NÂNG CAO KHẢ NĂNG PHÁT HIỆN XÂM NHẬP MẠNG SỬ DỤNG MẠNG CNN

Nguyễn Ngọc Điệp, Nguyễn Thị Thanh Thủy

Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Phát hiện xâm nhập mạng (NIDS) là vấn đề thu hút được sự quan tâm của người làm quản trị hệ thống mạng cũng như những người làm nghiên cứu an toàn hệ thống. Bài toán phát hiện xâm nhập mạng có thể được giải quyết thông qua việc phát hiện các hành vi truy nhập bất thường bằng cách sử dụng kỹ thuật học máy thông qua việc xây dựng mô hình dựa trên các thuật toán thống kê, học máy hay mạng nơ-ron nhân tạo. Tuy nhiên, các cuộc tấn công bảo mật ngày nay có xu hướng không thể đoán trước được. Việc xây dựng một hệ thống phát hiện xâm nhập mạng linh hoạt và hiệu quả có tỷ lệ báo động giả thấp và độ chính xác phát hiện cao trước các cuộc tấn công không xác định gặp rất nhiều thách thức. Bài báo này nghiên cứu áp dụng mạng CNN (mạng nơ-ron tích chập) cho mô hình phát hiện xâm nhập và so sánh hiệu năng với một số kỹ thuật học máy cơ bản khác trên cơ sở bộ dữ liệu NSL-KDD. Kết quả thực nghiệm cho thấy, với kết quả độ đo F1 là 98,2%, mô hình phát hiện xâm nhập dựa trên mạng CNN có hiệu năng vượt trội so với các mô hình học máy khác.

Từ khóa: phát hiện xâm nhập mạng, NSL-KDD, học sâu, CNN.

I. GIỚI THIỆU

Việc phát triển của các thiết bị tính toán và sự phổ biến của các ứng dụng mạng như thương mại điện tử, mạng xã hội, tính toán đám mây đã làm cho các vấn đề về an toàn thông tin càng trở nên phức tạp và cấp thiết. Hành vi xâm nhập hệ thống có thể được coi là các hành động cố gắng làm tổn hại các thuộc tính an toàn của hệ thống, bao gồm bí mật, toàn vẹn và sẵn sàng, bằng cách vượt qua các cơ chế hay biện pháp đảm bảo an toàn của hệ thống tính toán hay mạng. Nói cách khác, người tấn công cố gắng thực hiện các hành động để lấy được quyền truy nhập tới đối tượng mong muốn của mình và các hành động này xâm phạm đến các chính sách an ninh của hệ thống. Để ngăn ngừa hiệu quả các hành động trái phép, rõ ràng hệ thống cần nhận được sự hỗ trợ từ việc phát hiện và cảnh báo chính xác về các hoạt động gây tổn hại đến an toàn thông tin của hệ thống.

Việc phát hiện xâm nhập là quá trình xác định và đối phó với các hành vi xâm nhập nhằm vào các hệ thống tính toán hay mạng. Quá trình này được tiến hành dựa vào hệ thống phát hiện xâm nhập, thông qua việc giám sát các sự kiện xảy ra trong quá trình sử dụng hệ thống máy tính hay mạng và phân tích xem có dấu hiệu của việc xâm nhập hay không. Hệ thống phát hiện xâm nhập (IDS) có thể là hệ thống phần cứng hay phần mềm cho phép tự động hóa quá trình phát hiện hành vi xâm nhập và thông thường dựa trên hai phương pháp chính: dựa trên chữ ký và dựa trên bất thường. Phương pháp phát hiện xâm nhập dựa trên các dấu hiệu/chữ ký [6] là kỹ thuật căn bản của hệ thống phát hiện xâm nhập. Các dấu hiệu thường là các mô hình hay chuỗi ký tự tương ứng với các vụ tấn công hay mối đe dọa đã biết. Để phát hiện, IDS so sánh các mô hình với các sự kiện thu được để nhận biết việc xâm nhập. Phương pháp này còn được gọi là phương pháp dựa trên tri thức do sử dụng cơ sở tri thức về các hành vi xâm nhập trước đó. Rõ ràng, kỹ thuật này khó có thể phát hiện được các hành vi xâm nhập mới chưa có trong cơ sở tri thức của hệ thống cho dù có độ tin cậy và chính xác cao. Phương pháp phát hiện xâm nhập dựa trên bất thường [6] là phương pháp quan trọng trong hệ thống IDS. Sự bất thường được coi là sự khác biệt với hành vi đã biết bằng các lập hồ sơ các hành vi thông thường trên cơ sở việc theo dõi các hoạt động thường xuyên, các kết nối mạng, máy trạm hay người dùng qua một khoảng thời gian. Hệ thống phát hiện thực hiện việc so sánh các hồ sơ với các sự kiện quan sát được để nhận biết các vụ tấn công nghiêm trọng. Như vậy, phương pháp phát hiện dựa trên bất thường trang bị công cụ hữu hiệu cho người quản trị hệ thống để có thể chống chọi hiệu quả với các hình thức xâm nhập mới chưa được biết.

Bài toán phân biệt các hành vi truy nhập hay sử dụng các tài nguyên của hệ thống là một trong những bài toán tiêu biểu của kỹ thuật học máy [12]. Về cơ bản, các kỹ thuật học máy giúp xây dựng mô hình cho phép tự động phân loại các lớp hành vi sử dụng hệ thống dựa trên các đặc trưng của các hành vi này. Có thể kể tên một số kỹ thuật tiêu biểu như các kỹ thuật dựa trên cây quyết định C4.5 [9], máy véc-tơ tựa SVM [7], mạng nơ-ron [10].

Trong thời gian gần đây, mô hình học sâu đã có tác động sâu rộng đến ứng dụng mô hình học máy, đặc biệt trong lĩnh vực như nhận dạng tiếng nói, xử lý ảnh và xử lý ngôn ngữ tự nhiên [3, 4]. Đặc trưng nổi bật của mô hình học sâu là việc sử dụng khối lượng lớn dữ liệu so với cách tiếp cận

Tác giả liên hệ: Nguyễn Ngọc Điệp
Email: diepnn80@gmail.com
Đến tòa soạn: 10/2020, chỉnh sửa:11/2020 , chấp nhận đăng: 12/2020

truyền thống. Các mô hình sử dụng nhiều tham số cho phép khai thác các thông tin trong tập dữ liệu khổng lồ một cách hiệu quả hơn. Hiện nay, có nhiều nghiên cứu về phát hiện xâm nhập sử dụng kỹ thuật học sâu và phân tích các mô hình xây dựng dựa trên bộ dữ liệu KDD 99 [13] hay NSL-KDD [18] như [1, 5, 8, 11], tuy nhiên, rất ít trong số đó thể hiện hiệu quả sức mạnh của các kỹ thuật học sâu. Trong số các cách tiếp cận khác nhau trong học sâu, mạng nơ-ron tích chập (CNN) thể hiện khả năng vượt trội trong xử lý ảnh và nhiều lĩnh vực khác. Đây là một biến thể của mạng nơ-ron tiêu chuẩn, trong đó sử dụng các lớp tích chập và gộp thay thế cho các lớp ẩn được kết nối đầy đủ của một mạng nơ-ron truyền thống. Tuy nhiên, mặc dù mạng CNN thường cho thấy độ chính xác cao nhưng lại chưa được khai thác nhiều trong các hệ thống IDS. Bài báo này đề xuất một mô hình mạng CNN nhằm nâng cao độ chính xác và giảm mức độ cảnh báo sai trong các hệ thống phát hiện xâm nhập mạng. Ngoài ra, hiệu năng của mô hình CNN đề xuất sẽ được so sánh với một số kỹ thuật học máy cơ bản khác trên cơ sở bộ dữ liệu NSL-KDD.

Phần còn lại của bài báo được trình bày như sau: Phần 2 trình bày một số nghiên cứu về phát hiện xâm nhập. Phần 3 mô tả phương pháp phát hiện xâm nhập đề xuất dựa trên CNN. Phần 4 đưa ra các kết quả thực nghiệm, đánh giá mô hình trên tập dữ liệu NSL-KDD, và so sánh với các phương pháp khác. Cuối cùng là phần kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Phần này trình bày các nghiên cứu liên quan đến phát hiện xâm nhập mạng và phát hiện xâm nhập mạng sử dụng mạng nơ-ron học sâu.

A. Phát hiện xâm nhập mạng sử dụng học máy

IDS thường được phân loại thành hệ thống phát hiện dựa trên máy chủ (HIDS) và dựa trên mạng (NIDS). HIDS giám sát và phân tích thông tin máy chủ, ví dụ như các lệnh gọi hệ thống, tệp hệ thống quan trọng và tệp nhật ký. Trong khi đó, NIDS giám sát toàn bộ mạng bằng cách phân tích lưu lượng mạng, như lưu lượng truy cập, địa chỉ IP, cổng dịch vụ và việc sử dụng giao thức. Với sự phát triển của công nghệ mạng cũng như nhiều kiểu tấn công mới khó xác định, NIDS đối mặt với thách thức trong việc xử lý một lượng lớn dữ liệu, có thể đến từ nhiều nguồn tài nguyên khác nhau với môi trường mạng hay thay đổi. Đối với NIDS dựa trên phát hiện bất thường, khi hệ thống không được cập nhật thường xuyên, một số điểm bất thường có thể bị coi là lưu lượng truy cập bất thường. Do đó, với các biến thể trong các hành vi mạng, các IDS dựa trên bất thường phải được cập nhật liên tục và thích ứng với các môi trường mạng thường xuyên thay đổi. Nhiều phương pháp tiếp cận khác nhau đã được đề xuất trong IDS, tiêu biểu như các kỹ thuật học máy, giúp xây dựng mô hình cho phép tự động phân loại các lớp hành vi bất thường [7, 9, 10]. Tuy nhiên, các kỹ thuật này vẫn phải đối mặt với một số thách thức như số cảnh báo giả cao, độ chính xác phát hiện thấp trước các cuộc tấn công không xác định và không đủ khả năng phân tích.

B. Phát hiện xâm nhập mạng dựa trên học sâu

Một mạng nơ-ron đơn giản thường gặp là perceptron, thông thường chỉ có ba lớp (1 lớp đầu vào, 1 lớp ẩn, và 1 lớp đầu ra) phục vụ cho việc khai thác thông tin nhờ vào việc huấn luyện lớp ẩn và lớp đầu ra theo dữ liệu huấn luyện được cung cấp. Mỗi một nốt thuộc từng lớp trong mạng đều có kết nối đầy đủ với các nốt khác thuộc lớp kề với nó. Mạng này có thể được làm "sâu" thêm bằng cách bổ sung các lớp ẩn làm cho các đặc trưng của tập dữ liệu được biến đổi nhiều lần. Mỗi một lần biến đổi tương tự như một bước suy diễn mà có thể được biểu diễn một cách đơn giản bằng một chuỗi tính toán. Tương tự như các mạng nơ-ron khác, mạng nơ-ron perceptron đa lớp (MLP) [3] có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp. Các lớp ẩn sâu bên trong mạng có khả năng tổng hợp các đặc trưng từ các lớp trước đó, do đó cho phép mạng mô hình hóa được dữ liệu phức tạp hơn với số lượng các nút ít hơn các loại mạng nơ-ron khác.

Mạng perceptron nhiều lớp (MLP), mạng nơ-ron hồi quy và mạng nơ-ron học sâu tích chập là cách tiếp cận phổ biến hiện thời trong các mô hình học sâu. Nguyên nhân chủ yếu cho việc dùng mô hình học sâu chính là tính hiệu quả thực tế so với các cách tiếp cận khác. Hơn thế, mô hình học sâu còn cung cấp các kỹ thuật mới và tiên tiến về mặt lý thuyết như các biến thể của các thuật học. Sự thành công của mô hình học sâu cần phải kể đến sự phổ biến của tính toán hiệu năng cao sử dụng bộ xử lý đồ họa. Khi biểu diễn dưới dạng các ma trận véc-tơ, việc tính toán được tăng tốc nhờ phần cứng và thư viện đồ họa được tối ưu hóa. Kết quả huấn luyện và kiểm chứng mô hình được tiến hành một cách nhanh chóng và hiệu quả.

Mô hình học sâu đã được sử dụng cho việc phân biệt và phát hiện cách hành vi truy nhập trái phép. Các tác giả của nghiên cứu [1] sử dụng mạng nơ-ron hồi quy để tự động phân lớp dữ liệu truy nhập, chẳng hạn như các truy vấn http, bằng kỹ thuật học hồi quy thời gian thực. Sau đó, việc phân loại truy nhập tiếp theo sử dụng kỹ thuật SVM. Việc sử dụng kỹ thuật học thời gian thực giúp cho phương pháp đề xuất có khả năng áp dụng cho các hệ thống theo dõi thời gian thực và có thể mở rộng từng bước.

Các nghiên cứu [5, 11] sử dụng kiến trúc bộ nhớ dài-ngắn hạn (LSTM) cho mạng nơ-ron hồi quy để xây dựng mô hình phát hiện xâm nhập với tập dữ liệu thử nghiệm KDD 99 [13]. Các tác giả của [11] mở rộng kiến trúc LSTM bằng cách cho phép gán trọng số thích ứng giữa các phần tử trong mạng, cho phép các phần tử mạng chống lại trạng thái không mong muốn từ các đầu vào. Kết quả thu được khá khả quan với mức độ phát hiện đạt trên 90%. Tuy nhiên, nghiên cứu [5] chỉ sử dụng một phần của tập dữ liệu KDD 99 để làm dữ liệu huấn luyện.

Nghiên cứu [8] sử dụng kỹ thuật tự học (self-taught learning) của kỹ thuật học sâu để tiến hành phân loại xâm nhập và thử nghiệm trên tập dữ liệu NSL-KDD [18]. Về căn bản, quá trình phân loại trải qua hai giai đoạn. Giai đoạn đầu các đặc trưng sẽ được tự động nhận biết nhờ vào kỹ thuật tự động mã hóa (sparse auto-encoder), giai đoạn 2 sử

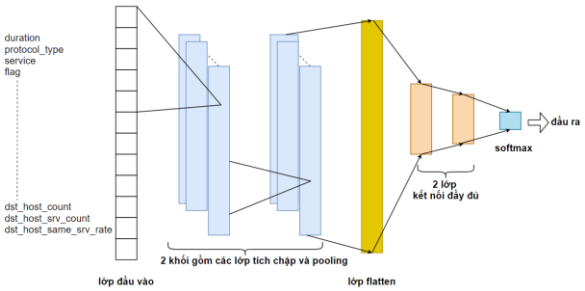
dụng kết quả của giai đoạn 1 để tiến hành phân loại bằng kỹ thuật hồi quy softmax.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Phần này trình bày đề xuất phương pháp phát hiện xâm nhập mạng sử dụng mạng nơ-ron tích chập CNN và tiền xử lý dữ liệu.

A. *Phát hiện xâm nhập mạng sử dụng CNN*

CNN là một biến thể của mạng nơ-ron, với mục đích chính là tự động học các biểu diễn đặc trưng phù hợp cho dữ liệu đầu vào. CNN có hai điểm khác biệt chính so với MLP, đó là chia sẻ trọng số và pooling. Mỗi lớp CNN có thể bao gồm nhiều nhân tích chập được sử dụng để tạo bản đồ đặc trưng (feature map) khác nhau. Mỗi vùng của các nơ-ron lân cận được kết nối với một nơ-ron của bản đồ đặc trưng của lớp tiếp theo. Hơn nữa, để tạo bản đồ đặc trưng, tất cả các vị trí không gian của đầu vào đều chia sẻ nhân. Sau một số lớp tích chập và pooling, một hoặc nhiều lớp được kết nối đầy đủ được sử dụng để phân loại [20]. Mô hình CNN có 2 tính chất quan trọng là tính bất biến (Location Invariance) và tính kết hợp (Compositionality). Lớp pooling sẽ đảm bảo tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling). Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các bộ lọc. Hai tính chất này cho phép CNN tạo ra mô hình với độ chính xác rất cao, do giống cách con người nhận biết các vật thể trong tự nhiên.



Hình 1. Kiến trúc mô hình CNN đề xuất

Mô hình CNN được đề xuất (Hình 1) là một chồng có nhiều lớp bao gồm: các lớp tích chập, max-pooling, dropout, lớp kết nối đầy đủ và softmax. Mỗi lớp tích chập bao gồm một tập các bộ lọc độc lập và có khả năng học khi phát hiện ra một số loại đặc trưng cụ thể ở một số vị trí không gian trong dữ liệu đầu vào. Trong lớp này, các phụ thuộc cục bộ về không gian được khai thác bằng cách đảm bảo sự ràng buộc trong kết nối nội bộ giữa các nút và các lớp liền kề. Mỗi nút chỉ được kết nối với một khu vực nhỏ trong khung đầu vào. Hàm ReLU được sử dụng để làm hàm kích hoạt cho bộ tích chập vì khả năng hoạt động tốt hơn các hàm kích hoạt khác trong hầu hết các tình huống. Để giảm kích thước không gian của bản đồ đặc trưng, mô hình sử dụng lớp pooling. Lớp pooling này giúp giữ lại các thông tin quan trọng nhất và đồng thời giúp làm giảm số lượng tham số và tính toán của mạng nơ-ron, do đó kiểm soát được vấn đề quá vừa dữ liệu khi huấn luyện. Các lớp cuối cùng là hai lớp kết nối đầy đủ và một lớp softmax. Lớp kết nối đầy đủ thực hiện suy diễn mức cao trong mạng nơ-ron bằng cách sử dụng các đặc trưng được trích xuất từ các lớp tích chập và lớp pooling để học các kết hợp phi tuyến của các đặc trưng. Lớp cuối cùng sofmax được sử dụng để dự đoán ra các lớp dựa trên tập huấn luyện. Mô hình CNN này sử dụng hai khối gồm các lớp tích chập và pooling. Một lớp dropout được sử dụng sau lớp kết nối đầy đủ thứ hai. Lớp dropout này có thể giúp giảm được sự quá vừa dữ liệu bằng cách tránh huấn luyện các nút trên tất cả dữ liệu huấn luyện, nhờ đó mạng nơ-ron học được nhiều đặc trưng tốt hơn [3].

Đối với các lớp tích chập, các lớp cao hơn thường sử dụng nhiều bộ lọc để xử lý các phần phức tạp hơn của dữ liệu đầu vào. Do đó, mô hình đề xuất sử dụng 32 bộ lọc cho lớp tích chập số 1 và 64 bộ lọc cho lớp tích chập số 2. Cả hai lớp đều sử dụng các tích chập có cùng độ rộng là 5 và độ trượt là 1. Các lớp max-pooling trong thử nghiệm cũng sử dụng độ rộng là 5. Kích thước của hai lớp được kết nối đầy đủ được thiết lập là 500. Và đối với lớp dropout, xác suất lựa chọn nút được thiết lập là 0,5.

Mạng nơ-ron CNN đề xuất được huấn luyện sử dụng các mini-batch với mỗi mini-batch có độ lớn là 32 và dữ liệu được phân nhóm theo phân bố mẫu dữ liệu từng lớp trong tập huấn luyện. Độ chính xác của mạng được tối ưu hóa sử dụng bộ tối ưu phổ biến là Adam, được cung cấp trong bộ thư viện Keras [14], với tham số learning rate là 0,001.

B. *Tiền xử lý dữ liệu*

Mạng nơ-ron học sâu nhận các giá trị đầu vào là các thuộc tính/đặc trưng của mỗi hành vi truy nhập hệ thống, các giá trị này bắt buộc là các giá trị kiểu số thực. Tuy nhiên, giá trị thuộc tính của các hành vi truy nhập thực tế có thể theo giá trị kiểu loại, dưới dạng chữ. Ví dụ như kiểu truyền tin đối với mỗi truy nhập có thể là: “tcp” hay “udp”. Khi đó, ta cần chuyển các giá trị dạng này sang kiểu số thực. Việc này có thể được thực hiện bằng cách sử dụng véc-tơ one-hot thường thấy trong xử lý ngôn ngữ tự nhiên. Một véc-tơ one-hot là một ma trận 1xN sử dụng để phân biệt mỗi từ trong bộ từ vựng với các từ khác. Véc-tơ chứa các giá trị 0 tại toàn bộ vị trí trừ một vị trí chứa giá trị 1 để nhằm xác định từ đó.

IV. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

Phần này trình bày về tập dữ liệu cho các thử nghiệm phát hiện xâm nhập dựa trên phát hiện bất thường, sử dụng kỹ thuật mạng CNN đã đề xuất và các bộ phân lớp khác, gồm: mạng nơ-ron perceptron đơn giản, máy véc-tơ tựa SVM (sử dụng kỹ thuật SVC), cây quyết định (sử dụng thuật toán CART), rừng ngẫu nhiên (Random Forest), phân loại giảm gradient ngẫu nhiên (SGD) và mạng MLP. Một phần nội dung khác trình bày các tham số cấu hình cho các thử nghiệm và phần cuối cùng trình bày về kết quả thực nghiệm cùng các phân tích đánh giá.

A. *Tập dữ liệu đánh giá*

Nghiên cứu này sử dụng tập dữ liệu NSL-KDD [4] trong các thực nghiệm. Đây là tập dữ liệu được tinh chỉnh của tập dữ liệu KDD 99 [13], trong đó các bản ghi trùng lặp được

loại bỏ và số lượng các bản ghi đủ lớn với tập huấn luyện và kiểm tra. Mỗi bản ghi bao gồm 41 thuộc tính thể hiện các đặc trưng khác nhau của luồng thông tin và được gán nhãn là tấn công hoặc bình thường. Các thuộc tính có thể được chia thành các nhóm liên quan đến kết nối mạng và lưu lượng mạng như dưới đây.

Các thuộc tính tiêu biểu về kết nối mạng:

- duration: thời gian kết nối
- protocol_type: kiểu giao thức, ví dụ tcp
- service: dịch vụ mạng sử dụng, ví dụ ftp
- flag: tình trạng kết nối bình thường hay lỗi, ví dụ SF

Các thuộc tính tiêu biểu về lưu lượng của trạm:

- dst_host_count: số kết nối có cùng địa chỉ trạm đích
- dst_host_srv_count: số kết nối có cùng địa chỉ cổng đích
- dst_host_same_srv_rate: tỷ lệ kết nối có cùng dịch vụ trong số các kết nối tới trạm đích

Ngoài các thuộc tính thể hiện thông tin trực tiếp về tình trạng kết nối hay lưu lượng, còn có các thuộc tính thể hiện thông tin chú giải mức cao như số lần đăng nhập không thành công (num_failed_logins), yêu cầu đăng nhập (is_host_login) hay thử chuyển chế độ đặc quyền (su_attempted).

Với thuộc tính kiểu hành vi trái phép, các hành vi xâm nhập được xếp vào 4 nhóm cơ bản như sau:

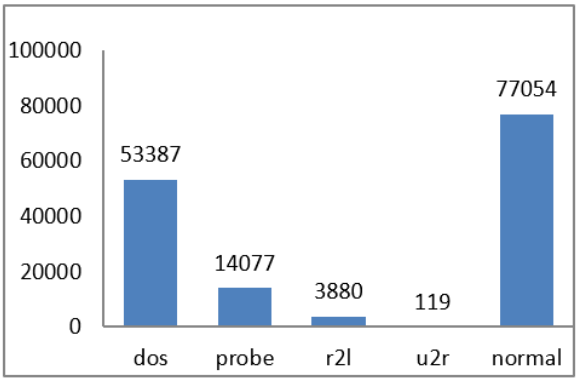
- dos: Nhóm tấn công từ chối dịch vụ
- probe: Giám sát hay thăm dò nhằm thu thập thông tin như quét cổng
- u2r: Truy nhập trái phép tới các tài khoản người dùng có đặc quyền
- r2l: Truy nhập trái phép từ máy ở xa, kẻ tấn công xâm nhập máy ở xa và lấy quyền truy nhập vào máy tính nạn nhân

Việc phân lớp các dạng hành vi tấn công xâm nhập cơ bản theo các hành vi trái phép được mô tả như trong Bảng I.

Bảng I. Phân loại các hành vi xâm nhập cơ bản

LOẠI	CHI TIẾT CÁC HÀNH VI XÂM NHẬP
dos	back, land, neptune, pod, smurf, teardrop, mailbomb, processtable, udpstorm, apache2, worm
probe	satan, ipsweep, nmap, portsweep, mscan, saint
r2l	guess_password, guess_passwd, ftp_write, imap, phf, multihop, warezmaster, warezclient, xlock, spy, xsnoop, snmpguess, snmpgetattack, httptunnel, sendmail, named
u2r	buffer_overflow, loadmodule, rootkit, perl, sqlattack, xterm, ps

Hình 2 dưới đây thể hiện số lượng các bản ghi ứng với từng loại hành vi truy nhập đã được phân loại trong toàn bộ tập dữ liệu, bao gồm dữ liệu để huấn luyện và dữ liệu kiểm tra. Phân chia giữa các hành vi bình thường (normal) và các hành vi xâm nhập (trái phép) tương đối cân xứng.



Hình 2. Phân bố của cách hành vi xâm nhập cơ bản và hành vi bình thường trên toàn tập dữ liệu NSL-KDD

Phân bố chi tiết về các dạng tấn công theo từng kiểu tấn công cụ thể và số lượng tương ứng trong tập dữ liệu huấn luyện và tập dữ liệu kiểm tra được cung cấp trong Bảng 2 và Bảng 3. Mỗi bảng thể hiện số các trường hợp ứng với hành vi xâm nhập cơ bản trong từng tập dữ liệu huấn luyện và kiểm tra. Ngoài ra, bảng này cũng cho biết số lượng các hành vi xâm nhập trái phép cụ thể được gán nhãn theo hành vi xâm nhập cơ bản. Dữ liệu trong các bảng này cho thấy có sự khác biệt lớn về số lượng giữa các dạng tấn công.

Về tổng thể, tập huấn luyện chứa gần 126.000 mẫu dữ liệu về 22 dạng tấn công/xâm nhập cụ thể. Trong khi đó tập kiểm tra chứa hơn 22.500 mẫu dữ liệu nhưng có tới 37 kiểu tấn công/xâm nhập. Sự khác biệt về kiểu tấn công giữa hai tập dữ liệu là thách thức đối với các mô hình phát hiện, đặc biệt là với các kiểu tấn công chưa được biết. Ngoài ra, sự

mất cân bằng giữa các dạng tấn công cũng là vấn đề khó khăn cho các kỹ thuật phân loại.

Bảng II. Dạng tấn công u2r và r2l theo hành vi trái phép trên hai tập dữ liệu.

u2r			
Huấn luyện	***52***	***Kiểm tra***	***67***
buffer_overflow	30	buffer_overflow	20
loadmodule	9	loadmodule	2
Perl	3	perl	2
rootkit	10	rootkit	13
		ps	15
		sqlattack	2
		xterm	13
r2l			
Huấn luyện	***995***	***Kiểm tra***	***2885***
ftp_write	8	ftp_write	3
guess_passwd	53	guess_passwd	1231
imap	11	imap	1
multihop	7	multihop	18
phf	4	phf	2
spy	2	named	17
warezclient	890	warezmaster	944
warezmaster	20	sendmail	14
		snmpgetattack	178
		snmpguess	331
		httptunnel	133
		xlock	9
		xsnoop	4

Bảng III. Dạng tấn công probe và dos theo hành vi trái phép trên hai tập dữ liệu.

probe			
Huấn luyện	***11656***	***Kiểm tra***	***2421***
Ipsweep	3599	ipsweep	141
Nmap	1493	nmap	73
Portsweep	2931	portsweep	157
Satan	3633	satan	735
		saint	319
		mscan	996
dos			
Huấn luyện	***45927***	***Kiểm tra***	***7460***
Back	956	Back	359
land	18	Land	7
neptune	41214	Neptune	4657
pod	201	Pod	41
smurf	2646	smurf	665
teardrop	892	teardrop	12
		processtable	685
		apache2	737
		mailbomb	293
		udpstorm	2
		worm	2

B. Các thiết lập thử nghiệm

1) Độ đo đánh giá và các tham số của bộ phân lớp

Độ chính xác tổng thể là một độ đo đơn giản thường được sử dụng trong các đánh giá phân loại, được tính bằng tỉ lệ giữa số phần tử được phân loại chính xác trên tổng số các phần tử. Tuy nhiên, để đánh giá hiệu năng một hệ thống phân loại xâm nhập với dữ liệu đầu vào không cân bằng giữa các lớp thì độ chính xác tổng thể không phải là một độ đo thực sự hiệu quả do ảnh hưởng của các lớp tới kết quả phân loại là không cân bằng [15,16]. Trong trường hợp này, độ đo đánh giá các mô hình phân loại được sử dụng trong các thử nghiệm là độ chính xác (precision), độ nhạy (recall), và F1 (trung bình điều hòa). Độ chính xác có thể xác định được số các dự đoán cho một nhãn lớp là dự đoán đúng thực sự, còn độ nhạy giúp xác định số nhãn lớp trong thực tế đã được dự đoán đúng. F1 là trung bình điều hòa của độ chính xác và độ nhạy. F1 giúp so sánh hiệu năng các mô hình phân lớp được dễ dàng theo tỉ lệ trung bình.

Các độ đo này được xác định theo công thức như sau:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

và

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

với TP (true positives) là tỉ lệ đo xác định số lần hệ thống phân loại vào đúng nhãn lớp và số lần thực tế nhãn lớp đó xuất hiện, FP (false positives) là tỉ lệ số lần hệ thống phân loại vào đúng nhãn lớp và số lần thực tế nhãn lớp đó không xuất hiện, FN (false negatives) là tỉ lệ số lần hệ thống phân loại vào nhãn lớp khác và số lần thực tế nhãn lớp đó xuất hiện.

Để đánh giá độ hiệu năng, phương pháp kiểm tra chéo 10 lần được áp dụng trong các thử nghiệm. Theo phương pháp này, dữ liệu được chia thành 10 phần, 9 phần được sử dụng để huấn luyện các mô hình học và phần còn lại được sử dụng để đánh giá. Quá trình được lặp lại cho tới khi tất cả các phần đều được đánh giá và khi đó các kết quả sẽ được lấy theo giá trị trung bình.

Mạng CNN đề xuất và mạng nơ-ron nhiều lớp sử dụng trong thử nghiệm được xây dựng và huấn luyện sử dụng bộ thư viện Keras [14], với các tham số đã mô tả trong phần 3. Mạng MLP sử dụng 4 lớp ẩn, số lượng nơ-ron trên các lớp đều là 60 phần tử. Hàm kích hoạt sử dụng cho lớp

ẩn là ReLU, và cho lớp đầu ra là hàm softmax. Bộ tối ưu sử dụng cho mạng là Adam, với tham số learning rate cũng là 0,001.

Ngoài ra, do các dữ liệu huấn luyện cho mô hình phát hiện xâm nhập mất cân bằng, để tăng độ chính xác, nghiên cứu này sử dụng các tham số để quy định trọng số cho các lớp trong các bộ phân lớp, tức là gán một giá trị phạt cho các lớp có số mẫu lớn hơn nhiều các lớp khác một cách hợp lý, giúp bộ phân lớp hoạt động hiệu quả hơn. Việc gán giá trị phạt cho mạng nơ-ron học sâu được thực hiện với tham số class_weight có trong các bộ thư viện Keras [14].

2) Cấu hình các bộ phân lớp khác

Để so sánh hiệu năng của mạng CNN đề xuất, nghiên cứu này triển khai các mô hình phân loại áp dụng các kỹ thuật: mạng nơ-ron sâu đa lớp, máy véc tơ tựa SVM (Support Vector Machine), cây quyết định, rừng ngẫu nhiên (Random Forest) và kỹ thuật phân loại giảm gradient ngẫu nhiên SGD (Stochastic Gradient Descent). Các kỹ thuật dựa trên cây quyết định hay rừng ngẫu nhiên là các kỹ thuật cơ bản và truyền thống của học máy. Tuy nhiên, hiệu năng của các kỹ thuật mới, nhất là SVM, khiến cho việc sử dụng tham chiếu hiệu năng SVM được nhiều người nghiên cứu quan tâm và thử nghiệm cho bài toán phân loại. Khi xét về việc biểu diễn kết quả thì cây quyết định hay rừng ngẫu nhiên dễ hiểu và dễ tiếp cận hơn đặc biệt với người dùng thông thường. Các tham số của các bộ phân lớp trên được xây dựng và huấn luyện tinh chỉnh tối ưu sử dụng hàm grid_search trong bộ thư viện Scikit-learn [17].

Chú ý rằng, mục tiêu của việc xây dựng mô hình phát hiện hành vi xâm nhập thường là tiến hành phân loại các hành vi của người dùng thành các nhóm bình thường (normal) và bất thường. Tuy nhiên trong nhiều trường hợp, các hành vi trong nhóm bất thường này cần được nhận biết chi tiết theo các hành vi tấn công/xâm nhập dos, probe, u2r và r2l (như phân loại trong tập dữ liệu NSL-KDD [18]). Nói cách khác, mô hình phát hiện đề xuất cần phân biệt chi tiết kiểu hành vi bất thường của người dùng chứ không chỉ dừng ở việc phân loại hành vi bình thường hay không. Như phân tích trong phần 4.1 về tập dữ liệu thử nghiệm, tập dữ liệu NSL-KDD [18] thể hiện hành vi thông thường và tấn công (bất thường) trên cả hai tập dữ liệu huấn luyện và kiểm thử tương đối cân bằng xét về số lượng bản ghi. Tuy nhiên, xem xét chi tiết các hành vi xâm nhập cơ bản (dos, probe, r2l, u2r) cho thấy dữ liệu về các dạng tấn công mất cân bằng nghiêm trọng, thể hiện ở số lượng các hành vi tấn công và các biểu hiện của các dạng tấn công.

Như phân tích về tập dữ liệu thử nghiệm NSL-KDD [18], số lượng mẫu hành vi thông thường và bất thường trên cả hai tập dữ liệu huấn luyện và kiểm thử tương đối mất cân bằng. Tuy nhiên, xem xét chi tiết các hành vi xâm nhập cơ bản (dos, probe, r2l, u2r) cho thấy dữ liệu về các dạng tấn công mất cân bằng nghiêm trọng, thể hiện ở số lượng các hành vi tấn công và các biểu hiện của các dạng tấn công. Để giải quyết vấn đề này, kỹ thuật phạt với các giá trị class_weight được áp dụng khi huấn luyện các mô hình phân lớp. Đối với các tham số khác để xây dựng và huấn luyện các bộ phân lớp, các giá trị mặc định được sử dụng đối với bộ thư viện Scikit-learn [17].

C. Kết quả thực nghiệm đối với các bộ phân lớp

Từ quan sát và tiến hành thử nghiệm, ta được các kết quả với thứ tự được trình bày trong các bảng và hình dưới đây lần lượt là: mô hình mạng CNN đề xuất, mạng MLP, cây quyết định CART và rừng ngẫu nhiên (Random Forest), máy véc-tơ tựa SVM và bộ phân lớp giảm gradient ngẫu nhiên SGD.

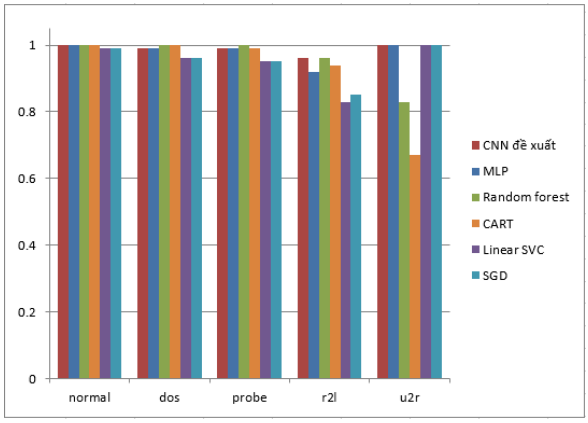
Bảng IV. Hiệu năng của các bộ phân lớp đo theo F1.

BỘ PHÂN LỚP	GIÁ TRỊ F1 TRUNG BÌNH THEO CÁC LỚP (%)
CNN đề xuất	**98,4**
MLP	96,2
Random Forest	95,6
CART	91,8
SVM	86,6
SGD	81,2

Bảng IV cho phép đánh giá trực tiếp các mô hình quan tâm bằng cách kết hợp cả hai yếu tố độ chính xác và độ nhạy nhờ vào độ đo F1 trung bình. Theo trình bày trong bảng, hiệu năng của các bộ phân lớp CNN đề xuất, MLP, Random Forest và CART là cao nhất với giá trị F1 đều lớn hơn 91%. Trong đó mạng CNN đề xuất có giá trị F1 cao nhất với 98,4%, theo sau là mạng MLP là 96,2%, bộ phân lớp Random Forest là 95,6% và cây quyết định CART là 91,8%. Các bộ phân lớp còn lại có giá trị F1 nhỏ hơn hẳn, và kém nhất là SGD với F1 chỉ đạt 81,2%. Kết quả này cho thấy, việc sử dụng số lượng các lớp ẩn phù hợp, mạng nơ-ron học sâu có khả năng học được nhiều đặc trưng hữu ích trong việc phát hiện chính xác các hành vi xâm nhập vào hệ thống thông tin.

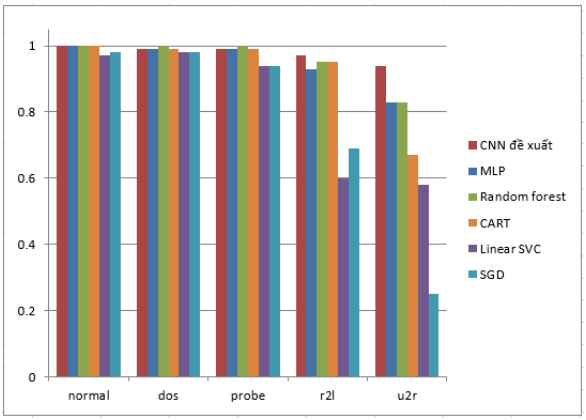
Đặc biệt, mạng CNN đề xuất có khả năng học nhiều đặc trưng tốt hơn, học mô hình phân loại chính xác hơn hẳn nhờ các đặc tính ưu việt so với các mạng nơ-ron thông thường khác. Đó là các tính chất về tính bất biến và tính kết hợp cục bộ, giúp cải thiện hiệu suất nhận dạng bất thường. Tính cục bộ trong các lớp tích chập bảo vệ mô hình khỏi sự ảnh hưởng của nhiễu đối với dữ liệu. Nhờ đó, các đặc trưng được trích xuất trong các lớp tích chập có khả năng chống nhiễu cao. Vì các đặc trưng mức thấp được trích xuất từ dữ liệu bất thường và bất thường có thể tương tự nhau, do đó các phương pháp học máy truyền thống khó phân loại được một cách chính xác. Tuy nhiên, CNN có thể xử lý những điểm tương đồng này bằng cách tạo ra các đặc trưng mức cao và phân biệt nhau, kết hợp từ các đặc trưng mức thấp đã có. Các lớp pooling gộp các giá trị đặc trưng tương tự từ các vị trí khác nhau lại với nhau và gán bằng một giá trị. Lớp pooling có thể kiểm soát việc phát hiện sự bất thường với phân bố khác nhau.

Kết quả trong bảng IV chỉ mô tả đánh giá tổng quan hiệu năng chung của các bộ phân lớp. Để xem xét cặn kẽ hơn khả năng phát hiện xâm nhập của từng bộ phân lớp, dưới đây sẽ trình bày phân tích chi tiết các giá trị về độ chính xác, độ nhạy và F1 đối với từng loại hành vi.



Hình 3. Độ chính xác của các mô hình phân lớp

Như trong Hình 3, khi phân biệt hành vi truy nhập thông thường của người dùng thì tất cả các kỹ thuật phân loại đều có hiệu năng tốt với tỷ lệ chênh lệch nhau khoảng 2% dao động từ (98% đến 100%). Cụ thể, các kỹ thuật truyền thống (không tính SGD) tốt hơn các kỹ thuật khác ngoại trừ MLP. Kết quả này một phần do số lượng dữ liệu về hành vi bình thường của người dùng chiếm phần lớn trong tập dữ liệu sử dụng. Mạng CNN vẫn cho độ chính xác thuộc nhóm đứng đầu trong từng loại tấn công.

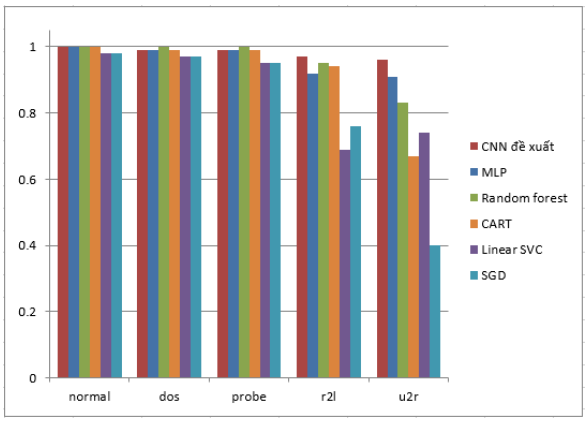


Hình 4. Độ nhạy (Recall) của các mô hình phân lớp

Vấn đề về hiệu năng xuất hiện khi xem xét kết quả phân loại các hành vi xâm nhập từ các thao tác truy nhập của người dùng. Các kỹ thuật truyền thống có độ chính xác rất cao nhất với các hành vi tấn công dos và probe (gần như 100%). Mạng CNN và MLP có kết quả rất sát (99%) với bộ phân loại Random Forest và ngang bằng với bộ phân loại cây quyết định CART. Các kỹ thuật phân loại còn lại dao động trong khoảng từ 93% đến 95%.

Điểm đáng chú ý nhất là độ chính xác khi phân loại hành vi tấn công u2r. Hành vi này chỉ chiếm tỷ lệ rất nhỏ trong toàn bộ tập dữ liệu với gần 120 trường hợp. Mạng CNN và MLP cho kết quả tốt nhất là 100% tương tự mô hình SVM và SGD. Bộ phân loại Random Forest và cây quyết định cũng có hiệu năng tương đối kém với độ chính xác lần lượt là 83% và 67%.

Hình 4 thể hiện độ nhạy của 6 mô hình được khảo sát. Mô hình sử dụng CNN là tốt nhất, sau đó là MLP cũng nằm trong nhóm đứng đầu như các mô hình CART và Random Forest. Hiệu năng tách biệt rõ rệt khi xác định hành vi tấn công r2l và u2r. Kết quả của kỹ thuật nơ-ron truyền thống và SVM chỉ xoay quanh giá trị 60%.



Hình 5. Giá trị F1 của các mô hình

Hình 5 cho phép đánh giá tổng thể các mô hình quan tâm theo độ đo F1. Mô hình CNN cho kết quả tốt nhất, sau đó là MLP cho kết quả tốt ngang bằng với mô hình sử dụng rừng ngẫu nhiên và cây quyết định khi xác định hành vi bình thường của người dùng. Các mô hình còn lại đứng sau với độ chênh lệch khoảng 2%. Khi xác định các hành vi truy nhập bất thường, mô hình CNN vẫn đạt kết quả tốt nhất, tuy nhiên mô hình dựa trên mạng học sâu là MLP kém hơn mô hình Random Forest một chút nhưng tốt hơn các kỹ thuật mạng nơ-ron truyền thống và SVM, vượt trội các mô hình khác khi xác định hành vi u2r.

Các kết quả của mạng CNN và MLP thử nghiệm trong nghiên cứu này không so sánh trực tiếp được với các nghiên cứu sử dụng mạng nơ-ron hồi quy một phần do sự khác biệt về cách đánh giá và chỉ số hiệu năng được công bố. Chỉ có nghiên cứu [11] cung cấp kết quả phân loại chi tiết về từng hành vi xâm nhập. Dù vậy, các kết quả của nghiên cứu này cung cấp thêm góc độ khác về hiệu năng của mạng nơ-ron học sâu cho bài toán phân loại hành vi người dùng hay phát hiện xâm nhập.

V. KẾT LUẬN

Bài báo này nghiên cứu việc sử dụng mạng CNN cho việc phát hiện hành vi xâm nhập mạng trái phép để đảm bảo an toàn cho hệ thống thông tin. Ngoài ra, hiệu năng của mô hình mạng CNN đề xuất được kiểm nghiệm với các mô hình sử dụng các kỹ thuật tiêu biểu khác bao gồm rừng ngẫu nhiên, cây quyết định, giảm gradient ngẫu nhiên, máy véc-tơ tựa SVM, và mạng MLP bằng tập dữ liệu NSL-KDD. Do đặc trưng của tập dữ liệu NSL-KDD, bài báo sử dụng phương pháp đánh giá kiểm tra chéo 10 lần trên toàn bộ tập dữ liệu nhằm đánh giá hiệu năng thuần túy của các kỹ thuật phân loại hành vi truy nhập. Kết quả cho thấy hiệu năng của kỹ thuật CNN thể hiện sự vượt trội so với các mô hình còn lại. Khi xác định chi tiết các hành vi xâm nhập, mô hình dựa trên CNN cũng vượt trội các kỹ thuật khác. Kết quả này đạt được là do các đặc tính ưu việt trong quá trình học đặc trưng của CNN, giúp mô hình có thể học được các đặc trưng tốt nhất để phân loại các tấn công.

LỜI CẢM ƠN

Nghiên cứu sinh được hỗ trợ bởi chương trình học bổng đào tạo tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup, mã số VINIF.2020.TS.94.

TÀI LIỆU THAM KHẢO

[1] Anyanwu, L.O., Keengwe, J. and Arome, G.A., 2010, April. Scalable intrusion detection with recurrent neural networks. In Information Technology: New Generations (ITNG), 2010 Seventh International Conference on (pp. 919-923). IEEE.

[2] Gao, N., Gao, L., Gao, Q. and Wang, H., 2014, November. An intrusion detection model based on deep belief networks. In Advanced Cloud and Big Data (CBD), 2014 Second International Conference on (pp. 247-252). IEEE.

[3] Na, Seung-Hoon. "Advanced Deep Learning." (2020).

[4] Berman, Daniel S., et al. "A survey of deep learning methods for cyber security." Information 10.4 (2019): 122.

[5] Kim, J., Kim, J., Thu, H.L.T. and Kim, H., 2016, February. Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. In Platform Technology and Service (PlatCon), 2016 International Conference on (pp. 1-5). IEEE.

[6] Liao, H.J., Lin, C.H.R., Lin, Y.C. and Tung, K.Y., 2013. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), pp.16-24.

[7] Lippmann, Richard. "An introduction to computing with neural nets." IEEE Assp magazine 4.2 (1987): 4-22.

[8] Niyaz, Q., Sun, W., Javaid, A.Y. and Alam, M., 2015. A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS), BICT-15 (Vol. 15, pp. 21-26).

[9] Salzberg, Steven L. C4. 5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning 16.3 (1994): 235-240.

[10] Schölkopf, B., and Smola A. J.. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: MIT Press, 2002.

[11] Staudemeyer, R.C., 2015. Applying long short-term memory recurrent neural networks to intrusion detection. 10 African Computer Journal, 56(1), pp.136-154.

[12] Tsai, C.F., Hsu, Y.F., Lin, C.Y. and Lin, W.Y., 2009. Intrusion detection by machine learning: A review. Expert Systems with Applications, 36(10), pp.11994-12000.

[13] S. Hettich, S.D. Bay, The UCI KDD Archive. Irvine, CA: University of California, Department of Information and Computer Science, http://kdd.ics.uci.edu, 1999.

[14] Joshi, Deepa, Shahina Anwarul, and Vidyanand Mishra. "Deep Leaning Using Keras." Machine Learning and Deep Learning in Real-Time Applications. IGI Global, 2020. 33-60.

[15] Tang, Y., Zhang, Y.-Q., Chawla, N. V, Krasser, S. (2009), SVMs modeling for highly imbalanced classification, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), IEEE. 39(1), p. 281–8.

[16] Veropoulos, K., Campbell, C., Cristianini, N., others. (1999), Controlling the sensitivity of support vector machines, Proceedings of the International Joint Conference on AI, p. 55–60.

[17] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[18] Tavallaee, Mahbod, et al. Nsl-kdd dataset. http://www. iscx. ca/NSL-KDD (2012).

[19] Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model. 160, 249–264.

[20] Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013, May). Deep convolutional neural networks for LVCSR. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8614-8618). IEEE.

ENHANCED NETWORK INTRUSION DETECTION USING CNN

Abstract: Network intrusion detection (NIDS) is a very attractive topic for both system administrators and security researchers. The problem of intrusion detection can be tackled by machine learning models, based on statistical algorithms or artificial neural networks, to identify abnormal behaviours from those of users accessing systems. However, security attacks tend to be unpredictable today. It is very difficult to build a flexible and effective NIDS with low false alarms and high detection accuracy against unknown attacks. This paper introduces a deep learning model based on Convolutional Neural Network (CNN) to detect intrusions and compare its performance with other machine learning techniques on NSL-KDD dataset. The experimental results of a 98.4% at F1 score show that the proposed CNN-based intrusion detection model could be a potential model for IDS systems.

Keywords: network intrusion detection, NSL-KDD, deep learning, CNN.

Nguyễn Ngọc Điệp. Nhận học vị Tiến sĩ năm 2017. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, an toàn thông tin, xử lý ngôn ngữ tự nhiên.

Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009 tại Hàn Quốc. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.