


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/Pw5horT4sdE>
- Link slides (dạng .pdf đặt trên Github):
https://github.com/ltn0tbug/CS2205.MAR2024/blob/main/An_Interpretable_DL_Based_IDS_using_Shapley_Value.Slide.pdf
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Huỳnh Thái Thi• MSSV: 230202032 	<ul style="list-style-type: none">• Lớp: CS2205.MAR2024• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 3• Link Github: https://github.com/ltn0tbug/CS2205.MAR2024/
--	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

Xây dựng hệ thống phát hiện xâm nhập dựa trên học sâu khả diễn giải sử dụng giá trị Shapley

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

An Interpretable Deep Learning-Based Intrusion Detection Systems using Shapley Value

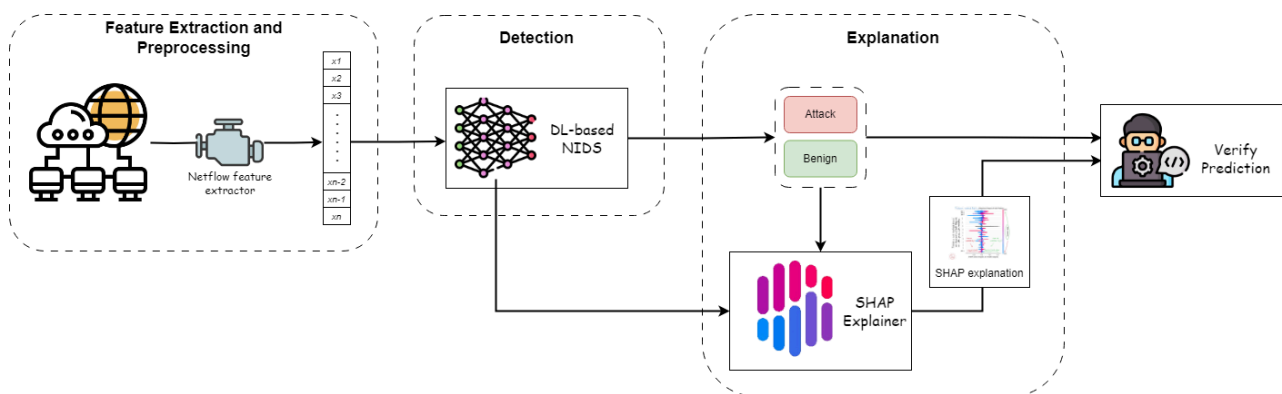
TÓM TẮT (Tối đa 400 từ)

Xu hướng áp dụng học sâu vào hệ thống phát hiện xâm nhập mạng đang nổi lên mạnh mẽ, hứa hẹn mang đến khả năng phòng chống tấn công mạng hiệu quả hơn. Tuy nhiên, các mạng nơ-ron cấu thành nên các mô hình học sâu được xem là “hộp đen” đối với cả người dùng và nhà phát triển trí tuệ nhân tạo, khiến việc diễn giải kết quả dự đoán trở nên khó khăn. Học máy khả giải xuất hiện như giải pháp tiềm năng, cung cấp các công cụ và kỹ thuật để “mở khóa” trí tuệ nhân tạo, giúp người dùng hiểu rõ cơ sở cho các quyết định của mô hình. Nhận thức được sự hạn chế của học máy khả giải trong lĩnh vực an ninh mạng, nghiên cứu này đề xuất một hệ thống phát hiện xâm nhập mạng dựa trên học sâu tích hợp thuật toán học máy khả giải dựa trên giá trị Shapley, qua đó nâng cao khả năng diễn giải cho hệ thống, hỗ trợ nhà nghiên cứu bảo mật đưa ra đánh giá chính xác và tự tin hơn.

GIỚI THIỆU (Tối đa 1 trang A4)

Xu hướng áp dụng học sâu vào hệ thống phát hiện xâm nhập mạng đang nổi lên mạnh mẽ, hứa hẹn mang đến khả năng phòng chống tấn công mạng hiệu quả hơn. Tuy nhiên, các mạng nơ-ron cấu thành nên các mô hình học sâu thường được xem là “hộp đen” bởi chúng gần như không có khả năng diễn giải quá trình suy luận cũng như kết quả mà chúng đưa ra theo cách mà con người có thể hiểu được. Tuy nhiên, khả năng diễn giải trên là đặc biệt quan trọng đối với các nhà phân tích bảo mật trong việc xác minh dự đoán của hệ thống nhằm có những phản hồi phù hợp.

Học máy khả giải xuất hiện như giải pháp tiềm năng, cung cấp các công cụ và kỹ thuật giúp người dùng hiểu rõ cơ sở cho các quyết định của mô hình. Tuy nhiên, việc ứng dụng học máy khả giải vào các hệ thống phát hiện xâm nhập sử dụng học sâu chỉ mang thiếu sót và còn nhiều hạn chế như chỉ phù hợp với một số mô hình nhất định hoặc độ chính xác còn thấp. Chính vì vậy, nghiên cứu này được tiến hành nhằm mục tiêu xây dựng một hệ thống phát hiện xâm nhập mạng dựa trên học sâu và được tăng cường khả năng diễn giải dựa trên giá trị Shapley thông qua bộ khung SHAP [2]. Giá trị Shapley được sử dụng là bởi đây là một thuật toán có cơ sở lý thuyết vững chắc, đảm bảo được tính chính xác cục bộ (local accuracy), tính thiếu (missingness) và tính nhất quán (consistency); và có thể diễn giải cho mọi loại mô hình trí tuệ nhân tạo hiện có (agnostic-model).



Hình 1: Mô hình đề xuất

Cụ thể, mô hình của chúng tôi như trong **Hình 1** sẽ có *Input* và *Output* như sau:

- *Input*: Lưu lượng mạng
- *Output*: Dự đoán đây là một lưu lượng mạng bình thường hay bất thường và diễn giải cho dự đoán trên dựa trên bộ khung SHAP.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Xây dựng các mô hình phát hiện xâm nhập dựa trên các mạng học sâu như CNN, LSTM, GRU, ... sử dụng bộ dataset NF-ToN-IoT [4].
- Diễn giải được dự đoán của mô hình phát hiện xâm nhập sau áp dụng bộ khung SHAP theo cách mà các nhà phân tích có thể hiểu được.

- So sánh ưu và nhược điểm của diễn giải từ bộ khung SHAP với các thuật toán học máy khả giải khác như LIME [1], Anchors [3],

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung 1: Thu thập và tiền xử lý dữ liệu

- Sử dụng bộ dữ liệu NF-ToN-IoT, bao gồm lưu lượng mạng được ghi lại từ các thiết bị IoT bị xâm nhập và lưu lượng mạng bình thường đã được trích xuất thuộc tính dựa trên giao thức NetFlow.
- Tiền xử lý dữ liệu để đảm bảo tính nhất quán và phù hợp với định dạng đầu vào của các mô hình học sâu. Các bước tiền xử lý có thể bao gồm:
 - Chuẩn hóa dữ liệu: biến đổi dữ liệu về cùng một thang đo.
 - Xử lý giá trị thiếu: thay thế các giá trị thiếu bằng giá trị trung bình, trung vị hoặc phương pháp khác phù hợp.
 - Xử lý ngoại lệ: loại bỏ hoặc điều chỉnh các giá trị bất thường.
 - Chia dữ liệu thành tập huấn luyện, tập kiểm tra và tập đánh giá.

Nội dung 2: Xây dựng mô hình học sâu

- Tham khảo nghiên cứu và thiết kế đi trước để xây dựng nên các mô hình phát hiện xâm nhập dựa trên các mạng học sâu như CNN, LSTM, GRU, ...
- Huấn luyện mô hình trên tập huấn luyện và đánh giá hiệu quả mô hình trên tập NF-ToN-IoT đã tiền xử lý.

Nội dung 3: Áp dụng bộ khung SHAP để diễn giải mô hình

- Sử dụng bộ khung SHAP để tính toán giá trị Shapley (thể hiện mức độ ảnh hưởng của từng thuộc tính đối với dự đoán của mô hình) cho mỗi thuộc tính đầu vào.
- Biểu diễn giá trị Shapley bằng các phương pháp trực quan như biểu đồ thanh,

biểu đồ nhiệt để giúp người dùng dễ dàng hiểu được.

Nội dung 4: So sánh độ chính xác của SHAP với các thuật toán học máy khả giải khác.

- Áp dụng các thuật toán học máy khả giải khác như LIME, Anchors vào mô hình phát hiện xâm nhập.
- So sánh độ chính xác của các thuật toán học máy khả giải trong việc giải thích dự đoán của mô hình.
- Đánh giá ưu và nhược điểm của từng thuật toán để lựa chọn phương pháp phù hợp nhất cho bài toán cụ thể.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Xây dựng được một hệ thống phát hiện xâm nhập mạng dựa trên học sâu và được tăng cường khả năng diễn giải bằng giá trị Shapley.
- Khả năng diễn giải của hệ thống giúp các nhà phân tích bảo mật có thể dễ dàng hiểu được cơ sở cho dự đoán của mô hình.
- So sánh được ưu và nhược điểm của các thuật toán học máy khả giải dựa trên giá trị Shapley so với một số thuật toán khác.

TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1]. Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144
- [2]. Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4765-4774
- [3]. Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018: 1527-1535
- [4]. Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, Marius Portmann: NetFlow Datasets for Machine Learning-based Network Intrusion Detection Systems. CoRR

abs/2011.09144 (2020)