máy đã được huấn luyện. Điều này có nghĩa là, các bên tham gia vào qui trình học liên kết chỉ chia sẻ mô hình sau khi được huấn luyện, mà không bắt buộc phải chia sẻ dữ liệu riêng tư của mình. Các trình phát hiện xâm nhập ứng dụng học liên kết trong các mạng SDN phân tán được đề xuất nhằm tận dụng nguồn dữ liệu cục bộ tại từng phân vùng, hệ thống mạng một cách hiệu quả mà vẫn đảm bảo được quyền riêng tư dữ liệu.

1.2. Mục tiêu, đối tượng và phạm vi nghiên cứu

1.2.1. Mục tiêu nghiên cứu

Luận án này thực hiện nghiên cứu và phát triển các giải pháp tăng cường khả năng phòng chống tấn công cho mạng khả lập trình (SDN) dựa trên việc xây dựng cơ chế đảm bảo an toàn,

tin cậy cho ứng dụng mạng thực hiện vai trò giám sát, phát hiện tấn công mạng và các tác nhân đe dọa nhắm vào hệ thống thông qua việc giải quyết 4 vấn đề trọng tâm như sau:

- Đảm bảo tính an toàn, tin cậy, bảo mật thông tin, bảo vệ quyền riêng tư dữ liệu trong mô hình liên kết phát hiện xâm nhập và săn tìm mối đe dọa cho hệ thống mạng bằng sơ đồ học liên kết cho việc cộng tác huấn luyện mô hình học máy trong các ứng dụng mạng.
- Ngăn chặn ảnh hưởng của tấn công đầu độc mô hình liên kết phát hiện xâm nhập bằng phương pháp xác định và loại bỏ các bản cập nhật mô hình độc hại từ các máy khách huấn luyện ác ý dựa trên chiến lược phân tích không gian tiềm ẩn của mô hình cập nhật thông qua biểu diễn lớp áp chót của mạng nơ-ron nhiều lớp.
- Triển khai kiểm tra, đánh giá và tăng cường tính bền vững của các ứng dụng phát hiện xâm nhập, săn tìm mối đe dọa được xây dựng trên các mô hình học máy đặt trong ngữ cảnh sơ đồ học liên kết.
- Đảm bảo tính tin cậy, an toàn và dễ mở rộng cho bộ khung kiểm soát truy cập các ứng dụng mạng kết nối với bộ điều khiển SDN để thực hiện các tác vụ quản lý, giám sát, điều phối, thu thập thông tin về trạng thái mạng SDN cho việc huấn luyện các ứng dụng IDS học máy.

Tóm tắt vấn đề nghiên cứu của đề tài được thể hiện trong **Hình 1.1**.

1.2.2. Đối tượng nghiên cứu

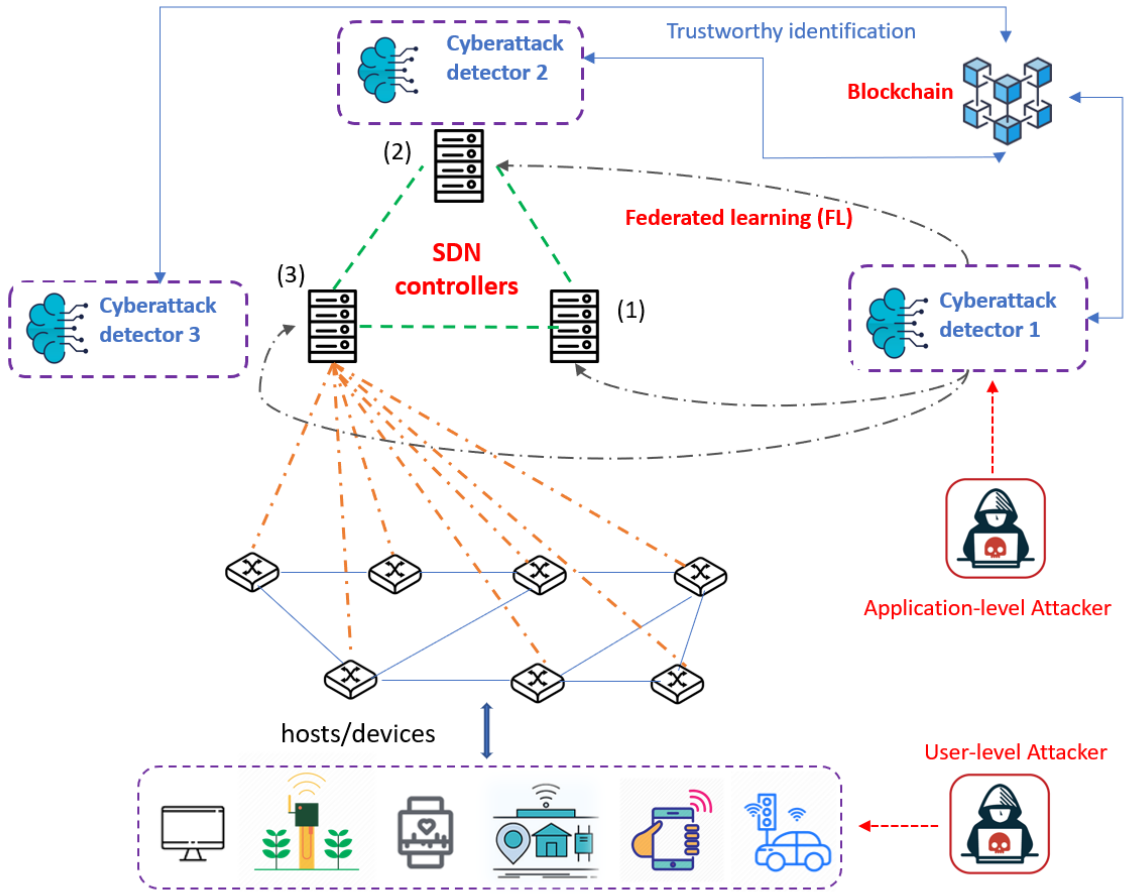
NCS thực hiện bài toán tăng cường tính an toàn, bảo mật, bền vững và xác thực của các ứng dụng liên kết phát hiện xâm nhập trong hệ thống mạng cộng tác SDN.

1.2.3. Phạm vi nghiên cứu

Luận án này giới hạn phạm vi nghiên cứu là các ứng dụng liên kết phát hiện xâm nhập dựa trên phân tích các luồng mạng hoạt động trong tầng ứng dụng của kiến trúc mạng SDN với sự tham gia của nhiều tổ chức, phân vùng mạng phân tán.

1.3. Những đóng góp chính của nghiên cứu

Nghiên cứu này tập trung vào việc thiết kế và xây dựng cơ chế tăng cường khả năng phòng chống tấn công trong mạng SDN thông qua cách tiếp cận đảm bảo tính bền vững, an toàn và tin cậy cho các ứng dụng điều phối an ninh điển hình như các ứng dụng điều khiển, giám sát và phát hiện tấn công mạng theo cơ chế liên kết. Các đóng góp chính của luận án được tóm tắt như sau:



Hình 1.1: *Mô hình đảm bảo an toàn cho các ứng dụng điều phối an ninh trong mạng SDN.*

- Cơ chế đảm bảo tính tin cậy, an toàn, đảm bảo quyền riêng tư trong việc khuyến khích các bên tham gia đóng góp vào quá trình huấn luyện mô hình học máy liên kết cho các ứng dụng phát hiện và săn tìm mối đe dọa trong mạng khả lập trình. Giải pháp được đề xuất dựa trên sự kết hợp giữa Mã hóa đồng cấu toàn phần (HE) và Riêng tư vi phân (DP) tích hợp với bộ khung blockchain. Cơ chế này giúp các ứng dụng IDS liên kết trở nên an toàn hơn trước nguy cơ tấn công riêng tư/suy diễn (privacy/inference attack).

- Cơ chế ngăn chặn tấn công đầu độc mô hình liên kết phát hiện xâm nhập trong các hệ thống mạng phân tán tham gia vào qui trình huấn luyện cộng tác. Cụ thể, dựa trên chiến lược phân tích không gian tiềm ẩn (latent space) của các bản cập nhật mô hình cục bộ, phương pháp đề xuất có thể ngăn chặn ảnh hưởng của tấn công đầu độc vào mô hình toàn cục, giúp việc huấn luyện các ứng dụng phát hiện xâm nhập dựa trên học liên kết được an toàn trước các tấn công phá hoại ác ý, mà không yêu cầu sử dụng một tập dữ liệu phụ trợ tại máy chủ để kiểm tra.

- Cơ chế đánh giá, nâng cao tính bền vững của các ứng dụng phát hiện xâm nhập, và săn tìm mối đe dọa dựa trên phương pháp học máy trước các mẫu đối kháng tiềm năng có thể trốn tránh sự phát hiện của hệ thống. Cụ thể, nghiên cứu đề xuất phương pháp sinh mẫu

đối kháng dựa trên việc đánh giá ràng buộc bảo tồn các thuộc tính chức năng của luồng lưu lượng mạng khi tạo nhiễu trên mẫu tấn công nguyên bản. Ngoài ra, các yếu tố đặc trưng của mạng SDN cũng được xem xét trong quá trình biến đổi mẫu tấn công để tạo biến thể luồng lưu lượng trốn tránh.

- Cơ chế kiểm soát truy cập tin cậy phi tập trung cho các ứng dụng mạng tham gia giám sát, điều phối, hay sử dụng trạng thái của các hệ thống mạng SDN, vốn được sử dụng trong tác vụ theo dõi, thu thập, và gán nhãn dữ liệu cho các ứng dụng phát hiện xâm nhập, săn tìm mối đe dọa dựa trên học máy. Cụ thể, cơ chế này được triển khai theo mô hình kiểm soát truy cập dựa trên vai trò tích hợp trên bộ khung blockchain nhằm đáp ứng tính linh hoạt trong xác thực liên miền và dễ dàng mở rộng khi kích thước của hệ thống cộng tác trở nên lớn hơn.

1.4. Cấu trúc luận án

Bố cục của luận án được trình bày theo cấu trúc sau:

- **Chương 1. Giới thiệu tổng quan**
- **Chương 2. Cơ sở lý thuyết**
- **Chương 3. Cơ chế học liên kết an toàn cho ứng dụng săn tìm mối đe dọa và phát hiện xâm nhập trong mạng khả lập trình**
- **Chương 4. Cơ chế ngăn chặn tấn công đầu độc cho ứng dụng liên kết phát hiện xâm nhập trong môi trường phân tán**
- **Chương 5. Cơ chế đánh giá tính bền vững của các trình phát hiện xâm nhập trong mạng khả lập trình**
- **Chương 6. Cơ chế xác thực và kiểm soát truy cập phi tập trung cho các ứng dụng trong mạng khả lập trình**
- **Chương 7. Kết luận và hướng phát triển**

CHƯƠNG 2. Cơ sở lý thuyết

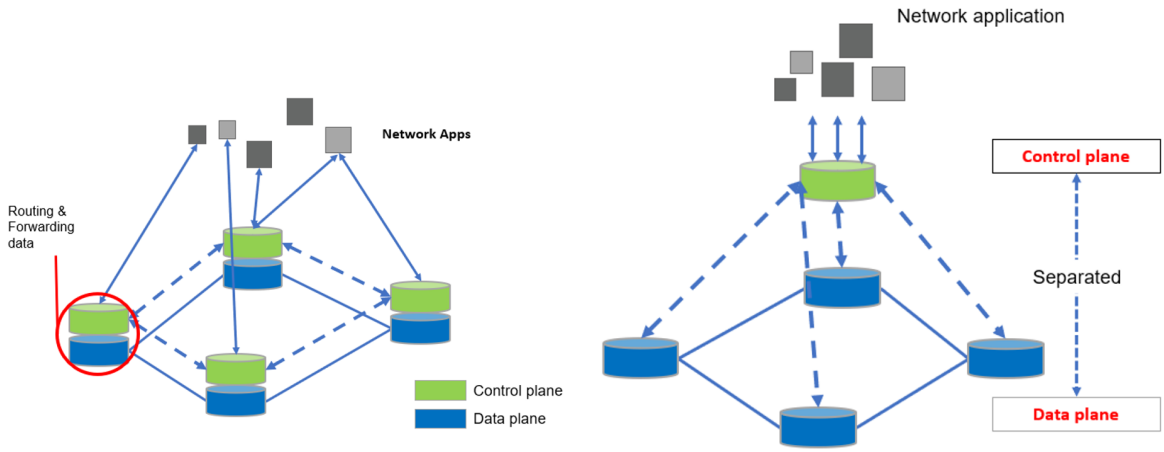
2.1. Mạng khả lập trình

Theo ONF (Open Networking Foundation – tổ chức phi lợi nhuận đang hỗ trợ việc phát triển SDN thông qua việc nghiên cứu các tiêu chuẩn mở phù hợp) thì Software Defined Network (SDN) là một kiểu kiến trúc mạng mới, linh động, dễ quản lý, chi phí hiệu quả, dễ thích nghi và rất phù hợp với nhu cầu mạng ngày càng tăng hiện nay. Kiến trúc này phân tách chức năng điều khiển mạng (Control Plane) và chức năng vận chuyển dữ liệu (Forwarding Plane hay Data Plane) trên mỗi thiết bị mạng trong các mô hình mạng truyền thống. Điều này cho phép việc điều khiển mạng SDN trở nên có thể lập trình được dễ dàng và cơ sở hạ tầng mạng độc lập với các ứng dụng và dịch vụ mạng. **Hình 2.1** thể hiện sự khác biệt giữa mô hình mạng truyền thống và mô hình mạng SDN nằm ở sự tách biệt hay tích hợp giữa chức năng điều khiển và chức năng vận chuyển dữ liệu mạng.

2.2. Hệ thống phát hiện xâm nhập, săn tìm mối đe dọa

Hệ thống phát hiện xâm nhập (Intrusion Detection System - IDS) là thiết bị hoặc phần mềm có nhiệm vụ giám sát lưu lượng mạng, các hành vi đáng ngờ và cảnh báo cho quản trị viên hệ thống. Mục đích của IDS là phát hiện và hỗ trợ ngăn chặn kịp thời các hoạt động gây hại tới hệ thống bao gồm các tấn công, xâm nhập từ bên ngoài hoặc truy cập trái phép vào hệ thống. Khi phát hiện các hoạt động bất thường, IDS sẽ đưa ra các cảnh báo (alert) để người quản trị đưa ra các quyết định ứng phó. Tuỳ theo cách triển khai và cấu hình, IDS có thể tự động ngăn chặn các hành vi xâm nhập khi ở chế độ phòng chống xâm nhập (Intrusion Prevention System - IPS mode).

Các mô hình IDS truyền thống phát hiện xâm nhập dựa trên dấu hiệu (Signature-Based IDS) thực hiện so sánh lưu lượng truy cập với cơ sở dữ liệu chứa các mẫu tấn công (gọi là dấu hiệu/chữ ký). Kiểu IDS này khó phát hiện ra những dạng tấn công mới do bị giới hạn bởi số lượng chữ ký đại diện cho lưu lượng các cuộc tấn công trong cơ sở dữ liệu. Để khắc phục các nhược điểm của IDS truyền thống, các thuật toán Học máy (Machine Learning - ML), hay học sâu (Deep Learning - DL) được áp dụng trong IDS để xác định và phân loại các mối đe dọa bảo mật. Các ML-IDS dùng phương pháp thống kê lưu lượng mạng trong các khoảng thời gian khác nhau để tạo nên một đường cơ sở (baseline) và dựa vào đó để phát hiện ra những hành vi đáng ngờ. Dạng IDS này thường sử dụng các kỹ thuật học máy để tạo ra một mô hình mô phỏng việc truy cập thông



(a) Kiến trúc mạng truyền thống *(b) Kiến trúc mạng SDN*

Hình 2.1: *So sánh sự khác biệt giữa kiến trúc mạng truyền thống và mạng SDN*

thường của người dùng mạng thông qua các dữ liệu đã ghi nhận trong lịch sử hoạt động hay quá trình huấn luyện. Do đó nếu có một truy cập bất thường, hệ thống ML-IDS sẽ đưa ra cảnh báo xâm nhập.

Có bốn trường hợp khi luồng dữ liệu truy cập cố gắng đi qua IDS. Hai trường hợp đầu tiên được mong đợi là lưu lượng truy cập bình thường đi qua và lưu lượng độc hại bị từ chối do bị phát hiện là tấn công, xâm nhập. Nhưng bên cạnh đó, sẽ có hai trường hợp luồng dữ liệu mạng có thể bị phân loại sai. Dương tính giả (false positive) là khi lưu lượng bình thường bị coi là độc hại và bị từ chối trước khi vào hệ thống. Trong khi đó, âm tính giả (false negative) là khi lưu lượng độc hại được coi là bình thường và được phép vào hệ thống. Các hệ thống trí tuệ nhân tạo đối kháng sẽ tập trung vào hai trường hợp cuối cùng bằng cách liên tục tạo ra lưu lượng tấn công đối kháng giả mạo là lưu lượng lành tính để đánh lừa IDS. Khi các cuộc tấn công ngày càng tinh vi và các cuộc tấn công mới liên tục xuất hiện, các hệ thống phòng thủ sau một thời gian hoạt động đặt ra yêu cầu cần phải được kiểm tra để thích ứng với các loại tấn công mới.

Theo sách trắng của hãng bảo mật nổi tiếng Kaspersky, khác với các lĩnh vực khác, lĩnh vực an toàn thông tin luôn chứng kiến sự biến đổi không ngừng của các dạng thức tấn công, các mẫu mã độc. Điều này bắt nguồn từ việc tin tặc, tác giả tạo ra mã độc luôn cố gắng thích nghi để trốn tránh các giải pháp phát hiện và ngăn chặn của hệ thống. **Hình 2.2** mô tả hiệu năng bị giảm sút đáng kể của các hệ thống bảo mật dựa trên học máy theo dòng thời gian nếu không được cập nhật liên tục, theo như nghiên cứu của hãng bảo mật Kaspersky.

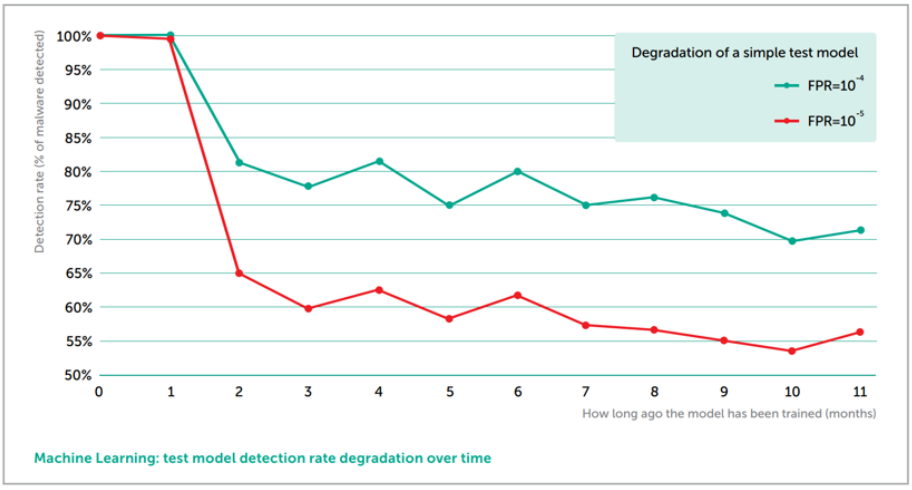
2.3. Học máy đối kháng

Trong học máy đối kháng (adversarial machine learning - AML), đối thủ luôn tìm kiếm cách để gây nhầm lẫn cho các mô hình học máy với việc đưa ra kết quả dự đoán sai lệch. Kẻ tấn công

này làm được việc này bằng cách hiệu chỉnh các dữ liệu đầu vào cho mô hình học máy trong quá trình huấn luyện (tấn công nhiễm độc – poisoning attack) hoặc trong quá trình suy luận, dự đoán (tấn công trốn tránh – evasion attack). Trong lĩnh vực an toàn thông tin, các biến thể mới được tạo ra cần phải bảo tồn được chức năng độc hại ban đầu của nó sau quá trình chỉnh sửa trên mẫu nguyên bản ban đầu. Điều này là sự khác biệt chính giữa bài toán phân loại hình ảnh và bài toán phát hiện các mối đe dọa trên không gian mạng; và là thách thức lớn nhất cho các bài toán tấn công đối kháng trong lĩnh vực an toàn thông tin. Để giải quyết vấn đề này, các mẫu đối kháng được tạo ra trong lĩnh vực an toàn thông tin phải thực hiện phương pháp biến đổi các thuộc tính của đối tượng mà không làm phá vỡ tính năng của chúng trên các phiên bản phát sinh mới.

2.4. Học máy liên kết

Học liên kết (Federated Learning - FL), là một phương pháp tiếp cận học máy có thể giúp các mô hình học máy được cập nhật từ nhiều nguồn dữ liệu khác nhau mà không yêu cầu chia sẻ dữ liệu huấn luyện trực tiếp. Thay vì đưa dữ liệu về một trung tâm huấn luyện tập trung, học liên kết cho phép việc huấn luyện mô hình diễn ra trên các thiết bị cục bộ hoặc các nút trong một mạng phân tán. Trong học liên kết, mô hình học máy ban đầu được tạo và phân phối đến các thiết bị hoặc nút trong mạng lưới các thực thể tham gia vào quá trình cộng tác huấn luyện. Sau đó, các mô hình này được huấn luyện bằng cách sử dụng dữ liệu cục bộ trên từng thiết bị mà không gửi dữ liệu quay lại trung tâm huấn luyện tập trung. Thay vào đó, chỉ có các thông số mô hình được gửi qua mạng để được tổng hợp và cập nhật. Quá trình tổng hợp thông tin và cập nhật mô hình trong học liên kết thường được thực hiện bằng cách sử dụng các thuật toán tối ưu hóa phân tán. Mô hình được cập nhật dựa trên sự kết hợp thông tin từ các mô hình cục bộ mà không tiết lộ dữ liệu cụ thể hoặc thông tin riêng tư từng cá nhân hoặc tổ chức tham gia.



Hình 2.2: *Hiệu năng giảm sút theo thời gian của các hệ thống bảo mật dựa trên phương pháp học máy*

Học liên kết mang lại một số lợi ích quan trọng. Đầu tiên, nó giảm thiểu việc chia sẻ dữ liệu nhạy cảm bằng cách giữ cho dữ liệu tại vị trí của nó, bảo vệ quyền riêng tư của người dùng. Thứ hai, học liên kết giúp giảm bớt băng thông mạng và thời gian truyền thông tin bằng cách chỉ gửi các thông số mô hình chứ không phải dữ liệu đầy đủ. Cuối cùng, học liên kết cho phép huấn luyện mô hình trên các thiết bị cục bộ, điều này có lợi khi dữ liệu không khả dụng hoặc có tính cục bộ cao, như trong các ứng dụng y tế hoặc IoT (Internet of Things).

2.5. Mã hóa đồng cấu

Mã hóa đồng cấu là một loại đặc biệt của mã hóa, có khả năng thực thi các phép toán trên dữ liệu đã mã hóa và cho ra kết quả giống như khi thực hiện phép toán trên dữ liệu ban đầu. Kết quả đầu ra là kết quả tính toán đã được mã hóa. Hiện nay có ba loại lược đồ mã hóa đồng cấu (HE scheme) khác nhau dựa vào các phép tính toán và số lượng các phép tính toán có thể thực hiện:

- Partially Homomorphic Encryption (PHE): chỉ hỗ trợ những phép tính của một loại (phép nhân hoặc phép cộng) với số lần hạn chế, như RSA và Paillier.
- Somewhat Homomorphic Encryption (SWHE): hỗ trợ các phép tính hạn chế (ví dụ: cộng hoặc nhân) lên đến một độ phức tạp nhất định, nhưng các phép tính này chỉ có thể được thực hiện một số lần nhất định.
- Fully Homomorphic Encryption (FHE): hỗ trợ bất cứ phép tính nào với số lần không hạn chế. Có 3 lược đồ FHE đã được phát triển và sử dụng nhiều nhất, đó là: BGV, BFV và CKKS. Trong đó, lược đồ CKKS là lược đồ thích hợp nhất cho các ứng dụng học máy bởi nó hỗ trợ phép cộng và nhân trên các số thực đã được mã hóa và cho ra các kết quả gần đúng.

2.6. Riêng tư vi phân

Riêng tư vi phân (Differential Privacy) là một phương pháp cung cấp một vài đảm bảo mang tính toán học về tính riêng tư của thông tin người dùng. Mục đích chính là giảm thiểu ảnh hưởng của bất cứ dữ liệu đơn lẻ nào đến kết quả tổng thể. Điều này có nghĩa là một người sẽ cho ra cùng suy luận về một dữ liệu đơn lẻ dù cho nó có hoặc không có mặt trong đầu vào của việc phân tích (xem **Hình 2.3**). Khi số lượng các phân tích trên dữ liệu gia tăng thì rủi ro tiết lộ thông tin người dùng càng lớn. Kết quả của việc thực hiện các phép tính toán có sử dụng DP miễn nhiễm với một số lượng lớn các cuộc tấn công về quyền riêng tư như: suy diễn lớp đại diện, suy diễn thành viên trong tập huấn luyện, suy diễn thuộc tính, suy diễn nhãn dữ liệu của tập huấn luyện.

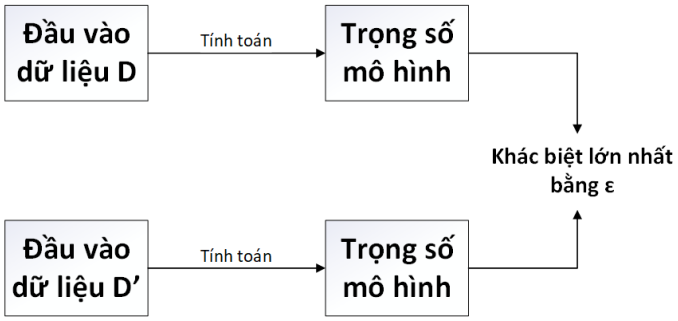
2.7. Chuỗi khối

Chuỗi khối (blockchain) là một công nghệ lưu trữ và truyền thông tin phi tập trung, xây dựng trên cơ sở dữ liệu phân tán. Đặc trưng của blockchain là sự kết hợp giữa tính an toàn, minh bạch và tính không thể thay đổi. Các thông tin được lưu trữ trong các khối dữ liệu liên kết với nhau thông qua mã xác nhận duy nhất và mối liên kết mã hóa, tạo nên một chuỗi có thứ tự và không thể sửa đổi. Hệ thống này loại bỏ sự phụ thuộc vào bên trung gian, ngăn chặn gian lận và tạo ra một cơ sở dữ liệu an toàn, minh bạch cho việc lưu trữ và truyền tải thông tin. Blockchain không chỉ được ứng dụng trong lĩnh vực tài chính mà còn mở ra nhiều tiềm năng trong các ngành công nghiệp khác, từ y tế đến quản lý chuỗi cung ứng và nhiều lĩnh vực khác.

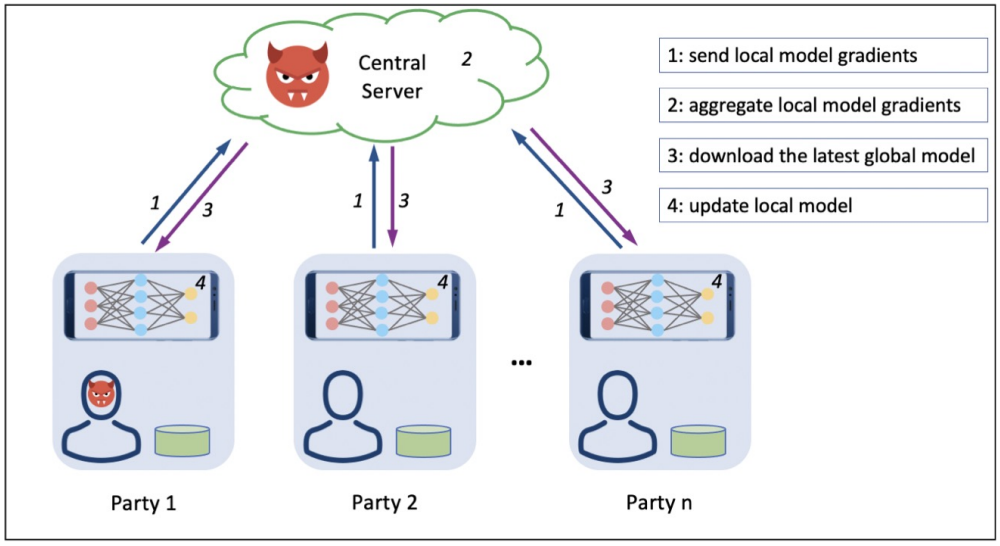
2.8. Tấn công đầu độc trong mô hình học liên kết

2.8.1. Tổng quan

Tấn công đầu độc (poisoning attack) là một kỹ thuật tấn công trong đó kẻ tấn công cố gắng thay đổi dữ liệu huấn luyện để làm sai lệch mô hình học máy. Khi mô hình được huấn luyện trên dữ liệu đã bị đầu độc, nó có thể cho ra kết quả dự đoán sai lệch với thiết kế ban đầu hoặc thực hiện hành động không mong muốn.



Hình 2.3: *Tổng quan ý tưởng kỹ thuật Differential Privacy (DP)*



Hình 2.4: *Tấn công đầu độc trong học liên kết*

Trong FL, tấn công đầu độc có thể được thực hiện bởi các bên tham gia cộng tác bằng cách thêm các dữ liệu độc hại vào tập dữ liệu huấn luyện hoặc sửa đổi các điểm dữ liệu đã có trong tập huấn luyện. Khi mô hình học máy được huấn luyện trên tập dữ liệu này, nó sẽ học cách phân loại dữ liệu độc hại như là một lớp dữ liệu bình thường, hoặc bị mất đi khả năng đưa ra dự đoán chính xác một mẫu dữ liệu bất kì. Tấn công đầu độc có thể gây ra những hậu quả nghiêm trọng cho hệ thống học máy, bao gồm việc giảm độ chính xác và độ tin cậy của mô hình, thậm chí có thể làm cho mô hình hoàn toàn vô dụng. Do đó, việc bảo vệ các hệ thống học máy trước tấn công đầu độc là rất quan trọng để đảm bảo an toàn và đáng tin cậy cho các ứng dụng học máy.

Để bảo vệ mô hình FL khỏi các cuộc tấn công đầu độc, cần áp dụng các biện pháp bảo mật, chẳng hạn như kiểm tra tính xác thực của người dùng, và xác thực mô hình cục bộ trước khi được sử dụng trong quá trình học. Ngoài ra, các phương pháp phát hiện và xử lý các điểm dữ liệu bất thường cũng có thể được áp dụng để phát hiện và ngăn chặn các tấn công đầu độc.

2.8.2. Các loại tấn công đầu độc (poisoning attack)

Trong học máy, có một số loại tấn công đầu độc (poisoning attacks) mà những kẻ tấn công có thể thực hiện để làm sai lệch quá trình huấn luyện và gây hại cho mô hình chung. Và dưới đây là một số loại tấn công đầu độc phổ biến trong học liên kết:

- Đầu độc dữ liệu (Data poisoning): Đây là loại tấn công phổ biến nhất trong machine learning. Kẻ tấn công sẽ chèn các dữ liệu sai lệch hoặc giả mạo vào tập huấn luyện để khiến mô hình học sai. Khi được huấn luyện trên các dữ liệu này, mô hình có thể trở nên không chính xác hoặc bị đánh lừa.
  - Làm nhiễu nhãn (Label poisoning): Loại tấn công này liên quan đến việc thay đổi nhãn

của các dữ liệu trong tập huấn luyện để khiến mô hình học sai. Kẻ tấn công có thể thay đổi nhãn của các dữ liệu để khiến mô hình học sai lầm.

  - Làm nhiễu đặc trưng (Feature poisoning): Loại tấn công này liên quan đến việc thay đổi các đặc trưng của các dữ liệu trong tập huấn luyện. Kẻ tấn công có thể thay đổi các đặc trưng để khiến mô hình học sai lầm hoặc bị đánh lừa.

- Đầu độc mô hình (Model poisoning): Loại tấn công này liên quan đến việc thay đổi mô hình học máy bằng cách chèn các mô hình giả mạo vào mô hình gốc. Một số cách thức phổ biến của loại tấn công này là đánh sập mô hình, phá vỡ tính toàn vẹn của mô hình hoặc thực hiện các thay đổi trên mô hình gốc để khiến nó trở nên không chính xác.

Ngoài ra, nếu xét trên mục tiêu của tấn công đầu độc, nó có thể được chia thành 2 loại chính, bao gồm Untargeted Attack và Targeted Attack. Sự khác biệt giữa chúng chủ yếu nằm ở mục tiêu và hậu quả cụ thể mà kẻ tấn công muốn gây ra. Chi tiết như sau:

- Untargeted attack (tấn công không nhắm mục tiêu cụ thể) là loại tấn công mà mục tiêu chính là làm giảm chất lượng tổng thể của mô hình học máy, không nhắm vào việc làm sai lệch mô hình trong việc dự đoán một lớp cụ thể hoặc một mẫu dữ liệu cụ thể. Kẻ tấn công muốn làm cho mô hình trở nên kém hiệu quả hoặc kém chính xác nhưng không tập trung vào mục tiêu cụ thể nào. Ví Dụ: Một kẻ tấn công thêm một lượng lớn dữ liệu nhiễu loạn vào quá trình học để làm giảm độ chính xác tổng thể của mô hình trên một loạt các tác vụ, mà không cố gắng gây ảnh hưởng đến một tác vụ hoặc lớp dữ liệu cụ thể nào.

- Targeted attack (tấn công nhắm mục tiêu cụ thể) là loại tấn công mà kẻ tấn công có một mục tiêu cụ thể, thường là làm cho mô hình học máy phát sinh ra dự đoán sai lệch cho một số mẫu dữ liệu cụ thể hoặc một lớp dữ liệu cụ thể. Mục tiêu của tấn công có thể là làm cho mô hình nhận diện nhầm một loại đối tượng là một loại khác hoặc kích hoạt một hành vi không mong muốn khi nhận dạng một mẫu dữ liệu nhất định. Ví Dụ: Trong một tấn công nhắm vào mô hình phân loại hình ảnh, kẻ tấn công có thể muốn mô hình luôn nhận diện hình ảnh chứa một biểu tượng nhất định (ví dụ, logo của một công ty) là thuộc một lớp cụ thể, bất kể nội dung thực sự của hình ảnh đó.

2.9. Biểu diễn lớp áp chót (Penultimate Layer Representation - PLR)

Penultimate layer representation (PLR) là biểu diễn lớp áp chót của một mạng nơ-ron. Nói cách khác, penultimate layer là lớp thứ hai tính từ lớp cuối cùng, và output là lớp cuối cùng. Lớp này có vai trò quan trọng trong việc trích xuất đặc trưng quan trọng của dữ liệu huấn luyện dưới dạng véc-tơ số. Ví dụ, trong mạng nơ-ron để phân loại ảnh, PLR có thể là một véc-tơ mô tả đặc

trưng của hình ảnh bao gồm màu sắc, hình dạng, đường viền, ... Khi đưa vào một mô hình máy học, PLR cung cấp cho mô hình thông tin quan trọng về dữ liệu để giúp mô hình học cách phân loại chính xác các đối tượng.

2.10. Thuật toán Centered Kernel Alignment (CKA)

Thuật toán Centered Kernel Alignment (CKA) được thiết kế để đo độ tương đồng giữa các biểu diễn bằng cách làm phẳng các ma trận nhân (kernel) tương ứng. CKA được sử dụng trong nhiều lĩnh vực khác nhau của máy học và trí tuệ nhân tạo, bao gồm các nghiên cứu tính toán về sự giống nhau của các mô hình học sâu, đánh giá sự tương đồng giữa các bộ lọc hình ảnh, và tìm kiếm các đặc trưng chung trong các tác vụ học khác nhau.

2.11. Bộ tự mã hóa (Autoencoder)

Bộ tự mã hóa (Autoencoder) là một mô hình học máy không giám sát (Unsupervised Machine Learning) được sử dụng để tự động học cách biểu diễn dữ liệu. Autoencoder có khả năng giảm chiều dữ liệu và khám phá các đặc trưng quan trọng từ dữ liệu ban đầu. Quá trình huấn luyện autoencoder nhằm tối thiểu hóa sai lệch giữa dữ liệu tái tạo và dữ liệu ban đầu. Mô hình cố gắng học cách biểu diễn dữ liệu quan trọng nhất trong quá trình nén và giải mã, từ đó tạo ra một biểu diễn nén của dữ liệu, hay còn gọi là latent space của dữ liệu. Autoencoder có nhiều ứng dụng trong lĩnh vực xử lý dữ liệu, nhưng phổ biến nhất là trong việc giảm chiều dữ liệu và trích xuất đặc trưng. Autoencoder cũng có thể được sử dụng trong các tác vụ như nén dữ liệu, phát hiện bất thường và tái tạo ảnh.

2.12. Cơ chế kiểm soát truy cập giữa bộ điều khiển và ứng dụng mạng

Các ứng dụng trong mạng khả lập trình sử dụng giao diện Northbound để kết nối với bộ điều khiển để thực hiện kết nối và sử dụng các thông tin cần thiết từ hệ thống mạng phục vụ mục đích được thiết kế của chính ứng dụng. Các ứng dụng này có thể thuộc về một bộ điều khiển hay từ bên ngoài. Tuy vậy, các ứng dụng mạng trong SDN vốn được phát triển để điều phối và triển khai các chức năng quản trị mạng, hay chức năng bất kì thông qua bộ điều khiển SDN có thể bị lạm dụng từ những tác nhân độc hại, đặc biệt là từ các ứng dụng từ các miền quản lý khác. Hiện tại, các ứng dụng như tường lửa (firewall), hệ thống phát hiện xâm nhập (IDS), cân bằng tải (load balancing),... có thể chạy đồng thời với bộ điều khiển để cài đặt các chính sách an ninh trong hệ thống mạng mà nó tham gia quản lý. Vấn đề thiếu cơ chế xác thực, quản lý các kết nối từ các ứng dụng mạng tới bộ điều khiển có thể gây ra sự mất an toàn cho cả hệ thống mạng. Khi đó, các hệ thống quản lý và kiểm tra xác thực, kiểm soát truy cập được đề xuất ở giao diện Northbound

trên bộ điều khiển SDN để khắc phục điểm yếu này.

CHƯƠNG 3. Cơ chế học liên kết an toàn cho ứng dụng săn tìm mối đe dọa và phát hiện xâm nhập trong mạng khả lập trình

3.1. Dẫn nhập

Dù việc sử dụng học liên kết mang lại lợi ích về giảm rủi ro về việc xâm phạm quyền riêng tư dữ liệu và tối ưu hiệu năng trên các máy chủ huấn luyện, thì mô hình này cũng đối mặt với những mối lo ngại về an toàn và bảo mật thông tin trong quá trình truyền tải và trao đổi mô hình huấn luyện giữa máy khách huấn luyện và máy chủ tổng hợp. Do đó, nghiên cứu này tập trung vào việc giới thiệu một phương pháp đảm bảo tính bảo mật và quyền riêng tư trong giai đoạn trao đổi mô hình huấn luyện giữa máy khách huấn luyện và máy chủ tổng hợp trung tâm trong ngữ cảnh các tổ chức tham gia cộng tác huấn luyện mô hình chung cho ứng dụng IDS. Hơn nữa, phương pháp cũng hướng đến sự minh bạch, khả năng kiểm tra trong việc hợp tác và quản lý đóng góp thông qua một nền tảng phi tập trung. Điều này đặc biệt quan trọng trong bối cảnh học liên kết, khi các tổ chức phải chia sẻ thông tin nhạy cảm như mô hình huấn luyện và gradient để thực hiện việc hợp tác mà không gây lộ thông tin quan trọng.

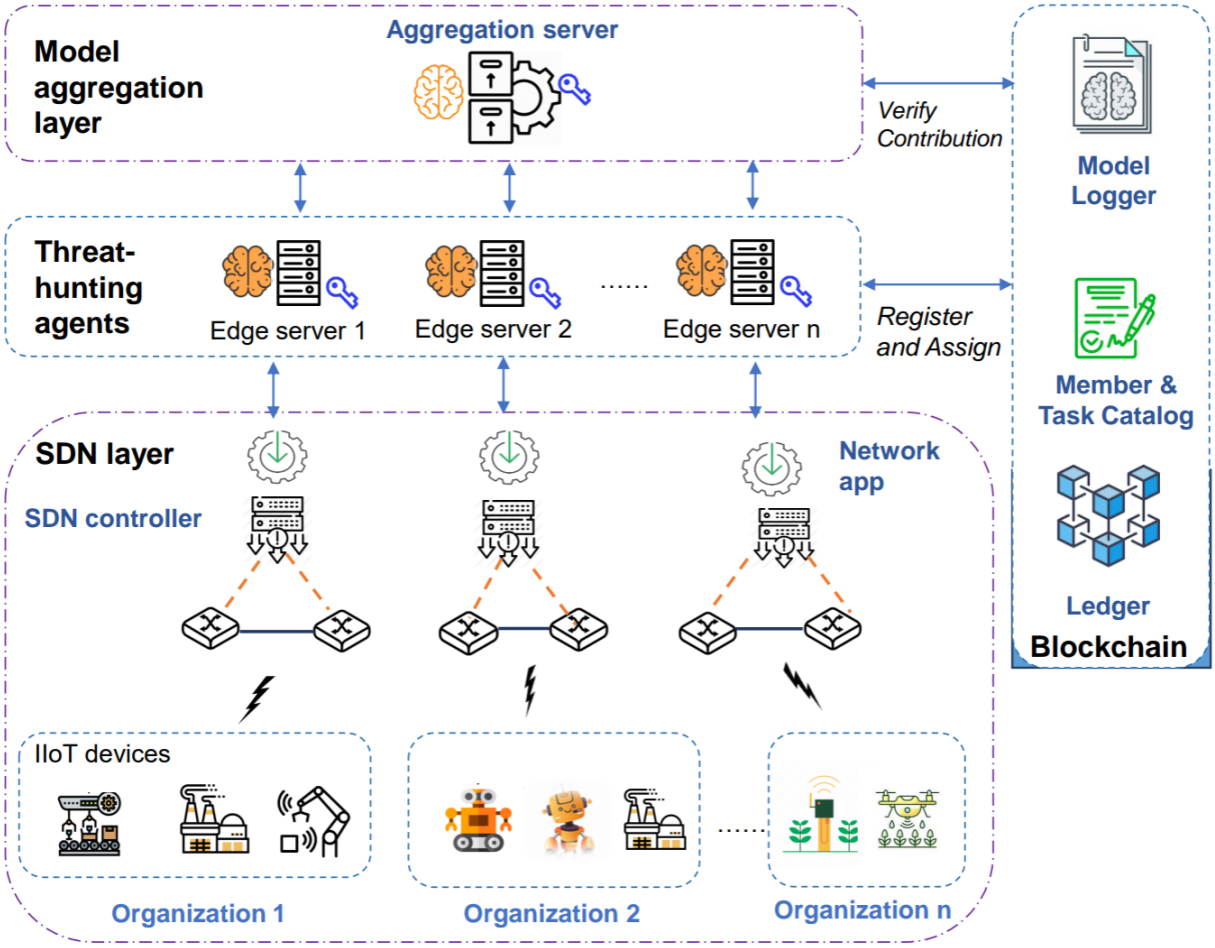
3.2. Kiến trúc tổng quan bộ khung FedChain-Hunter đề xuất

Để khuyến khích sự sẵn sàng và thúc đẩy sự cộng tác của việc chia sẻ thông tin giữa các chủ sở hữu dữ liệu, mô hình học liên kết được xem là một cơ chế huấn luyện các mô hình học máy mà vẫn đảm bảo quyền riêng tư dữ liệu. Trong luận án này, mô hình FedChain-Hunter được đề xuất để giải quyết các thách thức nêu trên trong việc xây dựng cơ chế cộng tác cho hệ thống, ứng dụng cảnh báo xâm nhập và săn tìm mối đe dọa. **Hình 3.1** biểu diễn kiến trúc chính của bộ khung FedChain-Hunter.

Bộ khung FedChain-Hunter bao gồm nhiều hệ thống mạng SDN thành phần (thuộc một hoặc nhiều tổ chức khác nhau) tận dụng cơ chế khả lập trình như một mô hình quản lý mạng để tạo điều kiện thuận lợi cho việc quan sát các sự kiện mạng bảo mật một cách linh hoạt. Mỗi mạng thành phần trong số đó được trang bị một tác nhân săn tìm mối đe dọa (threat hunting agent) đóng vai trò là cổng bảo mật (security gateway) trong mạng. Nó có thể là trình phát hiện phần mềm độc hại, trình phát hiện tấn công mạng, v.v. Để thực hiện điều đó, bộ điều khiển SDN thiết lập các quy tắc điều phối mạng cho các thiết bị chuyển mạch (switch) hỗ trợ giao thức OpenFlow để đưa ra các quyết định chuyển tiếp dữ liệu. Chính sách bảo mật này dễ dàng thay đổi và cập nhật theo yêu cầu của người quản trị mạng thông qua bộ điều khiển SDN. Các qui luật luồng

được thiết lập này chịu trách nhiệm chuyển hướng lưu lượng mạng đến một ứng dụng tìm kiếm mối đe dọa (threat hunting engine) để kiểm tra sự tồn tại của các hành động độc hại. Với sự hỗ trợ của nguyên tắc khả lập trình trong SDN, nó có thể hỗ trợ lọc và thu thập dữ liệu thô để xây dựng bộ dữ liệu huấn luyện cho mỗi mạng thành phần. Ngoài ra, các ứng dụng phát hiện xâm nhập, tìm kiếm mối đe dọa được triển khai bằng cách truy xuất mô hình toàn cục (global model) để giúp nâng cao khả năng phát hiện tấn công với tri thức mới được chia sẻ bởi các tổ chức khác nhau. Chú ý rằng, trong mỗi mạng, tập dữ liệu huấn luyện cho từng mô hình cục bộ (local model) được thu thập và tự gắn nhãn. Sau đó, dữ liệu được gắn nhãn của mỗi bên tham gia được sử dụng để huấn luyện cục bộ mô hình IDS trên máy chủ biên mà không để lộ dữ liệu cho các tổ chức và hệ thống mạng thành phần/ phân đoạn mạng khác.

Bên cạnh đó, nền tảng phi tập trung dựa trên chuỗi khối được sử dụng để quản lý những bên tham gia và những đóng góp của họ liên quan đến qui trình huấn luyện FL. Tất cả các hoạt động cho dịch vụ liên kết tìm kiếm mối đe dọa hay phát hiện xâm nhập được ghi lại trong chuỗi khối dưới dạng giao dịch để tổng hợp và đánh giá chất lượng. Chuỗi khối giúp đảm bảo tính minh bạch, bảo mật, tính bất biến và khả năng xác thực tính trung thực trong các đóng góp, kiểm toán trong việc thanh toán phần thưởng và phân phối lợi nhuận sau khi hoàn thành các nhiệm vụ FL.



Hình 3.1: *Kiến trúc tổng quan của bộ khung FedChain-Hunter.*

3.3. Các thành phần chính của bộ khung đề xuất

3.3.1. Mạng khả lập trình và các máy khách huấn luyện cục bộ

Thu thập và chuẩn hóa dữ liệu Ban đầu, mỗi tổ chức có ý định tham gia FedChain-Hunter cần chuẩn bị dữ liệu huấn luyện và tài nguyên phần cứng. Tại mạng của mỗi tổ chức/phân đoạn mạng tham gia, có một máy chủ Thu thập lưu lượng (Traffic Acquisition Server) chịu trách nhiệm thu thập dữ liệu lưu lượng mạng từ các thiết bị mạng được quan sát. Điều này được thực hiện bằng cách sử dụng cổng giám sát (mirroring port) trên thiết bị chuyển mạch hỗ trợ giao thức OpenFlow. Chính sách quan sát để thu thập lưu lượng mạng có thể dễ dàng thay đổi tương ứng với nhiều yêu cầu linh hoạt (loại dịch vụ, giao thức, thiết bị, v.v.) theo hướng dẫn của ứng dụng OpenFlow từ giao diện Northbound của bộ điều khiển SDN đến các thiết bị chuyển mạch hỗ trợ OpenFlow.

Máy khách huấn luyện cục bộ

Ngoài nhiệm vụ thu thập lưu lượng mạng, các máy chủ biên (edge server) chịu trách nhiệm huấn luyện các mô hình ML cục bộ mà không cần trao đổi dữ liệu nhạy cảm. Để bắt đầu, mô hình ML được gửi từ máy chủ tổng hợp đến từng máy chủ biên trong mọi mạng để bắt đầu giai đoạn huấn luyện cộng tác. Lưu ý rằng, mỗi tổ chức có thể có nhiều máy chủ biên để tham gia vào quá trình huấn luyện mô hình chung. Tuy nhiên, để đơn giản, nghiên cứu này giả định rằng mỗi tổ chức/ phân đoạn mạng chỉ có một máy chủ biên tham gia dưới vai trò của máy khách huấn luyện cục bộ.

3.3.2. Máy chủ tổng hợp

Trong sơ đồ huấn luyện FL, máy chủ tổng hợp được thiết lập để thu thập và tập hợp các bản cập nhật mô hình từ các máy khách huấn luyện tại các bên tham gia. Tuy nhiên, các bản cập nhật của máy khách huấn luyện, ví dụ: máy chủ biên, có thể bị khai thác hoặc thay đổi ngoài ý muốn bởi một người tham gia giả mạo hoặc máy chủ tổng hợp bị xâm nhập trong lược đồ FL. Để đạt được mục tiêu này, nghiên cứu này áp dụng Mã hóa đồng cấu đầy đủ dựa trên lược đồ CKKS (CKKS-FHE) và Riêng tư vi phân (DP) ở phía máy chủ cục bộ để bảo vệ quyền riêng tư dữ liệu cho các mô hình tốt hơn. Sẽ không an toàn nếu chỉ DP được sử dụng trong FL, vì chức năng của DP là ngăn chặn rò rỉ quyền riêng tư từ các cuộc tấn công suy luận thành viên hoặc suy luận gradient của mô hình huấn luyện cục bộ. Chi tiết về kỹ thuật CKKS-FHE và DP cho bộ khung FedChain-Hunter sẽ được trình bày cụ thể trong Mục 3.4.

3.3.3. Nền tảng quản lý sự tham gia và kiểm tra sự đóng góp dựa trên Blockchain

Để tham gia dịch vụ tìm kiếm mối đe dọa trong khuôn khổ FedChain-Hunter, mỗi máy chủ biên là đại diện của mạng cần phải đăng ký tư cách thành viên và ghi nhận các nhiệm vụ của họ cũng như mô hình ban đầu lên nền tảng chuỗi khối. Sau đó, mô hình và mô tả tác vụ được truyền đến máy chủ tổng hợp để gửi cài đặt ban đầu của mô hình đến các máy chủ biên của tổ chức. Khi quá trình tổng hợp mô hình kết thúc, tất cả các hành vi của các máy tham gia huấn luyện bao gồm các cập nhật mô hình của chúng được ghi lại dưới dạng đóng góp cho nhiệm vụ này. Nó có thể được sử dụng để xây dựng một cơ chế khuyến khích đóng góp cho mô hình học máy nhằm khuyến khích nhiều tổ chức tham gia vào giai đoạn huấn luyện của mô hình liên kết săn lùng mối đe dọa.

Để cung cấp hiệu suất tốt hơn về quyền riêng tư, khả năng mở rộng, thông lượng và thời gian trễ để xác minh các giao dịch được lưu giữ trong sổ cái phân tán, một chuỗi khối riêng với các thành viên là các tổ chức/doanh nghiệp, được gọi là chuỗi khối liên hiệp (consortium blockchain), được chọn để triển khai nền tảng quản lý phi tập trung trong bộ khung FedChain-Hunter.

3.4. Cơ chế tổng hợp an toàn và đảm bảo quyền riêng tư trong mô hình huấn luyện học liên kết cho ứng dụng phát hiện và săn tìm mối đe dọa

Phần này trình bày cơ chế mã hóa đồng cấu đầy đủ (Fully Homomorphic Encryption - FHE) và Riêng tư vi phân (Differential Privacy - DP) được sử dụng cho các mô hình cục bộ ở các máy chủ huấn luyện trong mô hình được đề xuất. Bộ khung FedChain-Hunter sử dụng hai cơ chế này cho qui trình huấn luyện của FL để cung cấp khả năng tổng hợp mô hình cục bộ an toàn và đảm bảo quyền riêng tư tại máy chủ tổng hợp trung tâm. Cụ thể, mỗi máy khách huấn luyện sẽ thêm nhiễu vào mô hình được huấn luyện ở phía cục bộ, sau đó mã hóa chúng trước khi gửi đến máy chủ tổng hợp.

3.5. Thực nghiệm và đánh giá hiệu năng nền tảng Blockchain

Hiệu suất về thời gian của nền tảng blockchain trong thực hiện các giao dịch cho quy trình huấn luyện học liên kết được phân tích đánh giá. Các giá trị này được thực nghiệm và lấy giá trị trung bình trong 20 lần.

Có thể thấy rằng tất cả các loại giao dịch đều có xu hướng tăng nhẹ về thời gian xử lý giao dịch trung bình khi số lượng nút máy khách tham gia quá trình huấn luyện mô hình tăng lên. Tuy nhiên, sự khác biệt là không đáng kể, đặc biệt là trong trường hợp 50 và 100 nút. Ngoài ra, việc

sử dụng 20 nút khác nhau để gửi các giao dịch là kịch bản duy nhất đạt được sự ổn định trong thời gian xử lý bất kể loại giao dịch hoặc số lượng giao dịch được yêu cầu từ các máy khách huấn luyện. Những kết quả hứa hẹn này chứng minh rằng mạng blockchain được triển khai trên cấu hình tối thiểu có thể đáp ứng các yêu cầu về hiệu suất của các yêu cầu FedChain-Hunter.

3.6. Thực nghiệm và đánh giá hiệu năng khung liên kết săn tìm mối đe dọa

Các thử nghiệm của nghiên cứu này về mô hình học liên kết được thực hiện trên các bộ dữ liệu khác nhau bao gồm Active Wiretap, ARP MitM, InSecLab-IDS-2021, InSDN và CICIDS-2017 một cách độc lập để so sánh và đánh giá cụ thể. Trong mỗi thử nghiệm, một tập dữ liệu được chia ngẫu nhiên cho 11 máy khách tham gia và huấn luyện trong 10 vòng liên tiếp.

3.6.1. Đánh giá hiệu năng của Differential Privacy trong FedChain-Hunter

Hai mô hình CNNGRU và LSTM được đánh giá trong các bối cảnh khác nhau, bao gồm cả mô hình được huấn luyện trong điều kiện có và không có DP. Trong đó, các kịch bản DP được thực hiện bằng cách lần lượt thêm nhiễu bằng cách sử dụng Phân phối chuẩn và Phân phối Gamma. Khi huấn luyện các mô hình tìm kiếm mối đe dọa có sử dụng DP, nghiên cứu này sử dụng hệ số nhiễu z để xác định lượng nhiễu được thêm vào trong quá trình huấn luyện. Hệ số nhiễu là tỷ lệ của độ lệch chuẩn so với định mức cắt (clipping norm). Lưu ý rằng, điều đó có nghĩa là nhiều nhiễu hơn dẫn đến sự riêng tư tốt hơn và hiệu năng của mô hình thấp hơn.

Theo kết quả thử nghiệm, việc sử dụng Phân phối Gamma và Phân phối chuẩn cho dịch vụ liên kết tìm kiếm mối đe dọa hỗ trợ DP đều duy trì ngưỡng Độ chính xác và điểm F1 có thể chấp nhận được sau gần 6 vòng huấn luyện. Nó chỉ ra rằng, với số vòng huấn luyện hạn chế, yêu cầu bảo vệ quyền riêng tư vẫn có tác động không đáng kể đến hiệu suất của các mô hình săn tìm mối đe dọa dựa trên FL bất kể việc áp dụng Cơ chế Gamma hay Cơ chế Gausian.

3.6.2. Đánh giá hiệu năng của lược đồ mã hóa đồng cấu CKKS so với lược đồ Paillier trong FedChain-Hunter

Để đánh giá thực nghiệm hiệu suất của HE trên sơ đồ săn tìm mối đe dọa liên kết, nghiên cứu này đo kích thước của cả hai mô hình CNNGRU và LSTM trước và sau khi mã hóa.

Hơn nữa, nhằm chỉ ra hiệu năng vượt trội của lược đồ CKKS trong FedChain-Hunter, chúng tôi cũng đã thực nghiệm kịch bản so sánh phương pháp mã hoá với lược đồ Paillier trong cùng điều kiện huấn luyện và kiểm thử với lược đồ CKKS. Bên cạnh hiệu năng của mô hình FedChain-Hunter khi sử dụng mã hoá đồng cấu mang lại, nghiên cứu này cũng so sánh kích thước dữ liệu

được truyền tải trong suốt quá trình huấn luyện giữa mô hình huấn luyện tập trung truyền thống (Centralized Learning) so với mô hình FedChain-Hunter sử dụng HE.

Kích thước truyền tải này hoàn toàn phụ thuộc vào kích thước của mô hình mà không phụ thuộc vào tập dữ liệu sử dụng. Trong các trường hợp FL áp dụng HE, kích thước dữ liệu truyền tải của bộ khung FedChain-Hunter sử dụng CKKS-HE và Paillier đều cho thấy sự gia tăng kích thước mô hình khi so với giải pháp FL không áp dụng cơ chế mã hóa mô hình. Cụ thể, bộ khung FedChain-Hunter sử dụng CKKS-HE và Paillier đã làm gia tăng đáng kể kích thước dữ liệu truyền tải khi huấn luyện với 11 máy khách lên gấp hơn 7.000 (bảy nghìn) lần so với giải pháp FL không áp dụng cơ chế mã hóa. Trong đó cơ chế mã hóa CKKS-HE tạo ra mô hình truyền tải sau khi áp dụng mã hóa có kích thước gấp 1,028 lần so với giải pháp mã hóa Paillier. Tuy nhiên, trong thực tế, kích thước dữ liệu huấn luyện các mô hình luôn được cập nhật liên tục nên kích thước truyền tải dữ liệu giữa máy chủ và máy khách của bộ khung FedChain-Hunter khi sử dụng CKKS-HE là không đáng kể so với lượng dữ liệu huấn luyện. Nó cũng không làm thay đổi lượng dữ liệu truyền tải giữa các máy khách và máy chủ khi kích thước tập dữ liệu huấn luyện gia tăng.

3.7. Thảo luận

Chương này đã trình bày phương pháp học liên kết để xây dựng các trình phát hiện xâm nhập, săn tìm mối đe dọa trong ngữ cảnh mạng khả lập trình. Cụ thể, nghiên cứu này đề xuất bộ khung FedChain-Hunter, một lược đồ bảo vệ quyền riêng tư và an toàn với khả năng kiểm tra tin cậy đối với các dịch vụ liên kết tìm kiếm mối đe dọa giữa nhiều tổ chức trong ngữ cảnh mạng SDN. Trong đó, mô hình liên kết săn lùng mối đe dọa được huấn luyện cục bộ với bộ dữ liệu riêng ngay trên các máy chủ biên của mỗi tổ chức. Riêng tư vi phân là kỹ thuật bổ sung nhiễu dựa trên phân phối Gaussian và Laplacian được đưa vào các tham số của mô hình được huấn luyện tại máy khách trước khi gửi đến máy chủ tổng hợp để tính toán mô hình toàn cục từ nhiều máy khách. Ngoài ra, Mã hóa đồng cấu đầy đủ (FHE) với sơ đồ CKKS được sử dụng để mã hóa tất cả các trọng số mô hình từ máy khách để đảm bảo tập hợp bảo vệ quyền riêng tư bền vững hơn. Hơn nữa, công nghệ chuỗi khối (blockchain) cũng được áp dụng trong bộ khung FedChain-Hunter để thiết lập một nền tảng phi tập trung nhằm quản lý danh tính và đóng góp từ các máy khách hợp tác trong quá trình học tập liên kết. Kết quả thử nghiệm trên 5 bộ dữ liệu chỉ ra rằng khung FedChain-Hunter có thể đạt được hiệu suất cao trong việc phát hiện mối đe dọa và bảo vệ quyền riêng tư, đồng thời khuyến khích các tổ chức khác nhau tham gia vào quá trình huấn luyện trên bộ dữ liệu nhạy cảm của họ với khả năng quản lý đóng góp tin cậy.

Tuy nhiên, do có sự tham gia của nhiều bên với nhiều hệ thống mạng khác nhau, nguy cơ về sự phá hoại mô hình chung đến từ những nhân tố tham gia vào qui trình học liên kết là một trong những thách thức tiếp theo cần xem xét và giải quyết. Các bên có ý đồ xấu có thể thực hiện tấn

công đầu độc mô hình toàn cục bằng cách gửi các bản cập nhật mô hình sai lệch từ các máy khách huấn luyện cục bộ nhằm điều chỉnh hay triệt tiêu khả năng hội tụ của mô hình tổng hợp. Do vậy, trong chương tiếp theo của luận án, chúng tôi trình bày phương pháp phát hiện và loại bỏ các mô hình độc hại được gửi lên từ máy khách huấn luyện để bảo vệ qui trình huấn luyện mô hình liên kết phát hiện xâm nhập khỏi các kẻ tấn công ác ý, phá hoại hiệu năng chung của mô hình.

CHƯƠNG 4. Cơ chế ngăn chặn tấn công đầu độc cho ứng dụng liên kết phát hiện xâm nhập trong môi trường phân tán

4.1. Dẫn nhập

Các hệ thống dựa trên FL vẫn dễ bị tấn công đầu độc, nguyên nhân xuất phát từ các bên nội bộ có thể làm giảm đáng kể hiệu suất mô hình. Cụ thể hơn, những cuộc tấn công này có thể được thực hiện thông qua các kỹ thuật khác nhau, chẳng hạn như đầu độc dữ liệu và đầu độc mô hình. Đầu độc dữ liệu xảy ra ở cấp độ dữ liệu, trong đó các máy khách độc hại đã đào tạo mô hình cục bộ của họ trên dữ liệu bị can thiệp một cách cố ý và sau đó gửi chúng đến máy chủ tổng hợp.

Do đó, các cách tiếp cận giải quyết rủi ro tấn công đầu độc gây ra đã được đề xuất bao gồm nhiều cơ chế và kỹ thuật phòng thủ, chẳng hạn như làm sạch dữ liệu, phát hiện ngoại lệ, tổng hợp mô hình mạnh mẽ,... để duy trì tính ổn định của mô hình dựa trên FL khi gặp các cuộc tấn công gây nhiễu mô hình. Mặc dù việc phát hiện đầu độc thông qua kiểm tra chất lượng dữ liệu, SecFedNIDS có thể vi phạm mục tiêu bảo tồn quyền riêng tư của FL, các phương pháp sử dụng thuật toán phát hiện ngoại lệ tỏ ra không hiệu quả trong các ngữ cảnh dữ liệu phân phối không đồng nhất. Ngoài ra, gần như tất cả các cơ chế phòng thủ trước đó được thực hiện trên không gian tham số mô hình, điều này làm tăng chi phí tính toán lớn trong quá trình đào tạo FL.

Trong thời gian gần đây, biểu diễn không gian tiềm ẩn (Latent space representation - LSR) đã nhận được sự chú ý đáng kể từ các nhà nghiên cứu trong việc tạo ra các cơ chế phòng thủ mới. Cụ thể, các nghiên cứu FedCC, FLARE đã áp dụng Biểu diễn Lớp Áp chót (PLR) để tiết lộ đặc trưng quan trọng trong các mô hình cập nhật, trong đó trọng số độc hại đi theo hướng tương tự so với các mô hình bất thường. Tuy nhiên, FLARE cần một tập dữ liệu phụ trợ phía máy chủ để thu được các vector PLR, điều này khó thu thập vì nó phải tuân theo phân phối giống như các tập dữ liệu đào tạo cục bộ. Ngoài ra, phương pháp trong FedCC trực tiếp trích xuất PLR từ mỗi cập nhật cục bộ mà không cần một tập dữ liệu chung, dẫn đến sự không chắc chắn trong các vector PLR khi áp dụng vào các các ngữ cảnh dữ liệu không đồng nhất. Cuối cùng, cả hai công trình trên chỉ tập trung vào các kỹ thuật tấn công đầu độc mô hình.

Để khắc phục các vấn đề trên, trong nghiên cứu này, chúng tôi đề xuất một bộ khung phòng thủ dựa trên không gian ẩn chống lại cả cuộc tấn công đầu độc dữ liệu và đầu độc mô hình, được gọi là Fed-LSAE, trong ngữ cảnh của hệ thống NIDS dựa trên FL. Cụ thể hơn, Fed-LSAE sử dụng Bộ mã hóa tự động - Auto Encoder (AE) để học biểu diễn không gian ẩn (LSR) của mỗi mô

hình PLR, giảm thiểu sự không ổn định của PLR như đã đề cập trong FedCC. Sau đó, mức độ tương đồng của mỗi LSR cục bộ và LSR toàn cục được tính bằng thuật toán CKA (Center Kernel Alignment) trước khi gom cụm chúng (điểm số CKA) thành các nhóm lành tính / tấn công. Cụm không độc hại sau đó sẽ được chuyển đến máy chủ trung tâm để tổng hợp theo phương pháp tổng hợp FedAvg.

4.2. Phương pháp ngăn chặn tấn công đầu độc ứng dụng IDS liên kết thông qua chiến lược phân tích không gian tiềm ẩn

Phần này cung cấp tổng quan về mô hình nguy cơ, mô tả kiến thức và khả năng của những kẻ tấn công có ý định phá hoại hệ thống. Một thiết kế chi tiết của phương pháp phòng thủ của chúng tôi cũng được trình bày, với mục tiêu xây dựng một hệ thống ứng dụng IDS dựa trên FL mạnh và an toàn.

4.2.1. Mô hình mối đe dọa

Để đảm bảo tính khách quan, chúng tôi giả định rằng số lượng bên thù trong hệ thống FL, được ký hiệu là m, luôn ít hơn một nửa số lượng tổng cộng của các máy khách. Cụ thể, m được đặt là 20% và 40% tương ứng. Các máy khách tham gia còn lại, bao gồm cả máy chủ, được coi là các bên đáng tin cậy trong quá trình huấn luyện mô hình toàn cục. Trong khi đó, các nút do kẻ tấn công kiểm soát liên tục thực hiện các cuộc tấn công nhiễu trên hệ thống FL bằng cách cập nhật các mô hình cục bộ độc hại của họ vào mọi thời điểm.

4.2.1.1. Kiến thức và mục tiêu của kẻ tấn công

Là một đại diện trong qui trình hợp tác, kẻ tấn công có hiểu biết sâu về kiến trúc toàn cục bao gồm thuật toán học, dữ liệu huấn luyện, các siêu tham số mô hình như kích thước lô, tốc độ học tập, bộ tối ưu hóa, v.v. Do đó, kẻ tấn công sẽ thực hiện các kỹ thuật tấn công nhiễu của họ một cách trong suốt. Mục tiêu chính của họ là làm hỏng nghiêm trọng độ chính xác và hiệu suất của hệ thống NIDS dựa trên FL bằng cách gửi trọng số mô hình độc hại đến máy chủ tập trung toàn cục. Nói cách khác, chỉ các kỹ thuật tấn công nhiễu không đích đến (untargeted poisoning techniques) được tập trung trong nghiên cứu này.

4.2.1.2. Khả năng của kẻ tấn công

- **Cho phép.** Kẻ tấn công có đầy đủ quyền kiểm soát quy trình huấn luyện cục bộ với bộ dữ liệu của riêng họ. Họ cũng có thể thay đổi một số siêu tham số của mô hình được lấy từ máy chủ toàn cục để đạt được hiệu quả cao nhất của các cuộc tấn công nhiễu.

- **Không được phép.** Trong cài đặt huấn luyện, việc mỗi người tham gia tuân theo các thuật toán đã đồng ý và không can thiệp vào dữ liệu huấn luyện hoặc quy trình huấn luyện của người tham gia khác. Hơn nữa, kẻ tấn công không thể phá hỏng giai đoạn tổng hợp của máy chủ tập trung.

4.2.1.3. Chiến lược của kẻ tấn công

- **Tấn công lật nhãn (Label Flipping - LF).** Đây là một loại cuộc tấn công nhiễu dữ liệu trong đó kẻ tấn công thay đổi cố ý các nhãn đúng của một phần dữ liệu huấn luyện để đánh lừa mô hình học máy trong quá trình huấn luyện. Kết quả là mô hình học được sẽ phân loại các mẫu kiểm tra vào các loại không chính xác. Vì nghiên cứu của chúng ta dựa trên các nhiệm vụ phân loại nhị phân, trong đó coi nhãn 0 là benign (bình thường) và nhãn 1 là tấn công, kẻ tấn công sẽ đảo tất cả các mẫu benign thành các mẫu tấn công và ngược lại.

- **Untargeted-Med.** Đây là một cuộc tấn công nhiễu mô hình được đề xuất trong bài báo nhằm phá vỡ cơ chế tổng hợp Coordinate-wise Median. Kẻ tấn công điều chỉnh các tham số của mô hình bằng cách sử dụng giá trị lớn nhất và nhỏ nhất để hướng các giá trị trung bình theo từng phần tử vào một hướng ngược lại.

4.2.2. Các thành phần của Fed-LSAE

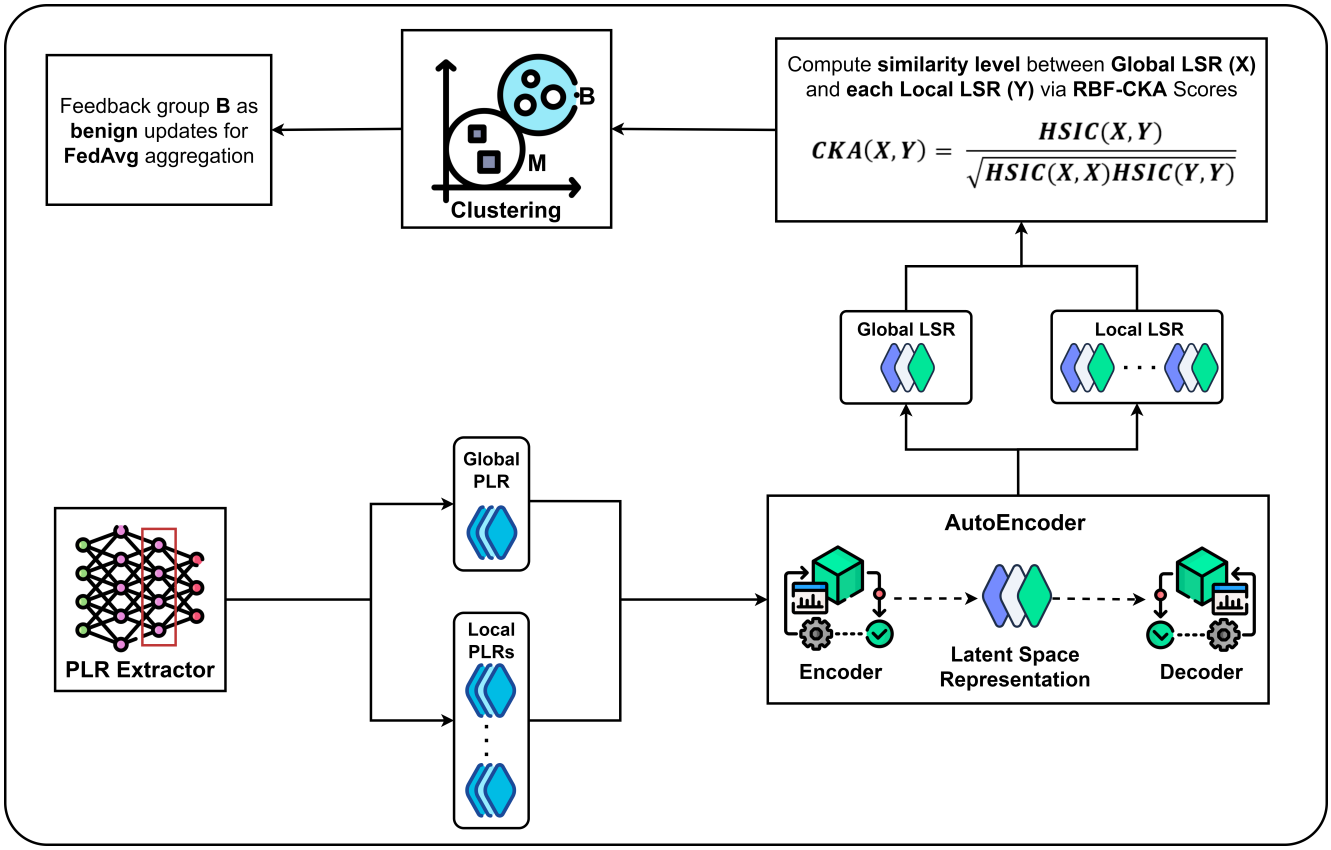
Dựa vào **Hình 4.1**, kiến trúc tổng quan của bộ khung liên kết phát hiện xâm nhập tích hợp cơ chế Fed-LSAE bao gồm các thành phần chính như sau:

Các tổ chức, thiết bị cộng tác. Họ là các đại diện cục bộ, huấn luyện các mô hình IDS học máy trên tập dữ liệu của họ trước khi gửi trọng số mô hình đến máy chủ tổng hợp để tính toán mô hình toàn cục.

Máy chủ trung tâm. Máy chủ có chức năng điều hành quá trình học liên kết với nhiệm vụ phân phối mô hình, tổng hợp mô hình toàn cục,... Trong luận án này, nó còn có thêm khả năng loại bỏ các mô hình tấn công đầu độc trước khi tổng hợp bằng việc tích hợp cơ chế phòng thủ Fed-LSAE của chúng tôi. Fed-LSAE sẽ được sử dụng tại máy chủ trung tâm để phát hiện và xóa bỏ các cập nhật độc hại khỏi quá trình tổng hợp mô hình toàn cục. Nó bao gồm 4 thành phần chính:

- *Bộ trích xuất PLR (PLR Extractor)*: Mô-đun này có trách nhiệm trích xuất chuỗi PLR của các đầu vào, đó là các mô hình cục bộ được cập nhật và mô hình toàn cục.

- *Mô hình Autoencoder (AE)*: Bằng cách đưa vào véc-tơ PLR vào AE, nó giúp trích xuất biểu diễn không gian tiềm ẩn (Latent space representation - LSR) chứa những đặc trưng quan

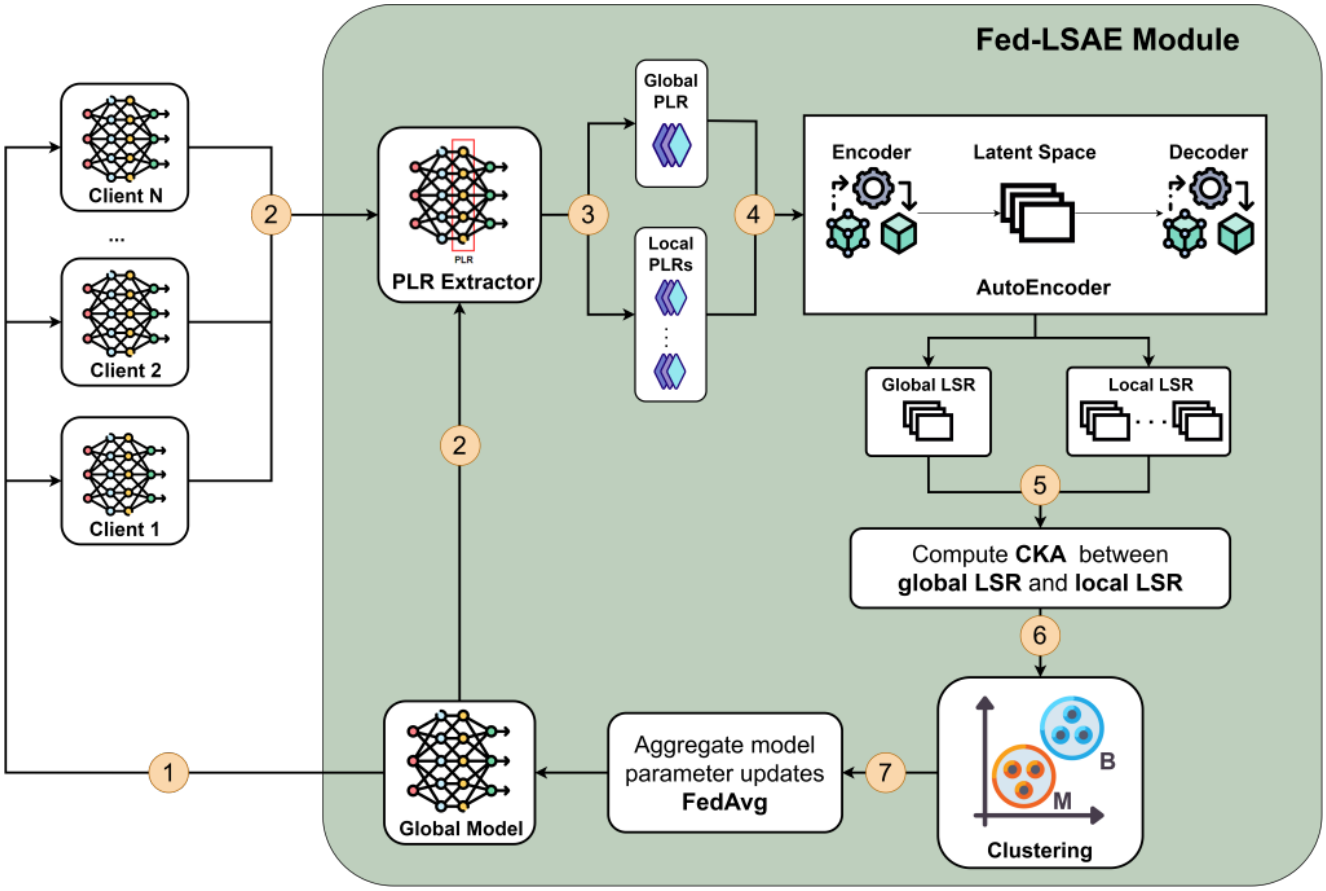


Hình 4.1: *Kiến trúc tổng quát của cơ chế Fed-LSAE trong ngăn chặn tấn công đầu độc mô hình học liên kết phát hiện xâm nhập*

trọng nhất của PLR. Để hoạt động tốt, thành phần AE sẽ được đào tạo trước ở phía máy chủ để nó học được các tính chất của PLR lành tính.

- *Thuật toán CKA*: Trong nghiên cứu này, chúng tôi sử dụng thuật toán CKA để đánh giá sự tương đồng giữa mỗi biểu diễn không gian tiềm ẩn (LSR) cục bộ và LSR toàn cục. Từ đó, chúng tôi có thể lọc ra các véc-tơ LSR độc hại có sự khác biệt rõ ràng với véc-tơ LSR toàn cục so với các véc-tơ còn lại. Ngoài ra, không như thuật toán Cô-sin (Cosine), CKA có thể đưa ra sự khác biệt rõ ràng hơn giữa một mô hình độc hại và một mô hình lành tính huấn luyện bởi dữ liệu non-IID khi so sánh độ tương đồng với một mô hình lành tính khác.

- *Thuật toán phân cụm (Clustering algorithm)*: Bằng việc gom nhóm các điểm CKA thành hai nhóm, chúng ta có thể chỉ ra rằng nhóm nhỏ hơn chính là nhóm các cập nhật độc hại với giả định rằng số lượng các kẻ tấn công đầu độc luôn nhỏ hơn một nửa tổng số các đại diện tham gia học cộng tác. Dựa vào đó, ta có thể lọc nhóm này khỏi quá trình tổng hợp của FL.



Hình 4.2: *Tổng quan quy trình, nguyên lý hoạt động của cơ chế Fed-LSAE*

4.2.3. Nguyên lý hoạt động của Fed-LSAE

4.3. Hiện thực, đánh giá

- Tiến hành nhiều kịch bản thực nghiệm khác nhau để cho thấy hiệu quả của việc phòng thủ chống lại các cuộc tấn công đầu độc thông qua phân tích chuyên sâu về hai bộ dữ liệu về các cuộc tấn công mạng IoT với các mô hình ML khác nhau.

- Đánh giá hiệu quả của Fed-LSAE so với các phương pháp bảo vệ tương tự trước đó, FedCC. Cụ thể, làm rõ khả năng vượt trội của Fed-LSAE trong việc nhận diện được đâu là mô hình lành tính được huấn luyện trên dữ liệu non-IID và đâu là mô hình độc hại thực sự.

4.4. Thảo luận

Chương này đề xuất một cơ chế tổng hợp mạnh mẽ để phát hiện và ngăn chặn các cuộc tấn công nhiễu đối với hệ thống phát hiện xâm nhập mạng được huấn luyện liên kết từ nhiều nguồn dữ liệu khác nhau. Nghiên cứu này đã khai thác một xu hướng mới bằng cách sử dụng chiến thuật phân tích không gian tiềm ẩn để tìm ra những bất thường trong các mô hình cập nhật của các tổ chức tham gia cộng tác. So với các nghiên cứu trước, Fed-LSAE cho thấy độ hiệu quả hơn mà

không yêu cầu một bộ dữ liệu bổ trợ hay kiến thức gì trước đó. Cụ thể hơn, chúng tôi sử dụng kết hợp của Autoencoder và kiểm tra không gian ẩn để tiết lộ các cập nhật độc hại từ các máy khách cục bộ. Kết quả thử nghiệm trên các bộ dữ liệu CIC-ToN-IoT, N-BaIoT và Edge-IIoTset đã chứng minh hiệu quả chất lượng cao của Fed-LSAE trong việc đánh bại các cuộc tấn công nhiễu không đích đến như đảo nhãn và untargeted-Med. Ngoài ra, Fed-LSAE của chúng tôi thể hiện hiệu suất tốt hơn so với FedCC trong trường hợp dữ liệu không đồng đều. Các kết quả đã chứng minh được các mục tiêu của từng kịch bản, từ đó cho thấy sự hiệu quả của mô hình Fed-LSAE khi có khả năng phát hiện các cuộc tấn công đầu độc poisoning attack tiên tiến trong ngữ cảnh IDS học liên kết. Giải pháp hứa hẹn sẽ mang lại một hướng tiếp cận huấn luyện bền vững các mô hình nhận diện mối đe dọa học máy bằng phương pháp FL, đặc biệt trong bối cảnh an ninh mạng hiện nay.

Như vậy, chương này đã trình bày một phương pháp phát hiện và ngăn chặn ảnh hưởng xấu của tấn công đầu độc mô hình liên kết phát hiện xâm nhập giữa các tổ chức tham gia trong khuôn khổ hợp tác xây dựng mô hình phát hiện và săn tìm mối đe dọa trong hệ thống mạng. Tuy vậy, hệ thống học liên kết được đề xuất vẫn gặp phải nguy cơ phá hoại và kém hiệu quả trong trường hợp đối mặt với các luồng dữ liệu tinh vi được phát sinh nhằm trốn tránh sự phát hiện của các ứng dụng IDS cục bộ. Lí do là, những kẻ phá hoại có thể tham gia vào qui trình huấn luyện học liên kết, hoặc tương tác với một trong những tổ chức tham gia để thăm dò và khai thác khả năng của các ứng dụng mục tiêu. Do đó, trong chương tiếp theo của luận án, chúng tôi trình bày phương pháp đánh giá tính bền vững của các ứng dụng IDS học máy trước các luồng dữ liệu mạng trốn tránh sự phát hiện. Phương pháp này có thể giúp cho các ứng dụng IDS cho hiệu năng tốt ngay cả ở trường hợp phân tích dữ liệu đầu vào được biến đổi tinh vi.

CHƯƠNG 5. Cơ chế đánh giá tính bền vững của các trình phát hiện xâm nhập trong mạng khả lập trình

5.1. Dẫn nhập

Chương này sẽ tập trung vào phát triển phương pháp đánh giá tính bền vững của các ứng dụng IDS học máy trước sự xuất hiện của các mẫu trốn tránh. Phương pháp này sẽ không chỉ đánh giá hiệu suất của mô hình trước các mẫu trốn tránh, mà còn xem xét khả năng chống lại chúng. Bằng cách đối mặt với các mẫu được thiết kế tinh vi để đánh lừa, mục tiêu của phương pháp này là xác định mức độ bền vững, kháng nhiễu của các mô hình IDS học máy. Từ đó, nghiên cứu này sẽ đóng góp vào việc cải thiện khả năng chống lại các mẫu trốn tránh của các hệ thống IDS học máy. Điều này sẽ tạo ra những sự cải tiến quan trọng trong việc bảo vệ hệ thống mạng khỏi các mối đe dọa tiềm tàng và luôn biến đổi liên tục một cách tinh vi.

5.2. Mô hình hóa mối đe dọa và giả định

5.2.1. Mô hình hóa mối đe dọa

Mô hình hóa mối đe dọa trong phạm vi nghiên cứu này được coi là khả năng tương tác của kẻ tấn công đối với các bộ phân loại được nhắm mục tiêu trong cuộc tấn công lẩn tránh cũng như khả năng xảy ra các kịch bản tấn công. Cụ thể, khả năng của kẻ tấn công có thể được phân thành 5 nhóm:

- Dữ liệu huấn luyện: quyền truy cập vào tập dữ liệu dành cho huấn luyện IDS, bao gồm Đọc, Ghi hoặc Không.
- Bộ thuộc tính: kiến thức về các thuộc tính mà các IDS dựa trên học máy sử dụng để phân tích trong hệ thống của chúng, có thể là một phần, toàn bộ hoặc không có.
- Mô hình phân loại/phát hiện: kiến thức về các mô hình ML được huấn luyện để phát hiện xâm nhập.
- Khả năng tương tác và nhận phản hồi từ mục tiêu: thu thập dự đoán do IDS học máy tạo ra cho mẫu đầu vào từ kẻ tấn công trốn tránh. Phản hồi nhận được có thể bị giới hạn, không giới hạn hoặc không có gì.

- Mức độ tuỳ chỉnh: bản chất của việc tạo mẫu đối kháng, có thể tùy chỉnh trong không gian vấn đề hoặc không gian thuộc tính.

Các cuộc tấn công đối kháng có thể được chia thành ba trường hợp dựa trên mức độ hiểu biết của kẻ tấn công về mục tiêu: mô hình hộp đen (black-box), hộp xám (gray-box) và hộp trắng (hộp trắng). Trong cách tiếp cận hộp đen, kẻ tấn công không biết gì về tập huấn luyện, cấu trúc và siêu tham số của mục tiêu nhưng có thể tương tác với thuật toán học máy để truy vấn dự đoán cho các đầu vào cụ thể. Trong trường hợp hộp xám, những kẻ tấn công không biết về các tham số của mô hình học máy, nhưng thuật toán bị tiết lộ. Trong khi đó, hộp trắng được cho là những kẻ tấn công biết mọi thứ về mô hình học máy, bao gồm cấu trúc, tham số và tập dữ liệu huấn luyện.

5.2.2. Giả định

Trong nghiên cứu này, chúng tôi tập trung vào các IDS hộp đen nơi kẻ tấn công có thể truy cập kết quả phân loại mà không cần biết gì về kiến trúc hoặc thuật toán của trình phát hiện tấn công dựa trên sự bất thường. Trong trường hợp đó, chúng tôi làm rõ một số giả định trong luận án này như sau:

- Kẻ tấn công không có quyền truy cập vào tập dữ liệu huấn luyện của ứng dụng IDS.
- Kẻ tấn công không có kiến thức về các thông số hoặc thuật toán học máy cơ bản.
- Kẻ tấn công biết thuộc tính được dùng trong ứng dụng IDS, thuận lợi trong việc tạo mẫu đối kháng.
- Kẻ tấn công có thể truy xuất tới thông tin phản hồi từ ứng dụng IDS mục tiêu trên từng bản ghi dữ liệu được gửi tới IDS.
- Thao tác dữ liệu để tạo mẫu đối kháng được thực hiện trong không gian thuộc tính, bao gồm các thuộc tính luồng mạng.

5.3. Phương pháp phát sinh biến thể luồng mạng trốn tránh và huấn luyện đối kháng

5.3.1. Tổng quan phương pháp phát sinh biến thể luồng mạng trốn tránh

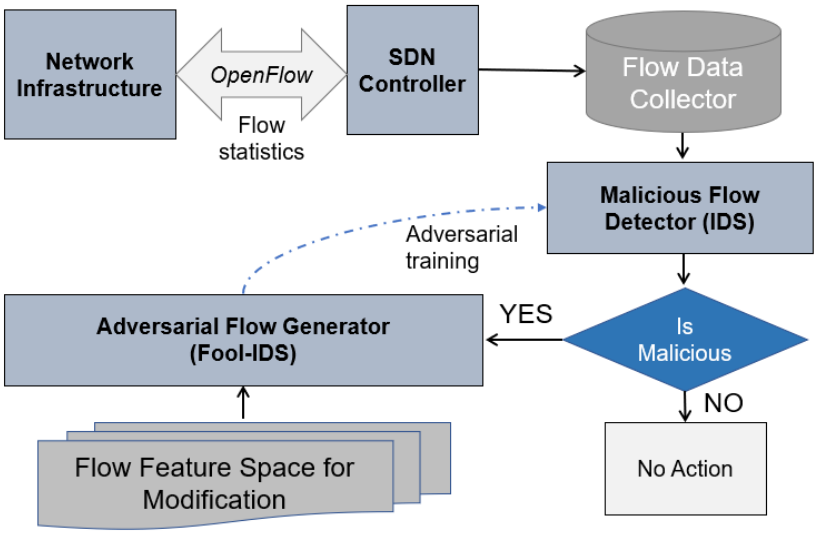
Trong phần này, chúng tôi sẽ giới thiệu bộ khung Fool-IDS với 2 cách tiếp cận chính sử dụng các mô hình tổng quát để tạo các mẫu trốn tránh hay mẫu đối kháng, được gọi là luồng độc hại để lẩn tránh sự phát hiện của ứng dụng IDS trong môi trường SDN. Hướng tiếp cận đầu tiên, sẽ dựa trên các biến thể của WGAN, bao gồm WGAN-GP và WGAN-GP TTUR. Cách tiếp cận thứ

hai là mô hình AdvGAN với số lượng mẫu đối kháng được tạo không phụ thuộc số lượng mẫu gốc, vì cách tiếp cận dựa trên GAN có thể tìm kiếm các mẫu đối kháng tiềm năng dựa trên sự phân bố của mẫu luồng thông thường.

Lưu ý rằng, phương pháp có thể sử dụng trong cả mạng truyền thống và mạng SDN. Điều đó có nghĩa là việc sửa đổi một số thuộc tính nhất định dựa trên các đặc trưng đã chọn trong một ngữ cảnh cụ thể. Thông thường sự khác biệt giữa mạng truyền thống và mạng SDN là không gian của các thuộc tính luồng được trích xuất. Có một số thuộc tính bị thiếu trong các mạng truyền thống so với mạng SDN vì bộ điều khiển của SDN cung cấp nhiều thuộc tính đặc biệt hơn. Việc sử dụng các thuộc tính luồng SDN ít tốn thời gian hơn so với phương pháp dựa trên gói tin. Trong SDN, bộ điều khiển được sử dụng để quản lý điều hướng luồng và truy xuất các số liệu thống kê luồng trong hạ tầng mạng thông qua giao thức OpenFlow. Thông tin về lưu lượng của luồng có thể thu được bằng hai phương pháp. Phương pháp đầu tiên là thiết lập một qui luật mới từ bộ điều khiển SDN đến Open vSwitch, sau đó thu thập và trích xuất các thuộc tính luồng từ Mirroring Server bằng công cụ CICFlowMeter. Phương pháp thứ hai là gửi các yêu cầu thống kê OpenFlow tới các thiết bị switch và thu thập các thông điệp phản hồi. Hai phương pháp này có thể kết hợp với nhau để tạo ra nhiều thuộc tính bao hàm các trạng thái mạng.

Ngoài ra, khả năng giám sát toàn bộ mạng của bộ điều khiển SDN có thể cung cấp các công cụ tốt hơn để trích xuất các thuộc tính luồng để phân tích trạng thái mạng. Do đó, nó cũng hỗ trợ một phần trong tác vụ tạo các mẫu luồng đối kháng từ quan điểm của kỹ sư đánh giá độ bền vững của ứng dụng trước tác nhân phá hoại, hay còn được gọi là nhiệm vụ của đội ngũ tấn công (red team) trong lĩnh vực an ninh mạng. Trong nghiên cứu này, các cuộc tấn công mạng chống lại mạng SDN được xem xét trên góc độ người dùng, trong khi hệ thống phát hiện hay phòng thủ dựa trên việc tận dụng các thuộc tính luồng mạng SDN. Quy trình tổng thể của phương pháp được mô tả ở **Hình 5.1**, bao gồm giai đoạn thu thập và tiền xử lý dữ liệu luồng, tạo luồng đối kháng và huấn luyện đối kháng. Đối với mỗi khoảng thời gian t được xác định trước, dữ liệu luồng được thu thập từ hạ tầng mạng thông qua bộ điều khiển SDN dưới sự hỗ trợ của giao thức OpenFlow. Sau khi tiền xử lý, các luồng này được đưa vào ứng dụng IDS để phân loại. Nếu kết quả được nhận dạng là luồng độc hại, thì luồng đó được sử dụng để tạo mẫu lưu lượng tấn công mới chứa nhiễu. Sau cùng, tất cả các mẫu đối kháng được thu thập để huấn luyện lại ứng dụng IDS. Chiến lược này có thể làm cho hệ thống phát hiện xâm nhập ít nhầm lẫn hơn với các mẫu lưu lượng tấn công đối kháng.

Đặt vấn đề, ta ký hiệu x là bản ghi luồng lưu lượng gốc và f là bộ phát hiện với kết quả đúng là nhãn y. Mục tiêu của kẻ tấn công là tính toán độ nhiễu nhỏ δ để làm cho đầu ra của bộ phát hiện sai lệch với nhãn dự đoán sai y' trong khi vẫn giữ lại các đặc tính của y từ mẫu gốc, sao cho $f(x + \delta) \neq f(x)$. Những luồng được tạo ra với độ nhiễu này được gọi là mẫu đối kháng. Lưu ý, không sửa đổi tập huấn luyện được sử dụng cho các mô hình phát hiện mà thao tác này chỉ được



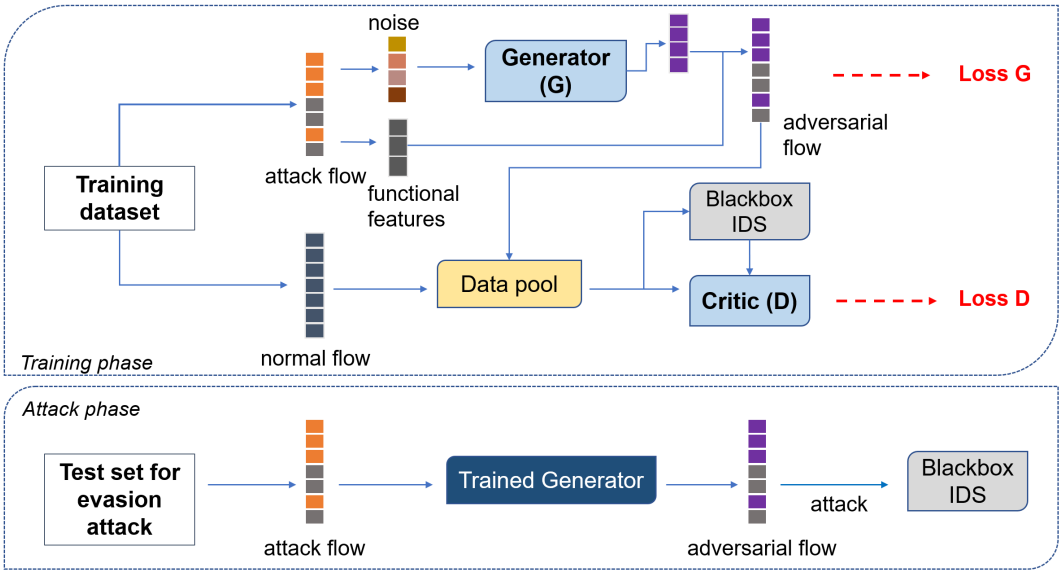
Hình 5.1: *Quy trình tạo ra luồng mạng đối kháng và chiến lược đào tạo với Fool-IDS trong môi trường SDN.*

thực hiện trong giai đoạn kiểm tra mô hình để tối ưu khả năng tránh né của lưu lượng tấn công đối kháng. Sau đó, các mẫu tự phát hiện được thu thập và sử dụng để duy trì tính bền vững của bộ phát hiện tấn công trong quá trình hoạt động.

Trong nghiên cứu này, chúng tôi sửa đổi các phần khác nhau của các thuộc tính phi chức năng để đánh giá hiệu năng trốn tránh của bộ khung. Ý tưởng chính là giữ lại các thuộc tính chức năng để không ảnh hưởng đến các chức năng độc hại ban đầu.

5.3.2. Tạo mẫu đối kháng bằng Wasserstein GAN

Kiến trúc tổng quan của bộ khung được đề xuất, có tên Fool-IDS, được mô tả trong **Hình 5.2**. Để huấn luyện GAN được ổn định hơn, nghiên cứu này sử dụng WGAN-GP. Nếu trình phân biệt



Hình 5.2: *Kiến trúc tổng quan của bộ khung Fool-IDS dựa trên WGAN-GP để trốn tránh tấn công.*

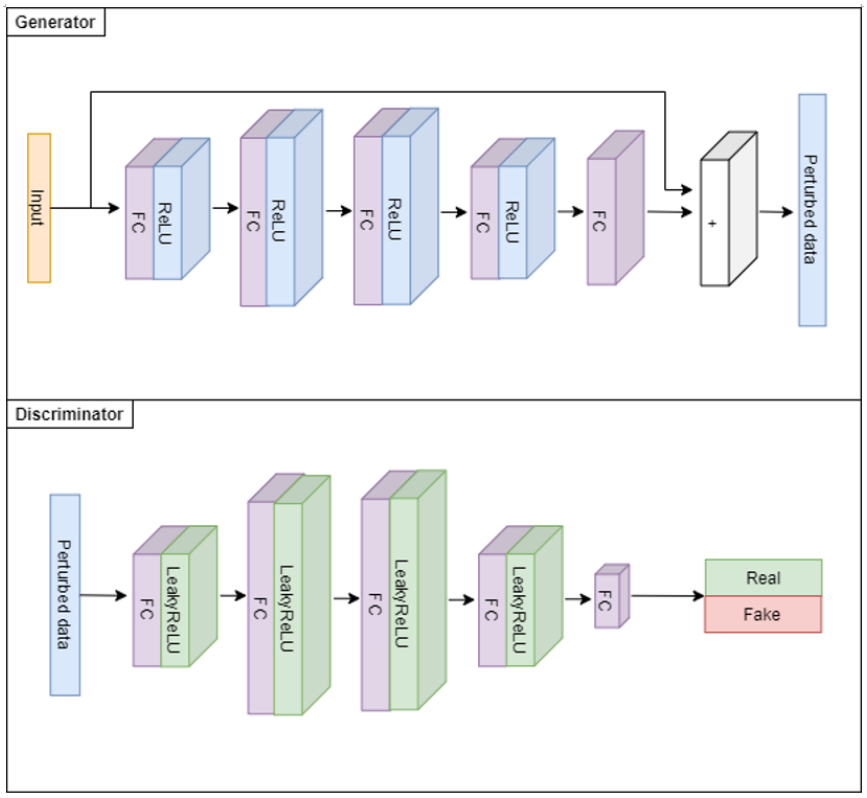
có hiệu năng phân loại và dự đoán tốt các mẫu lành tính và độc hại thì mô hình phân biệt áp dụng một khoảng Wasserstein để đưa ra điểm số chất lượng giữa mẫu gốc và mẫu đối kháng. Ngoài ra, để tránh giá trị mất mát Wasserstein (loss) quá lớn, WGAN sử dụng kỹ thuật cắt trọng số để duy trì ràng buộc Lipschitz. Tuy nhiên, việc sử dụng kỹ thuật này gây ra các vấn đề về tối ưu và chất lượng mẫu kém. Như một cải tiến mới, mô hình WGAN-GP khác với WGAN vì nó sử dụng độ phạt gradient thay vì cắt bớt trọng số để duy trì ràng buộc Lipschitz. Mô hình này được chứng minh là vượt trội so với kiến trúc GAN tương tự. Thêm vào đó, nghiên cứu này cũng sử dụng WGAN-GP TTUR để khai thác hiệu quả của việc khuyến khích trình phân biệt học nhanh hơn trình tạo sinh $\mathcal{G}$.

5.3.2.1. Trình IDS hộp đen mục tiêu

Trong mô hình được đề xuất, các ứng dụng black-box IDS mô phỏng một IDS mục tiêu trong các tình huống thực tế. Rõ ràng là chúng ta không thể biết cấu trúc thực sự của các ứng dụng IDS được sử dụng trong các hệ thống mạng hay các tổ chức khác nhau. Trong nghiên cứu này, các thuật toán ML được sử dụng để triển khai black-box IDS và cho rằng kẻ tấn công (trong trường hợp này là trình tạo sinh) không biết gì về black-box IDS ngoài số lượng thuộc tính đầu vào. Nhiệm vụ chính của black-box IDS là phân loại và gán nhãn các mẫu luồng bình thường và đối kháng (được tạo từ trình tạo sinh) và sau đó cung cấp đầu ra để trình phân biệt đánh giá chất lượng của các mẫu.

5.3.2.2. Trình tạo mẫu đối kháng

Trình tạo mẫu đối kháng đóng vai trò quan trọng trong các mô hình tổng quan như WGAN-GP. Cụ thể, nó thực hiện tổng hợp dữ liệu mới chưa nhìn thấy dựa trên phân phối dữ liệu hiện tại. Trình tạo sinh là nơi dữ liệu đối kháng được tạo, cập nhật liên tục và chuyển đổi trở thành các biến thể hoàn chỉnh, có khả năng đánh bại trình phân biệt. Mục tiêu của nghiên cứu này là đánh lừa các black-box IDS dựa trên trình tạo sinh $\mathcal{G}$ được hỗ trợ bởi phản hồi của trình phân biệt về chất lượng của các mẫu tổng hợp từ $\mathcal{G}$. Cấu trúc của trình tạo sinh bao gồm một mạng nơ-ron 5 lớp fully connected và 4 lớp đầu ra được kích hoạt bởi hàm ReLU mà không áp dụng batch normalization. Đầu vào của trình tạo sinh $\mathcal{G}$ bao gồm vector nhiễu và các thuộc tính phi chức năng được trích xuất từ bộ dữ liệu luồng tấn công đã được chuẩn hoá thành các giá trị trong phạm vi từ 0 đến 1. Trong quá trình huấn luyện, kích thước đầu vào của trình tạo sinh thay đổi linh hoạt dựa trên số lượng các thuộc tính phi chức năng đã chọn của mẫu lưu lượng. Lưu ý rằng, chỉ sửa đổi các thuộc tính phi chức năng của bản ghi tấn công trong các bản ghi được tạo để duy trì tính độc hại của chúng trong trường hợp thực tế. Tuy nhiên, có nhiều trường hợp sửa đổi số lượng khác nhau của thuộc tính này, có thể là 100%, 50%, 25%, hoặc 10% trong số tất cả thuộc tính phi chức năng có sẵn. Đầu ra của mô hình là một mẫu đối kháng có cùng kích thước với mẫu



Hình 5.3: *Cấu trúc trình tạo sinh và trình phân biệt trong AdvGAN.*

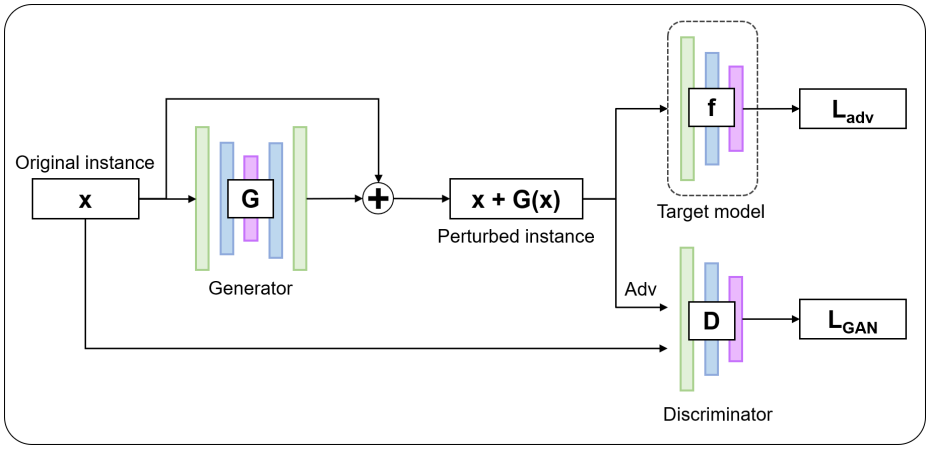
luồng tấn công đầu vào, các giá trị vector này được chuẩn hóa trong phạm vi từ 0 đến 1. Nó đảm bảo rằng các mẫu được tạo từ mô hình phải hợp lí.

5.3.2.3. Trình phân biệt

Trình phân biệt $\mathcal{D}$ chịu trách nhiệm bắt chước ứng dụng IDS hộp đen mục tiêu, tính toán chất lượng các mẫu được tạo và cung cấp độ đo mất mát để hỗ trợ huấn luyện trình tạo sinh. Kiến trúc mạng nơ-ron của trình phân biệt cũng có 5 lớp linear, nhưng sử dụng hàm kích hoạt Leaky ReLU. Lưu ý rằng, WGAN-GP không sử dụng batch normalization trong trình phân biệt. Đầu vào mô hình bao gồm các mẫu luồng bình thường và đối kháng đã được phân loại và gán nhãn trước đó bởi ứng dụng IDS hộp đen.

5.3.3. Tạo mẫu đối kháng bằng AdvGAN

Ngoài WGAN-GP và WGAN-GP TTUR, nghiên cứu này cũng áp dụng AdvGAN để tạo mẫu đối kháng. Cấu trúc bao gồm trình tạo sinh $\mathcal{G}$, trình phân biệt $\mathcal{D}$ và mô hình mục tiêu được mô tả chi tiết ở **Hình 5.3**.



Hình 5.4: *Quy trình huấn luyện trình tạo sinh của AdvGAN.*

5.4. Thực nghiệm

Nghiên cứu này tiến hành thử nghiệm trên hai bộ dữ liệu CICIDS2018 và InSDN để giả lập môi trường mạng, trong ngữ cảnh thông thường và SDN. Trong bộ dữ liệu CICIDS2018, nghiên cứu tập trung vào loại tấn công DoS để tạo mẫu đối kháng.Trong đó, phương pháp tạo mẫu đối kháng chỉ biến đổi các thuộc tính phi chức năng của các luồng tấn công DoS nguyên bản.

Để đánh giá hiệu suất của các cuộc tấn công trốn tránh đối với IDS, chúng tôi sử dụng 2 chỉ số là Detection Rate (DR) và Evasion Increase Rate (EIR). Tỷ lệ phát hiện DR, như được định nghĩa trong **Công thức (5.1)**, là tỷ lệ các cuộc tấn công được IDS phát hiện trên tổng số mẫu tấn công trong tập thử nghiệm. Số liệu này được đo trong trường hợp dữ liệu gốc (tỷ lệ phát hiện ban đầu O-DR) cũng như các mẫu đối kháng (tỷ lệ phát hiện đối kháng A-DR). Sự khác biệt giữa 2 DR này có thể phản ánh hiệu quả của các tấn công né tránh. Trong khi đó, EIR, được tính bằng **Công thức (5.2)**, cho biết sự gia tăng các mẫu tấn công đối kháng không bị phát hiện bởi IDS so với các mẫu ban đầu.

$$DR = \frac{\text{Số tấn công phát hiện đúng}}{\text{Tổng số tấn công}} \tag{5.1}$$

$$EIR = 1 - \frac{\text{Tỉ lệ phát hiện đối kháng A-DR}}{\text{Tỉ lệ phát hiện gốc O-DR}} \tag{5.2}$$

5.4.1. Tỷ lệ phát hiện mẫu đối kháng của IDS

Nhìn chung, các kết quả đã chỉ ra rằng các thuật toán ML dùng để phát hiện sự bất thường trong không gian mạng dễ bị ảnh hưởng bởi các mẫu đối kháng. Do đó, chúng ta cần thường xuyên cập nhật các điểm dữ liệu mới và các biến thể của các cuộc tấn công để có khả năng phát

hiện xâm nhập tốt hơn. Tuy nhiên, chúng ta cần xem xét sự giống nhau của các mẫu tấn công ban đầu và các mẫu đối kháng để đánh giá mức độ mạnh mẽ của IDS dựa trên ML với nhiều độ nhiễu khác nhau. Mục đích này có thể đạt được bằng cách phân tích trực quan hóa các mẫu luồng tấn công trước và sau khi thêm nhiễu vào các mẫu luồng tấn công ban đầu, cùng với khả năng lẩn tránh của các mẫu mới được phát sinh.

Kết quả thí nghiệm cho thấy hiệu suất chung của AdvGAN tốt hơn WGAN-GP TTUR trong việc tránh các mô hình ML trong hầu hết các trường hợp sửa đổi thuộc tính phi chức năng. Chỉ số EIR trong các tình huống tấn công các mô hình DT, LR, CNN, MLP và LSTM cho thấy rằng AdvGAN hoạt động tốt hơn ngay cả khi thay đổi số lượng nhỏ các thuộc tính phi chức năng. Tóm lại, việc áp dụng WGAN-GP TTUR và AdvGAN vượt trội so với SDN-GAN. Ngoài ra, AdvGAN tạo ra hiệu suất tránh IDS dựa trên ML ổn định hơn so với WGAN-GP TTUR.

5.4.2. Tái huấn luyện black-box IDS với mẫu đối kháng và lặp lại tấn công né tránh

Sau khi các black-box IDS được huấn luyện lại với các mẫu đối kháng được tạo bằng cách áp dụng chiến lược huấn luyện đối kháng (AT), kẻ tấn công sẽ huấn luyện lại bộ khung với phiên bản IDS mới để trốn tránh chúng. Sau khi thực hiện quá trình này, black-box IDS vẫn khó có thể phát hiện lại lưu lượng tấn công đối kháng. Để chứng minh kịch bản này, kiến trúc WGAN-GP được sử dụng để đào tạo các mẫu đối kháng và lặp lại cuộc tấn công sau khi nâng cấp IDS với các luồng mạng được tạo trên bộ dữ liệu CICIDS2018. Kết quả thực nghiệm chứng tỏ rằng cần phải liên tục kiểm tra và củng cố tính bền vững của IDS này cho các ứng dụng thực tế.

5.5. Thảo luận

Chương này đã trình bày phương pháp kiểm tra tính bền vững của các ứng dụng IDS học máy bằng phương pháp phát sinh mẫu luồng mạng đối kháng dựa trên kiến trúc GANs. Các kết quả thí nghiệm cho thấy, để đảm bảo hiệu năng phát hiện tấn công mạng trong điều kiện thực tế, các ứng dụng này cần được thực hiện kiểm tra và liên tục cập nhật các dấu hiệu xâm nhập từ lưu lượng luồng mạng trong hệ thống nội bộ và các tổ chức, hệ thống liên kết cộng tác khác. Để thu thập và trích xuất các thông tin từ các lưu lượng luồng mạng trong SDN, các ứng dụng mạng cần phải kết nối với bộ điều khiển mạng SDN để thực hiện thao tác điều phối, giám sát và rút trích thông tin cần thiết từ hệ thống. Nhằm gia tăng tính tin cậy, an toàn và bảo mật cho các thiết lập kết nối giữa bộ điều khiển và các ứng dụng mạng, cơ chế xác thực và kiểm soát truy cập các ứng dụng kết nối đến bộ điều khiển SDN là yêu cầu trọng yếu cần xem xét. Trong chương tiếp theo, chúng tôi trình bày một cách tiếp cận xác thực và kiểm soát truy cập phi tập trung cho các ứng dụng trong mạng SDN nhằm giải quyết những tồn tại hiện thời nhằm quản lý và điều phối các ứng dụng mạng một cách an toàn, bảo mật và hiệu quả.

CHƯƠNG 6. Cơ chế xác thực và kiểm soát truy cập phi tập trung cho các ứng dụng trong mạng khả lập trình

6.1. Dẫn nhập

Để thực hiện các tác vụ quản lý, điều phối an ninh trong một hệ thống mạng của mỗi tổ chức, các ứng dụng mạng được sử dụng như một phương tiện truy cập vào điều khiển để thực hiện các lệnh cấu hình, thiết lập cho hoạt động triển khai chính sách, cũng như giám sát trạng thái của hệ thống. Do đó, bộ điều khiển trong mạng SDN cần được bảo vệ bởi một cơ chế xác thực và kiểm soát truy cập từ các đối tượng ở lớp ứng dụng. Các cơ chế này thường được tích hợp trực tiếp trên mỗi bộ điều khiển và chỉ có tính hợp lệ trong nội tại phân vùng mạng được quản lý bởi một bộ điều khiển nhất định, khó hỗ trợ tính năng cộng tác, chia sẻ thông tin cũng như khả năng mở rộng khi hệ thống có kích thước lớn. Do đó, chương này sẽ trình bày mô hình xác thực và kiểm soát truy cập phi tập trung nhằm tăng cường tính tin cậy, bảo mật và dễ mở rộng cho các hệ thống cộng tác trong mạng SDN - hệ thống liên kết phát hiện xâm nhập và săn tìm mối đe dọa.

6.2. Các cách tiếp cận xác thực ứng dụng tại bộ điều khiển SDN

Các ứng dụng trong mạng khả lập trình sử dụng giao diện Northbound để kết nối với bộ điều khiển để thực hiện kết nối và sử dụng các thông tin cần thiết từ hệ thống mạng phục vụ mục đích được thiết kế của chính ứng dụng. Các ứng dụng này có thể thuộc về một bộ điều khiển hay từ bên ngoài. Tuy vậy, các ứng dụng mạng trong SDN vốn được phát triển để điều phối và triển khai các chức năng quản trị mạng, hay chức năng bất kì thông qua bộ điều khiển SDN có thể bị lạm dụng từ những tác nhân độc hại. Vấn đề thiếu cơ chế xác thực, quản lý các kết nối từ các ứng dụng mạng tới bộ điều khiển có thể gây ra sự mất an toàn cho cả hệ thống mạng. Trong bối cảnh đó, các hệ thống quản lý và kiểm tra xác thực, kiểm soát truy cập được đề xuất ở giao diện Northbound trên bộ điều khiển SDN để khắc phục điểm yếu này. Nhưng các hệ thống như vậy được triển khai ngay trên bộ điều khiển SDN sẽ dẫn đến quá tải trong xử lí do phải thực hiện nhiều tác vụ từ việc điều khiển tới theo dõi trạng thái của toàn bộ hệ thống mạng. Khi số lượng các thiết bị và kết nối bùng nổ trong mạng, bộ điều khiển không thể đáp ứng được các tác vụ về kiểm soát truy cập các ứng dụng và hạ tầng mạng. Ngoài ra, sự tập trung hóa của hệ thống xác thực sẽ gặp hạn chế trong vấn đề co giãn (scalability), phối hợp tin cậy (trust) khi số lượng các thiết bị mạng tăng cao ở từng phân vùng mạng. Do sự thiếu tin cậy, mỗi bộ điều khiển chỉ có khả năng thực hiện tác vụ điều khiển mạng thông qua ứng dụng của chính nó. Nếu một ứng

dụng SDN muốn gọi tới các tài nguyên mạng liên lĩnh vực (cross-domain), nó cần phải xác thực và phân cấp quyền hạn nhiều lần bởi nhiều bộ điều khiển khác nhau. Cụ thể hơn, khi một ứng dụng SDN được tích hợp sang một miền điều khiển khác, qui trình thực hiện kiểm soát truy cập lại được thực hiện lặp lại như một yêu cầu bắt buộc. Ngoài ra, vấn đề quản lý nhật ký các kết nối truy cập của các ứng dụng này tới các tài nguyên mạng cũng đóng vai trò quan trọng trong việc tính phí dịch vụ, phân tích và kiểm tra. Cơ chế này cũng đòi hỏi tính tin cậy cao và chống sai sót hay tránh phụ thuộc vào một thực thể tập trung. Các cách tiếp cận điển hình bao gồm:

- Giải pháp kiểm soát truy cập phụ thuộc vào bộ điều khiển
- Giải pháp kiểm soát truy cập độc lập với bộ điều khiển

6.3. Kiến trúc tổng quan mô hình đề xuất B-DAC

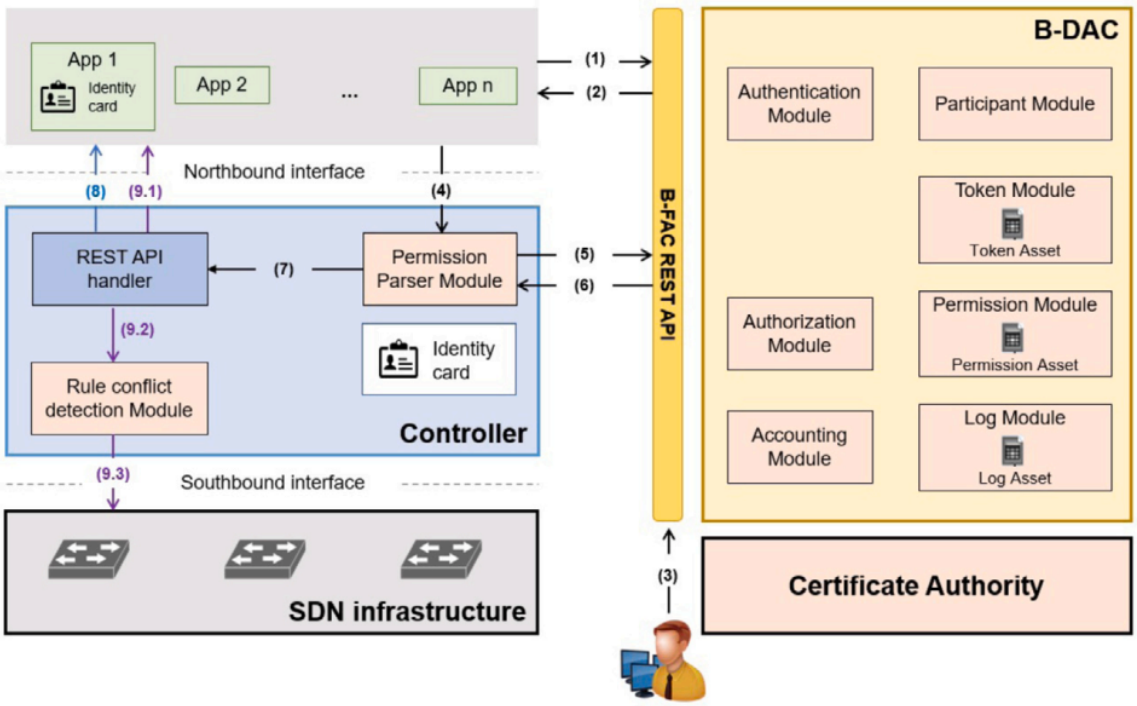
Tận dụng các đặc tính của blockchain, nghiên cứu này đề xuất một kiến trúc kiểm soát truy cập phi tập trung cho các ứng dụng mạng trong SDN, với tên gọi là B-DAC. Nguyên tắc hoạt động của bộ khung B-DAC độc lập với bộ điều khiển SDN để có thể dễ dàng thích ứng với các bộ điều khiển cụ thể khác nhau. Nó được thiết kế để trở thành một khuôn khổ phân quyền để xử lý các ứng dụng mạng trong việc đảm bảo hoạt động ổn định của bộ điều khiển. Cụ thể, hệ thống kiểm soát truy cập B-DAC chịu trách nhiệm xác thực, ủy quyền và giám sát các ứng dụng trong giao tiếp với bộ điều khiển. Ngoài ra, các quy tắc luồng được xác nhận hợp lệ để ngăn ngừa xung đột trước khi được cài đặt vào bộ chuyển mạch theo lệnh của bộ điều khiển thông qua các ứng dụng mạng. Bộ khung B-DAC bao gồm nhiều mô-đun để đảm bảo tính bảo mật của giao tiếp ứng dụng-bộ điều khiển, được mô tả trong **Hình 6.1**.

6.3.1. Các thực thể chính trong kiến trúc hệ thống

6.3.1.1. Thành viên tham gia

Trong hệ thống B-DAC, blockchain được tận dụng phục vụ mục đích kiểm soát truy cập theo mô hình AAA, trong đó các thành phần của mạng SDN được xem xét như là các đối tượng trong hệ thống blockchain. Bởi vì mục đích chính là bảo vệ an toàn cho các kết nối giữa bộ điều khiển và các ứng dụng mạng, các thực thể tham gia hệ thống blockchain bao gồm: Bộ điều khiển (controller), ứng dụng, và quản trị viên (administrator).

Mỗi thành viên tham gia được xác định bằng hai trường: thông tin mã định danh (identity) *id* và tên của thành viên *name*. Hơn nữa, đối với những thành viên tham gia là ứng dụng, thì nó có các trường bổ sung là *role* trong hồ sơ của thành viên tham gia, để lưu trữ vai trò của chúng được hệ thống B-DAC chỉ định trong SDN.



Hình 6.1: *Mô hình kiến trúc tổng quan của bộ khung B-DAC.*

Để thiết lập cơ chế quản lý độ tin cậy của ứng dụng dựa trên hành vi của nó, một trường thông tin bổ sung là Trust Index được thiết kế bên trong hồ sơ của ứng dụng. Giá trị của Trust Index sẽ được điều chỉnh giảm xuống một khi ứng dụng gửi các yêu cầu vượt quyền hay xung đột quy luật luồng (flow rules) đã cài đặt trước đó trong hạ tầng mạng.

6.3.1.2. Các thành phần cấu trúc chính

Hình 6.1 biểu diễn thành phần cấu trúc của hệ thống B-DAC dựa trên blockchain, bao gồm các thành phần (mô-đun) sau:

- Participant module: Trong B-DAC, Mô-đun quản lý thành viên (Participant module) đảm nhận nhiệm vụ quản lý các thông tin và yêu cầu của các thành viên tham gia (bộ điều khiển, ứng dụng) trong hệ thống B-DAC.
- Token module: quản lý cấp phát token truy cập cho các thực thể muốn giao tiếp với B-DAC.
- Permission module: Thành phần này đóng vai trò quản lý các quyền hạn tương ứng với việc truy cập vào các tài nguyên mạng trong SDN.
- Authentication module: Mô-đun xác thực đóng vai trò xác thực danh tính của các ứng dụng, bộ điều khiển tham gia cơ chế xác thực phi tập trung.
- Authorization module: Mô-đun xác thực, cấp quyền, phân quyền cho ứng dụng.

- Accounting module: Mô-đun kiểm tra, dùng cho thao tác kiểm tra nhật ký hoạt động trên hệ thống.

- Log module: thành phần này hỗ trợ Mô-đun Accounting trong việc ghi nhận log hệ thống B-DAC.

Để hỗ trợ tác vụ Xác thực, B-DAC sử dụng một token để chỉ cho phép các ứng dụng đã đăng ký mới có thể tương tác với hệ thống. Các token này được tạo và quản lý bởi Mô-dun Token (Token module). Thêm vào đó, Mô-đun Permission (Permission Module) được dùng để quản lý các đối tượng Quyền hạn (Permission) được định nghĩa trước của các ứng dụng. Các thông tin này được sử dụng trong tác vụ phân quyền ứng dụng (Authorization) sau khi ứng dụng đã được xác thực bởi hệ thống B-DAC.

Bên cạnh đó, cơ chế ghi nhật ký (logging) là một trong những tác vụ liên quan tới Kiểm tra (Accounting) bên trong một sơ đồ kiểm soát truy cập AAA. Bộ khung B-DAC cũng được thiết kế một mô-đun có tên là *Log Module* để thực hiện ghi lại những hành vi tác động tới những tài nguyên chính (asset) trên mạng blockchain. Các thay đổi này liên quan tới các giao dịch mà hệ thống blockchain đã xử lý cho các yêu cầu từ ứng dụng cũng như bộ điều khiển.

6.3.1.3. RESTful API

Để tăng cường tính linh động và độc lập của hệ thống kiểm soát truy cập với bất kỳ bộ điều khiển SDN nào, sơ đồ kiểm soát truy cập AAA được đặt bên ngoài bộ điều khiển. Cụ thể hơn, bộ khung B-DAC cung cấp các RESTful API cho bộ điều khiển cũng như ứng dụng OpenFlow có thể tương tác trực tiếp hệ thống B-DAC. Các thực thể tham gia có thể gửi các yêu cầu HTTP thông qua các REST API này để truy vấn hay cập nhật thông tin trong hệ thống mạng. Bên cạnh đó, để đảm bảo tính an toàn cho việc trao đổi thông tin, giao thức HTTPS được dùng cho tất cả các kết nối REST API này.

6.3.2. Định nghĩa chính sách

6.3.2.1. Nguyên tắc Request-based Permission

Bộ điều khiển cung cấp các API cho việc truy cập các tài nguyên mạng tại giao diện Northbound của mạng SDN. Mỗi API được phân biệt bởi một *URI* và một *HTTP method*. Các phương thức phổ biến của giao thức HTTP bao gồm POST, GET, PUT và DELETE tương ứng với thao tác tạo mới, đọc, cập nhật và xóa dữ liệu. Trong đó, URI được dùng để định vị tài nguyên mạng. Một cặp thông tin *URI* và *HTTP method* cung cấp khả năng thực hiện một tác vụ truy cập vào các tài nguyên trong mạng. Do đó, B-DAC định nghĩa trước tập hợp các quyền hạn tương ứng với các API được cung cấp bởi bộ điều khiển. Các thông tin chi tiết như *access token*, *content type* được

giữ trong trường *header* của một yêu cầu *HTTP*. Ứng với mỗi yêu cầu từ ứng dụng tới bộ điều khiển, các tham số của *URI* và *HTTP method* được trích xuất và xử lý bởi Mô-đun Permission Parser trong bộ điều khiển nhằm xác định các quyền hạn được yêu cầu cho việc thực thi yêu cầu.

6.3.2.2. Định nghĩa chính sách

Để quyết định chấp nhận hay từ chối một yêu cầu truy cập REST API từ một ứng dụng, một chính sách (policy) được khai báo và định nghĩa trước để ngăn chặn các yêu cầu bất hợp lệ tới tài nguyên mạng. Trong hệ thống B-DAC, các chính sách được phân thành 2 loại, bao gồm: Role Policy và Trust Policy.

Tất cả các tương tác của ứng dụng và bộ điều khiển, hay hành động của các quản trị viên đều được thực hiện qua các giao dịch (transaction) trên hệ thống blockchain.

6.4. Thực nghiệm và đánh giá

- Đánh giá thực nghiệm hiệu năng của các giao dịch trên hạ tầng blockchain: đánh giá mức độ tiêu thụ tài nguyên khi thực hiện các giao dịch trên blockchain tương ứng với các tác vụ kiểm soát truy cập trong hệ thống.
- Đánh giá tác động của cơ chế kiểm soát truy cập lên hiệu năng của bộ điều khiển: đo lường sự quá tải tăng thêm của bộ điều khiển được tích hợp hệ thống kiểm soát truy cập B-DAC, các thông số về mức độ sử dụng CPU và bộ nhớ được đo đạc khi thực hiện truy cập với số lượng yêu cầu khác nhau.

Các thí nghiệm cho thấy hiệu năng xử lý của B-DAC trên blockchain đáp ứng yêu cầu và ảnh hưởng không đáng kể với bộ điều khiển.

6.5. Thảo luận

Trong chương này, B-DAC - bộ khung kiểm soát truy cập phi tập trung được đề xuất để kiểm soát các yêu cầu truy cập từ các ứng dụng mạng trên giao diện Northbound của bộ điều khiển trong mạng khả lập trình. Bộ khung này được hiện thực và kiểm tra tính khả thi trên mô hình mẫu (prototype) dựa trên nền tảng Hyperledger Fabric nhằm tăng cường khả năng an toàn, bảo mật cho mạng trong các ngữ cảnh mạng có kích thước lớn, phân tán. Hệ thống B-DAC được thiết kế để đạt được các yếu tố: độc lập với bộ điều khiển, trong suốt xử lý với các ứng dụng, kiểm soát phi tập trung các yêu cầu từ ứng dụng. Điều này đảm bảo rằng tất cả các kết nối từ các ứng dụng khác nhau tới bộ điều khiển luôn luôn được kiểm tra và xác thực trước khi nó có thể truy cập hay thay đổi các tài nguyên trong hệ thống mạng. Ngoài ra, các chính sách an ninh được định nghĩa

và thực hiện trên từng loại tài nguyên khác nhau giúp cho việc kiểm soát truy cập của ứng dụng mịn hơn cho bộ điều khiển. Cơ chế độc lập với bộ điều khiển giúp các quản trị viên vừa ngăn chặn được các vấn đề truy cập trái phép, leo thang đặc quyền mà ít ảnh hưởng tới hiệu năng của bộ điều khiển, nơi điều phối trung tâm các hoạt động trong mạng.

CHƯƠNG 7. Kết luận và Hướng phát triển

7.1. Kết luận

Luận án đã trình bày các phương pháp nhằm gia tăng tính an toàn, bảo mật cho các hệ thống mạng khả lập trình thông qua ba khía cạnh đảm bảo tính an toàn, bền vững và xác thực của các ứng dụng IDS liên kết trong tác vụ phát hiện tấn công mạng và các mối đe dọa bảo mật. Kết quả đạt được của luận án được tóm tắt như sau:

- Đảm bảo tính an toàn, tin cậy, bảo mật thông tin và bảo vệ quyền riêng tư dữ liệu trong mô hình liên kết phát hiện xâm nhập và săn tìm mối đe dọa cho hệ thống mạng bằng cách sử dụng sơ đồ học liên kết để cộng tác huấn luyện mô hình học máy trong các ứng dụng mạng. Điều này đảm bảo rằng dữ liệu mạng được xử lý một cách an toàn và đáng tin cậy, đồng thời đảm bảo quyền riêng tư của người dùng và bảo vệ thông tin quan trọng khỏi các mối đe dọa mạng.

- Đảm bảo hiệu năng của mô hình liên kết phát hiện xâm nhập cho hệ thống mạng phân tán cộng tác bằng phương pháp ngăn chặn tấn công đầu độc mô hình toàn cục từ các máy khách huấn luyện cục bộ ác ý, phá hoại. Cơ chế này cho phép huấn luyện ứng dụng IDS hiệu quả hơn trong điều kiện tồn tại sự không tin cậy hoàn toàn giữa các thực thể tham gia chung vào qui trình huấn luyện liên kết giữa các hệ thống mạng.

- Triển khai kiểm tra, đánh giá và tăng cường tính bền vững của các ứng dụng phát hiện xâm nhập và săn tìm mối đe dọa được xây dựng trên các mô hình học máy trong ngữ cảnh sơ đồ học liên kết giữa các hệ thống mạng SDN. Việc kiểm tra và đánh giá nhằm đảm bảo hiệu suất và độ tin cậy của các ứng dụng phát hiện xâm nhập và săn tìm mối đe dọa trong môi trường mạng SDN. Đồng thời, việc tăng cường tính bền vững giúp đảm bảo rằng các ứng dụng này có khả năng phát hiện và ngăn chặn các tấn công mạng mới và tiến hóa liên tục để tránh sự nhận diện từ phía kẻ tấn công.

- Thiết kế và triển khai cơ chế kiểm soát truy cập phi tập trung cho ứng dụng mạng, nhằm đảm bảo tính tin cậy, an toàn và dễ mở rộng cho bộ khung kiểm soát truy cập các ứng dụng mạng kết nối với bộ điều khiển SDN để thực hiện các tác vụ quản lý, giám sát, điều phối và thu thập thông tin về trạng thái mạng SDN. Cơ chế xác thực và kiểm soát truy cập trong mạng SDN đảm bảo rằng chỉ những ứng dụng đã được cấp phép mới có thể truy cập vào tài nguyên mạng. Đồng thời, việc đảm bảo tính dễ mở rộng giữa các miền và khả năng liên kết

giữa các ứng dụng và bộ điều khiển trong các tổ chức khác nhau cho phép tài nguyên mạng chỉ được sử dụng đúng mục đích và ngăn chặn các nguy cơ xâm nhập và tấn công từ các ứng dụng không được ủy quyền.

7.2. Hướng phát triển

Bên cạnh những kết quả đạt được, kết quả của luận án vẫn tồn tại một số hạn chế cần được tiếp tục phát triển, giải quyết trong tương lai. Một số khía cạnh có thể phát triển tiếp theo, bao gồm:

- Đánh giá hiệu suất của các mô hình liên kết phát hiện mối đe dọa bằng cơ chế thực hiện mã hóa đồng cấu trên phương pháp học liên kết sử dụng kỹ thuật cập nhật logit thay vì cập nhật trọng số mô hình lên máy chủ tổng hợp trung tâm thông qua kỹ thuật chắt lọc tri thức (knowledge distillation).

- Ngoài ra, vấn đề phát hiện và ngăn chặn các bản cập nhật bị nhiễm độc theo phương thức tấn công có chủ đích (targeted poisoning attack, back door attack); và dữ liệu không phải iid (không độc lập và phân phối giống hệt nhau) trong học liên kết sẽ được xem xét để giải quyết sự sai lệch của phân phối dữ liệu từ các thực thể không đáng tin cậy cho các trường hợp phổ biến trong thế giới thực.

- Nghiên cứu áp dụng mô hình Trí tuệ nhân tạo khả diễn giải (Explainable AI - XAI) để nghiên cứu cách thức phát sinh và phát hiện và ngăn chặn tấn công đối kháng hiệu quả hơn do tận dụng được sự lí giải của mô hình ứng với kết quả dự đoán mà nó đưa ra cho dữ liệu đầu vào nhất định.

- Xây dựng một bộ khung blockchain và cơ chế đồng thuận trên các nút trong mạng blockchain có tích hợp quá trình kiểm tra chất lượng mô hình huấn luyện được đóng góp từ các bên tham gia là một trong những hướng nghiên cứu tiềm năng.

- Cải thiện tốc độ giao dịch và thông lượng xử lý (throughput) trong mạng blockchain cũng là một khía cạnh có thể được xem xét để nâng cao hiệu năng xử lý của toàn bộ hệ thống thông qua nghiên cứu áp dụng kỹ thuật phân mảnh (sharding) khi có yêu cầu mở rộng kích thước mạng lưới blockchain.

- Nghiên cứu chiến lược phòng thủ với khả năng kháng mẫu đối kháng nhằm gia tăng độ bền vững của các ứng dụng phát hiện xâm nhập, săn tìm mối đe dọa trên không gian mạng. Ngoài ra, hướng nghiên cứu tạo mẫu đối kháng dưới dạng mức độ dữ liệu khác luồng mạng cũng có thể được xem xét như ở mức độ gói tin, payload.

CÁC CÔNG TRÌNH KHOA HỌC ĐÃ CÔNG BỐ

Nghiên cứu sinh đã công bố 5 công trình khoa học với vai trò là tác giả chính tại các Tạp chí khoa học quốc tế liên quan tới nội dung nghiên cứu của luận án:

[TC1] Phan The Duy, Le Khac Tien, Nghi Hoang Khoa, Do Thi Thu Hien, Anh Gia-Tuan Nguyen, Van-Hau Pham, "DIGFuPAS: Deceive IDS with GAN and Function-Preserving on Adversarial Samples in SDN-enabled networks", Computers & Security, Vol.109, 2021. (SCIE-Q1, Impact Factor: 5.105).

[TC2] Phan The Duy, Hien Do Hoang, Do Thi Thu Hien, Anh Gia-Tuan Nguyen, Van-Hau Pham, "B-DAC: A decentralized access control framework on Northbound interface for securing SDN using blockchain", Journal of Information Security and Applications, Vol. 64, 2022. (SCIE-Q1, Impact Factor: 4.96).

[TC3] Phan The Duy, Nghi Hoang Khoa, Do Thi Thu Hien, Hien Do Hoang, Van-Hau Pham, "Investigating on the Robustness of Flow-based Intrusion Detection System against Adversarial Samples using Generative Adversarial Networks ", Journal of Information Security and Applications, Vol. 74, May 2023. (SCIE-Q1, Impact Factor: 3.8).

[TC4] Phan The Duy, Nguyen Huu Quyen, Nghi Hoang Khoa, Tuan-Dung Tran, Van-Hau Pham, "FedChain-Hunter: A Reliable and Privacy-Preserving Aggregation for Federated Threat Hunting Framework in SDN-based IIoT", Internet of Things, Vol. 24, December 2023. (SCIE-Q1, Impact Factor: 5.9).

[TC5] Phan The Duy, Do Thi Thu Hien, Tran Duc Luong, Nguyen Huu Quyen, Van-Hau Pham, "Fed-Evolver: An Automated Evolving Approach for Federated Intrusion Detection System using Adversarial Autoencoder in SDN-enabled Networks", Internet of Things, Vol. 28, December 2024. (SCIE-Q1, Impact Factor: 6.0).

Ngoài ra, NCS cũng đã công bố các công trình khoa học với vai trò là tác giả chính tại Kỷ yếu hội nghị khoa học quốc tế liên quan tới nội dung nghiên cứu của luận án:

[CT1] Phan The Duy, Huynh Nhut Hao, Huynh Minh Chu, Van-Hau Pham (2021). A Secure and Privacy Preserving Federated Learning Approach for IoT Intrusion Detection System. In: Yang, M., Chen, C., Liu, Y. (eds) Network and System Security. NSS 2021. Lecture Notes in Computer Science(), vol 13041. Springer, Cham.

[CT2] Phan The Duy, Tran Van Hung, Nguyen Hong Ha, Hien Do Hoang and Van-Hau Pham, "Federated learning-based intrusion detection in SDN-enabled IIoT networks," 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), 2021, pp. 424-429.

[CT3] Phan The Duy, Tuan-Dung Tran, Nguyen Duy Hoang, Hien Do Hoang, Van-Hau Pham, "DeFL-BC: Empowering Reliable Cyberattack Detection through Decentralized Federated Learning and Poisoning Attack Defense", The 2023 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF).