


## THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/I2Wf5ylsETU>
- Link slides: (dạng .pdf đặt trên Github):  
(ví dụ: <https://github.com/PhanNgocVuUIT/CS2205.MAR2024.git>)

<ul style="list-style-type: none"><li>• Họ và Tên: Phan Ngọc Vũ</li><li>• MSSV: 210201023</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS2205.APR2024</li><li>• Tự đánh giá (điểm tổng kết môn): 9/10</li><li>• Số buổi vắng: 1</li><li>• Số câu hỏi QT cá nhân: 3<ul style="list-style-type: none"><li>• Link Github: <a href="https://github.com/PhanNgocVuUIT/CS2205.MAR2024.git">https://github.com/PhanNgocVuUIT/CS2205.MAR2024.git</a></li></ul></li><li>• Link YouTube video: <a href="https://youtu.be/I2Wf5ylsETU">https://youtu.be/I2Wf5ylsETU</a></li></ul>
--	--

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

CẢI TIẾN BỘ DỮ LIỆU PHÂN LOẠI MÃ ĐỘC TRONG TỆP THỰC THI  
WINDOWS BẰNG THUẬT TOÁN DI TRUYỀN

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

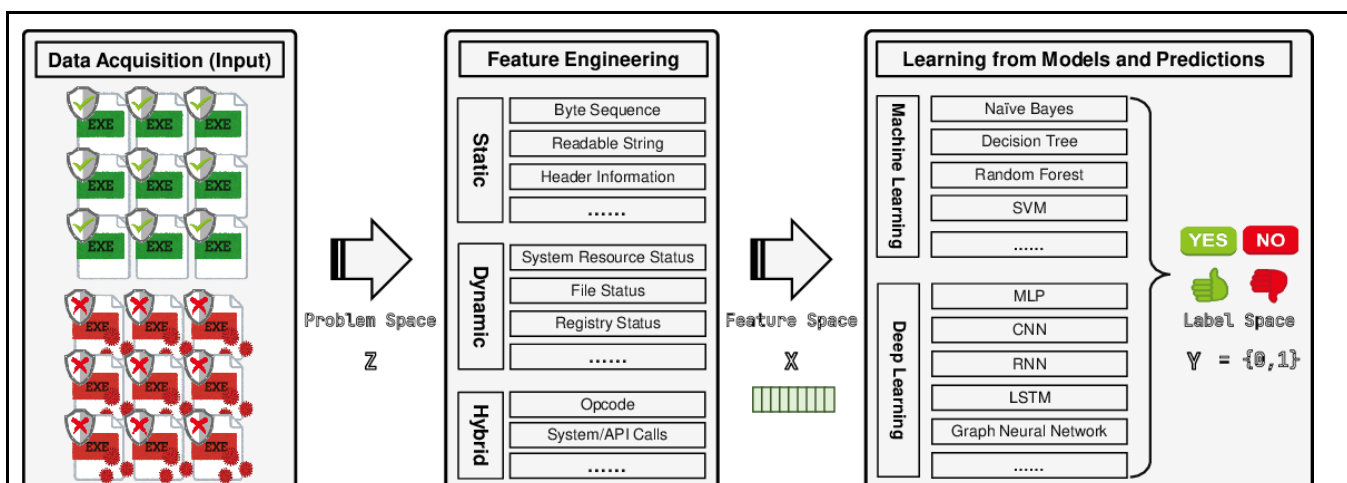
ENHANCING MALWARE CLASSIFICATION DATASETS FOR PORTABLE  
EXECUTABLE FILES VIA GENETIC ALGORITHMS

## TÓM TẮT

Việc phát hiện và ngăn chặn các tập tin Portable Executable chứa mã độc là rất quan trọng để bảo vệ hệ thống máy tính khỏi những nguy cơ an ninh mạng. Ngày nay, phân loại mã độc Portable Executable (PE) bằng các thuật toán học máy là một lĩnh vực nghiên cứu chủ yếu trong an ninh mạng. Các bộ dữ liệu hiện có như EMBER [1], SOREL-20M [2], và BODMAS [3] cung cấp nguồn lực quan trọng để phát triển mô hình học máy, nhưng chưa tối ưu trong việc lựa chọn và tinh chỉnh đặc trưng. Đề tài này giới thiệu một cải tiến mới là áp dụng thuật toán di truyền để chọn lọc và giảm số lượng đặc trưng của file PE nhằm giảm kích thước bộ dữ liệu, rút ngắn thời gian huấn luyện trong quá trình phân loại.

## GIỚI THIỆU (Tối đa 1 trang A4)

Trong bối cảnh nghiên cứu hiện đại, việc tạo ra các bộ dữ liệu mở như EMBER, SOREL-20M và BODMAS đã mang lại tiến bộ đáng kể trong phân loại mã độc PE, mở rộng nguồn lực cho việc huấn luyện mô hình học máy. Các bộ dữ liệu này đã thực hiện hướng tiếp cận như sau để trích xuất đặc trưng và tạo bộ dữ liệu:



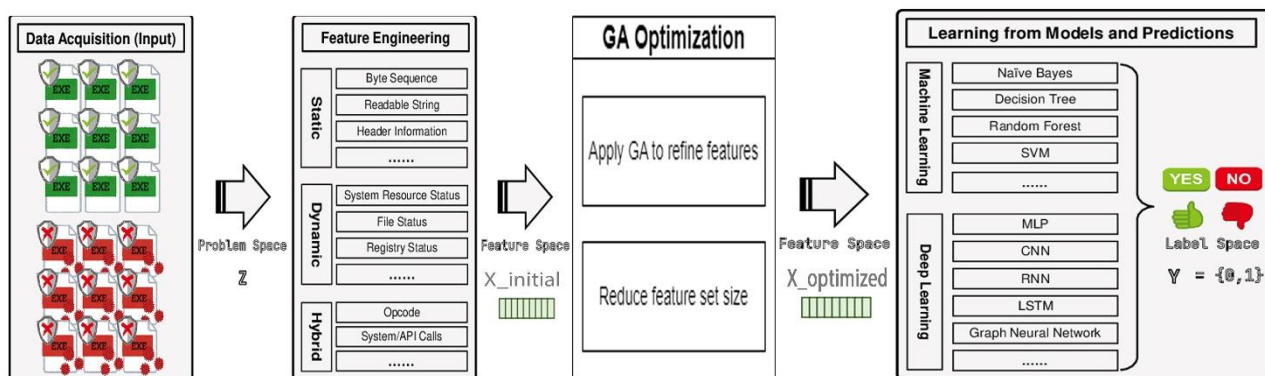
Hình 1: Quy trình phân trích xuất đặc trưng và dùng thuật toán học máy để phân loại mã độc [4]

Trong quy trình trên đây việc trích xuất đặc trưng xong và sử dụng luôn các đặc trưng đó để tạo bộ dữ liệu sẽ gặp 1 hạn chế là việc giảm kích thước dữ liệu lớn chưa được chú trọng đúng mức, dẫn đến số lượng đặc trưng quá lớn, tạo ra thách thức trong việc lưu trữ và xử lý dữ liệu. Bảng thống kê sau có được từ việc tổng hợp 3 công trình nghiên cứu đã được trích dẫn ở [1][2][3]:

Bảng 1: Số lượng đặc trưng và kích thước dữ liệu của EMBER, SOREL-20M và BODMAS

Bộ dữ liệu	Số lượng đặc trưng	Kích thước dữ liệu
EMBER	25.000	1.2 GB
SOREL-20M	20.000	2.3 GB
BODMAS	30.000	4.1 GB

Đề tài đề xuất phương pháp chọn lọc đặc trưng là sau khi trích xuất các đặc trưng từ tập tin PE, đề tài sẽ áp dụng thuật toán GA để chọn lọc lại, loại bỏ những đặc trưng không cần thiết, qua đó giảm dung lượng dữ liệu. Quy trình sẽ được cải tiến như sau:



Hình 2: Mô hình bổ sung thuật toán GA để cải tiến quy trình

Từ quy trình đề xuất ở trên đây, đề tài được đưa về bài toán như sau:

*Input:*

- Tập hợp các file Portable Executable (PE) là mã độc (malware).
- Tập hợp các file Portable Executable (PE) là mã lành (benign).

*Output:*

- Bộ dữ liệu đã được tinh chỉnh số lượng đặc trưng dùng để phân loại tập tin file Portable Executable (PE) là mã độc hay mã lành.

## MỤC TIÊU

- Tạo ra bộ dữ liệu UIT\_Pe: Ứng dụng thuật toán di truyền (GA) để giảm ít nhất 50% số lượng đặc trưng từ file PE, tạo ra bộ dữ liệu UIT\_Pe gọn nhẹ dùng để phân biệt giữa mã độc và mã lành.
- Huấn luyện và đánh giá hiệu suất của bộ dữ liệu UIT\_Pe bằng các mô hình RandomForestClassifier, GradientBoostingClassifier, và LogisticRegression trên bộ dữ liệu UIT\_Pe, với kỳ vọng cải thiện độ chính xác tối thiểu 10% so với bộ dữ liệu gốc và giảm thời gian huấn luyện xuống 20%.

## PHẠM VI

- Nghiên cứu tập trung vào việc phân tích và giảm kích thước bộ dữ liệu thông qua việc tối ưu hóa các đặc trưng tĩnh của file PE.
- Sử dụng các file PE bao gồm 1513 file malware từ bộ dữ liệu Bodmas và 1504 file benign từ Windows 11 để trích xuất đặc trưng, xây dựng bộ dữ liệu UIT\_Pe, rồi dùng UIT\_Pe để huấn luyện và đánh giá mô hình.

## ĐỐI TƯỢNG

- Các đặc trưng tĩnh trong file PE dành cho hệ điều hành Windows, với mục tiêu tìm ra những đặc trưng quan trọng nhất có ảnh hưởng đến việc phân loại mã độc và mã lành.

## **NỘI DUNG VÀ PHƯƠNG PHÁP**

### **Nội dung thực hiện:**

1. Thu thập và chuẩn bị dữ liệu:
  - Thu thập malware và benign từ nguồn Bodmas và Windows 11.
  - Tiền xử lý dữ liệu bao gồm kiểm tra tính toàn vẹn của các tệp, loại bỏ các tệp hỏng, và chuẩn hóa các tệp thành định dạng phù hợp để sử dụng trong quá trình trích xuất đặc trưng.
2. Trích xuất đặc trưng từ file PE:
  - Phân tích cấu trúc của file PE để hiểu các thành phần như headers, sections,...
  - Trích xuất các đặc trưng tĩnh : header, sections, imports, ... từ các tập tin PE.
3. Lựa chọn đặc trưng bằng thuật toán di truyền (GA):
  - Áp dụng thuật toán di truyền để lựa chọn các đặc trưng quan trọng và loại bỏ các đặc trưng ít quan trọng mà không ảnh hưởng đến quá trình phân loại, giúp giảm kích thước của bộ dữ liệu.
4. Huấn luyện và đánh giá bộ dữ liệu thu được trên các mô hình học máy:
  - Sử dụng các mô hình học máy như RandomForest, LogisticRegression, và GradientBoostingClassifier để huấn luyện trên bộ dữ liệu đã được tối ưu hóa.
  - Đánh giá hiệu suất của các mô hình dựa trên các chỉ số như độ chính xác, điểm F1, và ROC AUC.

### **Phương pháp thực hiện:**

Đề tài sẽ sử dụng các phương pháp và công cụ cụ thể sau để thực hiện các bước trên:

- Phân tích và tiền xử lý dữ liệu: Sử dụng thư viện pefile và LIEF để phân tích cấu trúc và trích xuất dữ liệu từ file PE. Tiến hành tiền xử lý dữ liệu bao gồm chuẩn bị dữ liệu và loại bỏ các tệp không hợp lệ.
- Lựa chọn đặc trưng bằng thuật toán di truyền (GA): Sử dụng thuật toán di truyền để tìm ra tập hợp các đặc trưng tối ưu dựa trên các tiêu chí đánh giá đã xác định trước đó.
- Huấn luyện và đánh giá mô hình: Sử dụng Jupyter Notebook trên môi trường Anaconda để thực hiện quá trình huấn luyện và đánh giá mô hình trên dữ liệu đã được tối ưu hóa.

## **KẾT QUẢ MONG ĐỢI**

Đề tài được thực hiện với kỳ vọng sẽ đạt được các kết quả sau:

- Giảm kích thước bộ dữ liệu: Thông qua áp dụng thuật toán GA, dự kiến sẽ giảm được 50% số lượng đặc trưng, từ đó giảm 30% dung lượng lưu trữ và rút ngắn 20% thời gian xử lý cho việc phân loại.

- Hiệu suất: Các mô hình học máy, sau khi được huấn luyện trên bộ dữ liệu đã được tối ưu hóa, dự kiến sẽ đạt độ chính xác cao hơn 5% so với trước, với thời gian huấn luyện giảm 20%.

### **TÀI LIỆU THAM KHẢO**

- [1] Hyrum S. Anderson, Phil Roth: EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. ArXiv 2018: [arxiv.org/abs/1804.04637](https://arxiv.org/abs/1804.04637).
- [2] Richard Harang, Ethan M. Rudd: SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection. Conference on Applied Machine Learning for Information Security 2021.
- [3] Limin Yang, Arridhana Ciptadi, Ihar Laziuk, Ali Ahmadzadeh, Gang Wang: BODMAS: An Open Dataset for Learning based Temporal Analysis of PE Malware. University of Illinois at Urbana-Champaign and Blue Hexagon.
- [4] Xiang Ling et al.: Adversarial Attacks against Windows PE Malware Detection: A Survey of the State-of-the-Art. Computers & Security, December 23, 2021. DOI:10.1016/j.cose.2023.103134