**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

---------



**BÁO CÁO ĐỒ ÁN**

**PHÂN LOẠI MÃ ĐỘC PDF BẰNG MÁY HỌC**

**Sinh viên thực hiện:**

Trịnh Tuấn Tú – 21522747

Nguyễn Ngọc Thanh Sang - 21522544

**Giảng viên:**

TS. Nguyễn Tấn Cầm

Thành phố Hồ Chí Minh, tháng 11 năm 2023

**BÁO CÁO TÓM TẮT**

**1. Tiêu đề báo cáo: PHÂN LOẠI MÃ ĐỘC PDF BẰNG MÁY HỌC**

**2.** **Danh sách thành viên**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ tên** | **Ghi chú** |
| 21522747 | Trịnh Tuấn Tú |  |
| 21522544 | Nguyễn Ngọc Thanh Sang | Nhóm trưởng |

**3.** **Nội dung chi tiết**

***Abstract***

***Nội dung 1: Giới thiệu***

***Nội dung 2: Các nghiên cứu liên quan***

***Nội dung 3: Đề xuất hệ thống phân loại***

- Đề mục 1: Giới thiệu dataset và tiền xử lý dữ liệu

- Đề mục 2: Ứng dụng các phương pháp máy học và huấn luyện mô hình

- Đề mục 3: Kiểm tra và đánh giá mô hình

***Nội dung 5: Phân tích kết quả***

***Nội dung 6: Kết luận và nhận xét***

***Tài liệu tham khảo***

**PHÂN LOẠI MÃ ĐỘC PDF BẰNG MÁY HỌC**

TS. Nguyễn Tấn Cầm

Khoa học và Kỹ thuật Thông tin, Trường Đại học Công Nghệ Thông Tin, TP.Hồ Chí Minh

Nguyễn Ngọc Thanh Sang, Trịnh Tuấn Tú

Khoa học và Kỹ thuật Thông tin, Trường Đại học Công Nghệ Thông Tin, TP Hồ Chí Minh

**Abstract:** Hiện nay, Portable document format (PDF) là loại định dạng file được sử dụng rất phổ biến. Chính vì vậy, các hacker đã tạo ra nhiều phương pháp đưa các vector lây nhiễm vào các tập tin PDF. Thông thường, hacker sẽ ẩn mã độc nhúng vào trong tập tin pdf nhằm truy cập vào máy tính của nạn nhân để khai thác thông tin. Điều này đã tạo nên sự đe dọa bảo mật nghiêm trọng. Vì vậy, các nhà nghiên cứu đã đưa nhiều phương pháp phát hiện, phân loại PDF độc hại và PDF bình thường. Trong bài báo này, chúng tôi đề xuất các mô hình máy học nhằm phân loại các tập tin PDF độc hại. Mô hình được đề xuất sử dụng thuật toán XGBoost và AdaBoost được train và test trên tập dữ liệu CIC-PDF Malware 2022. Qua đánh giá, mô hình dự đoán XGBoost đạt độ chính xác 99,54%. Do đó, mô hình này có thể sử dụng để phát hiện các tập tin pdf độc hại rất hiệu quả.

**Keywords:** mã độc PDF; machine learning; phân loại; AdaBoost; XGBoost; an ninh mạng.



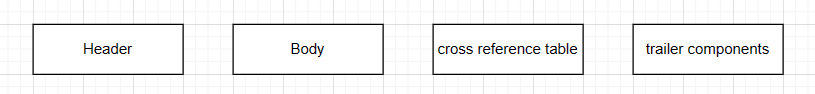
**1. Giới thiệu:**

Ngày nay, dựa vào tốc độ phát triển cực kì nhanh chóng của công nghệ thông tin, con người đã xây dựng và phát triển nhiều thành tựu to lớn. Song, cũng dựa vào công nghệ thông tin mà con người đã tạo ra các tác động tiêu cực gây ảnh hưởng đến xã hội. Một trong số đó là việc tấn công an ninh mạng, mang đến sự đe dọa lớn đến an ninh thông tin, gây ra nhiều tác động tiêu cực cho xã hội.

Các đối tượng xấu đã sử dụng rất nhiều phương pháp tạo ra nhiều phần mềm độc hại đe dọa đến sự an toàn thông tin. Việc phát hiện và phân loại các phần mềm độc hại luôn là một thách thức không hề nhỏ. Nhiều phần mềm, hệ thống ngăn chặn các phần mềm độc hại ra đời thì các phần mềm độc hại càng tinh vi, khả năng che giấu hành vi ngày càng phức tạp hơn.

Một trong số đó là ẩn các đoạn mã độc và nhúng vào tập tin PDF – tập tin cực kì phổ biến được con người sử dụng rất nhiều trên toàn thế giới. Các kẻ tấn công đã sử dụng nhiều kỹ thuật làm giảm khả năng phát hiện và chống lại các phần mềm PDF độc hại.

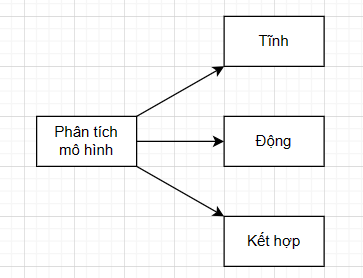
Các nhà nghiên cứu đã đưa ra nhiều phương pháp để phát hiện cũng như ngăn chặn các phần mềm PDF độc hại. Các phương pháp sử dụng máy học được đánh giá rất cao trong việc ngăn chặn các phần mềm độc hại. Tuy nhiên, nhiều kỹ thuật đối nghịch đã gây khó khăn trong việc phát hiện và sự đe dọa này vẫn chưa được khắc phục hoàn toàn.



*Hình 1: Cấu trúc của một tệp PDF.*

Hình 1 là cấu trúc chính của 1 tập tin PDF, bao gồm header chứa thông tin về phiên bản của PDF và các thông số khác như kích thước trang và mã hóa. Body chứa nội dung chính của tài liệu, bao gồm văn bản, hình ảnh, bảng, và các đối tượng khác. Cross reference table nhằm theo dõi vị trí của các đối tượng trong tệp PDF, giúp nhanh chóng truy cập đến chúng. Trailer components chứa các thông tin cuối cùng về tệp PDF, bao gồm đối chiếu với bảng quy chiếu và chỉ định vị trí của phần bắt đầu của nội dung.

Các phương pháp phân tích, phát hiện PDF độc hại thường sử dụng phương pháp tĩnh, động và kết hợp được mô tả như hình 2. Các phương pháp tĩnh phân tích kiểm tra mẫu mà không run code và phụ thuộc vào các thuộc tính như là code structure. Phương pháp động run code để quan sát hành vi của đối tượng. Phương pháp kết hợp chống lại mã độc pdf hiệu quả nhưng phức tạp và tốn nhiều thời gian.

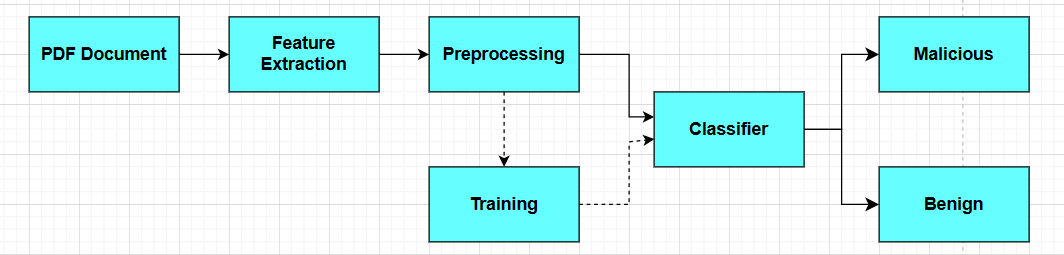


*Hình 2: Các phương pháp phân tích*

Có nhiều phương pháp đã được đề xuất và sử dụng, tuy nhiên trong bài nghiên cứu này, chúng tôi đề xuất mô hình sử dụng thuật toán XGBoost được train và test trên tập dữ liệu CIC-Evasive-PDFMal 2022. Qua huấn luyện và đánh giá, mô hình dự đoán đạt độ chính xác 99,54%. Vì vậy, mô hình này có thể phát hiện cá phần mềm PDF độc hại rất hiệu quả.

**2. Các nghiên cứu liên quan:**

Phân loại tập tin PDF bất kỳ là có hại hay lành tính là mục tiêu chính của việc dùng máy học để nhận dạng tài liệu độc hại. Các phương pháp học máy có giám sát thường được sử dụng. Các phương pháp đó hoạt động bằng cách kiểm tra dữ liệu được lấy từ nội dung hoặc cấu trúc của tài liệu. Hình 2 là mô tả quy trình chung cách hoạt động của các phương pháp máy học.



*Hình 3: Quy trình chung của các phương pháp máy học*

Để đảm bảo dự đoán chính xác, bộ phân loại chọn thuật toán học tốt nhất để huấn luyện và sửa đổi nhằm thu được các tham số được tối ưu. Bởi vì chất lượng của các đặc trưng có thể có tác động đến hiệu suất dự đoán nên việc trích xuất đặc trưng là rất quan trọng [2]. Một phương pháp tích hợp để phát hiện phần mềm độc hại sử dụng các phương pháp tĩnh và động đã được giới thiệu bởi [3]. Việc kết hợp các phương pháp tĩnh và động đã cải thiện độ chính xác nhận dạng so với việc sử dụng các phương pháp tĩnh hoặc động riêng biệt. Theo kết quả nghiên cứu, phương pháp học máy SVM là phương pháp phân loại hiệu quả nhất. Ngoài những cải thiện về tỷ lệ False Positive và False Negative, Random Forest cũng cải thiện độ chính xác. Kết quả phân loại cho thấy phương pháp động vượt trội hơn, có độ chính xác cao hơn so với các phương pháp tĩnh [4].

Trong [5] đã đề xuất một hệ thống phát hiện phần mềm độc hại nhúng hoàn toàn mới dựa trên các kỹ thuật phát hiện bất thường về thống kê. Đây là phương pháp phát hiện phần mềm độc hại dựa trên sự bất thường đầu tiên để xác định vị trí lây nhiễm trong tệp bị nhiễm.

Trong [6], các tác giả giới thiệu một phương pháp máy học để phát hiện mạnh mẽ các phần mềm độc hại. Phương pháp được đề xuất dựa trên các thành phần được lấy từ cấu trúc và siêu dữ liệu của tài liệu. Nghiên cứu chứng minh sự phù hợp của một số thuộc tính tài liệu nhất định để nhận dạng phần mềm độc hại và khả năng phục hồi của các tính năng này trước các loại vi-rút mới bằng cách sử dụng bộ dữ liệu thực tế.

Trong [7], các nhà nghiên cứu đã nghĩ ra một phương pháp để xác định một tập hợp các tính năng được trích xuất bằng các công cụ hiện có và rút ra một bộ tính năng mới từ việc cải thiện khả năng phát hiện mã độc PDF và kéo dài tuổi thọ của các công nghệ phân tích và phát hiện. Các tính năng dẫn xuất được đánh giá bằng hàm bao bọc sử dụng ba thuật toán học có giám sát cơ bản [8] (Random Forest, C5.0 Decision Trees và SVM) và deep neural network chuyển tiếp nguồn cấp dữ liệu để xác định tầm quan trọng của tính năng.

Nghiên cứu [9] chứng minh khả năng chống chịu yếu hơn của thuật toán phân loại K-nearest neighbor (KNN) điển hình trong môi trường đối nghịch bằng cách sử dụng phương pháp tấn công gradient-descent để thay đổi các mẫu độc hại trong bộ thử nghiệm nhằm tránh bị mô hình phân loại phát hiện.

Bằng cách sử dụng phương pháp Gradient-descent, SVM đơn giản được sử dụng trong [10]. Các biện pháp chống lại cuộc tấn công này là đặt ngưỡng cho từng đặc trưng được xét. Điều này cho phép phương pháp được đề xuất thực tế có thể ngăn chặn tất cả các cuộc tấn công Gradient-descent.

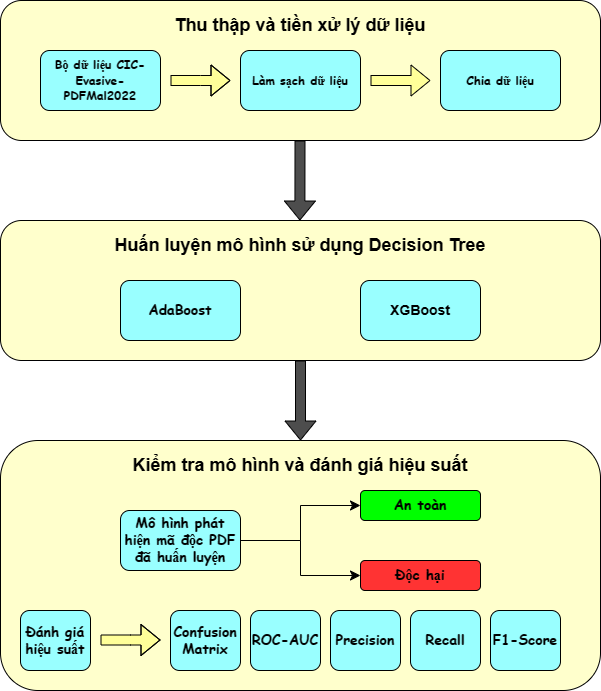
Các tác giả [11] đưa ra những phân tích chuyên sâu về nội dung và cấu trúc JavaScript của tệp PDF. Họ đã tạo ra một bộ tính năng phong phú trong JavaScript bao gồm các tính năng nội dung như tên đối tượng, từ khóa và chuỗi có thể đọc được cũng như các tính năng về cấu trúc và siêu dữ liệu như kích thước tệp, phiên bản, phương pháp mã hóa và từ khóa.

Ở [12] đã đưa ra một số khái niệm hướng dẫn lựa chọn đặc trưng nhằm giảm khả năng lẩn trốn mà vẫn giữ được độ chính xác cao. Các hướng dẫn này được tuân theo để trích xuất các đặc trưng và huấn luyện bộ phân loại hai giai đoạn. Nó cũng có thể phân biệt lỗ hổng được sử dụng trong các tệp độc hại.

Một mô hình mới cho các đặc trưng có độ bền giới hạn đã được đề xuất trong [13]. Mô hình được đề xuất duy trì độ chính xác 99,74% và tỷ lệ False positive 0,56% trong khi đạt được độ chính xác cao (VRA) được xác định lần lượt là 99,68% và 85,28% cho các thuộc tính chèn và xóa tương ứng. Trong [14] các tác giả đã sử dụng mô hình Hidost với phương pháp học tập tích cực và phân tích thỏa thuận chung làm tiêu chí để chọn các mẫu nghi ngờ, cho phép chúng tôi sử dụng tập huấn luyện nhỏ hơn và cải thiện hiệu suất phân loại.

**3. Đề xuất hệ thống phân loại**

Trong nghiên cứu này, chúng tôi đề xuất mô hình Decision Tree dùng để phân tích các tệp PDF nhằm phân loại các tệp PDF thành lành tính hoặc độc hại. Hình 4 cung cấp kiến trúc toàn diện cho hệ thống của chúng tôi từ những bước xử lý đầu tiên.



*Hình 4: Kiến trúc tổng quan của mô hình đề xuất*

*3.1. Giới thiệu dataset và tiền xử lý dữ liệu*

Trong nghiên cứu này, chúng tôi sử dụng bộ dataset CIC-Evasive-PDFMal 2022. Bộ dataset này có sẵn được cung cấp sẵn trên Kaggle. Bộ dataset bao gồm 10023 mẫu, các nhãn được gán là 1 và 0 (1: độc hại/Malicious, 0: lành tính/Benign); được phân phối dưới dạng 4468 mẫu lành tính và 5555 mẫu độc hại được thể hiện như hình 5. Ngoài ra, CIC-Evasive-PDFMal 2022 có 31 đặc trưng, bao gồm 12 đặc trưng chung và 19 đặc trưng chứa cấu trúc được trích xuất từ mỗi tệp PDF [57]. Các đặc trưng được trích xuất bao gồm:

* Đặc trưng tổng quát (General Features): Colors, EmbededFiles, Encrypt, MetadataSize, ObjStm, Pages, PdfSize, Stream, TitleCharaters, Trailer, XrefLength, isEncrypted.
* Đặc trưng cấu trúc (Structural Features): Images, Text, Header, Obj, Endobj, Endstream, Xref, StartXref, PageNo, JS, JavaScript, AA, OpenAction, Acroform, JBIG2Decode, RichMedia, Launch, EmbeddedFile, XFA.



*Hình 5: Số lượng mẫu*

A screen shot of a computer

Description automatically generated

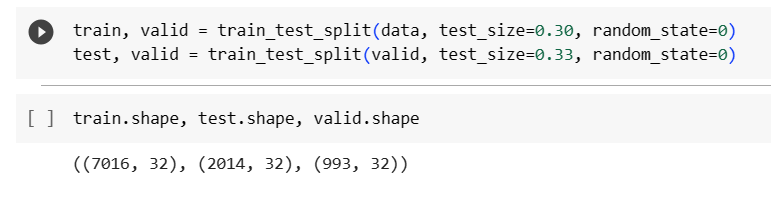
*Hình 6: Các đặc trưng tổng quát.*

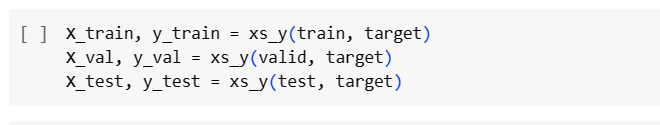
A screen shot of a computer

Description automatically generated

*Hình 7: Các đặc trưng cấu trúc.*

Chúng tôi đã tiến hành tiền xử lý dữ liệu chẳng hạn như xóa các mẫu dữ liệu thiếu, xóa bỏ những đặc trưng không cần thiết hoặc trùng lặp/có lỗi. Cuối cùng, dữ liệu được chia thành 3 phần: train (70%), valid (10%) và test (20%) như hình 6.





*Hình 8: Chia dataset*

*3.2 Ứng dụng các phương pháp máy học và huấn luyện mô hình*

Chúng tôi lựa chọn mô hình decision tree bởi vì nó là một phương pháp học có giám sát phi tham số được sử dụng để thực hiện các nhiệm vụ phân loại và hồi quy. Sử dụng hai thuật toán XGBoost và AdaBoost để train mô hình.

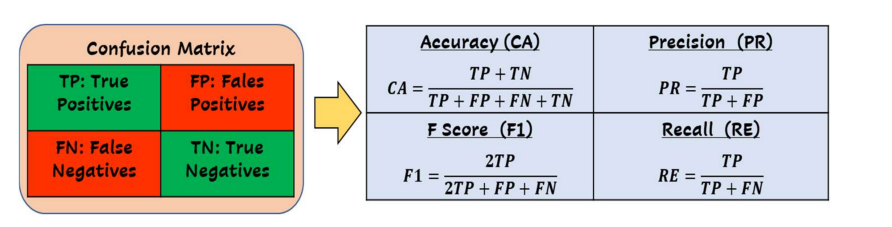
AdaBoost là một thuật toán máy học được thiết kế để cải thiện hiệu suất của các mô hình yếu để tạo ra một mô hình mạnh. AdaBoost là một trong những thuật toán phổ biến được sử dụng cho các bài toán phân loại.

XGBoost xây dựng cây quyết định một cách tuần tự, được tối ưu hóa để giảm thiểu hàm mất mát (loss function), hỗ trợ nhiều hàm mất mát như Regression Loss, Logistic Loss. Vì vậy, nó phù hợp cho mô hình này.

*3.3 Kiểm tra và đánh giá mô hình*

Kiểm tra mô hình là phần rất quan trọng để hiểu rõ tính chính xác, hiệu suất của mô hình. Từ đó đưa ra các điểm mạnh, điểm yếu nhằm cải thiện mô hình.

Sử dụng ROC-AUC, Accuracy, Precision, Recall, F1-score để kiểm tra mô hình. Đưa ra ma trận nhầm lẫn (Hình 9) nhằm đánh giá mô hình sau khi kiểm thử.

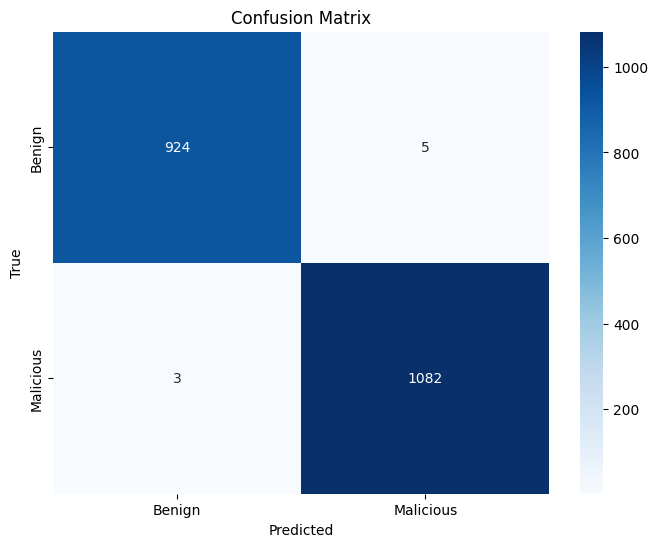


*Hình 9: Ma trận nhầm lẫn*

**4. Phân tích kết quả**

Mô hình XGBoost có Accuracy là 99,60%, Precision là 99,54%, Recall đạt 99,72% , ROC-AUC đạt 99,59%, F1-Score đạt 99.63%.

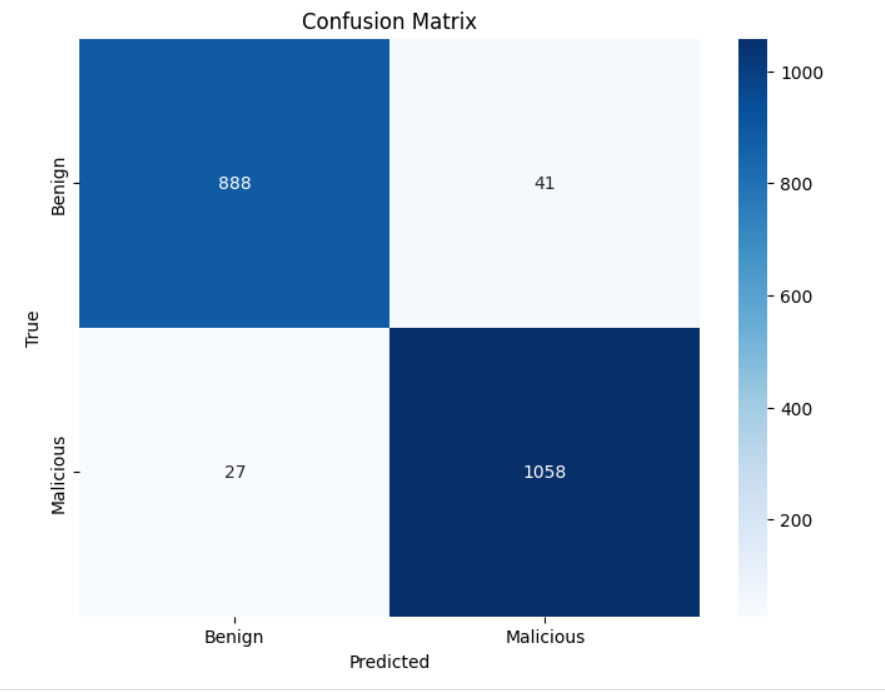
Hình 10 tóm tắt các chỉ số đánh giá hiệu suất tiêu chuẩn cho mô hình XGBoost



*Hình 10: Ma trận nhầm lẫn của XGBoost*

Mô hình AdaBoost có Accuracy là 96.62%, Precision là 96.26%, Recall đạt 97.51% , ROC-AUC đạt 96.54%, F1-Score đạt 96.88%.

Hình 11 tóm tắt các chỉ số đánh giá hiệu suất tiêu chuẩn cho mô hình AdaBoost.



*Hình 9: Ma trận nhầm lẫn của AdaBoost*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mô hình | ROC-AUC | Accuracy | Precision | Recall | F1-score |
| XGBoost | 99,59% | 99,6% | 99,54% | 99,72% | 99,63% |
| AdaBoost | 96,54% | 96,62% | 96,26% | 97,51% | 96,88% |

Bảng so sánh hai mô hình

Với kết quả trên, mô hình Decision Tree sử dụng thuật toán XGBoost đạt kết quả chính xác hơn, hiệu suất cao hơn AdaBoost.

**5. Kết luận và nhận xét**

Nhu cầu ngày càng tăng vọt, thúc đẩy sự phát triển của công nghệ. Việc sử dụng tài liệu kỹ thuật số đã trở nên phổ biến và PDF, một định dạng tập tin rất quen thuộc, được ưa chuộng trên toàn thế giới. Tuy nhiên, với sự phổ biến này cũng đi kèm với nguy cơ lợi dụng các tập tin PDF để thực hiện các cuộc tấn công. Hacker thường ẩn và chèn mã độc vào tập tin PDF để tận dụng máy tính của nạn nhân.

Do đó, trong bài viết này, chúng tôi giới thiệu một hệ thống có khả năng phân loại tập tin PDF lành tính hoặc độc hại. Hệ thống này đã được huấn luyện và đánh giá trên một bộ dữ liệu (CIC-Evasive-PDFMal2022) thông qua hai phương pháp là XGBoost và AdaBoost. Kết quả thu được với độ chính xác đáng kể.

**Tài liệu tham khảo**

[1] Qasem Abu Al-Haija, Ammar Odeh and Hazem Qattous, PDF Malware Detection Based on Optimizable Decision Trees , accessed on 19 September 2022

[2] Wu, Y.X.; Wu, Q.B.; Zhu, J.Q. Data-driven wind speed forecasting using deep feature extraction and LSTM. IET Renew. Power Gener. 2019, 13, 2062–2069.

[3] Shijo, P.; Salim, A. Integrated static and dynamic analysis for malware detection. Procedia Comput. Sci. 2015, 46, 804–811.

[4] Al-Haija, A.Q. Top-Down Machine Learning-Based Architecture for Cyberattacks Identification and Classification in IoT Communication Networks. Front. Big Data 2022, 4, 782902.

[5] Shafiq, M.Z.; Khayam, S.A.; Farooq, M. Embedded malware detection using Markov n-grams. In Proceedings of the International

Conference on Detection of Intrusions and Malware and Vulnerability Assessment, Paris, France, 10–11 July 2008.

[6] Smutz, C.; Stavrou, A. Malicious PDF detection using metadata and structural features. In Proceedings of the 28th Annual Computer Security Applications Conference, Orlando, FL, USA, 3–7 December 2012.

[7] Falah, A.; Pan, L.; Huda, S.; Pokhrel, S.R.; Anwar, A. Improving malicious PDF classifier with feature engineering: A data-driven approach. Future Gener. Comput. Syst. 2021, 115, 314–326.

[8] Al-Haija, Q.A.; Nasr, K.A. Supervised Regression Study for Electron Microscopy Data. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 2661–2668.

[9] Li, K.; Gu, Y.; Zhang, P.; An, W.; Li, W. Research on KNN algorithm in malicious PDF file classification under adversarial environment. In Proceedings of the 2019 4th International Conference on Big Data and Computing, Guangzhou, China, 10–12 May 2019.

[10] Cuan, B.; Damien, A.; Delaplace, C.; Valois, M. Malware detection in pdf files using machine learning. In Proceedings of the SECRYPT 2018-15th International Conference on Security and Cryptography, Porto, Portugal, 26–28 July 2018.

[11] Kang, A.R.; Jeong, Y.S.; Kim, S.L.; Woo, J. Malicious PDF detection model against adversarial attack built from benign PDF containing javascript. Appl. Sci. 2019, 9, 4764.

[12] He, K.; Zhu, Y.; He, Y.; Liu, L.; Lu, B.; Lin, W. Detection of Malicious PDF Files Using a Two-Stage Machine Learning Algorithm. Chin. J. Electron. 2020, 29, 1165–1177.

[13] Chen, Y.; Wang, S.; She, D.; Jana, S. On training robust {PDF} malware classifiers. In Proceedings of the 29th USENIX Security Symposium (USENIX Security 20), Berkeley, CA, USA, 12–14 August 2020; pp. 2343–2360.

[14] Li, Y.; Wang, X.; Shi, Z.; Zhang, R.; Xue, J.; Wang, Z. Boosting training for PDF malware classifier via active learning. Int. J. Intell. Syst. 2022, 37, 2803–2821.