Phát hiện và phân loại Botnet IoT bằng cách sử dụng thuật toán học máy

Phạm Văn Quân1 , Ngô Văn Úc1 , Đỗ Phúc Hào2,3 , Nguyễn Năng Hùng Văn4

[1] Đại học Đông Á, Đà Nẵng, Việt Nam

[2] Đại học Viễn thông Bang Bonch-Bruevich Saint-Petersburg, Saint-Petersburg, Liên bang Nga

[3] Đại học Kiến trúc Đà Nẵng, Đà Nẵng, Việt Nam

[4] Đại học Bách khoa Đà Nẵng, Đà Nẵng, Việt Nam

Tác giả liên hệ: Nguyễn Năng Hùng Vân, Email: nguyenvan@dut.udn.vn

Ngày nhận bài: xxx, ngày sửa chữa: xxx, ngày đăng: xxx

Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1221

Tóm tắt: Bài nghiên cứu học thuật này đề cập đến thách thức quan trọng và phức tạp trong việc phát hiện và phân loại các mạng botnet Internet vạn vật (IoT) thông qua việc sử dụng các thuật toán học máy. Nghiên cứu tập trung vào việc tiến hành phân tích tỉ mỉ và thao tác dữ liệu mạng botnet IoT, với sự nhấn mạnh cụ thể vào bộ dữ liệu IoT-23 đư ợc thừa nhận rộng rãi. Mục đích chính là sử dụng các thuật toán học máy đư ợc sử dụng rộng rãi và đư ợc công nhận rộng rãi, bao gồm Cây quyết định (DT), k-Láng giềng gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cư ờng độ dốc eXtreme (XGBoost), với mục đích phân loại và phát hiện botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23. Bằng cách triển khai các thuật toán này, bài báo tìm cách nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng trong lĩnh vực phát hiện và phân loại mạng botnet IoT. Việc thực hiện phân tích so sánh, đối chiếu các kết quả thu đư ợc từ các thuật toán đa dạng, sẽ cung cấp những hiểu biết vô giá về giá trị và hạn chế tư ơ ng ứng của chúng, từ đó cho phép các nhà nghiên cứu và ngư ời thực hành đư a ra quyết định sáng suốt liên quan đến thuật toán phù hợp nhất để đạt đư ợc sự phát hiện và phân loại mạng botnet IoT thành công.

Từ khóa: Internet vạn vật, Học có giám sát, Phát hiện xâm nhập, IoT Botnet, Học máy.

Tiêu đề: Phát hiện và phân loại Botnet IoT bằng thuật toán học máy

Tóm tắt: Bài nghiên cứu học thuật này đề cập đến thách thức quan trọng và phức tạp trong việc phát hiện và phân loại các mạng botnet Internet vạn vật (IoT) thông qua việc sử dụng các thuật toán học máy. Nghiên cứu tập trung vào việc tiến hành phân tích tỉ mỉ và thao tác dữ liệu mạng botnet IoT, với sự nhấn mạnh cụ thể vào bộ dữ liệu IoT-23 đư ợc thừa nhận rộng rãi. Mục đích chính là sử dụng các thuật toán học máy đư ợc sử dụng rộng rãi và đư ợc công nhận rộng rãi, bao gồm Cây quyết định (DT), k-Láng giềng gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cư ờng độ dốc eXtreme (XGBoost), với mục đích phân loại và phát hiện botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23.
Bằng cách triển khai các thuật toán này, bài báo tìm cách nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng trong lĩnh vực phát hiện và phân loại mạng botnet IoT. Việc thực hiện phân tích so sánh, đối chiếu các kết quả thu đư ợc từ các thuật toán đa dạng, sẽ cung cấp những hiểu biết vô giá về giá trị và hạn chế tư ơ ng ứng của chúng, từ đó cho phép các nhà nghiên cứu và ngư ời thực hành đư a ra quyết định sáng suốt liên quan đến thuật toán phù hợp nhất để đạt đư ợc sự phát hiện và phân loại mạng botnet IoT thành công.

Từ khóa: Internet vạn vật, Học có giám sát, Phát hiện xâm nhập, IoT Botnet, Học máy.

I. GIỚI THIỆU

Khái niệm Internet of Things (IoT) ban đầu đư ợc đề xuất bởi K.Ashton (1999) như một giải pháp cho số lư ợng ngày càng tăng của các thiết bị yêu cầu kết nối internet. Cơ quan An ninh mạng và thông tin của Liên minh châu Âu định nghĩa IoT là một hệ sinh thái vật lý không gian mạng bao gồm các cảm biến và bộ truyền động đư ợc kết nối với nhau để tạo điều kiện quá trình tạo ra quyết định. Đáng chú ý, IoT là một công nghệ quan trọng trong Công nghiệp 4.0 [1]. IoT công nghiệp (IIoT) đại diện cho việc triển khai IoT chuyên biệt kết nối các động cơ và linh kiện công nghiệp để nâng cao năng suất và hiệu suất của các hoạt động công nghiệp. IIoT đạt đư ợc mục tiêu này bằng cách cung cấp khả năng giám sát theo thời gian thực, quản lý hiệu quả và kiểm soát các quy trình công nghiệp, tài sản và thời gian hoạt động. Hơ n nữa, IIoT nhằm mục đích giảm chi phí hoạt động trong khi

nâng cao hiệu quả tổng thể [2].

Tuy nhiên, việc phát triển các hệ thống IoT đư a ra những thách thức bảo mật đáng kể. Các thiết bị IoT thư ờng có khả năng quản lý tài nguyên và bảo mật hạn chế. Bên cạnh việc phát triển các hệ thống IoT, cũng cần có các biện pháp bảo mật hiệu quả để bảo vệ chống lại các mối đe dọa mạng và vi phạm dữ liệu [3].

Ngăn chặn các cuộc tấn công và xâm nhập khác nhau và bảo vệ dữ liệu là rất cần thiết. Hiện nay, có rất nhiều hệ thống phát hiện xâm nhập (IDS) đã đư ợc phát triển để giải quyết vấn đề này. IDS có nhiệm vụ phát hiện và báo cáo xâm nhập hệ thống. Ngoài ra, IDS ngăn chặn phần mềm độc hại khi tấn công và duy trì hiệu suất trong khi tấn công. IDS là một thành phần quan trọng của hệ thống bảo mật hiện đại[4].

Mạng botnet IoT là một mạng gồm các thiết bị IoT bị xâm nhập do các tác nhân độc hại kiểm soát để thực hiện các hoạt động độc hại. Các mạng botnet này khai thác các lỗ hổng trong thiết bị IoT để có quyền truy cập trái phép và tạo ra một mạng lư ới lớn đư ợc kết nối với nhau gồm các thiết bị bị xâm nhập. Các tính năng chính của botnet IoT bao gồm tính đa dạng của thiết bị, khai thác lỗ hổng, cơ sở hạ tầng chỉ huy và kiểm soát, tấn công DDoS, khả năng gửi thư rác và lừa đảo, khai thác tiền điện tử cũng như đánh cắp dữ liệu và vi phạm quyền riêng tư . Để giải quyết các mối đe dọa do mạng botnet IoT gây ra, điều quan trọng là phải tăng cư ờng bảo mật thiết bị IoT, triển khai các cơ chế xác thực mạnh, thư ờng xuyên cập nhật phần mềm, phân đoạn mạng hoạt động và hư ớng dẫn ngư ời dùng về các biện pháp bảo mật tốt nhất.

Động lực đằng sau việc tạo ra một bài báo học thuật tập trung vào việc xác định và phân loại các mạng botnet Internet vạn vật (IoT) thông qua các kỹ thuật máy học phát sinh từ nhu cầu cấp thiết phải giải quyết bối cảnh nguy hiểm leo thang do các mạng ác ý này gây ra. Không thể phủ nhận sự phổ biến của các thiết bị IoT đư ợc kết nối với nhau đã khuếch đại tiềm năng tấn công mạng botnet, do đó gây nguy hiểm cho sức khỏe của cả ngư ời dùng cá nhân và cơ sở hạ tầng thiết yếu. Bằng cách khai thác sức mạnh của các phư ơ ng pháp học máy, chúng ta có thể nâng cao trình độ của mình trong việc nhận biết và phân loại các mạng botnet IoT, từ đó tạo điều kiện thuận lợi cho việc thực hiện các biện pháp chủ động để ngăn chặn và giảm bớt hậu quả nguy hiểm của chúng. Việc sử dụng các thuật toán máy học cho phép chúng tôi xem xét kỹ lư ỡng khối lư ợng dữ liệu lư u lư ợng truy cập mạng đáng kể, tạo điều kiện thuận lợi cho việc xác định các mẫu thông thư ờng và các điểm bất thư ờng, đồng thời phân biệt hành vi thông thư ờng của thiết bị IoT với các hoạt động của mạng botnet. Bằng cách đắm mình trong lĩnh vực tìm hiểu này, chúng ta có thể đóng góp vào việc xây dựng các cơ chế phòng thủ mạnh mẽ và hiệu quả, từ đó củng cố tính bảo mật của hệ sinh thái IoT.

Mục tiêu chính của tài liệu nghiên cứu học thuật này là giải quyết thách thức then chốt trong việc phát hiện và phân loại các botnet IoT thông qua việc sử dụng các thuật toán máy học. Nghiên cứu nhấn mạnh đáng kể vào các phân tích tỉ mỉ và thao tác dữ liệu mạng botnet IoT, tập trung cụ thể vào bộ dữ liệu IoT-23 nổi tiếng.
Mục đích chính là triển khai các thuật toán học máy đư ợc sử dụng rộng rãi và đư ợc công nhận rộng rãi, cụ thể là Cây quyết định (DT), k-Láng giềng gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cư ờng độ dốc eXtreme (XGBoost), với mục đích phân loại và phát hiện các botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23. Việc áp dụng các thuật toán này nhằm mục đích nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng khi áp dụng cho nhiệm vụ phát hiện và phân loại mạng botnet IoT. Việc thực hiện phân tích so sánh, đặt cạnh các kết quả thu đư ợc từ các thuật toán đa dạng, sẽ mang lại những hiểu biết vô giá về các điểm mạnh và hạn chế tư ơ ng ứng của chúng, cho phép các nhà nghiên cứu và các nhà thực hành đư a ra quyết định sáng suốt liên quan đến thuật toán phù hợp nhất để phát hiện và phân loại thành công các botnet IoT.

II. CÔNG TRÌ NH LIÊN QUAN

Trong những năm gần đây, nghiên cứu về phát hiện xâm nhập IoT đã đư ợc chú ý nhiều hơ n. Nghiên cứu lại về chủ đề này đang trở nên cần thiết hơ n. Dựa trên các kết quả tìm kiếm trư ớc đó, có nhiều giải pháp để phát triển hệ thống phát hiện xâm nhập IoT. Việc hiểu rõ những điểm mạnh cũng như những vấn đề của các nghiên cứu trư ớc giúp ngư ời nghiên cứu tìm ra hư ớng nghiên cứu mang lại lợi ích lớn cho việc xây dựng hệ thống IDS. Do đó, các nghiên cứu trư ớc đây là nền tảng trong việc cung cấp kiến thức cho sự phát triển trong tư ơ ng lai.

Sự phổ biến của các thiết bị IoT đã mở ra một mối đe dọa ngày càng leo thang do các mạng botnet IoT gây ra, mang đến cho tội phạm mạng những con đư ờng mới để dàn dựng các cuộc tấn công quy mô lớn. Nhiệm vụ cấp thiết là phát hiện và phân loại các mạng botnet này có tầm quan trọng đặc biệt trong việc hạn chế rủi ro đang phát triển. Tận dụng khả năng của các kỹ thuật máy học đã nổi lên như một chiến lư ợc mạnh mẽ để tăng cư ờng các biện pháp an ninh mạng bằng cách tạo điều kiện thuận lợi cho việc xác định các mẫu đặc biệt, điểm bất thư ờng và hành vi xấu ăn sâu trong lư u lư ợng mạng mê cung của các botnet IoT. Điều này cho thấy tính tất yếu của việc triển khai các thuật toán máy học để giải quyết những thách thức phức tạp vốn có trong các mạng botnet IoT.

J. Hajji et al. [5] đã tiến hành một nghiên cứu để áp dụng các thuật toán học máy không giám sát, bao gồm K-means, PCA và Autoencoder, để phát hiện lư u lư ợng mạng bất thư ờng. MỘT. Rahim và cộng sự. [6] đã sử dụng phân tích thống kê để xác định các tính năng quan trọng nhất và sau đó áp dụng một số thuật toán học máy, chẳng hạn như DT, RF và Naive Bayes (NB), để phân loại lư u lư ợng thông thư ờng và độc hại.

SMZ Hồi giáo et al. [7] đã nghĩ ra các mô hình tính trung bình và xếp chồng dựa trên Support Vector Machine (SVM), RF và

Thuật toán tăng cư ờng độ dốc. Y. Li và cộng sự. [8] đã phát triển một mô hình đóng gói từ bốn mô hình học máy dựa trên DT, KNN, Hồi quy logistic (LR) và RF để phân loại lư u lư ợng mạng là bình thư ờng hoặc độc hại.

PH Do [9] đã đề xuất một phư ơ ng pháp trích xuất đặc trư ng bằng cách chia các tập đặc trư ng thành các lớp khác nhau, tiếp theo là ứng dụng thuật toán học máy cho các tập đặc trư ng ở các lớp cổng nạp. Những nghiên cứu này đã sử dụng một loạt các thuật toán, bao gồm một số kết hợp, để đạt đư ợc độ chính xác cao trong các nhiệm vụ phân loại của chúng.

Một số nhà nghiên cứu đã khám phá tiềm năng của các mô hình học sâu để phát hiện hành vi bất thư ờng trong lư u lư ợng mạng. Alotaibi et al. [10] đã phát triển mô hình dựa trên Mạng thần kinh chuyển đổi (CNN) để phân loại lư u lư ợng mạng là bình thư ờng hoặc độc hại. Lý và cộng sự. [11] đã sử dụng Bộ nhớ dài hạn ngắn hạn (LSTM) và Mạng nơ -ron dày đặc để phân loại lư u lư ợng mạng là bình thư ờng hoặc độc hại. Tư ơ ng tự, Abdallah et al. [12] đã sử dụng mô hình học sâu dựa trên CNN và LSTM.

Kiani et al. [13] đã đề xuất Mạng thần kinh Deep Autoencoder để tìm hiểu hành vi bình thư ờng của mạng IoT
lư u lư ợng truy cập và sử dụng nó để phát hiện hành vi bất thư ờng trong thực tế thời gian. Ngoài ra, Rasool et al. [14] đã thử nghiệm với các mô hình học chuyển đổi như VGG16, ResNet50 và InceptionV3. Những nghiên cứu này cho thấy tiềm năng của các mô hình học sâu trong việc phát hiện sự bất thư ờng trong lư u lư ợng mạng.

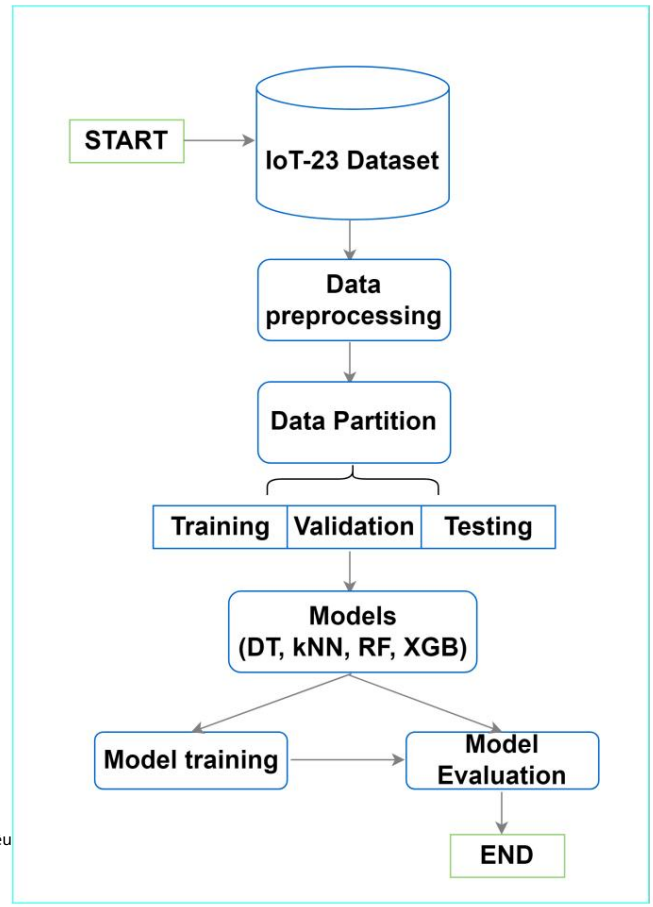
Bộ dữ liệu IoT-23 [15], đư ợc phát hành vào năm 2020, đã trở thành tâm điểm cho các nhà nghiên cứu về bảo mật IoT và máy học. Nó đã đư ợc sử dụng rộng rãi để phát triển và đánh giá các kỹ thuật phát hiện và giảm thiểu lây nhiễm phần mềm độc hại IoT. Một nghiên cứu đáng chú ý đã kết hợp các thuật toán máy học với các phư ơ ng pháp học sâu để cải thiện độ chính xác của việc phát hiện phần mềm độc hại bằng bộ dữ liệu. Một dự án khác đã sử dụng bộ dữ liệu để tạo ra một phát hiện xâm nhập phù hợp
hệ thống cho các thiết bị IoT, tập trung vào việc phát hiện các hoạt động bất thư ờng và độc hại theo thời gian thực. Ngoài ra, những ngư ời tìm kiếm lại đã sử dụng bộ dữ liệu IoT-23 để đánh giá nhiều phư ơ ng pháp trích xuất tính năng và thuật toán phân loại để phát hiện phần mềm độc hại IoT. Sự sẵn có của bộ dữ liệu này đã góp phần đáng kể vào những tiến bộ trong nghiên cứu bảo mật IoT và phát triển các thuật toán và hệ thống hiệu quả để bảo vệ các thiết bị IoT khỏi các hoạt động độc hại.
Việc khám phá bộ dữ liệu IoT-23 đang diễn ra tiếp tục tăng cư ờng tính bảo mật và khả năng phục hồi của hệ sinh thái IoT đang mở rộng.

Mỗi nghiên cứu này trình bày các phư ơ ng pháp khác nhau để phát hiện và phân loại các cuộc tấn công mạng và tất cả đều cho kết quả tốt. Tuy nhiên, một số nghiên cứu chư a đề cập đến hiệu suất của mô hình cũng như thời gian thực hiện mô hình, một số nghiên cứu khác chỉ trình bày về phát hiện hoặc phân loại.

III. PHƯ Ơ NG PHÁP

Phần này sẽ bắt đầu với phần trình bày bộ dữ liệu sẽ đư ợc sử dụng trong phân tích tiếp theo. Các kỹ thuật tiền xử lý sau đó sẽ đư ợc áp dụng để đảm bảo tính chính xác và độ tin cậy của dữ liệu. Sau đó, chúng ta sẽ đi sâu vào quá trình lựa chọn dữ liệu, trực quan hóa, để làm mờ và chia tách. Phân tích của chúng tôi sẽ kết thúc bằng việc đánh giá hiệu quả của các thuật toán đư ợc sử dụng, kèm theo phân tích so sánh các kết quả.

Luồng thực hiện chính của nghiên cứu này đư ợc minh họa trong Hình 1.



Hình 1. Mô hình đề xuất

Để đánh giá hiệu quả của các mô hình máy học để phân loại IoT-Botnet, trư ớc tiên chúng tôi thực hiện xử lý trư ớc dữ liệu. Tiền xử lý dữ liệu giúp nâng cao hiệu quả của quá trình phân loại. Tiếp theo, chúng tôi chia tập dữ liệu thành các tập huấn luyện, kiểm tra và xác thực tư ơ ng ứng. Cuối cùng, chúng tôi triển khai các mô hình máy học phổ biến như cây quyết định, k-láng giềng gần nhất (kNN), rừng ngẫu nhiên và XGBoost (XGB) để đánh giá hiệu quả của từng mô hình.

1. Tập dữ liệu

Bộ dữ liệu IoT-23, được phát hành vào tháng 1 năm 2020, là một trọng tâm đáng kể của việc tìm kiếm lại trong các lĩnh vực bảo mật IoT và học máy. Các nhà nghiên cứu đã sử dụng rộng rãi bộ dữ liệu này để phát triển và đánh giá các kỹ thuật khác nhau cho việc phát hiện và giảm thiểu lây nhiễm phần mềm độc hại IoT. Bộ dữ liệu bao gồm các lưu lượng truy cập mạng được ghi lại từ một mảng đa dạng gồm 23 Internet of Things (IoT) riêng biệt thiết bị, thiết bị mở rộng như phích cắm thông minh, máy ảnh, khóa thông minh và bộ điều nhiệt thông minh, trong số những thứ khác. Những cái này lưu lượng truy cập được chụp một cách tỉ mỉ trong một môi trường được kiểm soát, trong đó mỗi thiết bị được kết nối với một mạng Wi-Fi tách biệt và tiếp xúc với một phạm vi kịch bản tấn công, bao gồm các cuộc tấn công vũ phu, Mirai tấn công botnet, tấn công tiêm nhiễm, v.v.

Bảng 1. Nội dung lưu lượng của bộ dữ liệu IoT-23 theo các cuộc tấn công được thực hiện.

Tên tấn công	Chảy
Part-Of-A-Horizontal-PortScan	213, 852, 924
Okiru	47, 381, 241
Okiru-Attack	13, 609, 479
DDoS	19, 538, 713
C&C-Heart Beat	33,673
C&C	21,995
Tấn công	9398
C&C-	888
C&C-Heart Beat Attack	883
C&C-Tải xuống tệp	53
C&C-Torii	30
Tập tin tải về	18
Tải xuống tệp C&C-Heart Beat	11
Tấn công Part-Of-A-Horizontal-PortScan	5
C&C-Mirai	2

Siêu dữ liệu chi tiết liên quan đến mỗi lần chụp được tích hợp trong tập dữ liệu, bao gồm thông tin như loại thiết bị, phiên bản phần sụn và loại tấn công. Một mô tả toàn diện về phân phối lưu lượng truy cập được cung cấp trong Bảng 1, trong khi một danh sách phức tạp, trình bày mô tả của tất cả các tính năng, được trình bày trong Bảng 2.

Để hỗ trợ phát hiện lưu lượng độc hại, phòng thí nghiệm hình cầu Strato đã phát triển nhãn cho các loại khác nhau của luồng mạng, dựa trên phân tích của họ về các bản chụp phần mềm độc hại. Các nhãn được sử dụng để phát hiện luồng độc hại là Tại tack, C&C, DDoS, FileDownload, HeartBeat, Mirai, Okiru, PartOfAHorizontalPortScan và Torii. Nhãn "Tấn công" được chỉ định cho các luồng cố gắng khai thác lỗ hổng dịch vụ, chẳng hạn như brute-force đăng nhập telnet hoặc tiêm một lệnh trong tiêu đề của yêu cầu GET. Sự "lành tính" nhãn chỉ ra rằng không có hoạt động độc hại hoặc đáng ngờ đã được phát hiện. Nhãn "C&C" chỉ ra rằng người bị nhiễm thiết bị được kết nối với máy chủ CC, trong khi "DDoS" nhãn chỉ ra rằng thiết bị bị nhiễm đang tham gia trong một cuộc tấn công từ chối dịch vụ phân tán. "Tệp xuống tải" đã được gán cho các kết nối liên quan đến tải tệp xuống thiết bị bị nhiễm.

Bảng 2. Các tính năng của bộ dữ liệu IoT-23.

Tính năng	Tên	Mô tả
trường-ts	Luồng thời gian bắt đầu	
uid		ID duy nhất
id.orig-h		Nguồn Địa chỉ IP
id.orig-p		Cổng nguồn
id.resp-h		Địa chỉ IP đích
id.resp-p		cảng đích
dịch		giao thức
vụ proto		Loại dịch vụ (http, dns, v.v.)
khoảng thời gian		Tổng thời lượng lưu lượng
orig-bytes		Byte giao dịch nguồn-đích
resp-bytes		Byte giao dịch nguồn đích
liên kết trạng thái		trạng thái kết nối
local-orig		Nguồn địa chỉ địa phương
local-resp		Địa chỉ đích
resp-pkts		gói đích
orig-ip-bytes	Dòng byte nguồn	
lịch sử	Lịch sử của các gói nguồn	
byte bị thiếu	Thiếu byte trong quá trình giao dịch	
gói nguồn	orig-pkts	
resp-ip-bytes	Luồng byte đích	
nhãn	Tên kiểu tấn công	

Nhãn "Heart-Beat" được gán cho các kết nối đã được sử dụng để theo dõi máy chủ bị nhiễm bởi máy chủ C&C. Các Nhãn "Mirai" được gán cho các kết nối thể hiện đặc điểm của một cuộc tấn công botnet Mirai, trong khi "Okiru" nhãn đã được sử dụng cho các kết nối có mẫu tương tự với mạng botnet Okiru ít phổ biến hơn. "PartOfAHorizontal PortScan" đã được gán cho các kết nối được sử dụng cho quét cổng ngang để thu thập thông tin để biết thêm các cuộc tấn công. Cuối cùng, nhãn "Torii" chỉ ra các kết nối thể hiện các đặc điểm của một cuộc tấn công botnet Torii.

Tính khả dụng của bộ dữ liệu IoT-23, có nhãn chụp phần mềm độc hại và lưu lượng truy cập thiết bị IoT thực, đã phát một vai trò quan trọng trong việc thúc đẩy nghiên cứu về bảo mật IoT. Nó có phục vụ như một nguồn tài nguyên quý giá để phát triển và đánh giá các thuật toán học máy, phát hiện xâm nhập hệ thống và các giải pháp bảo mật khác nhằm mục đích bảo vệ thiết bị IoT khỏi các hoạt động độc hại. Các nhà nghiên cứu tiếp tục để khám phá và mở rộng dựa trên những hiểu biết được cung cấp bởi Bộ dữ liệu IoT-23 để tăng cường hơn nữa tính bảo mật và khả năng phục hồi của hệ sinh thái IoT đang mở rộng nhanh chóng.

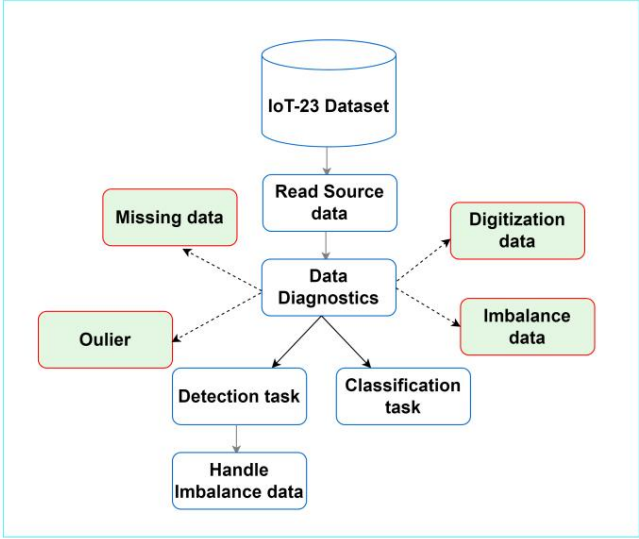
2. Tiền xử lý

Giai đoạn tiền xử lý dữ liệu đóng một vai trò quan trọng trong chuẩn bị dữ liệu cho đào tạo mô hình, bao gồm nhiều

các bư ớc và cân nhắc. Quá trình này, như đư ợc mô tả trong Hình 2, là điều cần thiết để đảm bảo chất lư ợng dữ liệu và giải quyết những thách thức cụ thể có thể phát sinh trong quá trình phân tích.

Bư ớc đầu tiên trong tiền xử lý dữ liệu liên quan đến việc đọc nguồn dữ liệu, cho phép truy cập vào tập dữ liệu để kiểm tra và thao tác thêm. Sau đó, trọng tâm chuyển sang đánh giá chất lư ợng dữ liệu thông qua toàn diện kiểm tra và chẩn đoán. Bư ớc quan trọng này nhằm xác định và giải quyết các vấn đề tiềm ẩn, chẳng hạn như thiếu dữ liệu, ngoại lệ, hoặc nhu cầu chuyển đổi số.

Để giải quyết những thách thức này, các giải pháp phù hợp đư ợc đư a ra cho từng vấn đề cụ thể gặp phải, đồng thời đảm bảo tác động tối thiểu đến tập dữ liệu tổng thể. Cách tiếp cận này cho phép can thiệp có mục tiêu và bảo vệ tính toàn vẹn của dữ liệu.



Hình 2. Các bư ớc tiền xử lý dữ liệu

Hơ n nữa, điều quan trọng là phải thừa nhận rằng sự khác biệt nhiệm vụ yêu cầu kỹ thuật xử lý dữ liệu phù hợp. Đặc biệt, mất cân bằng dữ liệu nổi lên như một mối quan tâm nổi bật do sự khác biệt đáng kể về số lư ợng hồ sơ trên các lớp khác nhau. Để giảm thiểu vấn đề này, các chiến lư ợc khác biệt đư ợc thông qua dựa trên nhiệm vụ hiện tại.

Sau khi kiểm tra và đánh giá tỉ mỉ bộ dữ liệu IoT 23, ngư ời ta xác định rằng bộ dữ liệu này không có của bất kỳ điểm dữ liệu bị thiếu. Tuy nhiên, một sự hiện diện nhỏ của điểm dữ liệu ngoại lệ đã đư ợc quan sát. Hơ n nữa, việc phân tích mang lại một cái nhìn tổng quan về tập dữ liệu, tiết lộ thông tin như tổng số 3.000.779 bản ghi, bao gồm 20 tính năng riêng biệt. Bảng 3 cung cấp một đại diện toàn diện của các loại dữ liệu liên quan với mỗi tính năng.

Để xử lý các tính năng với các loại dữ liệu không phải là số, một quá trình phân loại đã đư ợc tiến hành, kết quả là chia các đặc điểm này thành hai nhóm riêng biệt. Các nhóm đầu tiên bao gồm các loại dữ liệu không phải là số không thể dễ dàng chuyển đổi thành định dạng số. Cái này nhóm bao gồm các tính năng như uid, id.orig_h, id.resp_h, local_orig và local_resp. Ngư ợc lại, nhóm thứ hai bao gồm các loại dữ liệu không phải là số có thể thuận tiện đư ợc chuyển thành dạng số. Nhóm này bao gồm các tính năng như proto, dịch vụ, conn_state, lịch sử, thời lư ợng, orig_bytes và resp_bytes. Để tạo thuận lợi cho quá trình chuyển đổi này, mỗi giá trị riêng biệt trong các tính năng nói trên là đư ợc gán một giá trị số nguyên duy nhất, bắt đầu từ 0. Lớp LabelEncoder từ thư viện sklearn đã đư ợc sử dụng cho Mục đích này. Sau đó, với việc xử lý kiểu dữ liệu hoàn thành trên tập dữ liệu gốc, thử nghiệm đã đư ợc đư ợc thực hiện trên 15 tính năng đư ợc chọn như một phần của phần tiếp theo Phân tích.

Bảng 3. Loại dữ liệu của các tính năng IoT-23

STT	Tên tính năng	Kiểu dữ liệu
1	trư ờng-ts	Float
2	uid	Sự vật
3	id.orig-h	Sự vật
4	id.orig-p	Trôi nổi
5	id.resp-h	Sự vật
6	id.resp-p	Trôi nổi
7	dịch	Sự vật
[illegible]	vụ proto	Sự vật
9	khoảng thời gian	Sự vật
10	byte gốc	Sự vật
11	byte tư ơ ng ứng	Sự vật
12	trạng thái kết nối	Sự vật
13	local-orig	Sự vật
14	local-resp	Sự vật
15	resp-pkts	Trôi nổi
16	orig-ip-byte	Trôi nổi
17	lịch sử	Trôi nổi
		Sự vật
18	byte bị bỏ lỡ	Float
19	gói gốc	Trôi nổi
20	resp-ip-byte	Trôi nổi
21	nhãn	Sự vật

Để căn chỉnh loại dữ liệu cột nhãn với yêu cầu nghiên cứu của chúng tôi, nó là bắt buộc để thực hiện một quá trình phân nhánh. Bộ phận này phục vụ một mục đích quan trọng, đặc biệt là trong bối cảnh các nhiệm vụ phát hiện của chúng tôi. Trong khoảng này khuôn khổ, hồ sơ mang nhãn lành tính sẽ đư ợc quy cho giá trị 0, trong khi tất cả các bản ghi khác sẽ đư ợc gán giá trị của 1. Trong lĩnh vực nhiệm vụ phân loại, các lớp thể hiện bản ghi hạn chế và chia sẻ các loại tấn công giống hệt nhau sẽ đư ợc hợp nhất thành các thực thể lớp cố kết. sự phức tạp chi tiết của lư ợc đồ ghi nhãn này đư ợc phác thảo tỉ mỉ trong Bảng 4, cung cấp một cái nhìn tổng quan toàn diện về nhiệm vụ phân loại và phát hiện.

Bảng 4. Nhãn chi tiết cho nhiệm vụ đa phân loại

Tên nhãn	Số nhãn
Đang ở	0
C&C-Heart Beat	1
C&C	1
C&C-	1
C&C-Heart Beat Attack	1
C&C-Tải xuống tệp	1
C&C-Torii	1
Tập tin tải về	1
Tải xuống tệp C&C-Heart Beat	1
C&C-Mirai	1
Tấn công	2
DDoS	3
Okiru	4
Okiru-Attack	4
Part-Of-A-Horizontal-PortScan	5

Trong bối cảnh môi trư ờng mạng, sự xuất hiện các sự kiện diễn ra với tốc độ đáng kinh ngạc, đòi hỏi hoạt động dữ liệu nhanh chóng, liên tục và tự động. Việc thu thập dữ liệu đư ợc sắp xếp liền mạch bởi các hệ thống IoT, encom truyền các thông số cần thiết và sau đó đư ợc phân phối theo kênh vào các đư ờng ống xử lý tự động và các mô hình đư ợc đào tạo để tạo kết quả thời gian thực. Mục tiêu cơ bản của những các hành động năng động xoay quanh việc phát hiện kịp thời và ngăn chặn sự bất thư ờng của mạng, đảm bảo ứng phó kịp thời và can thiệp.

Trong ngữ cảnh của các nhiệm vụ phát hiện, nơ i phân biệt giữa các lớp độc hại và lành tính là rất quan trọng, việc lấy mẫu kỹ thuật đư ợc sử dụng để giải quyết sự mất cân đối đáng kể. Điều này liên quan đến việc điều chỉnh cẩn thận sự phân phối của nhóm độc hại để đảm bảo đại diện tối ư u.

Đối với các nhiệm vụ phân loại, trong đó nhiều lớp với một số lư ợng bản ghi hạn chế có liên quan, một phư ơ ng pháp hợp nhất ap đư ợc triển khai để giảm sự mất cân bằng dữ liệu. Cái này hợp nhất tạo điều kiện cho các đại diện cân bằng hơ n của các lớp khác nhau, cải thiện quá trình phân loại.

3. Phư ơ ng pháp phân tích

a) Cây quyết định

Trong lĩnh vực phân loại botnet IoT, cây quyết định đại diện cho một máy học nền tảng và hiệu quả kỹ thuật. Những cây này cung cấp một khuôn khổ minh bạch và toàn diện cho việc ra quyết định, thúc đẩy các thuộc tính riêng biệt đư ợc tìm thấy trong dữ liệu lư u lư ợng truy cập mạng IoT. Việc sử dụng rộng rãi chúng trong lĩnh vực an ninh mạng thể hiện năng lực của họ trong việc phát hiện và phân loại mạng botnet IoT.

Về cốt lõi, nguyên tắc củng cố cây quyết định xoay quanh việc phân chia lặp đi lặp lại các bộ dữ liệu, đư ợc hư ớng dẫn bởi các tính năng khác nhau, để tạo ra một cấu trúc giống như cây. Mỗi nút bên trong cấu trúc này biểu thị một quyết định dựa trên trên một tính năng cụ thể, trong khi mỗi nút lá tư ơ ng ứng với một kết quả phân loại dứt khoát. Xây dựng quyết định cây đòi hỏi phải lựa chọn tỉ mỉ những thông tin hữu ích nhất tính năng và việc xác định các tiêu chí chia nhỏ tối ư u tại mỗi nút, nhằm mục đích tối đa hóa sự khác biệt giữa các lớp riêng biệt.

Cây quyết định sở hữu những lợi thế đáng chú ý trong bối cảnh phân loại mạng botnet IoT, xuất phát từ khả năng của chúng để nắm bắt các mối quan hệ phức tạp và các tư ơ ng tác tính năng. Tính linh hoạt của họ trong việc xử lý cả phân loại và số các tính năng làm cho chúng phù hợp với nhiều loại dữ liệu lư u lư ợng truy cập mạng. Hơ n nữa, cây quyết định thể hiện mạnh mẽ khi đối mặt với sự không hoàn hảo của dữ liệu trong thế giới thực, cung cấp các giá trị còn thiếu và ngoại lệ một cách hiệu quả.

Yếu tố khả năng diễn giải đóng một vai trò quan trọng trong lĩnh vực phân loại mạng botnet IoT với cây quyết định. Các những con đư ờng vốn có của cấu trúc cây mang lại giá trị vô giá hiểu biết sâu sắc về các đặc điểm và hành vi của các mạng botnet IoT. Thông qua kiểm tra cẩn thận các quy tắc quyết định, bảo mật các nhà phân tích có thể đạt đư ợc sự hiểu biết sâu sắc hơ n về các đặc điểm và kiểu phân biệt liên quan đến các các loại hoạt động của botnet.

Thuật toán cây quyết định đư ợc sử dụng rộng rãi và phổ biến loại mô hình học tập có giám sát có hiệu quả trong việc giải quyết cả vấn đề phân loại và hồi quy. cấu trúc của nó bao gồm các nút đại diện cho các biến và các nhánh biểu diễn mối quan hệ giữa các biến này. tại nút gốc, thuật toán xem xét toàn bộ tập dữ liệu và thông qua một loạt các quyết định nhị phân dựa trên đầu vào các biến, nó phân vùng đệ quy dữ liệu thành các biến nhỏ hơ n tập hợp con. Các phân vùng này tạo thành các nút bên trong cây và các lá tư ơ ng ứng với giá trị dự đoán của mục tiêu biến cho mỗi tập hợp con.

Để xây dựng cây quyết định, thuật toán tuân theo cách tiếp cận từ trên xuống để chọn thuộc tính tốt nhất để phân chia dữ liệu dựa trên một tiêu chí như thu đư ợc thông tin, chỉ số Gini, hoặc entropy. Sau khi dữ liệu đư ợc chia nhỏ, thuật toán đệ quy áp dụng quy trình tư ơ ng tự cho từng tập hợp con kết quả cho đến khi đáp ứng tiêu chí dừng, chẳng hạn như cây đư ợc xác định trư ớc độ sâu, số lư ợng phiên bản tối thiểu trên mỗi lá hoặc không có gì khác cải thiện hiệu suất dự đoán.

Để cập nhật và tính toán các giá trị nút trong cây quyết định, hai công thức đư ợc sử dụng: thư ớc đo tạp chất và tiêu chí để lựa chọn cách phân chia tốt nhất. thư ớc đo tạp chất định lư ợng mức độ đồng nhất hoặc không đồng nhất của biến mục tiêu trong một nút, trong khi tiêu chí cho chọn cách phân chia tốt nhất sẽ xác định thuộc tính nào cung cấp mức tăng thông tin cao nhất hoặc tạp chất thấp nhất sau khi tách nút.

Entropy = (1, 2, xác suất) hài lòng phân bổ
3, . . . , = 1 =1 1

$$( ) = \sum_{=1} \times \text{nhật ký} \quad (1)$$

Mức tăng thông tin là một công thức kiểm tra nút mang lại lượng thông tin nút đó còn lại sau khi chuyển sang nút con.

$$\text{Tăng } ( ) = ( ) - \sum_{=1} \times ( ) \quad (2)$$

b) Rừng ngẫu nhiên (RF)

Random Forest là một thuật toán học máy nổi bật đã thể hiện hiệu quả rõ rệt trong lĩnh vực phân loại mạng botnet IoT. sử dụng của nó vì một kỹ thuật học tập đồng bộ cho phép kết hợp nhiều cây quyết định, đạt đến đỉnh cao trong các dự đoán vừa mạnh mẽ vừa chính xác. Trong lĩnh vực an ninh mạng, Random Forest đã thu hút được sự chấp nhận đáng kể nhờ sự tinh thông trong việc xác định và phân loại các mạng botnet IoT.

Trọng tâm của thuật toán Rừng ngẫu nhiên nằm ở việc xây dựng một tập hợp bao gồm các cây quyết định. Mỗi cây quyết định được huấn luyện trên một tập con ngẫu nhiên của dữ liệu đào tạo, kết hợp một tập hợp con ngẫu nhiên các tính năng tại mỗi nút. Bằng cách giới thiệu tính ngẫu nhiên trong cả quy trình lựa chọn dữ liệu và tính năng, Random Forest tăng cường tính đa dạng trong mô hình, hạn chế hiệu quả các rủi ro liên quan đến trang bị thừa. Sự đa dạng hóa này trao quyền cho thuật toán để nắm bắt các khía cạnh khác nhau và các mẫu phức tạp bên trong dữ liệu lưu lượng truy cập mạng IoT, từ đó nâng cao hiệu suất phân loại.

Trong suốt giai đoạn đào tạo, các cây quyết định riêng lẻ trong Rừng ngẫu nhiên được xây dựng một cách tỉ mỉ thông qua phân vùng dữ liệu đệ quy. Việc phân vùng này được thực hiện dựa trên các ngưỡng tính năng riêng biệt, với mục đích chính là giảm thiểu tạp chất hoặc tối đa hóa mức thu được thông tin ở mỗi lần phân tách. Do đó, một sự sắp xếp có thứ bậc của các nút và các lá trở thành hiện thực, phục vụ như một đại diện tập thể của các mẫu thu được và các mối quan hệ qua lại có trong dữ liệu.

Khi đạt đến giai đoạn dự đoán, Random Forest kết hợp các kết quả đầu ra được tạo bởi từng cây quyết định riêng lẻ thông qua cơ chế bỏ phiếu, được điều chỉnh cụ thể để phân loại. Mỗi cây quyết định trình bày phiếu bầu của nó cho lớp được dự đoán và lớp tích lũy được đa số phiếu bầu cuối cùng được coi là dự đoán cuối cùng. Bằng cách tận dụng cách tiếp cận dựa trên tập hợp này, Random Forest đã giảm thiểu thành công các sai lệch hoặc sai sót tiềm ẩn vốn có trong các cây quyết định riêng lẻ, từ đó củng cố độ bền và độ chính xác tổng thể của phân loại.

Random Forest bao gồm một số lợi thế đáng chú ý trong lĩnh vực phân loại mạng botnet IoT. Nó xử lý thành thạo dữ liệu nhiều chiều và phức tạp, khiến nó đặc biệt phù hợp để phân tích các tính năng và mẫu đa diện vốn có trong lưu lượng mạng IoT. Ngoài ra, Random Forest vượt trội trong việc ước tính tầm quan trọng của tính năng, cung cấp thông tin chi tiết vô giá về các yếu tố then chốt góp phần vào các hoạt động của mạng botnet.

c) K-Láng giềng gần nhất (KNN)

Thuật toán k-Láng giềng gần nhất (k-NN) là một kỹ thuật học máy linh hoạt và quan trọng được sử dụng trong phân loại mạng botnet IoT. Được đánh giá cao về tính đơn giản, k NN trình bày một cách tiếp cận trực quan để phân biệt và phân loại các mạng botnet IoT dựa trên dữ liệu lưu lượng mạng của chúng.

Về cốt lõi, thuật toán k-NN cố gắng gán một lớp cho một điểm dữ liệu mới bằng cách đánh giá các lớp của k hàng xóm gần nhất của nó trong không gian đối tượng. Đánh giá này dựa trên các số liệu khoảng cách như Euclidean hoặc Manhattan khoảng cách, tạo thuận lợi cho việc xác định k hàng xóm gần nhất từ tập dữ liệu huấn luyện. Nhãn lớp của phiên bản mới sau đó sẽ xuất hiện thông qua quy trình bỏ phiếu đa số được tiến hành giữa những người hàng xóm được chọn này.

Trong bối cảnh phân loại mạng botnet IoT, thuật toán k-NN mang lại rất nhiều lợi thế. Trước hết, nó thể hiện khả năng đóng gói các mối quan hệ phức tạp và các mẫu phi tuyến tính phổ biến trong dữ liệu lưu lượng mạng. Khả năng thích ứng này trao quyền cho k-NN để đáp ứng các biểu hiện đa dạng của các hoạt động mạng botnet IoT và các hoạt động liên quan của chúng. đặc trưng. Hơn nữa, k-NN thể hiện sự thành thạo trong việc xử lý cả các tính năng phân loại và số, tạo điều kiện thuận lợi cho ứng dụng của nó đối với dữ liệu không đồng nhất gặp phải trong môi trường IoT. Cuối cùng, tính đơn giản vốn có của k-NN không cần thiết phải đào tạo rõ ràng, khiến việc triển khai và hiểu nó trở nên đơn giản.

Tuy nhiên, thuật toán k-NN có chứa các phép lặp giới hạn nhất định. Khi kích thước của tập dữ liệu mở rộng, độ phức tạp tính toán của việc xác định các hàng xóm gần nhất tăng lên đáng kể. Do đó, thời gian dự đoán kéo dài, đặc biệt là trong không gian nhiều chiều. Hơn nữa, hiệu suất của k-NN xoay quanh việc lựa chọn đúng tham số k, vì một lựa chọn không phù hợp có thể dẫn đến trang bị thiếu hoặc thừa.

d) Tăng cường độ dốc cực cao (XGBoost)

Extreme Gradient Boosting (XGBoost) là một thuật toán học máy đáng gờm đã nhận được sự hoan nghênh đáng chú ý trong lĩnh vực phân loại mạng botnet IoT sự. Là một phương pháp học tập đồng bộ, XGBoost kết hợp liền mạch nhiều mô hình dự đoán yếu, điển hình là cây quyết định, để xây dựng một bộ phân loại chính xác và mạnh mẽ.
Hiệu suất vượt trội và khả năng giải quyết các vấn đề phức tạp

Các nhiệm vụ phân loại mạng botnet IoT đã định vị XGBoost là một lựa chọn ư a thích trong lĩnh vực an ninh mạng.

Chức năng cơ bản của thuật toán XGBoost nằm trong quy trình đào tạo lặp đi lặp lại của nó, trong đó các mô hình dự đoán yếu đư ợc tích hợp tuần tự vào tập hợp trong khi đồng thời giảm thiểu lỗi dự đoán tổng thể. Mỗi mô hình tiếp theo tập trung vào việc nắm bắt các lỗi còn sót lại của các mô hình trư ớc đó, dần dần tinh chỉnh các khả năng dự đoán của XGBoost. Bằng cách kết hợp khéo léo các kỹ thuật tăng cư ờng độ dốc và chính quy hóa, XGBoost cân bằng một cách khéo léo độ phức tạp của mô hình và khả năng tổng thể hóa, đạt đến đỉnh cao về hiệu suất vư ợt trội.

Trong bối cảnh phân loại mạng botnet IoT, XGBoost mang lại một số lợi thế đáng chú ý. Nó dễ dàng điều hư ớng các thách thức đặt ra bởi dữ liệu tân đa chiều và không đồng nhất thư ờng gặp trong môi trư ờng IoT, từ đó tạo điều kiện kết hợp các tính năng và mẫu đa dạng nội tại với các hoạt động của mạng botnet. Ngoài ra, XGBoost vư ợt trội trong việc nắm bắt các mối quan hệ và tư ơ ng tác phức tạp giữa các tính năng này, nâng cao khả năng phân loại chính xác các botnet IoT. Ngoài ra, XGBoost seam tích hợp một cách dễ dàng các kỹ thuật chính quy giúp giảm thiểu hiệu quả nguy cơ trang bị quá mức, tăng cư ờng độ bền và khả năng khái quát hóa của nó.

Tuy nhiên, XGBoost gặp phải những thách thức về khả năng diễn giải. Là một tập hợp các cây quyết định, việc làm sáng tỏ những đóng góp riêng lẻ của từng cây trong mô hình XGBoost có thể rất khó khăn. Tuy nhiên, các kỹ thuật tập trung vào phân tích tầm quan trọng của tính năng có thể đư ợc sử dụng để hiểu rõ hơ n về tầm quan trọng tư ơ ng đối của các tính năng trong quá trình phân loại.

Extreme Gradient Boosting (XGBoost) nổi lên như một thuật toán mạnh mẽ và đáng khen ngợi trong lĩnh vực phân loại mạng botnet IoT. Thông qua phư ơ ng pháp học tập đồng bộ, đư ợc kết hợp với các kỹ thuật tăng cư ờng độ dốc và điều chỉnh độ dốc, XGBoost nắm bắt một cách thành thạo các mối quan hệ phức tạp và xử lý dữ liệu nhiều chiều một cách khéo léo.
Bất chấp những thách thức tiềm ẩn liên quan đến khả năng dự đoán lẫn nhau, hiệu suất kiên quyết và khả năng phân loại chính xác các botnet IoT của XGBoost đã nhấn mạnh vai trò không thể thiếu của XGBoost trong lĩnh vực an ninh mạng.

4. Đánh giá hiệu suất

Khi đánh giá các mô hình thuật toán đư ợc mô tả trư ớc đó, các kỹ thuật đa dạng đã đư ợc sử dụng để đánh giá độ chính xác của kết quả và rút ra kết quả toàn diện cho từng mô hình. Nghiên cứu này liên quan đến một số khái niệm cơ bản, chẳng hạn như TP, TN, FP và FN. TP biểu thị số lư ợng dư ơ ng tính thực sự đã đư ợc xác định chính xác, trong khi TN là số lư ợng âm tính thực sự đã đư ợc xác định chính xác xác định. FP biểu thị số trư ờng hợp dư ơ ng tính thực tế đã bị phân loại nhầm là âm tính, trong khi FN biểu thị số trư ờng hợp âm tính bị phân loại nhầm là dư ơ ng tính.

a) Độ chính xác (PRE)

Độ chính xác là thư ớc đo hiệu suất đư ợc sử dụng để đánh giá hiệu quả của mô hình bằng cách xác định tỷ lệ các trư ờng hợp tích cực đư ợc xác định chính xác. Biện pháp này có thể đư ợc tính toán bằng toán học thông qua việc sử dụng một công thức, có tính đến số lư ợng dư ơ ng tính thật và dư ơ ng tính giả. Cụ thể, độ chính xác đư ợc tính bằng cách chia số lư ợng dư ơ ng tính thực cho tổng số dư ơ ng tính thật và dư ơ ng tính giả.

$$= /(+) \quad (3)$$

b) Độ chính xác (ACC)

Độ chính xác là thư ớc đo hiệu suất đư ợc sử dụng để đánh giá hiệu quả của mô hình bằng cách xác định tỷ lệ dự đoán chính xác trong tổng số dự đoán. Biện pháp này có thể đư ợc tính toán bằng toán học thông qua việc sử dụng một công thức, xem xét số lư ợng dư ơ ng tính thực và âm tính thực. Cụ thể, độ chính xác đư ợc tính bằng cách chia tổng số dư ơ ng thực và âm thực cho tổng số dự đoán.

$$= (TP + TN)/(TP + TN + FP + FN) \quad (4)$$

c) Điểm nhớ lại (RE)

Điểm thu hồi là chỉ số hiệu suất đư ợc sử dụng để đánh giá hiệu quả của mô hình trong việc xác định chính xác các trư ờng hợp tích cực thực tế. Biện pháp này có thể đư ợc tính toán bằng toán học thông qua việc sử dụng một công thức, kết hợp số lư ợng dư ơ ng tính thực và âm tính giả. Cụ thể, điểm thu hồi đư ợc tính bằng cách chia số lư ợng dư ơ ng tính thực cho tổng số dư ơ ng tính thật và âm tính giả.

$$= /(+) \quad (5)$$

d) Điểm F1 cho lớp nhị phân

Điểm F1 là thư ớc đo hiệu suất thu đư ợc bằng cách tính trung bình cả điểm chính xác và điểm thu hồi. Biện pháp này đư ợc sử dụng rộng rãi để đánh giá hiệu quả tổng thể của một mô hình vì nó cung cấp một cái nhìn toàn diện về các kết quả dư ơ ng tính giả và âm tính giả. Cụ thể, điểm F1 đư ợc tính là giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi. Công thức tính điểm F1 tính đến cả số lư ợng kết quả dư ơ ng tính thật, kết quả dư ơ ng tính giả và âm tính giả, từ đó đư a ra đánh giá cân bằng hơ n về hiệu suất của mô hình.

$$1 \quad \text{điểm} = 2 \times \frac{\times}{+} \qquad (6)$$

e) Điểm F1 cho Đa lớp

Điểm F1 [16] đư ợc công nhận rộng rãi và phổ biến số liệu đư ợc sử dụng để đánh giá hiệu suất của các thuật toán phân loại nhiều lớp. Nó phục vụ như là một toàn diện thư ớc đo có tính đến cả độ chính xác và thu hồi, do đó cho phép đánh giá cân bằng về phân loại của hiệu quả tổng thể.

Trong bối cảnh phân loại nhiều lớp trong IoT vấn đề botnet, điểm số F1 cung cấp những hiểu biết có giá trị về khả năng của các thuật toán để phân loại chính xác các trư ờng hợp trên các lớp khác nhau. Bằng cách xem xét sự xuất hiện của dư ơng tính giả (các trư ờng hợp đư ợc phân loại sai) và âm tính giả (các trư ờng hợp bị bỏ lỡ) cho mỗi lớp, điểm F1 cung cấp một đánh giá mạnh mẽ về hiệu suất của thuật toán.

Việc tính điểm F1 liên quan đến việc xác định trung bình điều hòa của độ chính xác và thu hồi cho mỗi lớp. Cách tiếp cận này đánh giá hiệu quả trạng thái cân bằng giữa việc xác định chính xác các trư ờng hợp tích cực (độ chính xác) và nắm bắt tất cả các trư ờng hợp tích cực (thu hồi). F1 cao hơ n điểm biểu thị sự cân bằng vư ợt trội giữa độ chính xác và nhớ lại, qua đó thể hiện năng lực của bộ phân loại trong xác định chính xác các trư ờng hợp trên tất cả các lớp.

Tóm lại, điểm F1 đóng vai trò then chốt như một số liệu quan trọng trong việc đánh giá hiệu suất của các thuật toán phân loại nhiều lớp trong vấn đề IoT-botnet. Nó tạo điều kiện cho việc đánh giá toàn diện bằng cách xem xét độ chính xác và thu hồi cho từng lớp, do đó cho phép đánh giá toàn diện về hiệu quả tổng thể của bộ phân loại trong việc phân loại chính xác các trư ờng hợp trên nhiều lớp.

IV. KẾT QUẢ VÀ THẢO LUẬN

1. Phân loại nhị phân

Bảng 5 trình bày một phân tích toàn diện về các kết quả đánh giá thu đư ợc từ phân loại nhị phân trong vấn đề mạng botnet IoT. Bốn thuật toán học máy riêng biệt, cụ thể là Cây quyết định (DT), Rừng ngẫu nhiên (RF), k-Gần nhất Neighbor (KNN) và Extreme Gradient Boosting (XGB), đã đư ợc sử dụng để phân biệt hiệu suất của họ bằng cách sử dụng khác nhau số liệu.

Bảng 5. Một số thư ớc đo phân loại nhị phân

đánh giá DT	RF	KNN	XGB
ACC	99,9%	89,5%	99,6% 99,8%
TRƯ ỚC	100,0%	88,0%	100,0% 100%
NỐT RỀ	100,0%	86,0%	99,0% 100%
F1	100,0%	87,0%	100,0% 100%
Thời gian	5,9	40.2	58.2 99,4

Về độ chính xác (ACC), thuật toán Cây quyết định trư ng bày độ chính xác vô song, đạt đư ợc một đặc biệt tỷ lệ chính xác 99,9%. Theo sát gót, Extreme Gra dient Boosting đã thể hiện độ chính xác ấn tư ợng của 99,8%. k-Nearest Neighbor thể hiện độ chính xác đáng khen ngợi ở mức 99,6%, trong khi Random Forest đạt đư ợc mức đáng nể độ chính xác 89,5%.

Độ chính xác (PRE), đo lư ờng khả năng chính xác phân loại các trư ờng hợp tích cực trong số tất cả các trư ờng hợp đư ợc dự đoán tích cực, mang lại kết quả đáng chú ý. cả quyết định Tree và Extreme Gradient Boosting đạt đư ợc hiệu quả hoàn hảo điểm chính xác 100,0%. k-Nearest Neighbor cũng thể hiện độ chính xác tuyệt vời, phù hợp với số điểm hoàn hảo của 100,0%. Mặc dù thấp hơ n một chút, Random Forest đã thực hiện đáng khen ngợi với số điểm chính xác là 88,0%.

Việc đánh giá Thu hồi (RE), định lư ợng tỷ lệ phần trăm của các trư ờng hợp tích cực đư ợc phân loại chính xác trong số tất cả trư ờng hợp tích cực thực tế, tiết lộ kết quả mẫu mực. Cả hai Cây quyết định và Extreme Gradient Boosting xuất sắc với điểm thu hồi hoàn hảo là 100,0%. k-Hàng xóm gần nhất đã chứng minh tỷ lệ thu hồi đáng khen ngợi là 99,0%, trong khi Random Forest đạt đư ợc tỷ lệ thu hồi đáng nể là 86,0%.

Điểm F1, đại diện cho giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi, cung cấp đánh giá tổng thể về hiệu suất của bộ phân loại. Cả cây quyết định và cực trị Gradient Boosting đạt đư ợc điểm số F1 hoàn hảo là 100,0%. k-Nearest Neighbor thể hiện hiệu suất tuyệt vời với điểm F1 là 100,0%, trong khi Random Forest đạt đư ợc điểm số điểm đáng nể là 87,0%.

Xét về thời gian tính toán, thuật toán Cây quyết định nổi lên là hiệu quả nhất, chỉ cần một 5,9 giây để xử lý. Rừng ngẫu nhiên theo sau với thời gian xử lý là 40,2 giây, trong khi k-Hàng xóm gần nhất yêu cầu 58,2 giây. Tăng cư ờng độ dốc cực cao thể hiện thời gian xử lý lâu nhất, đạt 99,4 giây.

2. Đa phân loại

Bảng 6 minh họa việc đánh giá toàn diện và kết quả phân tích liên quan đến đa phân loại trong bối cảnh của vấn đề IoT-botnet. Nghiên cứu có sự tham gia của đánh giá bốn thuật toán học máy riêng biệt, cụ thể là Cây quyết định (DT), Rừng ngẫu nhiên (RF), k-Gần nhất Neighbor (KNN) và Extreme Gradient Boosting (XGB), sử dụng một loạt các chỉ số hiệu suất.

Số liệu chính về độ chính xác (ACC) cho thấy hiệu suất đáng kể đạt đư ợc bởi cả Quyết định Các thuật toán Tree và Extreme Gradient Boosting, đạt đư ợc tỷ lệ chính xác cao 99,9%. Theo sát, k-Gần nhất Neighbor thể hiện độ chính xác đáng khen ngợi ở mức 99,8%,

trong khi Random Forest thể hiện độ chính xác giảm nhẹ là 78,2%.

Bảng 6. Một số thư ớc đo của đa phân loại

Đánh giá	DT	RF	KNN	XGB
ACC	99,9%	78,2%	99,8%	99,9%
TRƯ ỚC	99,9%	59,6%	99,7%	100,0%
NỐT RỀ	99,7%	46,1%	98,7%	99,8%
F1	99,6%	49,8%	99,2%	99,9%
Thời gian	11,5	206,1	85,2	956.2

Độ chính xác (PRE), đánh giá khả năng phân loại chính xác các phiên bản trong mỗi lớp, nhấn mạnh điểm số chính xác đặc biệt đạt đư ợc bởi các thuật toán Cây quyết định (99,9%) và Tăng cư ờng độ dốc cực độ (100,0%). k-Nearest Neighbor thể hiện độ chính xác đáng khen ngợi là 99,7%, trong khi Random Forest hiển thị độ chính xác tư ơ ng đối thấp hơ n là 59,6%.

Chỉ số thu hồi (RE), phản ánh độ nhạy trong việc phân loại chính xác các phiên bản trong mỗi lớp, cho thấy các kết quả nổi bật đối với thuật toán Cây quyết định (99,7%) và Tăng cư ờng độ dốc cực cao (99,8%). k-Hàng xóm gần nhất có tỷ lệ thu hồi là 98,7%, trong khi Random Forest đạt tỷ lệ thu hồi là 46,1%, cho thấy hiệu suất tư ơ ng đối thấp hơ n ở khía cạnh này.

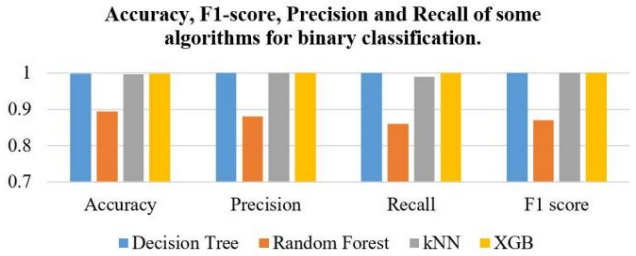
Điểm số F1, đóng vai trò đánh giá toàn diện về cả độ chính xác và khả năng thu hồi, càng làm nổi bật hiệu suất vư ợt trội của các thuật toán Cây quyết định (99,6%) và Tăng cư ờng độ dốc cực cao (99,9%). k-Nearest Neighbor đạt đư ợc số điểm F1 đáng nể là 99,2%, trong khi Random Forest mang lại số điểm tư ơ ng đối thấp hơ n là 49,8%.

Về mặt thời gian tính toán, Cây quyết định al gorithm đã thể hiện thời gian xử lý ngắn nhất là 11,5 giây, cho thấy tính hiệu quả của nó. k-Hàng xóm gần nhất yêu cầu 85,2 giây, trong khi Khu rừng ngẫu nhiên tiêu tốn thời gian xử lý lâu hơ n là 206,1 giây. Mặt khác, thuật toán Extreme Gradient Boosting thể hiện nhu cầu tính toán cao nhất, với thời gian xử lý là 956,2 giây.

Những phát hiện của đánh giá này làm nổi bật hiệu suất vư ợt trội của các thuật toán Cây quyết định và Tăng cư ờng độ dốc cực cao trên nhiều chỉ số, bao gồm độ chính xác, độ chính xác, thu hồi và điểm F1. Ngoài ra, thuật toán Cây quyết định nổi bật về hiệu quả tính toán của nó. Tuy nhiên, điều quan trọng là phải xem xét các yêu cầu cụ thể và mức độ ư u tiên của ứng dụng khi chọn thuật toán phù hợp nhất, có tính đến việc xem xét cân bằng độ chính xác, độ chính xác, khả năng thu hồi và hiệu quả tính toán. Kết quả của nghiên cứu này cung cấp những hiểu biết có giá trị về các đặc tính hiệu suất của các thuật toán học máy khác nhau để phân loại đa dạng trong các vấn đề về botnet IoT.

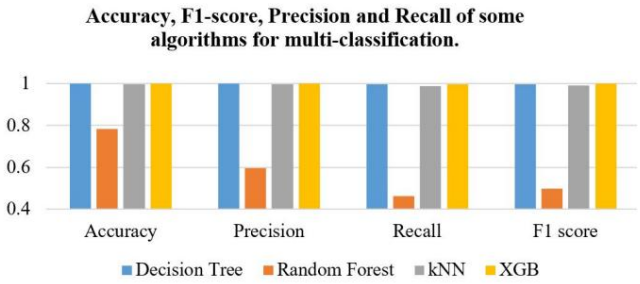
3. Đánh giá mô hình

Để xác định sự phù hợp của một mô hình nhất định đối với một vấn đề cụ thể, các kỹ thuật đánh giá hiệu suất đư ợc sử dụng. Một loạt các phư ơ ng pháp có thể đư ợc sử dụng, bao gồm như ng không giới hạn ở các phép đo độ chính xác, độ chính xác, thu hồi và điểm F1. Những phư ơ ng pháp này phục vụ như đáng tin cậy các chỉ số về khả năng của một mô hình để giải quyết vấn đề ban đầu một cách hiệu quả.



Hình 3. Accuracy, F1-score, Precision và Recall của một số thuật toán phân loại nhị phân

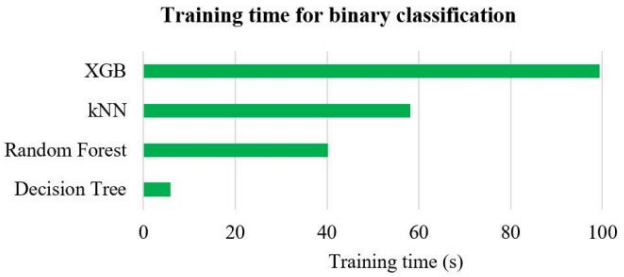
Đánh giá hiệu suất của các thuật toán DT, RF, KNN và XGB để phân loại nhị phân đư ợc trình bày trong Hình 3, cung cấp tổng quan toàn diện về độ chính xác, độ chính xác, thu hồi và số liệu điểm F1. Các kết quả thể hiện rõ ràng hiệu suất vư ợt trội đư ợc thể hiện bởi cả thuật toán DT và XGB, vư ợt qua các đối tác tư ơ ng ứng của chúng. Ngoài ra, về thời gian đào tạo, thuật toán Cây quyết định nổi lên như một tùy chọn hiệu quả hơ n, do đó mang lại hiệu suất tổng thể vư ợt trội cây chùy.



Hình 4. Accuracy, F1-score, Precision và Recall của một số thuật toán cho đa phân loại

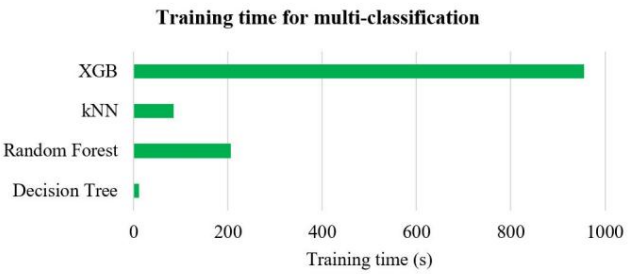
Bài toán đa lớp đư ợc đánh giá để đánh giá hiệu suất của bốn thuật toán riêng biệt, cụ thể là Cây quyết định, Rừng ngẫu nhiên, KNN và XGB, như đư ợc mô tả trong Hình 4. Đánh giá bao gồm phân tích toàn diện về độ chính xác, độ chính xác, khả năng thu hồi và chỉ số điểm F1.
Đáng chú ý, kết quả chứng minh rằng thuật toán Cây quyết định vư ợt qua các thuật toán khác về các chỉ số hiệu suất này, thể hiện hiệu suất vư ợt trội.

Hình 5 trình bày thời gian đào tạo của các thuật toán cho bài toán phân loại nhị phân. Kết quả cho thấy thuật toán XGB có thời gian đào tạo lâu hơ n đáng kể so với các thuật toán khác. Ngư ợc lại, thuật toán cây quyết định nổi lên là hiệu quả nhất về mặt thời gian đào tạo, vư ợt trội so với tất cả các thuật toán khác về mặt này.



Hình 5. Thời gian đào tạo để phân loại nhị phân

Trong bài toán đa phân loại, thời gian đào tạo của các thuật toán đư ợc trình bày trong hình 6. Kết quả minh họa rằng thuật toán XGB mất nhiều thời gian hơ n để đào tạo so với các thuật toán khác, với thời gian đào tạo gần gấp 100 lần so với thuật toán cây quyết định. Mặt khác, thuật toán cây quyết định thể hiện hiệu suất tốt nhất về thời gian đào tạo.



Hình 6. Thời gian đào tạo đa phân loại

V. KẾT LUẬN

Tóm lại, bài nghiên cứu học thuật này đã thực hiện nhiệm vụ quan trọng là phát hiện và phân loại các botnet IoT bằng thuật toán học máy. Nghiên cứu đã nhấn mạnh đáng kể vào việc phân tích và thao tác kỹ lư ỡng dữ liệu mạng botnet IoT, với trọng tâm cụ thể là bộ dữ liệu IoT-23 đư ợc đánh giá cao. Bằng cách triển khai các thuật toán học máy đư ợc sử dụng rộng rãi và đư ợc công nhận rộng rãi như Cây quyết định (DT), k-Láng giềng gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cư ờng độ dốc cực đại (XGBoost), mục tiêu là phân loại và phát hiện các botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23.

Việc ứng dụng các thuật toán này nhằm nâng cao hiểu biết của chúng tôi về hiệu suất và hiệu quả của chúng trong lĩnh vực phát hiện và phân loại mạng botnet IoT.

Thông qua phân tích so sánh các kết quả thu đư ợc từ các thuật toán đa dạng này, ngư ời ta đã thu đư ợc những hiểu biết có giá trị, làm sáng tỏ những điểm mạnh và hạn chế tư ơ ng ứng của chúng. Những hiểu biết sâu sắc như vậy trang bị cho các nhà nghiên cứu và ngư ời thực hành kiến thức cần thiết để đư a ra quyết định sáng suốt khi chọn thuật toán phù hợp nhất để phát hiện và phân loại mạng botnet IoT thành công. Cái này nghiên cứu đóng góp đáng kể cho lĩnh vực này bằng cách cung cấp đánh giá toàn diện về hiệu suất của các thuật toán học máy này, từ đó tạo điều kiện phát triển các phư ơ ng pháp phân loại và phát hiện linh hoạt và hiệu quả hơ n.

Bằng cách giải quyết thách thức then chốt trong việc phát hiện và phân loại mạng botnet IoT, nghiên cứu này có ý nghĩa mở rộng sang lĩnh vực tăng cư ờng các biện pháp an ninh mạng trong hệ sinh thái IoT. Những phát hiện và hiểu biết sâu sắc thu đư ợc từ nghiên cứu này có khả năng cung cấp thông tin cho việc phát triển các biện pháp chủ động nhằm ngăn chặn và giảm thiểu tác động bất lợi của các cuộc tấn công mạng botnet IoT. Ngư ời ta dự đoán rằng công việc này sẽ truyền cảm hứng cho những ngư ời đam mê nghiên cứu sâu hơ n và thúc đẩy sự hợp tác trong lĩnh vực này, cuối cùng dẫn đến sự tiến bộ của các hệ thống IoT an toàn hơ n và bảo vệ các cơ sở hạ tầng quan trọng.

Nghiên cứu này mở ra những con đư ờng đầy hứa hẹn cho việc khám phá và tiến bộ trong tư ơ ng lai trong lĩnh vực phát hiện và phân loại mạng botnet IoT. Đầu tiên, cần phải thực hiện thăm dò và đánh giá toàn diện các thuật toán học máy bổ sung để mở rộng danh sách các tùy chọn có sẵn nhằm phát hiện và phân loại hiệu quả các botnet IoT. Việc kết hợp các thuật toán và kỹ thuật mới hơ n có khả năng nâng cao hiệu suất tổng thể và độ chính xác của quy trình phát hiện. Ngoài ra, việc khám phá các phư ơ ng pháp học tập đồng bộ, trong đó nhiều thuật toán đư ợc kết hợp đồng thời, đư a ra một con đư ờng hấp dẫn để có khả năng mở khóa các kết quả đư ợc cải thiện và khung phát hiện mạnh mẽ hơ n.

NGƯ ỜI GIỚI THIỆU

[1] Williams, P., Dutta, IK, Daoud,. H, và Bayoumi, M.: Một cuộc khảo sát về bảo mật trong internet vạn vật tập trung vào tác động của các công nghệ mới nổi. Elsevier. (2022).

[2] Khan, WZ, Rehman, MH, Zangoti, HM, Afzal, MK, Armi, N., và Salah, K.: Internet vạn vật trong công nghiệp: Những tiến bộ gần đây, hỗ trợ công nghệ và thách thức mở. Elsevier. (2020).

[3] Vitorino, J., Andrade, R., Prac¸a, I., Sousa, O., và Maia, E.: Phân tích so sánh các kỹ thuật học máy của Ma chine để phát hiện xâm nhập IoT

sự. Nền tảng và Thực hành Bảo mật (trang 191- 207). Springer. (2022).

[4] Haq, NF, Onik, AR, Hridoy, MAK, Rafni, M., Shah, FM và Farid, DM: Ứng dụng các phư ơ ng pháp học máy trong Hệ thống phát hiện xâm nhập: Khảo sát. Bài báo đư ợc đăng trên Tạp chí Quốc tế về Nghiên cứu Tiên tiến về Trí tuệ Nhân tạo (IJARAI), Tập 4 Số 3. (2015).

[5] Hajji, J., Khalily, M., Moustafa, N., và Nelms, T. IoT-23: Bộ dữ liệu để phân tích lư u lư ợng mạng IoT. lò xo. (2019).

[6] Rahim, A., Razzaque, MA, Hasan, R. và Hossain, MF Bảo mật mạng IoT hiệu quả thông qua lựa chọn tính năng và kỹ thuật máy học. IEEE. (2020).

[7] Islam, SMZ, Bhuiyan, MZH và Hasan, R. Sự kết hợp của các mô hình học máy để phát hiện xâm nhập trong mạng IoT bằng Bộ dữ liệu IoT-23. IEEE. (2020)

[8] Li, Y., Qiu, L., Chen, Y. và Chen, Y. Hệ thống phát hiện xâm nhập dựa trên Ensemble cho các mạng IoT sử dụng Bộ dữ liệu IoT-23. IEEE. (2020)

[9] PH Do, TD Dinh, DT Le, VD Pham, L. Myrova và R. Kirichek, "An Efficient Fea-ture Extraction Method for Attack Classification in IoT Networks," 2021 13 International Congress on Ultra Modern Viễn thông and Control Systems and Workshop (ICUMT)

[10] Alotaibi, F., Al-Qaness, MA, Abunadi, A. và Al ghazzawi, MA Một phư ơ ng pháp tiếp cận Ap-learning sâu để phát hiện xâm nhập trong mạng IoT bằng cách sử dụng Bộ dữ liệu IoT-23. IEEE. (2020).

[11] Li, J., Hu, C., Yang, K., Zhang, X. và Lu, J. Một hệ thống phát hiện xâm nhập IoT dựa trên IoT-23 sử dụng Deep Learning. IEEE. (2020).

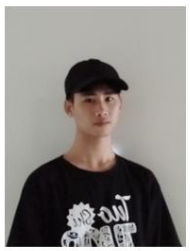
[12] Abdallah, A., Khalil, I., Al-Emadi, N., Almohaimeed, A., và Kim, H. Phát hiện Botnet IoT thời gian thực bằng cách sử dụng Deep Learning trên Bộ dữ liệu IoT-23. IEEE. (2020)

[13] Kiani, AT, Abbas, RA, Abbasi, AZ và Khan, MK Phát hiện sự bất thư ờng dựa trên học sâu cho các mạng IoT bằng cách sử dụng Bộ dữ liệu IoT-23. IEEE. (2020)

[14] Rasool, S., Saeed, S., Farooq, F., và Madani, A. Nghiên cứu so sánh các phư ơ ng pháp học tập chuyển đổi để phát hiện phần mềm độc hại IoT bằng bộ dữ liệu IoT-23. IEEE. (2021).

[15] Sebastian Garcia, Agustin Parmisano, và Maria Jose Erquiaga. (2020). IoT-23: Tập dữ liệu đư ợc gắn nhãn có lư u lư ợng truy cập mạng IoT độc hại và lành tính (Phiên bản 1.0.0) [Tập dữ liệu]. Zenodo. http://doi.org/10.5281/zenodo.4743746 [16] Stoian, NA Máy học để phát hiện bất thư ờng trong mạng IoT : Phân tích phần mềm độc hại trên IoT-23 bộ dữ liệu. EEMCS: Kỹ thuật Điện, Toán học và Khoa học Máy tính. (2020)

Phạm Văn Quân hiện đang là sinh viên năm cuối chuyên ngành Khoa học dữ liệu và Trí tuệ nhân tạo trư ờng Đại học Đông Á. Lĩnh vực nghiên cứu của ông bao gồm DS, ML, AI và ứng dụng của nó trong các lĩnh vực khác nhau như Tài chính, Mạng và Xử lý ngôn ngữ tự nhiên.



Ngô Văn úc trở thành sinh viên Đại học Đông Á năm 2020, chuyên ngành Trí tuệ nhân tạo và Khoa học dữ liệu. Lĩnh vực nghiên cứu của anh bao gồm học máy, học sâu, khoa học dữ liệu, trí tuệ nhân tạo, xử lý hình ảnh và các ứng dụng của chúng

Đỗ Phúc Hào nhận bằng Thạc sĩ Khoa học máy tính của Đại học Đà Nẵng - Đại học Bách khoa năm 2017. Anh hiện là Tiến sĩ Khoa học máy tính. sinh viên khoa Mạng Truyền thông và Truyền số liệu tại

Đại học Viễn thông Bang Bonch-Bruevich Saint-Peterburg, Nga.
Lĩnh vực nghiên cứu của anh bao gồm Trí tuệ nhân tạo, Học máy và ứng dụng của nó trong các lĩnh vực khác nhau như mạng, chuỗi khối.

Nguyễn Năng Hùng Vân nhận bằng Tiến sĩ. bằng Khoa học Máy tính của Đại học Đà Nẵng, Việt Nam, năm 2021. Ông hiện là giảng viên tại Đại học Đà Nẵng - Đại học Khoa học và Công nghệ. Lĩnh vực nghiên cứu của anh bao gồm Trí tuệ nhân tạo, Học máy, Đại số hình học, Máy tính

Kết nối mạng.