


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
(ví dụ: <https://www.youtube.com/watch?v=AWq7uw-36Ng>)
- Link slides (dạng .pdf đặt trên Github):  
(ví dụ: <https://github.com/mynameuit/CS2205.APR2023/TenDeTai.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"><li>• Họ và Tên: Tô Trọng Nghĩa</li><li>• MSHV: 220202019</li></ul>  | <ul style="list-style-type: none"><li>• Lớp: CS2205.CH1702</li><li>• Tự đánh giá (điểm tổng kết môn): 10/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 3</li><li>• Số câu hỏi QT của cả nhóm: 15</li><li>• Link Github:<br/><a href="https://github.com/TTNghiaUIT/220202019_CS2205_Final_Project">https://github.com/TTNghiaUIT/220202019_CS2205_Final_Project</a></li></ul> |
|--|---|

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

HỆ THỐNG PHÁT HIỆN MÃ ĐỘC CÓ KHẢ NĂNG DIỄN GIẢI DỰA TRÊN HỌC TỔNG HỢP

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

AN EXPLAINABLE MALWARE DETECTION SYSTEM BASED ON ENSEMBLE LEARNING

## TÓM TẮT (*Tối đa 400 từ*)

Đảm bảo an ninh cho hệ thống thông tin trong thời đại hiện nay là một yêu cầu cực kỳ quan trọng do sự phức tạp và khối lượng ngày càng gia tăng của mã độc và các mối đe dọa. Tuy nhiên, các phương pháp tiếp cận dựa trên chữ ký (signature-based approaches) truyền thống gặp rất nhiều hạn chế khi không thể xác định được những phần mềm độc hại chưa biết hay cũng như dễ bị tác động bởi các mẫu đối kháng (adversarial samples). Để vượt qua những thách thức này, việc ứng dụng trí tuệ nhân tạo (Artificial Intelligence) và học máy (Machine Learning) trong việc phát hiện phần mềm độc hại đã trở thành một hướng đi triển vọng để cải thiện độ chính xác và hiệu quả của hệ thống phòng vệ. Trong nghiên cứu này, chúng tôi đề xuất một hệ thống phát hiện phần mềm độc hại có khả năng giải thích các kết quả trả về với tên gọi MalDEX. MalDEX sẽ tận dụng việc trích xuất thuộc tính từ các tệp Portable Executable (PE) trên hệ điều hành Windows và sử dụng một loạt bộ phân loại, bao gồm các thuật toán đơn (Single Algorithms), các thuật toán tổng hợp đồng nhất (Homogeneous Ensemble Algorithms) và các thuật toán tổng hợp không đồng nhất (Heterogeneous Ensemble Algorithms), để xác định sự hiện diện của phần mềm độc hại. Để đánh giá chất lượng mô hình và giải quyết vấn đề tính diễn giải, chúng tôi áp dụng phương pháp Kernel SHAP. Phương pháp này sẽ giúp hệ thống của chúng tôi có khả năng giải thích toàn diện quá trình đưa ra quyết định của mô hình học máy bằng cách sử dụng một hàm hồi quy tuyến tính có trọng số để tính toán tầm quan trọng của từng đặc trưng đầu vào.

## GIỚI THIỆU (*Tối đa 1 trang A4*)

Việc đảm bảo an toàn cho các hệ thống thông tin luôn là một yêu cầu cần thiết và cấp thiết trong xã hội hiện nay. Số lượng và độ phức tạp của các mối đe dọa và chương trình độc hại ngày càng tăng, điều này dẫn đến những thách thức trong việc nghiên cứu và phát triển các giải pháp giải quyết. Việc áp dụng trí tuệ nhân tạo vào các giải pháp này là tất yếu phải diễn ra khi trí tuệ nhân tạo có thể giải phóng sức lao động và giải quyết được nhiều vấn đề phức tạp.

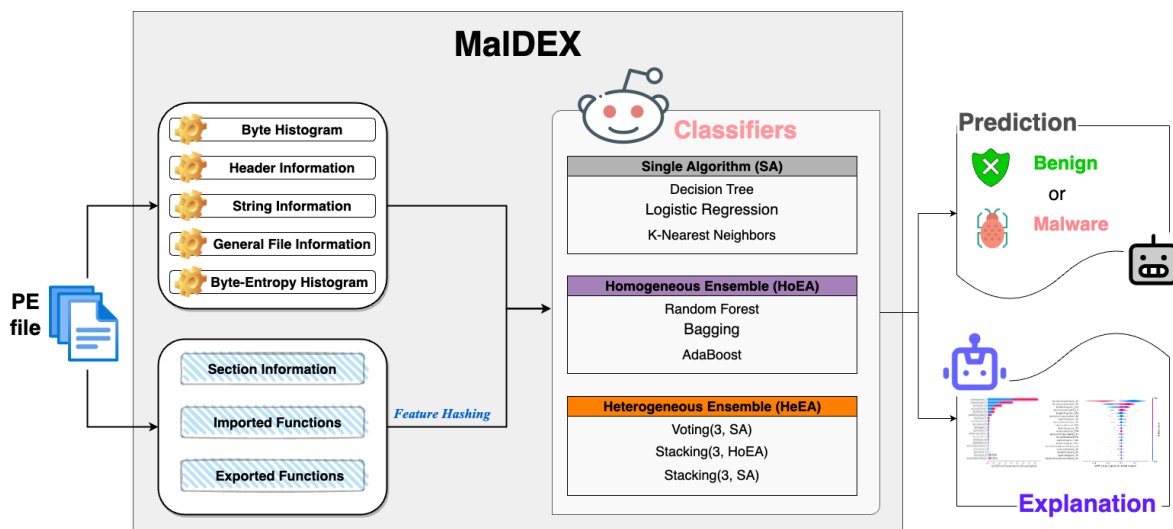
Rất nhiều công trình nghiên cứu trong việc phát hiện lỗi hỏng hay mã độc áp dụng trí tuệ nhân tạo đã được triển khai và đạt nhiều thành công nhất định [1, 2, 3]. Tuy nhiên, chúng vẫn còn gặp nhiều hạn chế trong việc giải quyết các vấn đề liên quan đến tính tin cậy của các kết quả trả về.

Để giải quyết vấn đề này, trong một số nghiên cứu khác như [4, 5, 6] đã ứng dụng thành công các phương pháp diễn giải (Explainable Artificial Intelligence - XAI) vào mô hình học máy. Tuy nhiên, các mô hình diễn giải này lại thiếu tính thực tế khi sử dụng các bộ dữ liệu đã được xử lý sẵn, ít đặc trưng và chỉ phù hợp với một số mô hình nhất định. Với số lượng ít các đặc trưng như vậy, liệu rằng mô hình học máy có thể đưa ra lời giải thích hợp lý cho các quyết định của nó và sử dụng lại dữ liệu đó trong các nghiên cứu trong tương lai hay không?

Dựa trên những động lực này, chúng tôi đề xuất một hệ thống phát hiện phần mềm độc hại mới với tên gọi MalDEX có khả năng diễn giải, tức với đầu vào là tệp PE trên Windows, hệ thống trả về kết quả về loại của mẫu này (lành tính hay độc hại), kèm theo đó là lời giải thích về lý do đưa ra quyết định đó (mức độ ảnh hưởng của các đặc trưng trong mẫu đưa vào). Ngoài ra, để tăng tính chính xác và đa dạng của các kết quả trả về, MalDEX còn ứng dụng các thuật toán học tổng hợp vào trình phân loại để tăng tính chính xác và sự đa dạng cho các kết quả trả về. Cụ thể:

**Input:** một chương trình thực thi định dạng Portable Executable trên hệ điều hành Windows.

**Output:** kết quả dự đoán về loại của tệp (lành tính hay độc hại) và bảng mô tả tầm quan trọng của các đặc trưng đối với kết quả dự đoán đó.



Hình 1. Mô hình tổng quan hệ thống đề xuất MalDEX

## MỤC TIÊU

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

- Xây dựng mô-đun để trích xuất các thuộc tính từ mẫu chương trình thực thi định dạng PE trên hệ điều hành Windows.
- Xây dựng các trình phân loại mã độc dựa trên các đặc trưng đã được trích xuất trên (phân loại các mẫu lành tính và độc hại) sử dụng 3 thuật toán chính bao gồm các thuật toán đơn, các thuật toán tổng hợp đồng nhất và các thuật toán tổng hợp không đồng nhất.
- Áp dụng phương thức diễn giải Kernel SHAP cho tất cả các bộ phân loại nhằm mục đích tính toán tầm quan trọng của từng đặc trưng đầu vào (được trích xuất ở trên) đối với kết quả dự đoán (lành tính hay độc hại).

## NỘI DUNG VÀ PHƯƠNG PHÁP

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

**Nội dung 1:** Xây dựng mô-đun trích xuất các đặc trưng cho hệ thống MalDEX

**Phương pháp:**

- Tìm hiểu cấu trúc và đặc điểm của chương trình thực thi định dạng PE trên hệ điều hành Windows.
- Tìm hiểu và nghiên cứu các phương pháp trích xuất các đặc trưng của tệp PE để sử dụng cho các trình phân loại.

- Nghiên cứu và thiết kế mô-đun trích xuất các đặc trưng cho hệ thống MalDEX.

## **Nội dung 2: Xây dựng các trình phân loại mã độc cho hệ thống MalDEX**

### **Phương pháp:**

- Tìm hiểu và nghiên cứu các thuật toán, phương pháp sử dụng thuật toán học tổng cho các trình phân loại mã độc.
- Nghiên cứu và thiết kế các trình phân loại mã độc dựa trên các đặc trưng đã được trích xuất ở nội dung 1.
- Đánh giá hiệu năng của các trình phân loại trong việc phát hiện các chương trình độc hại. So sánh, đánh giá hiệu năng của các trình phân loại sử dụng thuật toán tổng hợp với các trình phân loại sử dụng thuật toán đơn.

## **Nội dung 3: Áp dụng phương pháp diễn giải cho các trình phân loại**

### **Phương pháp:**

- Tìm hiểu, nghiên cứu các phương pháp tăng khả năng giải thích của các mô hình học máy, đặc biệt là Kernel SHAP.
- Nghiên cứu và xây dựng các trình phân loại trong MalDEX sử dụng Kernel SHAP để giải thích các kết quả trả về.
- Đánh giá và nhận xét các kết quả giải thích của các bộ phân loại.

## **KẾT QUẢ MONG ĐỢI**

- Hệ thống MalDEX có khả năng đưa ra các loại phân loại đúng với các đầu vào tương ứng cùng với hình ảnh biểu thị tầm quan trọng của các đặc trưng cho quyết định đó.
- Bản báo cáo trình bày đánh giá hiệu năng của các thuật toán tổng hợp trong việc phân loại các chương trình thực thi định dạng PE.
- Bài báo khoa học về hệ thống phát hiện mã độc MalDEX có khả năng diễn giải dựa trên học tổng hợp.

## **TÀI LIỆU THAM KHẢO (Định dạng DBLP)**

[1]. Gueltoum Bendiab, Stavros Shiaeles, Abdulrahman Alruban, Nicholas Kolokotronis: IoT Malware Network Traffic Classification using Visual Representation and Deep Learning. NetSoft 2020: 444-449

- [2]. Jan Stiborek, Tomás Pevný, Martin Reháč: Multiple instance learning for malware classification. *Expert Syst. Appl.* 93: 346-357 (2018)
- [3]. Irina Baptista, Stavros Shiaeles, Nicholas Kolokotronis: A Novel Malware Detection System Based on Machine Learning and Binary Visualization. *ICC Workshops 2019*: 1-6
- [4]. Francesco Paolo Caforio, Giuseppina Andresini, Gennaro Vessio, Annalisa Appice, Donato Malerba: Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems. *DS 2021*: 385-400
- [5]. Basim Mahbooba, Mohan Timilsina, Radhya Sahal, Martin Serrano: Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model. *Complex.* 2021: 6634811:1-6634811:11 (2021)
- [6]. Rafa Alenezi, Simone A. Ludwig: Explainability of Cybersecurity Threats Data Using SHAP. *SSCI 2021*: 1-10