TỔNG QUAN VỀ KỸ THUẬT BẢO MẬT ĐA TẦNG CHO HỆ THỐNG IOT TÍCH HỢP AI

Nguyễn Đăng Hiếu[1], Đào Xuân Phúc[2*]
Email: phucdx@hou.edu.vn. ORCID: 0009-0000-6217-8603

Ngày tòa soạn nhận được bài báo: 15/01/2026
Ngày phản biện đánh giá: 17/03/2026
Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1175

Tóm tắt: *Trong bối cảnh hệ sinh thái Internet vạn vật (IoT) phát triển nhanh chóng với quy mô hàng tỷ thiết bị, vấn đề bảo mật đang trở thành một thách thức mang tính sống còn đối với các hệ thống thông tin truyền thống. Sự hạn chế về tài nguyên tính toán và tính phân tán cao của các nút mạng biên đã làm suy yếu hiệu quả của các cơ chế phòng thủ cổ điển. Các kỹ thuật trí tuệ nhân tạo (AI) và học máy (ML), tiêu biểu như Random Forest, mạng Nơ-ron tích chập (CNN), mạng bộ nhớ dài - ngắn (LSTM), Autoencoder và BiLSTM, đang được áp dụng ngày càng sâu rộng trong nỗ lực nâng cao tính an toàn, linh hoạt và khả năng tự thích ứng của các hệ thống IoT. Bài viết tổng quan này tiến hành phân tích chi tiết kiến trúc phân lớp của hệ thống IoT hiện đại, đánh giá toàn diện các lỗ hổng bảo mật tại từng phân tầng, và khảo sát các phương pháp tiếp cận AI/ML tiên tiến đang được triển khai. Đặc biệt, nghiên cứu đi sâu vào các xu hướng công nghệ mới nổi định hình tương lai bảo mật IoT, bao gồm học liên kết (Federated Learning - FL) giúp bảo vệ quyền riêng tư, AI có khả năng giải thích (XAI) nhằm tăng tính minh bạch, các cơ chế bảo vệ tính toàn vẹn của mô hình AI, và kiến trúc phòng thủ đa tầng. Dựa trên các phân tích chuyên sâu, tác giả đề xuất một mô hình kiến trúc bảo mật IoT tích hợp AI đa tầng, đồng thời vạch ra các định hướng nghiên cứu chiến lược nhằm đáp ứng yêu cầu thực tiễn trong kỷ nguyên công nghiệp 4.0.*

Từ khóa: *internet vạn vật, trí tuệ nhân tạo, học máy, hệ thống phát hiện xâm nhập, học liên kết, AI có thể giải thích*

I. Đặt vấn đề

Sự tiến hóa vượt bậc của công nghệ Internet vạn vật (IoT) đã và đang tạo ra những chuyển dịch kiến tạo sâu sắc trong cách thức con người và máy móc tương tác với thế giới vật lý. Thông qua việc kết nối hàng tỷ thiết bị nhúng thông minh, hệ sinh thái IoT đóng vai trò hạt nhân

[1] Trường Đại học Kỹ thuật - Hậu cần Công an Nhân dân, Hà Nội, Việt Nam
[2] Khoa Điện - Điện tử, Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

trong các lĩnh vực trọng yếu như y tế từ xa, giao thông thông minh, tự động hóa công nghiệp 4.0 và quản trị năng lượng (Ahmed, 2022; Statista, 2024). Theo các báo cáo phân tích thị trường, quy mô thiết bị IoT toàn cầu dự kiến sẽ vượt ngưỡng 29,4 tỷ thiết bị vào năm 2030, tạo ra một không gian dữ liệu khổng lồ và liên tục (Statista, 2024). Tuy nhiên, chính sự gia tăng theo cấp số nhân về quy mô, đặc tính phân tán diện rộng và tính dị thể phức tạp của các thiết bị này đã vô tình mở rộng bề mặt tấn công mạng lên mức chưa từng có, đặt ra những bài toán hóc búa cho công tác quản trị rủi ro an toàn thông tin (PaloAlto, 2024; Nguyen, 2023).

Các kịch bản tấn công mạng hiện đại không còn giới hạn ở các phương thức đơn lẻ, mà có xu hướng tiến hóa thành các cuộc tấn công đa diện, có chủ đích (APT), khai thác đồng thời nhiều bề mặt dễ tổn thương. Các lỗ hổng trải dài từ thiết bị đầu cuối với tài nguyên hạn chế, các giao thức truyền dẫn mạng, cho đến các giao diện lập trình ứng dụng (API) và thậm chí là các kỹ nghệ xã hội nhắm vào con người (Wei, 2024; Hassan, 2024). Các cuộc tấn công điển hình như mạng botnet Mirai khai thác lỗ hổng mật khẩu mặc định của thiết bị IoT đã chứng minh sức tàn phá khủng khiếp của việc lợi dụng các thiết bị bảo mật yếu kém để thực hiện tấn công từ chối dịch vụ phân tán (DDoS) quy mô lớn (Antonakakis, 2017). Trong bối cảnh này, các phương thức bảo mật truyền thống đang bộc lộ những điểm yếu chí mạng, thiếu tính linh hoạt và gần như bất lực trước các biến thể tấn công zero-day (Laghari, 2024; Raeisi, 2024).

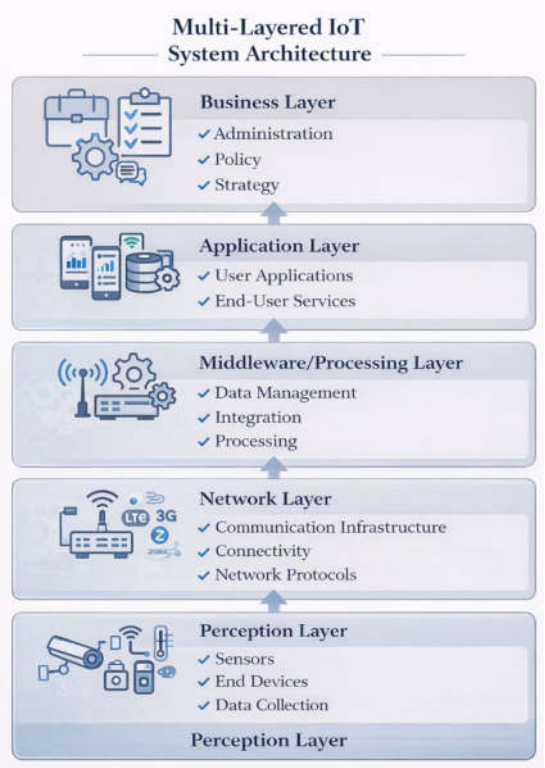
Để vượt qua những giới hạn cốt lõi của bảo mật truyền thống, trí tuệ nhân tạo (AI) và học máy (ML) đã nổi lên như một hệ hình mới, cung cấp khả năng tạo ra một "lớp phòng thủ thông minh". Nhờ sức mạnh trong việc xử lý lượng dữ liệu khổng lồ và nhận diện các mẫu hình phi tuyến, AI/ML cho phép hệ thống tự động thích nghi, học hỏi từ môi trường và đưa ra các quyết định phản ứng theo thời gian thực trước các mối đe dọa bất thường (Kikissagbe & Adda, 2024; Soofi, 2024). Bất chấp những tiềm năng to lớn, tiến trình hội nhập AI vào không gian IoT vẫn đối mặt với những rào cản kỹ thuật khắt khe. Yêu cầu lớn nhất nằm ở việc triển khai các mô hình tính toán cường độ cao trên các thiết bị biên (edge devices) vốn bị giới hạn nghiêm ngặt về năng lượng, bộ nhớ và khả năng xử lý (Compunnel, 2024).

Nhằm hệ thống hóa lại các bước tiến công nghệ trong lĩnh vực này, bài báo thực hiện một nghiên cứu tổng quan chuyên sâu về việc ứng dụng AI/ML trong kiến trúc bảo mật IoT. Điểm khác biệt của nghiên cứu này là việc phân loại các kỹ thuật học máy không chỉ dựa trên thuật toán, mà còn gắn chặt với từng phân tầng cụ thể trong kiến trúc hệ thống IoT hiện đại.

II. Cơ sở lý thuyết và kiến trúc hệ thống IoT

Hệ thống IoT hiện đại thường được thiết kế và triển khai dựa trên nguyên lý kiến trúc phân tầng. Việc phân tách này không chỉ giúp trừu tượng hóa các chức năng phần cứng/phần mềm để dễ dàng quản trị, mà còn cho phép áp dụng các chính sách bảo mật vi mô một cách hiệu quả. Mặc dù kiến trúc 3 lớp kinh điển từng được sử dụng rộng rãi, sự bùng nổ của điện toán biên và các dịch vụ đám mây phức tạp đã thúc đẩy giới nghiên cứu đề xuất các mô hình kiến trúc mở rộng, tiêu biểu là kiến trúc 5 lớp (Gupta, 2024; Hassan, 2024). Mô hình này bổ sung thêm lớp xử

lý trung gian và lớp quản lý dịch vụ, đặc biệt thiết yếu cho các hệ thống yêu cầu độ trễ thấp, phản ứng thời gian thực và đảm bảo tính riêng tư như y tế từ xa hay xe tự lái (Soofi, 2024).



Hình 1: Mô hình kiến trúc phân lớp IoT hiện đại (Nguồn tham khảo:Gupta, 2024)

Sự phân chia thành các lớp chuyên biệt đồng nghĩa với việc phát sinh các véctơ tấn công và lỗ hổng bảo mật mang tính đặc thù cho từng tầng phân rã như sau:

Lớp cảm nhận (Perception Layer): Đây là tuyến đầu của hệ thống, bao gồm các cảm biến, cơ cấu chấp hành và các vi điều khiển cấp thấp. Lớp này cực kỳ nhạy cảm trước các rủi ro bảo mật vật lý (tháo dỡ, sao chép firmware), tấn công nhiễu sóng, giả mạo nút mạng, hoặc bắt gói tin do không đủ khả năng chạy các giao thức mã hóa phức tạp (Wei, 2024; Laghari, 2024).

Lớp mạng (Network Layer): Đóng vai trò cầu nối dữ liệu, lớp này sử dụng đa dạng các giao thức truyền thông (Zigbee, LoRaWAN, 5G, Wifi). Bề mặt tấn công tại đây chủ yếu tập trung vào các nỗ lực đánh sập băng thông thông qua tấn công từ chối dịch vụ (DDoS), tấn công chèn ép định tuyến, tấn công định tuyến giả mạo và tấn công trung gian (Gupta, 2024; Netgear, 2024).

Lớp xử lý trung gian (Middleware/ Edge Layer): Nơi tập kết và tiền xử lý dữ liệu từ hàng ngàn thiết bị đầu cuối. Điểm yếu của phân tầng này nằm ở nguy cơ leo thang đặc quyền, lây nhiễm mã độc tĩnh/động trong quá trình luân chuyển tệp dữ liệu, và các lỗ hổng liên quan đến ảo hóa trên các trạm biên (Hassan, 2024; Bitdefender, 2024).

Lớp ứng dụng (Application Layer): Cung cấp dịch vụ trực tiếp cho người dùng cuối thông qua các giao diện web, di động. Lỗ hổng thường xuất phát từ việc thiếu cơ chế xác thực đa yếu tố, phân quyền lỏng lẻo, rò rỉ API, và các lỗi phần mềm truyền thống như SQL Injection hay XSS (Hassan, 2024).

Lớp quản lý dịch vụ (Business/ Management Layer): Tầng cao nhất quyết định logic kinh doanh và chiến lược vận hành. Tấn công vào lớp này thường là các cuộc tấn công tinh vi nhắm vào logic quản trị, thay đổi chính sách hệ thống hoặc khai thác dữ liệu tập trung quy mô lớn, gây hậu quả thảm khốc cho toàn bộ mạng lưới (Netgear, 2024; Bitdefender, 2024).

Việc ánh xạ chính xác các bề mặt tấn công này vào kiến trúc phân lớp là nền tảng cốt lõi để triển khai các mô hình AI/ ML một cách có mục tiêu. Các nghiên cứu học thuật gần đây liên tục khẳng định rằng AI không chỉ là một công cụ phân tích phụ trợ, mà phải được nhúng trực tiếp như một lớp phòng thủ tự thân tại mọi điểm nút mạng (Soofi, 2024; Hassan, 2024).

III. Phương pháp nghiên cứu

Nghiên cứu này tiếp cận việc đánh giá AI/ML thông qua lăng kính chức năng bảo mật. Các thuật toán học máy từ cơ bản đến học sâu được phân tách dựa trên khả năng của chúng trong việc giải quyết ba bài toán cốt lõi gồm phát hiện xâm nhập, phát hiện mã độc và phân tích hành vi bất thường. Việc áp dụng AI cho phép hệ thống vượt qua các ràng buộc của phương pháp dựa trên chữ ký, chuyển đổi sang phương pháp phân tích hành vi có khả năng đối phó với các cuộc tấn công Zero-day (Gueriani, 2024; Ferdous, 2024). Tuy vậy, bài toán tối ưu hóa tài nguyên phần cứng để duy trì khả năng suy diễn tốc độ cao tại thiết bị biên vẫn là trọng tâm của các phương pháp luận hiện nay.

3.1. Phát hiện xâm nhập

Các hệ thống phát hiện xâm nhập dành cho IoT đã chứng kiến một sự dịch chuyển công nghệ mạnh mẽ từ các luật heuristic cứng nhắc sang các mô hình mạng nơ-ron sâu. Các kiến trúc như mạng nơ-ron tích chập (CNN), mạng bộ nhớ dài-ngắn (LSTM), hay mô hình lai thể hiện sự vượt trội trong việc khai thác các đặc trưng không gian và thời gian từ dòng lưu lượng mạng cực lớn (Mahanipour & Khamfroush, 2024; Diro & Chilamkurti, 2018). Ví dụ, một mô hình CNN khi được tối ưu hóa với các kỹ thuật lựa chọn đặc trưng tự động có thể đạt độ chính xác lên tới 98.1% trên bộ dữ liệu chuẩn CICIDS2017 (Mahanipour & Khamfroush, 2024). Dựa vào kiến trúc phân tầng, AI-IDS có thể được cấu trúc như sau:

Tại lớp mạng: Các mô hình LSTM được triển khai tại các Gateway để phân tích chuỗi các gói tin mạng theo thời gian thực, từ đó nhận diện các cuộc tấn công DDoS hay quét cổng với độ trễ tối thiểu (Mahanipour & Khamfroush, 2024).

Tại lớp xử lý trung gian: Áp dụng các thuật toán học tăng cường Sâu để liên tục tối ưu hóa các chính sách chặn lọc lưu lượng, giúp ngăn chặn sự lây lan của các botnet tinh vi lẩn khuất trong các cụm thiết bị biên (Gueriani, 2024).

Tại lớp quản lý: Sử dụng mạng nơ-ron hồi quy để đối chiếu và phát hiện những sai lệch trong luồng hành vi điều khiển từ máy chủ trung tâm xuống thiết bị, giảm thiểu rủi ro từ các cuộc tấn công nhắm vào hạ tầng quản trị.

3.2. Phát hiện phần mềm độc hại (Malware Detection)

Bối cảnh IoT với sự đa dạng của các kiến trúc vi xử lý tạo điều kiện cho malware biến hình dễ dàng lẩn trốn. Các thuật toán học máy truyền thống kết hợp với trích xuất đặc trưng tĩnh và động đã được chứng minh là cực kỳ hiệu quả. Theo nghiên cứu của Ferdous 2024, các mô hình rừng ngẫu nhiên (Random Forest - RF) và máy học véc-tơ hỗ trợ (SVM) tỏ ra rất mạnh mẽ khi hoạt động với các bộ dữ liệu đặc trưng mã nguồn IoT. Trong khi đó, các mạng phức tạp như Deep Belief Network (DBN) có khả năng tự động học các biểu diễn phi tuyến từ chuỗi mã byte của phần mềm độc hại. Chiến lược phân bổ tại các lớp:

Lớp cảm nhận: Triển khai mô hình SVM hoặc Random Forest dung lượng nhỏ trực tiếp trên firmware thiết bị để kiểm tra tính toàn vẹn và phân tích mã độc tĩnh trước khi thực thi.

Lớp trung gian: Triển khai DBN hoặc các hộp cát (sandboxes) sử dụng AI để quan sát hành vi động của các tập tin

khả nghi khi chúng được truyền tải giữa đám mây và biên (Ferdous, 2024).

Lớp ứng dụng: AI đóng vai trò như một lớp bảo vệ thời gian chạy (runtime protection), kiểm duyệt các API call và ngăn chặn các đoạn script độc hại tiêm nhiễm vào quy trình hoạt động của thiết bị.

3.3. Phát hiện hành vi bất thường (Anomaly Detection)

Trái ngược với việc tìm kiếm các mẫu tấn công đã biết, phát hiện bất thường tập trung vào việc mô hình hóa trạng thái "bình thường" của hệ thống và gióng lên hồi chuông cảnh báo trước mọi độ lệch chuẩn. Các thuật toán học không giám sát như K-means, DBSCAN và đặc biệt là Autoencoder (AE) đóng vai trò xương sống trong phương pháp này. Việc sử dụng Autoencoder trong mô hình học tự giám sát liên kết tại biên không chỉ cho phép thiết bị học thuộc các hành vi cục bộ một cách độc lập mà còn giảm thiểu việc trao đổi dữ liệu thô, qua đó tăng cường bảo mật dữ liệu (Gelenbe, 2024a,b).

Tại lớp cảm nhận: Sử dụng các thuật toán gom cụm (Clustering) tiêu tốn ít năng lượng để khoanh vùng và cách ly các dữ liệu cảm biến bất thường, phòng ngừa tấn công tiêm dữ liệu sai lệch (Hassan, 2024).

Tại lớp xử lý trung gian: Các bộ giải mã tự động đóng vai trò nén và tái tạo dữ liệu, mọi gói dữ liệu không thể tái tạo hoàn chỉnh sẽ bị gán cờ bất thường (Gelenbe, 2024a).

Tại lớp ứng dụng: Ứng dụng mô hình chuỗi mạng nơ-ron BiLSTM hoặc Transformer dạng nhẹ để lập hồ sơ thói quen người dùng, ngay lập tức vô hiệu hóa các phiên đăng nhập giả mạo, dù chúng sử dụng đúng thông tin xác thực (Gelenbe, 2024b).

IV. Kết quả và thảo luận

4.1. So sánh đánh giá các kỹ thuật AI/ML

Để cung cấp một cái nhìn khách quan về tính khả thi khi áp dụng AI vào IoT, Bảng 1 trình bày một sự so sánh chi tiết giữa các thuật toán tiêu biểu. Các tiêu chí đánh giá được thiết lập dựa trên các yêu cầu khắt khe của môi trường IoT thực tiễn: độ chính xác, độ trễ suy diễn, ngân sách năng lượng, tài nguyên phần cứng yêu cầu, và khả năng giải thích được (Explainability) - một yếu tố ngày càng được coi trọng trong đánh giá rủi ro thuật toán (Gueriani, 2024; Ferdous, 2024; Gelenbe, 2024a).

Các mô hình học máy kinh điển (SVM, RF) với yêu cầu tài nguyên thấp và tính minh bạch cao phù hợp hoàn hảo cho các nút mạng biên ở lớp cảm nhận. Ngược lại, sức mạnh phân tích luồng của học sâu (CNN, LSTM) lại thể hiện tại các Gateway hoặc bộ điều khiển trung tâm ở lớp mạng, dù phải đánh đổi bằng chi phí tính toán cao (Mahanipour & Khamfroush, 2024; Ferdous, 2024). Điều này khẳng định luận điểm rằng việc phòng thủ hệ thống IoT cần phải dựa trên một kiến trúc lai, kết hợp đan xen các thuật toán khác nhau ở từng tầng cụ thể.

4.2. Các thách thức hiện hữu và rào cản kỹ thuật

Sự kết hợp giữa AI và IoT tạo ra một hệ thống phòng thủ tiên tiến, nhưng bản thân sự kết hợp này cũng tiềm ẩn các rủi ro nội tại nghiêm trọng:

Bảng 1: So sánh các kỹ thuật AI/ML ứng dụng trong bảo mật IoT

Kỹ thuật	Đặc tính kỹ thuật	Chính xác	Độ trễ	Tài nguyên	Tính giải thích	Phân lớp & Ứng dụng chính
Random Forest	Xử lý phi tuyến tốt / Kém với dữ liệu chuỗi thời gian	Trung bình	Trung bình	Thấp	Trung bình	Cảm nhận/Ứng dụng: Phát hiện mã độc ngoại tuyến
SVM	Phân tách tốt với dữ liệu ít / Rất khó mở rộng quy mô lớn	Cao	Trung bình	Trung bình	Cao	Cảm nhận/Ứng dụng: Phân loại mã độc, bất thường
DBN	Biểu diễn đặc trưng sâu / Huấn luyện rất tốn thời gian	Cao	Cao	Cao	Thấp	Trung gian: Phân tích hành vi, mã độc
CNN	Khai thác hiệu quả dữ liệu ma trận / Phụ thuộc vào NPU/GPU	Rất cao	Trung bình	Cao	Thấp	Mạng/Trung gian: IDS dựa trên lưu lượng mạng
LSTM	Tối ưu cho dữ liệu chuỗi dài / Nguy cơ quá khớp, tính toán chậm	Cao	Cao	Cao	Thấp	Mạng/Quản lý: IDS theo thời gian thực
Autoencoder	Học không giám sát xuất sắc / Nhạy cảm với nhiễu dữ liệu	Trung bình	Trung bình	TB - Cao	Thấp	Trung gian/Ứng dụng: Phát hiện hành vi bất thường
BiLSTM	Mô hình hóa chuỗi ngữ cảnh 2 chiều / Rất nặng về tính toán	Cao	Cao	Cao	Trung bình	Ứng dụng: Giám sát truy cập bất thường
K-means / DBSCAN	Thuật toán nhẹ, không cần gán nhãn / Kém với không gian phức tạp	Thấp	Thấp	Rất thấp	Cao	Cảm nhận/Ứng dụng: Phát hiện bất thường cục bộ
Transformer nhẹ	Khả năng xử lý chuỗi song song / Yêu cầu tinh chỉnh mạng cho biên	Cao	Trung bình	TB - Cao	Cao	Ứng dụng: Phân tích hành vi người dùng

Nguy cơ chống lại mô hình AI (Adversarial Machine Learning): Mô hình AI rất dễ bị đánh lừa bởi các cuộc tấn công đối kháng hoặc đầu độc dữ liệu. Kẻ tấn công có thể chèn các nhiễu vi mô vào lưu lượng mạng khiến mô hình IDS nhận diện sai lệch, biến AI trở thành mắt xích yếu nhất trong hệ thống (Aouedi, 2024; Xu, 2024). Việc ứng dụng các cơ chế thủy vân mô hình (Model Watermarking) đang được đẩy mạnh để xác minh tính toàn vẹn của mô hình (Zhang, 2025). Khủng hoảng niềm tin và tính giải thích: Đặc tính "hộp đen"của học sâu cản trở các nhà quản trị mạng hiểu được lý do tại sao một cảnh báo được đưa ra. Thiếu hụt công cụ AI có thể giải thích làm giảm khả năng truy vết sự cố pháp y và gây khó khăn trong việc đáp ứng các tiêu chuẩn tuân thủ pháp lý toàn cầu (Baral, 2024). Quyền riêng tư dữ liệu: Việc tập trung hóa luồng dữ liệu thô từ hàng tỷ thiết bị IoT về đám mây để huấn luyện mô hình vi phạm nghiêm trọng các quy định về quyền riêng tư, đòi hỏi sự thay đổi trong phương pháp luận huấn luyện mô hình (Uddin, 2024; Compunnel, 2024).

4.3. Hướng phát triển và Kiến trúc đề xuất

Để vượt qua các rào cản trên, tương lai của bảo mật AI-IoT định hình qua ba trụ cột công nghệ chiến lược:

Tối ưu hóa Mô hình và TinyML: Kỹ thuật lượng tử hóa và cắt tỉa trọng số cho phép thu nhỏ các mô hình học sâu khổng lồ xuống kích thước vài Kilobyte để chạy trực tiếp trên vi điều khiển. Ví dụ, một mô hình MobileNetV2 siêu nhẹ có thể phát hiện DDoS trên bo mạch Raspberry Pi Zero W với độ trễ dưới 300ms, đạt độ chính xác trên 91.8% (Schmitt, 2023).

Học liên kết: Chuyển đổi mô hình học máy từ tập trung sang phân tán. FL cho phép thiết bị chỉ chia sẻ trọng số mô hình lên máy chủ thay vì gửi dữ liệu thô, giải quyết triệt để bài toán quyền riêng tư. Theo nghiên cứu (Aouedi, 2024), FL kết hợp với Học liên tục giúp hệ thống tự động cập nhật tri thức mới, giảm 65% rủi ro rò rỉ dữ liệu.

Hình 2: Mô hình kiến trúc bảo mật đa tầng đề xuất

Trí tuệ nhân tạo có thể giải thích: Tích hợp các thuật toán giải thích như SHAP, LIME, hoặc Grad-CAM vào hệ thống IDS. Các kỹ thuật này giúp biểu diễn trực quan mức độ đóng góp của từng đặc trưng vào quyết định cảnh báo của AI, hỗ trợ đắc lực cho quản trị viên (Baral, 2024).

Dựa trên các phân tích công nghệ, tác giả đề xuất một kiến trúc bảo mật đa tầng tích hợp AI cho hệ thống IoT, được minh họa tại hình 2

Phân bổ logic của kiến trúc đề xuất:

Lớp cảm biến (Biên siêu nhỏ): Nhúng công nghệ TinyML kết hợp với các thuật toán K-means hoặc SVM nhẹ để liên tục thực hiện xác thực hành vi thiết bị dựa trên đặc điểm tiêu thụ năng lượng và thói quen gửi gói tin, loại bỏ thiết bị bị xâm phạm ngay từ phần cứng (Aouedi., 2024; Schmitt, 2023). Lớp mạng (Gateway): Triển khai các hệ thống IDS dựa trên phần cứng chuyên dụng chạy thuật toán CNN hoặc LSTM lượng tử hóa, chặn đứng các cuộc bão lưu lượng (DDoS) và phân tích các gói tin bất thường ở tốc độ mạng (Mahanipour & Khamfroush, 2024).

Lớp điều phối (Fog/Edge Nodes): Đóng vai trò hạt nhân trong mạng học liên kết (FL). Các trạm biên sẽ điều phối quá trình tổng hợp trọng số từ các thiết bị trực thuộc, đảm bảo mô hình AI của toàn hệ thống không ngừng được nâng cấp mã hóa mà không xâm phạm quyền riêng tư (Aouedi, 2024).

Lớp Giám sát Đám mây (Central Cloud): Sử dụng hạ tầng điện toán khổng lồ để đối chiếu (correlation) toàn cảnh về các mối đe dọa dai dẳng (APT). Tầng này bắt buộc áp dụng công nghệ XAI để cung cấp báo cáo suy luận học thuật rõ ràng cho chuyên gia phân tích sự cố

(SOC Analysts) (Baral, 2024). Kiến trúc này không chỉ coi AI là một công cụ phân tích rời rạc, mà biến nó thành các mạch máu của một cơ thể tự miễn dịch, có khả năng phát hiện sớm, cách ly tức thì và tự học hỏi qua thời gian.

V. Kết luận

Bài viết đã hoàn thành việc hệ thống hóa và phân tích một cách toàn diện các ứng dụng của trí tuệ nhân tạo và học máy trong kiến trúc bảo mật Internet vạn vật (IoT). Bằng việc tiếp cận theo mô hình phân lớp mạng, nghiên cứu đã chỉ rõ rằng không tồn tại một giải pháp AI đơn lẻ nào đủ sức chống đỡ các dạng thức tấn công đa diện hiện nay. Mọi lớp từ thiết bị vật lý, hạ tầng truyền tải đến nền tảng đám mây đều mang theo những lỗ hổng chí mạng riêng biệt.

Đóng góp quan trọng của nghiên cứu là việc định vị và đề xuất một mô hình bảo mật AI- IoT đa tầng và phân tán, nơi sự giao thoa của TinyML, học sâu, học liên kết và AI có thể giải thích tạo ra một màng lưới phòng ngự chủ động. Việc tối ưu hóa mô hình học máy cho môi trường ràng buộc tài nguyên, đồng thời bảo đảm tính toàn vẹn quyền riêng tư dữ liệu sẽ là tiêu chuẩn tối thượng cho thế hệ IoT tương lai. Bài báo hy vọng cung cấp một khung tham chiếu chiến lược (framework) giá trị dành cho các nhà nghiên cứu, kiến trúc sư hệ thống và kỹ sư an toàn thông tin trong hành trình xây dựng một hệ sinh thái kết nối vạn vật thực sự thông minh, kiên cường và bền vững.

Tài liệu tham khảo

Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., ... & Jha, S. (2022). A survey of COVID-19 contact tracing apps. *IEEE Access, 8*, 134577-134601.

Aouedi, O., Piamrat, K., Parrein, B., & Kamal, A. E. (2024). A survey on intelligent Internet of Things: Applications, security, privacy, and future directions. *IEEE Communications Surveys & Tutorials, 27*(2), 1238-1292.

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., et al. (2017). Understanding the Mirai botnet. In *Proceedings of the 26th USENIX Security Symposium* (pp. 1093-1110).

Baral, S., Pahl, C., & Helmer, S. (2024). An adaptive end-to-end IoT security framework using explainable AI and LLMs. *arXiv preprint arXiv:2409.13177*.

Bitdefender. (2024). *The 2024 IoT security landscape report*. Bitdefender Research Publications.

Compunnel. (2024). *Addressing 2024's IoT security challenges within compliance frameworks*. Compunnel Technical Report.

Diro, A. A., & Chilamkurti, N. (2018). Distributed attack detection scheme using deep learning approach for IoT. *Future Generation Computer Systems, 82*, 761-768.

Ferdous, J., Islam, R., Mahboubi, A., & Islam, M. Z. (2024). A survey on ML techniques for multi-platform malware detection. *Sensors, 25*(4), 1153.

Gelenbe, E., Domanska, J., Kadioglu, Y., & Domanski, A. (2024). DISFIDA: Distributed self-supervised federated intrusion detection algorithm. *Internet of Things, 28*, 101340.

Gueriani, A. K., Mazari, H., & Cherif, A. (2024). Deep reinforcement learning for intrusion detection in IoT: A survey. *ResearchGate*.

Gupta, P., Verma, D. K., & Gupta, A. (2024). Unveiling the layered architecture of IoT. In *Handbook of research on AI-enabled smart healthcare systems* (pp.

150-175). IGI Global.

Hassan, A., Nizam-Uddin, N., & Quddus, A. (2024). Navigating IoT security: Architecture, attacks, and AI-driven solutions. *Computers, Materials & Continua, 81*(3), 3499-3599.

Kikissagbe, B. R., & Adda, M. (2024). Machine learning-based intrusion detection methods in IoT systems. *Electronics, 13*(18), 3601.

Laghari, A. A., Li, H., Khan, A. A., & Shoulin, Y. (2024). Internet of Things applications: Security trends and challenges. *Discover Internet of Things, 4*(1), 1-15.

Nguyen, T. T., Tran, N. Q., & Pham, V. C. (2023). Federated learning for privacy-preserving IoT security. *IEEE Internet of Things Journal, 10*(12), 10582-10595.

Netgear. (2024). *The 2024 IoT security landscape report*. NETGEAR Security Team.

Statista. (2024). *Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030*. Statista Research Department

Schmitt, M. (2023). Securing the digital world: Protecting smart infrastructures with AI-enabled IDS. *arXiv:2401.01342*.

Soofi, A. A., Tahir, M., & Raza, N. (2024). Securing the Internet of Things: A review of security challenges and AI solutions. *FUJEAS, 4*(2), 112-128.

Wei, Z., Wei, Q., Geng, Y., & Yang, Y. (2024). A survey on IoT security: Vulnerability detection and protection. In *Proceedings of the International Conference on Artificial Intelligence of Things and Computing (AITC)* (pp. 1-8).

Zhang, et al. (2025). Digital watermarking for virtual physically unclonable function data concealment and authentication. *Scientific Reports*.

A COMPREHENSIVE SURVEY OF MULTI-LAYERED SECURITY TECHNIQUES FOR AI-INTEGRATED IOT SYSTEMS

Nguyen Dang Hieu[1], Dao Xuan Phuc[2]

Abstract: *In the rapidly evolving landscape of the Internet of Things (IoT) ecosystem, characterized by billions of connected devices, security has emerged as a critical challenge for traditional information systems. The inherent resource constraints and highly distributed nature of edge nodes have significantly undermined the efficacy of classical defense mechanisms. Artificial Intelligence (AI) and Machine Learning (ML) techniques, notably Random Forest, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTMs), Autoencoders, and BiLSTM, are increasingly being deployed to enhance the safety, flexibility, and autonomous responsiveness of IoT architectures. This comprehensive survey meticulously analyzes the layered architecture of modern IoT systems, critically evaluates the security vulnerabilities inherent at each stratum, and reviews prominent, state-of-the-art AI/ML methodologies currently in deployment. Furthermore, the research delves into emerging technological trends shaping the future of IoT security, including privacy-preserving Federated Learning (FL), Explainable AI (XAI) for enhanced transparency, model integrity protection mechanisms, and multi-tiered defense architectures. Grounded in profound analytical insights, the author proposes a multi-layered AI-integrated IoT security architecture spanning from the edge to the cloud, while simultaneously charting strategic research directions to address practical demands in the Industry 4.0 era.*

Keywords: *internet of things, artificial intelligence, machine learning, intrusion detection system, federated learning, explainable AI*

[1] People's Police University of Technology and Logistics, Hanoi, Vietnam
[2] Faculty of Electric and Electronic Engineering, Hanoi Open University, Hanoi, Vietnam