

ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐẠI HỌC CÔNG NGHỆ



Nguyễn Thanh Huyền

**PHÁT TRIỂN CÔNG CỤ TRÍCH XUẤT
VÀ ĐÁNH GIÁ ĐẶC TRƯNG PHỤC VỤ
PHÁT HIỆN MÃ ĐỘC TRONG TÀI LIỆU
PDF**

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công Nghệ Thông Tin

Hà Nội, 2022

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Thanh Huyền

PHÁT TRIỂN CÔNG CỤ TRÍCH XUẤT
VÀ ĐÁNH GIÁ ĐẶC TRƯNG PHỤC VỤ
PHÁT HIỆN MÃ ĐỘC TRONG TÀI LIỆU
PDF

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công Nghệ Thông Tin

Cán bộ hướng dẫn: TS. Lê Đình Thanh

Hà Nội, 2022

Tóm tắt

Trong những năm gần đây, PDF hay viết tắt của Portable Document Format, đã trở thành một định dạng tiêu chuẩn để lưu trữ và trao đổi tài liệu nhờ tính linh hoạt và những chức năng hữu ích. Việc sử dụng rộng rãi đã tạo ra những lo ngại về sự an toàn của tài liệu PDF với người dùng. Tuy nhiên, các đặc tính của PDF đã thúc đẩy tin tặc khai thác nhiều loại lỗ hổng khác nhau, vượt qua các biện pháp bảo vệ an toàn, từ đó khiến tài liệu PDF trở thành một trong những phương tiện tấn công mã độc hiệu quả, trong các cuộc tấn công quy mô lớn đến các cuộc tấn công có chủ đích. Do đó việc phát hiện các tệp PDF độc hại là rất quan trọng trong bảo mật thông tin. Một số kỹ thuật phân tích sẽ được đề xuất trong tài liệu, để trích xuất các đặc trưng giá trị cho phép phân biệt phần mềm độc hại hay lành tính. Từ những kỹ thuật đánh giá đặc trưng thông thường có thể gặp nhiều hạn chế, các kỹ thuật dựa trên học máy đã được áp dụng phục vụ việc phát hiện tự động tệp PDF độc hại từ một tập các mẫu huấn luyện. Mô hình đã được chứng minh hiệu quả hơn khi kết hợp nhiều đặc trưng khác nhau. Tuy nhiên, bản thân mô hình này phải đối mặt với nhiều thách thức khi gặp phải các cuộc tấn công né tránh - khi mà những mẫu độc hại được biến đổi trông lành tính. Từ đó ở phần sau, khóa luận giới thiệu một thuật toán học máy mục tiêu tạo nên mô hình phân loại không có dương tính giả, đảm bảo mang đến một tường lửa lọc những tệp chắc chắn độc hại, còn những tệp sạch sẽ được xử lý nâng cao ở những giai đoạn tiếp theo.

Lời cảm ơn

Lời đầu tiên cho phép em được gửi lời cảm ơn tới Khoa Công Nghệ Thông Tin - Trường Đại học Công Nghệ ĐHQG Hà Nội đã tạo điều kiện thuận lợi cho em được học tập, nghiên cứu và thực hiện đề tài tốt nghiệp này.

Em cũng xin được bày tỏ lòng biết ơn sâu sắc tới thầy Lê Đình Thanh đã tận tình hướng dẫn, đóng góp những ý kiến giúp em hoàn thành khóa luận tốt nghiệp một cách tốt nhất.

Em cũng vô cùng biết ơn các thầy cô trong trường tận tình giảng dạy, truyền thụ cho em những kiến thức và kỹ năng quan trọng làm hành trang vững chắc trên con đường học vấn của bản thân.

Chúc mọi người luôn luôn mạnh khỏe và gặt hái được nhiều thành công trong cuộc sống.

Lời cam đoan

Em xin cam đoan rằng khóa luận tốt nghiệp này không sao chép từ bất kỳ ai, tổ chức nào khác. Toàn bộ nội dung được trình bày trong tài liệu đều là cá nhân em qua quá trình học tập, tìm hiểu và nghiên cứu mà hoàn thiện. Mọi tài liệu tham khảo đều được ghi chép lại và trích dẫn hợp pháp. Nếu lời cam đoan là sai sự thật thì em xin chịu mọi trách nhiệm và hình thức kỷ luật theo quy định từ phía nhà trường.

Mục lục

1	Kiến thức nền tảng	3
1.1	Định dạng tệp PDF	3
1.1.1	Các đối tượng tệp	3
1.1.2	Cấu trúc vật lý	4
1.1.3	Cấu trúc logic	7
1.2	Một số thuật toán phân loại	10
1.2.1	Cây quyết định	10
1.2.2	Random Forest	11
1.2.3	Kỹ thuật boosting và thuật toán LightGBM	12
1.3	Phát hiện tài liệu PDF độc hại	13
2	Kết hợp đặc trưng thống kê và cấu trúc cây trong phát hiện PDF độc hại	16
2.1	Tập dữ liệu	16
2.2	Trích xuất đặc trưng	17
2.2.1	Đặc trưng thống kê	17
2.2.2	Đặc trưng cấu trúc cây	21
2.3	Phân loại và đánh giá kết quả	24
2.3.1	Xếp hạng đặc trưng theo độ quan trọng	24
2.3.2	So sánh kết quả khi điều chỉnh tham số và số lượng đặc trưng	25
2.3.3	So sánh kết quả khi kết hợp các loại đặc trưng	29
3	Xây dựng mô hình phân loại Zero False Positive	30
3.1	Đặt vấn đề	30
3.2	Bộ phân loại Zero False Positive	30
3.2.1	Thuật toán phân loại lặp với mục tiêu $FP = 0$	30
3.2.2	Thực nghiệm và đánh giá kết quả	30
3.3	Phương hướng xử lý các tệp được gán nhãn sạch	32

4	Kết luận	33
4.1	Kết quả đạt được	33
4.2	Hạn chế và khó khăn	34
4.3	Hướng nghiên cứu tương lai	34
A	Phụ lục chứng minh	35
A.1	Zero FP	35
A.2	Tập giảm lược S luôn có thể phân tách được	36
A.3	FN giảm dần khi FP bằng 0	36
	Tài liệu tham khảo	38

Danh sách hình vẽ

1	Biểu đồ các lỗ hổng của trình đọc PDF Acrobat Reader theo các năm ¹	1
1.1	Cấu trúc vật lý của tệp PDF	5
1.2	Một bảng xref trong tài liệu PDF	6
1.3	Một trailer trong tài liệu PDF	7
1.4	Một trailer trong tài liệu PDF	8
1.5	Mô hình phân cấp các đối tượng bên trong cấu trúc logic của tài liệu PDF	9
1.6	Đối tượng catalog trong tài liệu PDF	9
1.7	Sơ đồ cấu trúc cây trong thuật toán Cây quyết định	10
1.8	Quy trình phân loại dựa trên thuật toán Random Forest [17]	12
1.9	So sánh các kỹ thuật được sử dụng trong mô hình sử dụng phương pháp học kết hợp	13
1.10	Phân loại các đặc trưng được sử dụng trong phát hiện tài liệu PDF độc hại	14
2.1	Minh họa cấu trúc cây và đường dẫn cấu trúc của tệp PDF	22
2.2	Các cặp đường dẫn cấu trúc và giá trị chưa chuẩn hóa	22
2.3	Các cặp đường dẫn cấu trúc và giá trị sau khi được chuẩn hóa về dạng số	23
2.4	Biểu đồ 50 đặc trưng có độ quan trọng cao nhất trong các đặc trưng thống kê và đường dẫn cấu trúc.	25
2.5	Biểu đồ so sánh kết quả điều chỉnh tham số thuật toán Light-GBM qua các lần lặp trên tập kiểm chứng	26
2.6	Biểu đồ so sánh mô hình khi sử dụng số lượng đặc trưng tăng dần từ tập đặc trưng thống kê.	27
2.7	Biểu đồ so sánh kết quả của mô hình khi sử dụng số lượng đặc trưng tăng dần từ tập đặc trưng cấu trúc cây	28
2.8	Biểu đồ so sánh kết quả của mô hình khi sử dụng số lượng đặc trưng tăng dần từ sự kết hợp các tập đặc trưng thống kê và đặc trưng cấu trúc cây.	28

2.9	Biểu đồ so sánh kết quả của mô hình khi sử dụng hai loại đặc trưng độc lập và khi kết hợp các đặc trưng thống kê và đặc trưng cấu trúc cây	29
-----	--	----

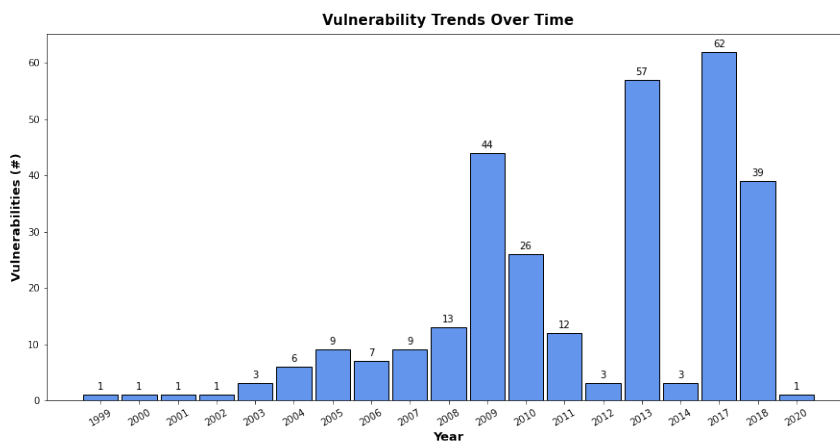
Danh sách bảng

2.1	Thống kê số lượng tệp dữ liệu có chứa javascript.	17
2.2	Phân bố tập huấn luyện, kiểm chứng và tập kiểm thử	17
2.3	Danh sách 67 đặc trưng thống kê và mô tả	18
3.1	Kết quả của thuật toán phân loại Zero False Positive trên tập dữ liệu đã trích xuất	31
3.2	Kết quả thuật toán phân loại Zero False Positive trên tập dữ liệu Hidost (1)	31
3.3	Kết quả thuật toán phân loại Zero False Positive trên tập dữ liệu Hidost (2)	31

Danh sách các đoạn mã

Mở đầu

Những kẻ tấn công bằng nhiều cách có thể cài đặt các phần mềm độc hại trên máy nạn nhân để có được quyền truy cập, thực hiện các hành vi xâm phạm dữ liệu cá nhân hay dữ liệu nhạy cảm, phá hoại hệ thống hoặc sử dụng vào những mục đích xấu khác. Người dùng thường có thể dễ dàng nhận ra mối đe dọa từ những tệp thực thi, ngoài ra nhận thức về mối nguy hiểm từ các tài liệu Microsoft Office cũng ngày càng tăng cao. Tuy nhiên, bỏ qua sự phức tạp về định dạng cấu trúc tệp PDF, người dùng có xu hướng coi các tệp PDF là tài liệu vô hại. Hiện nay, lợi dụng các lỗ hổng bảo mật, các trình đọc tài liệu PDF đang là mục tiêu của các kẻ tấn công mạng. Điển hình với trình đọc PDF Acrobat Reader, từ năm 1999 tới nay, đã có tới 298 lỗ hổng được phát hiện theo thống kê của CVE Details ²(Hình 1).



Hình 1: Biểu đồ các lỗ hổng của trình đọc PDF Acrobat Reader theo các năm³

Trong khóa luận này, tôi sẽ trình bày một số phương pháp để phát hiện các tệp PDF độc hại, trong đó các đặc trưng được trích xuất từ tài liệu PDF

³<https://www.cvedetails.com>

sẽ được sử dụng để xây dựng bộ phân loại học máy tự động phát hiện tài liệu PDF độc hại, thông qua các chương sau:

Chương 1 Kiến thức nền tảng về định dạng tài liệu PDF, một số thuật toán học máy, từ đó đưa ra một quy trình để phát hiện phần mềm độc hại.

Chương 2 Thực hiện trích xuất đặc trưng trên tập dữ liệu gồm khoảng 22000 tệp PDF, từ đó xây dựng bộ phân loại học máy và đánh giá kết quả khi kết hợp các đặc trưng khác nhau.

Chương 3 Đề xuất phương hướng phát hiện tài liệu PDF độc hại thông qua hai giai đoạn, với giai đoạn đầu xây dựng một thuật toán học máy không gây phân loại dương sai và giai đoạn hai gồm các phương hướng xử lý nâng cao với các tệp được gán nhãn sạch.

Kết luận Đánh giá kết quả đạt được, hạn chế và khó khăn gặp phải, và cuối cùng đề xuất các hướng mới để phát triển đề tài.

Chương 1

Kiến thức nền tảng

Chương này gồm ba kiến thức nền tảng. Đầu tiên trình bày sơ bộ về các đối tượng và cách cấu trúc một tài liệu PDF, từ đó là bước đệm để hiểu cách phần mềm đọc hại được nhúng vào tệp cũng như là cách để có thể xác định các bất thường trong cấu trúc một tệp PDF. Tiếp theo giới thiệu về một số thuật toán học máy được sử dụng khi xây dựng mô hình phân loại tự động các tệp PDF độc hại. Phần cuối cùng đưa ra cách tiếp cận và quy trình để có thể phát hiện tệp độc hại.

1.1 Định dạng tệp PDF

Tệp pdf là một trong những định dạng phổ biến nhất để lưu trữ tài liệu trên toàn thế giới. Năm 1993, PDF được Adobe Systems định nghĩa và đến năm 2008, PDF đã trở thành một tiêu chuẩn mở được phát hành dưới chuẩn ISO 32000-1.

1.1.1 Các đối tượng tệp

Cấu trúc của tệp PDF bao gồm các đối tượng liên kết với nhau trong một cấu trúc cây. Những đối tượng này được chia thành ba loại, gồm có các đối tượng đơn giản, các đối tượng ghép và đối tượng độc lập.

Đối tượng đơn giản

- Boolean: lưu giá trị true/false.
- Number: là đối tượng số.
- String: là các chuỗi ký tự với kích thước tối đa 65535 bytes.

- Name: là các đối tượng từ khóa được định nghĩa sẵn trong tiêu chuẩn PDF, theo sau bởi ký tự '/', ví dụ '/kids' xác định các đối tượng con, '/filter' chỉ định bộ lọc được sử dụng,...
- Indirect reference: là các tham chiếu tới các đối tượng độc lập.

Đối tượng ghép

- Array: mảng 1 chiều các đối tượng được sắp xếp trong cặp ký tự '[]'.
- Dictionary: giống như một danh sách lưu trữ các cặp từ khóa - giá trị, được đặt trong cặp ký tự '«»'. Từ khóa phải là một đối tượng Name còn giá trị thì có thể là bất kể đối tượng nào. Mỗi từ khóa chỉ xuất hiện 1 lần duy nhất trong một đối tượng dictionary.
- Stream: đối tượng stream giống như một đối tượng string, gồm một chuỗi các bytes, nhưng khác ở việc không giới hạn về độ dài. Do đó đối tượng stream có khả năng chứa một lượng lớn dữ liệu như hình ảnh, các thông tin mô tả trang,... Stream được đặt trong cặp từ khóa 'stream endstream'.

Đối tượng độc lập

Các đối tượng độc lập có định danh là các cặp (number object, generation object) ứng với số hiệu đối tượng và thế hệ của đối tượng. Các đối tượng khác có thể tham chiếu tới qua định danh này. Các đối tượng độc lập được định nghĩa bên trong cặp từ khóa 'obj endobj'.

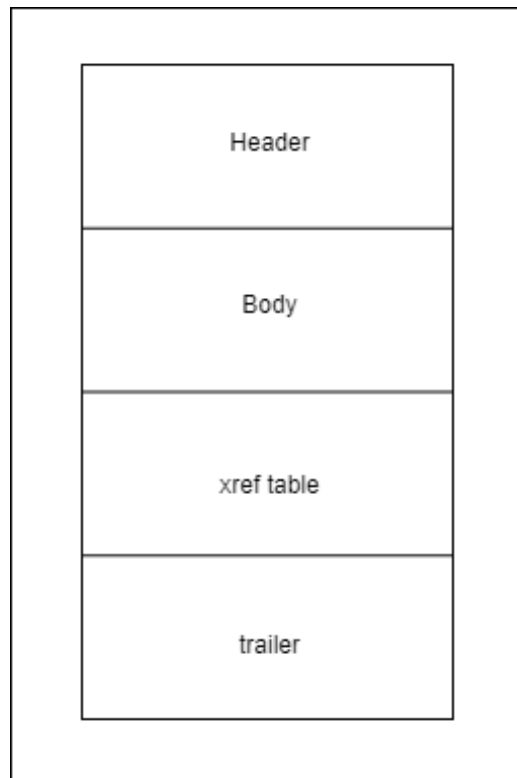
1.1.2 Cấu trúc vật lý

Cấu trúc vật lý của PDF xác định cách đối tượng được lưu trữ trong tệp, cách mà chúng được truy cập và cập nhật. Một tệp PDF cơ bản sẽ được xây dựng bằng bốn yếu tố: header, body, bảng tham chiếu xref table và trailer. Hình 1.1 thể hiện cấu trúc này.

Cấu trúc ban đầu của PDF sẽ được thay đổi sau mỗi lần cập nhật, bằng cách thêm các sự thay đổi vào cuối của tệp, được gọi là cập nhật tăng dần.

Header

Header là dòng đầu tiên trong tệp PDF, xác định phiên bản tệp, được viết dưới định dạng '%PDF-version'. Ví dụ '%PDF-1.7' xác định phiên bản PDF hiện tại là 1.7.



Hình 1.1: Cấu trúc vật lý của tệp PDF

Ngoài ra, nếu tệp PDF chứa dữ liệu nhị phân, header sẽ chứa một dòng comment gồm ít nhất 4 ký tự nhị phân giúp xác định được cần xử lý dữ liệu tệp dưới dạng văn bản hay nhị phân.

Body

Phần body của tệp chứa tất cả các đối tượng lưu trữ những dữ liệu sẽ hiển thị cho người dùng, được xác định từ sau phần header đến khi xuất hiện từ khóa 'xref'. Những dữ liệu trong body sau đó có thể được chỉnh sửa, và phần cập nhật sẽ được lưu trữ ở cuối tệp.

Xref Table

Xref table hay còn được gọi là cross-reference table là một bảng tham chiếu tới các đối tượng độc lập trong tài liệu nhằm mục đích cho phép truy cập ngẫu nhiên vào các đối tượng này mà không cần đọc toàn bộ tài liệu để định vị một đối tượng cụ thể.

Bảng có thể chứa một hoặc nhiều phần. Ban đầu, toàn bộ bảng chỉ bao

gồm một phần duy nhất. Một phần tương ứng sẽ được thêm vào nếu tệp được thay đổi.

```
xref
0 1
0000000023 65535 f
3 1
0000025324 00000 n
21 4
0000025518 00002 n
0000025632 00000 n
0000000024 00001 f
0000000000 00001 f
36 1
0000026900 00000 n
```

Hình 1.2: Một bảng xref trong tài liệu PDF

Ở trên hình 1.2, cho thấy một ví dụ về bảng tham chiếu trong tệp PDF, gồm có 4 phần con, bắt đầu bởi dòng chứa hai giá trị số với giá trị đầu là số thứ tự của đối tượng đầu tiên, giá trị thứ hai là số lượng đối tượng trong phần con này. Ví dụ ở phần thứ 3, giá trị '21 4' cho thấy có 4 đối tượng trong phần con, được đánh số bắt đầu từ 21 đến 24. Theo sau đó là các dòng đại diện cho các đối tượng, xác định bởi một chuỗi gồm 20 byte lưu trữ địa chỉ của đối tượng (offset), số thể hệ của đối tượng và một cờ 'f' hoặc 'n'. Cờ 'f' nghĩa là đối tượng vẫn tồn tại trong tệp nhưng được đánh dấu không nên được sử dụng, ngược lại, cờ 'n' xác định đối tượng đang được sử dụng.

Trailer

Trailer PDF được lưu ở cuối tệp, chỉ định vị trí của bảng tham chiếu và các đối tượng đặc biệt khác, bắt đầu bởi từ khóa 'trailer' (Hình 1.3).

Trong đối tượng dictionary của trailer lưu trữ các giá trị:

- Size: xác định số lượng đối tượng trong bảng tham chiếu (có đếm cả các đối tượng trong các phần mới cập nhật). Giá trị này phải là 1 số nguyên, và không thể là một đối tượng tham chiếu.

```

trailer
<<
/Size 22                % số đối tượng
/Root 2 0 R
>>
startxref
24212                    % offset của trailer
%%EOF

```

Hình 1.3: Một trailer trong tài liệu PDF

- Prev: xác định địa chỉ của bảng tham chiếu được cập nhật trước đó được sử dụng khi có nhiều bảng tham chiếu.
- Root: tham chiếu tới một đối tượng gốc trong mô hình cấu trúc logic của PDF.

Cập nhật tăng dần

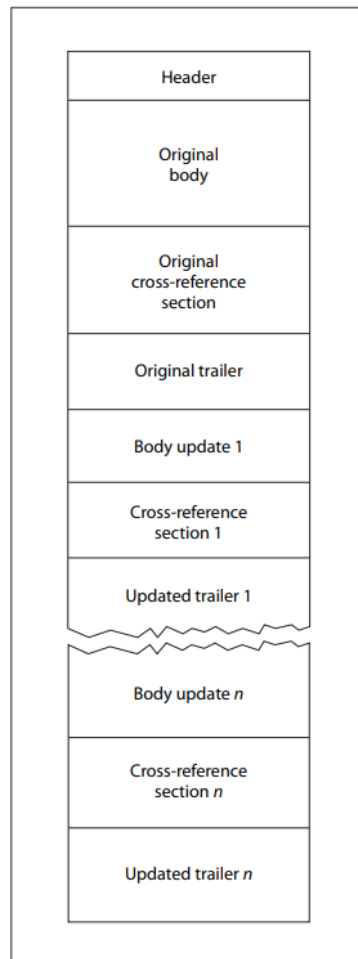
PDF được thiết kế với khả năng cập nhật tăng dần, tức là có thể cập nhật bằng cách thêm các đối tượng thay đổi vào cuối tệp PDF mà không cần phải viết lại toàn bộ tệp, mục tiêu cập nhật nhanh chóng các thay đổi nhỏ vào trong 1 tệp lớn.

Có thể thấy trong hình 1.4, tệp PDF vẫn chứa các phần header, body, xref table và trailer ban đầu. Tiếp theo sẽ có các body, xref table và trailer được thêm vào cuối tệp. Các bảng tham chiếu được bổ sung sẽ chỉ chứa các đối tượng đã bị thay đổi, thay thế hoặc xóa. Các đối tượng bị xóa sẽ vẫn nằm trong tệp nhưng được đánh dấu với cờ 'f'. Trong phần trailer được thêm mới sẽ có một cặp từ khóa '\Prev' cho biết vị trí của bảng tham chiếu trước đó và cần phải kết thúc bằng ký tự kết thúc "%%EOF"

1.1.3 Cấu trúc logic

Cấu trúc logic của tài liệu PDF chỉ ra cách các đối tượng liên kết với nhau và được tổ chức sao cho các trình đọc PDF có thể hiểu và xử lý tài liệu một cách thích hợp. Các đối tượng này được xây dựng dựa trên một mô hình phân cấp dạng cây, với nút gốc là một đối tượng catalog (Hình 1.5).

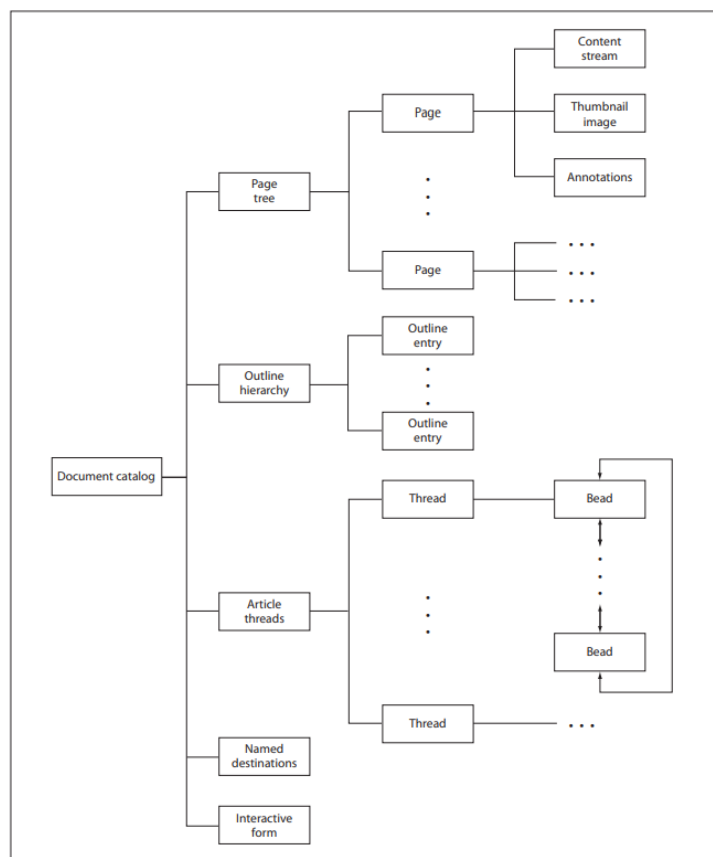
Catalog cung cấp thông tin chung về loại đối tượng nào sẽ được hiển thị trong tệp PDF, nội dung của tài liệu và một số hướng dẫn cụ thể cho các trình đọc PDF về cách hiển thị PDF lên màn hình sau khi được mở. Hình 1.6 là một ví dụ về một đối tượng catalog trong tệp PDF, với từ



Hình 1.4: Một trailer trong tài liệu PDF

khóa 'PageMode' xác định cách tệp sẽ được hiển thị, 'OpenAction' chỉ định hành động sẽ được thực thi ngay sau khi mở tệp, và 'Pages' tham chiếu tới nút gốc của một cây các trang trong tài liệu PDF.

Page Tree Các trang của tài liệu PDF được liên kết thông qua một cây các trang - định nghĩa thứ tự truy cập. Sử dụng cấu trúc cây cho phép trình đọc nhanh chóng mở các trang trong tài liệu chứa hàng nghìn trang. Nút gốc của cây được chỉ định thông qua tham chiếu /Pages trong đối tượng Catalog.



Hình 1.5: Mô hình phân cấp các đối tượng bên trong cấu trúc logic của tài liệu PDF

```

1 0 obj
<< /Type /Catalog
  /Pages 2 0 R
  /PageMode /UseOutlines
  /Outlines 3 0 R
  /OpenAction 11 0 R
>>
endobj

```

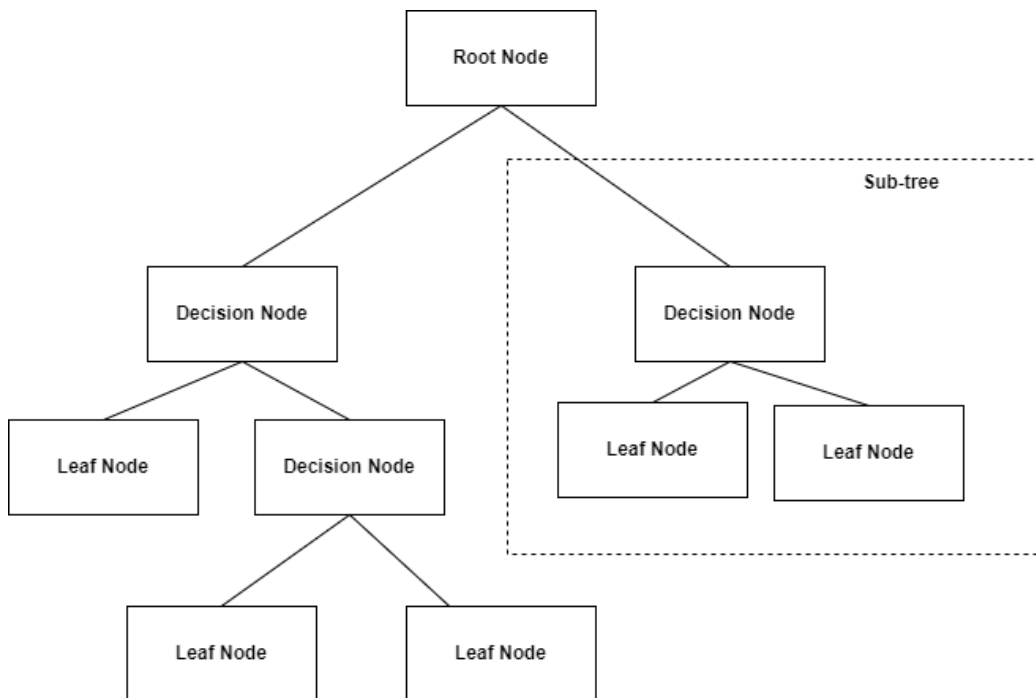
Hình 1.6: Đối tượng catalog trong tài liệu PDF

1.2 Một số thuật toán phân loại

Ngoài việc hiểu cấu trúc tài liệu PDF phục vụ cho việc phân tích và trích xuất những đặc trưng có giá trị thì việc huấn luyện một bộ phân loại hai lớp để phát hiện các mẫu độc hại với độ chính xác cao nhất có thể là vô cùng quan trọng. Thông thường, những mô hình phân loại sẽ được áp dụng vào tập dữ liệu một cách độc lập, từ đó xem xét mô hình nào hoạt động tốt nhất. Ở phạm vi khóa luận này, một số thuật toán phân loại được sử dụng trong các thử nghiệm, bao gồm Cây quyết định, Random Forest và LightGBM.

1.2.1 Cây quyết định

Cây quyết định là một thuật toán học máy có giám sát được sử dụng rộng rãi. Một cây quyết định sử dụng mô hình cấu trúc cây để biểu thị các dự đoán xuất phát từ một loạt các điều kiện phân tách dựa trên các đặc trưng, bắt đầu với nút gốc và kết thúc bằng một quyết định đưa ra bởi nút lá (Hình 1.7).



Hình 1.7: Sơ đồ cấu trúc cây trong thuật toán Cây quyết định

Nút gốc là nút bắt đầu của cây quyết định

Nút quyết định các nút sau nút gốc, chứa những điều kiện để phân chia tới các nút tiếp theo.

Nút lá nút kết thúc, chứa một nhãn duy nhất ứng với kết quả cuối cùng.

Cây con là các phần nhỏ của đồ thị biểu diễn cây quyết định, theo sau bởi nút gốc hoặc một nút quyết định

Cây quyết định phân loại bằng cách duyệt mẫu từ trên xuống dưới bắt đầu từ gốc của cây, tại mỗi nút quyết định, điều kiện sẽ được kiểm tra và mẫu sẽ được phân phối xuống các nút tiếp theo, cho đến khi đến một nút lá. Nhãn của nút lá sẽ chính là kết quả phân loại cuối cùng cho mẫu đó.

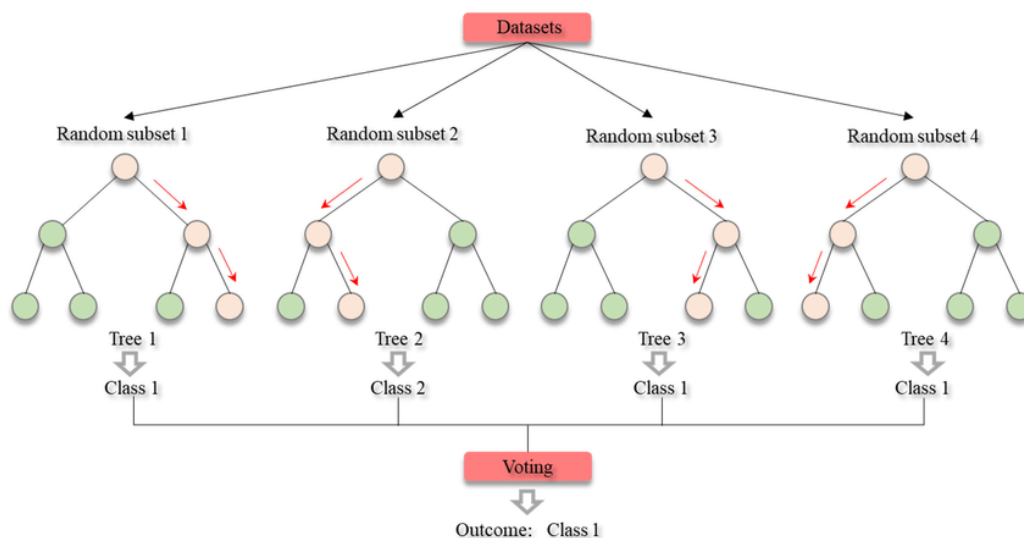
Trên thực tế, thuật toán đã được sử dụng trong rất nhiều công bố về phân loại tệp PDF độc hại: Davide Maiorca và cộng sự [8], Himanshu Pareek và cộng sự [9]. Ngoài ra một số công cụ phát hiện PDF độc hại như Hidost [13], Lux0R [4] hay PDF Malware Slayer [7] cũng sử dụng thuật toán này trong mô hình phân loại.

1.2.2 Random Forest

Random Forest là một thuật toán học kết hợp, bằng cách xây dựng rất nhiều các mô hình cây quyết định tạo thành rừng, mang lại một kết quả tốt hơn khi chỉ sử dụng một cây quyết định đơn lẻ. Mỗi cây quyết định trong rừng sẽ được học độc lập dựa trên một tập con ngẫu nhiên từ tập huấn luyện, và chỉ sử dụng một tập con các đặc trưng ngẫu nhiên của dữ liệu. Khi thực hiện phân loại, mẫu kiểm thử sẽ được phân loại song song trên các cây và trung bình cộng của các kết quả sẽ cho ra nhãn phân loại cuối cùng (Hình 1.8).

Trong thuật toán Cây quyết định, khi xây dựng cây với một độ sâu tùy ý, có thể gây nên việc mô hình khớp hoàn toàn với dữ liệu tập huấn luyện, nhưng lại mang lại kết quả kém trên các tập kiểm thử (hiện tượng overfitting). Các cây trong thuật toán Random Forest chỉ dùng một phần dữ liệu huấn luyện để học, do đó sẽ không gặp overfitting nhưng có thể gây ra underfitting - khi số tập dữ liệu không đủ hoặc số lượng đặc trưng được sử dụng quá ít, dẫn tới kết quả xấu trên cả tập huấn luyện lẫn tập kiểm thử. Tuy nhiên, kết quả của thuật toán Random Forest là sự tổng hợp từ tất cả các cây quyết định, nhờ đó, các cây sẽ bổ sung cho nhau và mang lại một kết quả phân loại tốt hơn.

Thuật toán Random Forest đã được sử dụng trong các công cụ phát hiện mã độc như PDFRate [12], PDF Malware Slayer [7] và LuxOR [4].



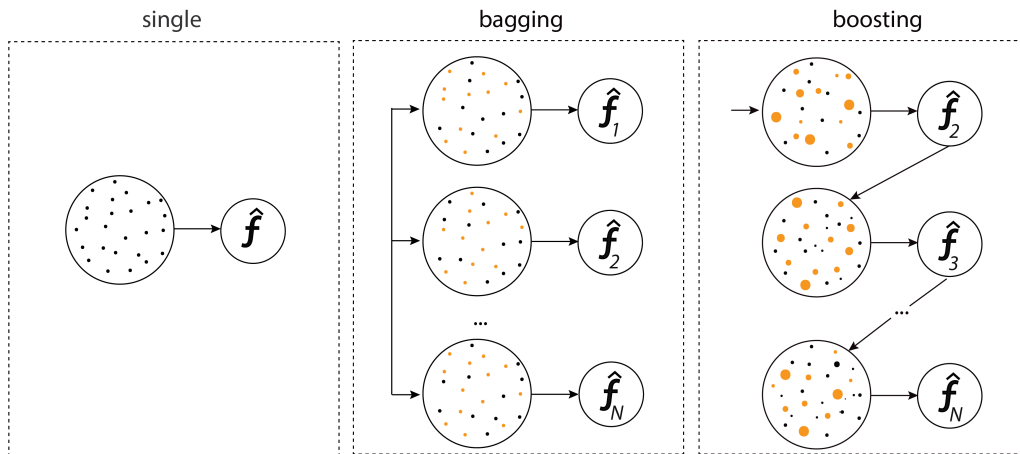
Hình 1.8: Quy trình phân loại dựa trên thuật toán Random Forest [17]

1.2.3 Kỹ thuật boosting và thuật toán LightGBM

Nhận thức được điểm yếu khi chỉ sử dụng một mô hình đơn lẻ như Cây quyết định, thuật toán Random Forest đưa ra một cải tiến mới khi kết hợp sử dụng mô hình trên nhiều phần dữ liệu khác nhau, huấn luyện độc lập và song song với nhau để đưa ra một kết quả tốt nhất. Đây là một ví dụ điển hình cho kỹ thuật bagging trong phương pháp học kết hợp. Ngoài ra, với phương pháp học kết hợp, một kỹ thuật được gọi là boosting cũng sẽ thực hiện xây dựng một chuỗi các mô hình trên những tập dữ liệu con khác nhau nhưng hoạt động học sẽ diễn ra một cách tuần tự. Theo đó, mô hình sau trong quá trình huấn luyện sẽ rút kinh nghiệm từ những lỗi của mô hình trước đó bằng cách cập nhật trọng số (Hình 1.9). Đầu ra của mô hình cuối cùng trong chuỗi sẽ là kết quả cuối cùng khi thực hiện phân loại.

Trong phạm vi của khóa luận này, tôi đề xuất sử dụng một thuật toán boosting được phát triển bởi Microsoft là LightGBM [2]. LightGBM viết tắt của Light Gradient Boosting Machine, là một mô hình xử lý thuật toán học kết hợp tăng cường. Mô hình này cũng sử dụng Cây quyết định để xây dựng các mô hình thành phần. Sau mỗi lần thực hiện huấn luyện trên một mô hình với một tập dữ liệu con, thuật toán sẽ cập nhật tham số của mô hình theo hướng giảm của đạo hàm hàm mất mát.

Với nhiều kỹ thuật và cơ chế phức tạp, thuật toán LightGBM mang lại một bộ phân loại mạnh mẽ, tốc độ tính toán được tối ưu, do đó cần ít thời gian xử lý hơn so với các thuật toán cùng loại khác trên các tập huấn luyện



Hình 1.9: So sánh các kỹ thuật được sử dụng trong mô hình sử dụng phương pháp học kết hợp

có kích thước lớn [2].

1.3 Phát hiện tài liệu PDF độc hại

Phát hiện các tệp độc hại có rất nhiều cách tiếp cận, tuy nhiên, có hai giai đoạn chính phục vụ mục đích này

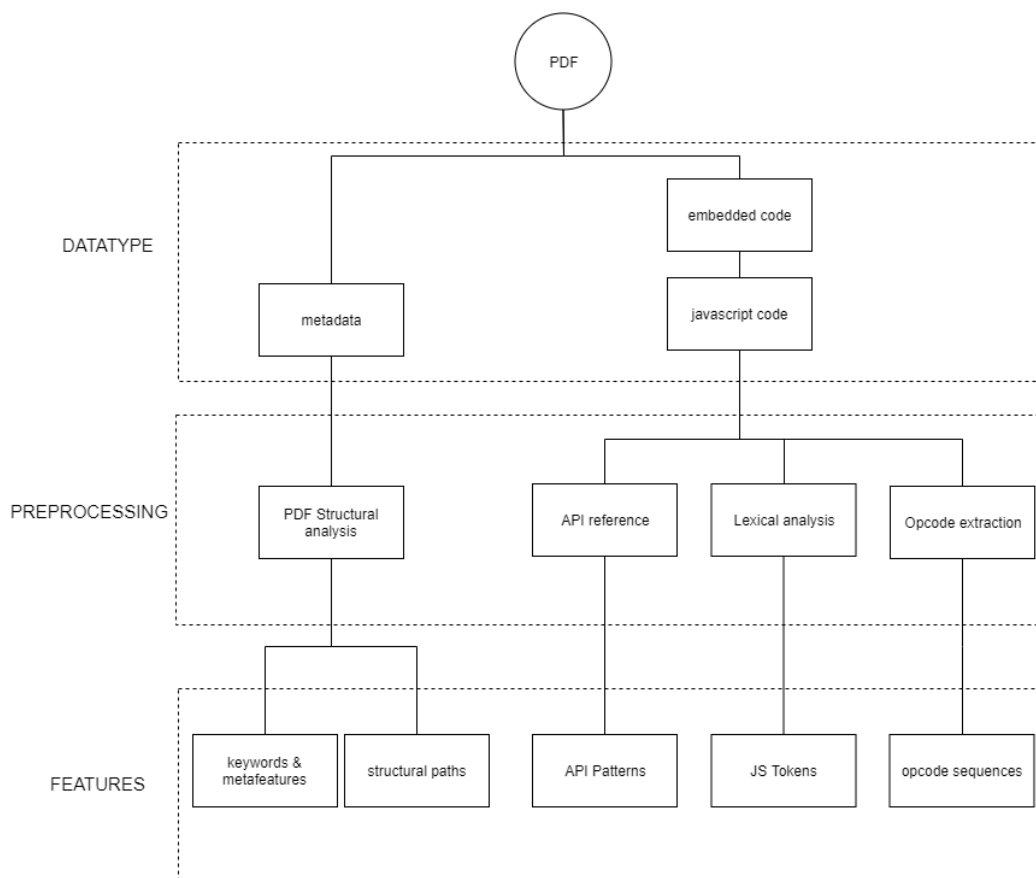
- Trích xuất đặc trưng
- Phân tích đặc trưng và đưa ra quyết định.

Trong giai đoạn trích xuất đặc trưng, các tệp PDF được xử lý và phân tích để trích xuất ra các đặc trưng cần thiết. Những đặc trưng này sẽ được thống kê và được đưa qua một bộ lọc để chọn ra những đặc trưng quan trọng trong giai đoạn Phân tích đặc trưng. Cuối cùng, một mô hình học máy sẽ được áp dụng để quyết định xem tệp đầu vào là sạch hay độc hại.

Mô hình 1.10 mô tả các đặc trưng được đề xuất để phát hiện tài liệu độc hại, và tổ chức chúng theo mô hình phân loại phân cấp. Mỗi cấp độ sẽ tương ứng với từng giai đoạn trong hoạt động trích xuất đặc trưng. Mô hình được giản lược từ mô hình được giới thiệu bởi Michele Elingiusti và cộng sự trong công bố năm 2018 [5].

Loại dữ liệu

Ở cấp đầu tiên, đại diện cho loại dữ liệu được trích xuất từ tài liệu PDF



Hình 1.10: Phân loại các đặc trưng được sử dụng trong phát hiện tài liệu PDF độc hại

- Siêu dữ liệu (metadata) là tất cả những dữ liệu có thể được trích xuất từ dữ liệu “thô” của tệp PDF từ những mô tả, các mối liên hệ và cấu trúc bên trong tài liệu. Các siêu dữ liệu bao gồm các từ khóa xuất hiện trong tệp, số lượng các ký tự kết thúc tệp, tác giả, đường dẫn cấu trúc,...
- Embedded code gồm các đoạn mã hay các tệp được nhúng vào trong tài liệu PDF, có thể kể đến một số tệp thực thi hay một số mã shellcode, diễn hình và xuất hiện nhiều nhất là các mã Javascript.

Tiền xử lý

Tiền xử lý cho thấy các kỹ thuật xử lý được sử dụng ứng với từng loại dữ liệu đầu vào, phục vụ cho mục đích cụ thể là đưa ra những đặc trưng hữu

ích trong việc phát hiện tài liệu độc hại. Kết quả của quá trình xử lý này là một tập các đặc trưng - được liệt kê ở cấp cuối cùng. Các đặc trưng này sẽ là đầu vào của các mô hình học máy khi thực hiện phân loại tệp độc hại.

- Phân tích cấu trúc của tệp PDF nhằm mục tiêu trích xuất các đặc trưng thống kê như từ khóa hay đường dẫn cấu trúc.
- Tham chiếu API mã Javascript: Quá trình xử lý và nhận diện các API được sử dụng trong mã javascript, bằng phương pháp phân tích tĩnh hoặc phân tích động. Cách tiếp cận này đã được đề xuất trong LuxOr [4]. Đầu ra của quá trình này là các mẫu API được sử dụng nhiều trong các tài liệu độc hại hay các tài liệu lành tính.
- Phân tích từ vựng: các đoạn mã javascript thường bị xáo trộn và làm phức tạp hóa, điều này yêu cầu một dạng biểu diễn trừu tượng hóa để loại bỏ các chi tiết không cần thiết và cô lập những đoạn mã thực sự liên quan để dễ dàng trong việc nhận diện và phát hiện bất thường. Một quy trình phân tích từ vựng mã nguồn sẽ hỗ trợ tự động hóa quá trình trừu tượng hóa như vậy. Cách tiếp cận này đã được đề xuất bởi Vatamanu và cộng sự [16] sử dụng công cụ SpiderMonkey ¹ để trích xuất ra các javascript token.
- Trích xuất opcode: một kỹ thuật tấn công được sử dụng trong các tài liệu PDF độc hại là xây dựng các đoạn mã shellcode trong thời gian thực thi bằng cách sao chép các chuỗi opcode ẩn giấu trong các biến. Do đó, một phương pháp phát hiện được đề xuất thực hiện phân tích động để xác định các biến có thể chứa các chuỗi opcode độc hại hoặc đáng ngờ. PDF Scrutinizer [11] xây dựng một phương pháp phỏng đoán để xác định vị trí nào cần để phân tích. MDScan [14] chọn các chuỗi cần tìm shellcode bằng cách quan sát các chuỗi được tạo ra trong thời gian chạy, quét các vùng nhớ của các chuỗi mới được cấp phát.

¹<https://spidermonkey.dev/>

Chương 2

Kết hợp đặc trưng thống kê và cấu trúc cây trong phát hiện PDF độc hại

Ở chương trước, tôi đã giới thiệu về quy trình phát hiện một tài liệu PDF độc hại, gồm hai phần chính là trích xuất đặc trưng và phân loại dữ liệu. Chương hai này sẽ tiến hành thử nghiệm, từ cài đặt công cụ trích xuất đặc trưng trên tập dữ liệu thu thập được đến xây dựng mô hình học máy mang lại độ chính xác cao khi phân loại một tệp sạch hay độc hại. Các đặc trưng được trích xuất đều dựa trên phân tích tĩnh tệp PDF, gồm hai loại đặc trưng thống kê và đường dẫn cấu trúc. Phần cuối của chương sử dụng mô hình học máy để đánh giá kết quả khi kết hợp các đặc trưng trên.

2.1 Tập dữ liệu

Tập dữ liệu bao gồm 22046 tệp PDF riêng biệt, với tỉ lệ tệp độc là 58%, tệp sạch là 42%, không gây sự chênh lệch quá lớn trên hai nhãn. Trong đó số lượng tài liệu có nhúng javascript chiếm 75% (Bảng 2.1). Toàn bộ tập dữ liệu là kết quả của quá trình thu thập và loại bỏ tệp trùng lặp từ một số nguồn như Contagio [10], Evasive-PDFMal2022 [3] và VirusTotal ¹. Với VirusTotal, thực hiện truy vấn để tìm kiếm các tệp PDF độc hại có nhúng javascript và được gán nhãn CVE có độ nguy hiểm cao. Một số CVE được thu thập trong các tài liệu PDF độc như CVE-2007-5659, CVE-2008-2992, CVE-2009-4324, CVE-2010-0188,...

Các tệp PDF sau khi được xử lý qua quá trình trích xuất đặc trưng, sẽ

¹<https://www.virustotal.com/>

Bảng 2.1: Thống kê số lượng tệp dữ liệu có chứa javascript.

	contain js	no contain js	total
malware	12183	584	12767
benign	4326	4953	9279

Bảng 2.2: Phân bổ tập huấn luyện, kiểm chứng và tập kiểm thử

	Samples
Training set	15431
Validation set	4409
Test set	2205

được phân chia thành các tập dữ liệu phục vụ quá trình huấn luyện và kiểm định. Với mô hình yêu cầu một tập validation phục vụ cho quá trình điều chỉnh tham số hay lựa chọn bộ đặc trưng tốt, sự phân chia sẽ theo tỉ lệ 7:2:1 với 70% là tập huấn luyện, 20% là tập kiểm chứng, 10% là tập kiểm thử (Bảng 2.2).

2.2 Trích xuất đặc trưng

2.2.1 Đặc trưng thống kê

Các đặc trưng thống kê là các đặc trưng được trích xuất từ quá trình phân tích cú pháp, cấu trúc của tệp PDF, mang ý nghĩa thống kê như đếm số lượng từ khóa xuất hiện trong tài liệu, phân tích tương quan giữa các giá trị, số lượng bộ lọc được sử dụng...

Trình đọc PDF sử dụng các từ khóa được nhúng trong tài liệu, theo sau bởi ký tự '/' để hiểu những hành động cần thực thi, do đó tập hợp những từ khóa có thể là các chỉ báo hiệu quả về hành vi của tài liệu. Một số từ khóa có thể có liên quan chặt chẽ tới một số lỗ hổng hoặc hành vi độc hại, điển hình như '/JS' hay '/Javascript' chỉ ra có đoạn mã javascript được nhúng trong tài liệu, những đoạn mã này có độ nguy hiểm cao bởi xuất hiện trong hầu hết các cuộc tấn công mã độc; hay '/URI': cho phép truy cập một nguồn từ xa, bằng cách này, kẻ tấn công có thể chuyển hướng người dùng tới một trang web độc hại.

Tài liệu PDF hỗ trợ một số các bộ lọc phục vụ nén, mã hóa dữ liệu. Những kẻ tấn công có thể lợi dụng bộ lọc để ẩn các đoạn mã độc hại và che giấu hành vi. Do đó, việc xác định những bộ lọc nào được sử dụng, những bộ lọc nào thường xuất hiện trong tệp độc hại hay tệp sạch sẽ mang lại những

đặc trưng có giá trị trong việc phân loại tệp.

Ngoài ra đặc trưng thống kê cũng bao gồm các metafeatures - các đặc trưng phản ánh phần nào các thuộc tính của siêu dữ liệu, ví dụ như số lượng các từ viết hoa trong trường tác giả, hay tỉ lệ của số trang trên kích thước tài liệu.

Thông qua việc kết hợp sử dụng hai công cụ PeePDF ² và PDFiD ³, tôi thu lại được tập các đặc trưng thống kê được lựa chọn trích xuất. Bộ đặc trưng này gồm có 67 đặc trưng (Bảng 2.3).

Bảng 2.3: Danh sách 67 đặc trưng thống kê và mô tả

Tên đặc trưng	Mô tả
file_size	Kích thước tệp
pdfid_check_isPDF	Kiểm tra xem tệp có phải là tài liệu PDF hay không
is_linearized	Đặc trưng này dùng để kiểm tra tệp có là Linearized PDF hay không
is_encrypted	Kiểm tra xem tệp có bị mã hóa hay không
pdf_version	Phiên bản tài liệu PDF
num_updates	Số lần tệp PDF được cập nhật
num_objects	Số lượng đối tượng xuất hiện trong tệp
num_streams	Số lượng các đối tượng stream
peepdf_num_errors	Số lượng lỗi khi phân tích cú pháp tệp qua công cụ PeePDF
pdfid_error_occured	Kiểm tra tệp có bị lỗi khi phân tích qua công cụ PDFiD hay không
bad_pdf_header	Kiểm tra sự không hợp lệ của header tệp PDF
num_comments	Số lượng comment trong tệp
num_indirect_objects	Số lượng đối tượng tham chiếu trong tệp
num_compressed_objects	Số lượng các đối tượng được nén
num_encode_streams	Số lượng stream được mã hóa
num_decode_streams_error	Số lượng lỗi xuất hiện khi thực hiện giải mã những stream bị mã hóa
num_error_objects	Số lượng các đối tượng lỗi
num_object_streams	Số lượng các stream chứa đối tượng
num_xref_streams	Số lượng xref stream (từ phiên bản 1.5, tệp PDF cho phép bảng tham chiếu được lưu dưới dạng stream với từ khóa /ObjStm)

²<https://github.com/jesparza/peepdf>

³<https://www.kali.org/tools/pdfid/>

num_objects_contain_js	Số lượng đối tượng chứa đoạn mã javascript
num_%%EOF	Số lần xuất hiện của từ khóa đánh dấu sự kết thúc của tệp PDF
char_after_last_EOF	Số lượng ký tự sau từ khóa kết thúc cuối cùng trong tệp PDF
stream_entropy_less_than_2	Kiểm tra xem entropy của tệp có nhỏ hơn 2 hay không (entropy tính toán sự phân bố ngẫu nhiên của các bytes trong chuỗi)
total_file_entropy	Giá trị entropy của toàn bộ tệp
stream_entropy	Giá trị entropy bên trong các đối tượng stream
non_stream_entropy	Giá trị entropy của các chuỗi bên ngoài stream
obj	Số lần xuất hiện của từ khóa 'obj', từ khóa xác định sự bắt đầu của một đối tượng
endobj	Số lần xuất hiện của từ khóa 'endobj', từ khóa xác định sự kết thúc của một đối tượng
stream	Số lần xuất hiện của từ khóa 'stream', từ khóa xác định sự bắt đầu của một stream
endstream	Số lần xuất hiện của từ khóa 'endstream', từ khóa xác định sự kết thúc của một stream
xref	Số lần xuất hiện của từ khóa 'xref', từ khóa xác định sự bắt đầu của bảng tham chiếu tới vị trí các đối tượng trong tệp PDF
trailer	Số lần xuất hiện của từ khóa 'trailer', từ khóa xác định trailer của tệp PDF
startxref	Số lần xuất hiện của từ khóa 'startxref', từ khóa xác định vị trí bắt đầu của bảng tham chiếu.
/Page	Số lần xuất hiện của từ khóa 'Page', xác định một trang của tệp PDF
/Encrypt	Số lần xuất hiện của từ khóa 'Encrypt'
/ObjStm	Số lần xuất hiện của từ khóa 'ObjStm'
/JS	Số lần xuất hiện của từ khóa 'JS', xác định đoạn mã javascript
/JavaScript	Số lần xuất hiện của từ khóa 'JavaScript', tương tự như từ khóa 'JS'
/AA	Số lần xuất hiện của từ khóa 'AA', xác định các hành động bổ sung, các sự kiện kích hoạt một hành vi nào đó

/OpenAction	Số lần xuất hiện của từ khóa 'OpenAction', xác định hành động sẽ được thực thi ngay sau khi tệp PDF được mở
/AcroForm	Số lần xuất hiện của từ khóa 'AcroForm', xác định các phần tử biểu mẫu tương tác
/Names	Số lần xuất hiện của từ khóa 'Names'
/RichMedia	Số lần xuất hiện của từ khóa 'RichMedia', thường chứa embedded flash
/Launch	Số lần xuất hiện của từ khóa 'Launch', được sử dụng để mở một thực thể bên ngoài, có thể là một tệp thực thi.
/XFA	Số lần xuất hiện của từ khóa 'XFA'
/JBIG2Decode	Số lần xuất hiện của từ khóa 'JBIG2Decode'
/U3D	Số lần xuất hiện của từ khóa 'U3D'
/PRC	Số lần xuất hiện của từ khóa 'PRC'
/Flash	Số lần xuất hiện của từ khóa 'Flash'
/SubmitForm	Số lần xuất hiện của từ khóa 'SubmitForm'
/ImportData	Số lần xuất hiện của từ khóa 'ImportData'
Colors >2 ²⁴	Số lượng màu có kích thước lớn hơn 2 ²⁴ (hơn 3 bytes)
combination_js_and_action	Kiểm tra sự xuất hiện đồng thời của các đoạn mã javascript và các hành động khởi chạy
num_EmbeddedFile	Số lượng tệp được nhúng trong tài liệu PDF
num_URI	Số lần xuất hiện của từ khóa 'URI', đường dẫn tới một nguồn từ xa, có thể là một trang web
num_ASCIIHexDecode	Số lần sử dụng thuật toán giải mã ASCIIHexDecode
num_ASCII85Decode	Số lần sử dụng thuật toán giải mã ASCII85Decode
num_LZWDecode	Số lần sử dụng thuật toán giải mã LZWDecode
num_FlateDecode	Số lần sử dụng thuật toán giải mã FlateDecode
num_RunLengthDecode	Số lần sử dụng thuật toán giải mã RunLengthDecode
num_CCITTFaxDecode	Số lần sử dụng thuật toán giải mã CCITTFaxDecode

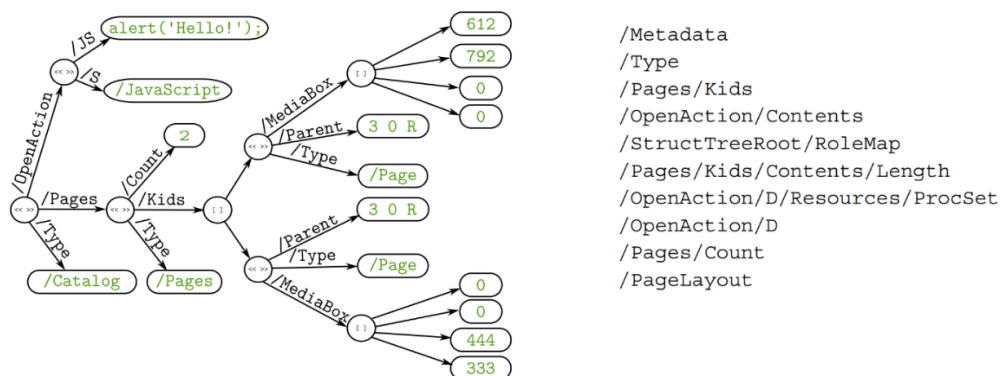
num_JBIG2Decode	Số lần sử dụng thuật toán giải mã JBIG2Decode
num_DCTDecode	Số lần sử dụng thuật toán giải mã DCTDecode
num_JPXDecode	Số lần sử dụng thuật toán giải mã JPXDecode
num_Crypt	Số lượng stream bị mã hóa
num_stream_filters	Số lượng bộ lọc được sử dụng
num_unknown_stream_filters	Số lượng bộ lọc được sử dụng nhưng không trong danh sách các bộ lọc được hỗ trợ của tài liệu PDF

2.2.2 Đặc trưng cấu trúc cây

Như đề cập ở mục 1.1.3, các đối tượng của tệp PDF được tổ chức phân cấp dạng cây. Do đó việc xem xét cách sắp xếp các đối tượng trên cấu trúc cây như vậy có thể tiết lộ những thông tin giá trị để phát hiện tài liệu độc hại. Phương pháp tiếp cận này được đề xuất bởi Nedim Srndic và Lavel Laskov [13]. Ở phương pháp này, mỗi đường dẫn cấu trúc tới lá trong cây sẽ đại diện cho một đặc trưng, tập hợp các đặc trưng sau đó sẽ được xử lý thông qua một kỹ thuật gọi là Hợp nhất đường dẫn cấu trúc để hợp nhất những đường dẫn tương tự nhau.

Đường dẫn cấu trúc

Đường dẫn cấu trúc trong cây cấu trúc PDF được định nghĩa là một chuỗi các cạnh bắt đầu từ nút gốc Catalog và kết thúc bởi một đối tượng đại diện cho nút lá. Ví dụ theo hình 2.1, một đường dẫn cấu trúc bắt đầu bởi nút gốc đi qua cạnh /Pages và /Count để tới nút lá có giá trị là 2, được biểu diễn dưới dạng chuỗi '/Pages/Count'.



Hình 2.1: Minh họa cấu trúc cây và đường dẫn cấu trúc của tệp PDF

Để tạo ra một đặc trưng, các đường dẫn sẽ trở thành tên, còn giá trị sẽ được xác định dựa trên giá trị của nút lá ứng với đường dẫn đó. Liên quan đến việc xử lý các dữ liệu ở nút lá mà không phải dạng số (dữ liệu dạng số bao gồm số nguyên, số thực, boolean), thì các dữ liệu kiểu chuỗi sẽ được thay thế bởi một giá trị cố định. Một mảng các số thì được thay thế bằng trung bình cộng của mảng đó. Với các đường dẫn xuất hiện nhiều lần, tham chiếu tới nhiều giá trị, thì sẽ hợp nhất thành một đường dẫn duy nhất có giá trị là một mảng các giá trị trên. Hình 2.2 và 2.3 mô phỏng lại cách thuật toán chuẩn hóa các giá trị về dạng số như vừa đề cập.

<pre> /OpenAction/JS: alert('Hello!'); /OpenAction/S: /JavaScript /Pages/Count: 2 /Pages/Kids/MediaBox: 0 0 612 792 0 0 333 444 /Pages/Kids/Parent: 3 0 R 3 0 R /Pages/Kids/Resources: ... /Pages/Kids/Type: /Page /Page /Pages/Type: /Pages /Type: /Catalog </pre>	
Keys	Values

Hình 2.2: Các cặp đường dẫn cấu trúc và giá trị chưa chuẩn hóa

/OpenAction/JS:1.0	1.0
/OpenAction/S:1.0	1.0
/Pages/Count:2	2.0
/Pages/Kids/MediaBox:0 0 612 792 0 0 333 444	166.5
/Pages/Kids/Parent:1.0 1.0	1.0
/Pages/Kids/Resources:...	...
/Pages/Kids/Type:1.0 1.0	1.0
/Pages/Type:1.0	1.0
/Type:1.0	1.0
Structural Multimap	Feature Vector

Hình 2.3: Các cặp đường dẫn cấu trúc và giá trị sau khi được chuẩn hóa về dạng số

Hợp nhất đường dẫn cấu trúc

Một tệp PDF có cấu trúc tương đối phức tạp, do đó có thể trích xuất được một lượng lớn các đường dẫn và có nhiều đường dẫn có cấu trúc tương tự nhau. Do đó, một tập các quy tắc được xây dựng để hợp nhất các đường dẫn tương đương [ref]

Lựa chọn đặc trưng

Một số đường dẫn có thể xảy ra rất ít trong tập dữ liệu được trích xuất, việc giữ những đặc trưng này làm tăng kích thước của không gian đầu vào mà không cải thiện được độ chính xác của bộ phân loại. Do đó, việc lựa chọn đặc trưng cần được thực hiện để hạn chế tác động của những đặc trưng ít xuất hiện. Ở trong phạm vi cài đặt thí nghiệm sử dụng công cụ Hidost, lựa chọn đặc trưng dựa trên việc đặt ngưỡng cho sự phổ biến của đặc trưng trong tập dữ liệu.

Công cụ Hidost

Công cụ Hidost ⁴ là công cụ được phát triển dựa theo ý tưởng trên, thực hiện trích xuất các đặc trưng đường dẫn cấu trúc của một tệp PDF, ngoài ra còn hỗ trợ các tệp SWF. Đầu vào của công cụ là một tập các tệp PDF, đầu ra là một tập các tệp dữ liệu đã được trích xuất tương ứng với từng tệp

⁴<https://github.com/srndic/hidost>

đầu vào. Trong quá trình cài đặt, với số lượng mẫu hơn 20000 tệp, tôi đã tiến hành đặt ngưỡng 900, tức các đặc trưng đường dẫn xuất hiện trong ít nhất 900 tệp sẽ được sử dụng trong tập đặc trưng cuối cùng. Kết quả sau khi thực thi, có 315 đặc trưng được trích xuất, phục vụ cho quá trình huấn luyện xây dựng mô hình học máy ở các phần sau.

2.3 Phân loại và đánh giá kết quả

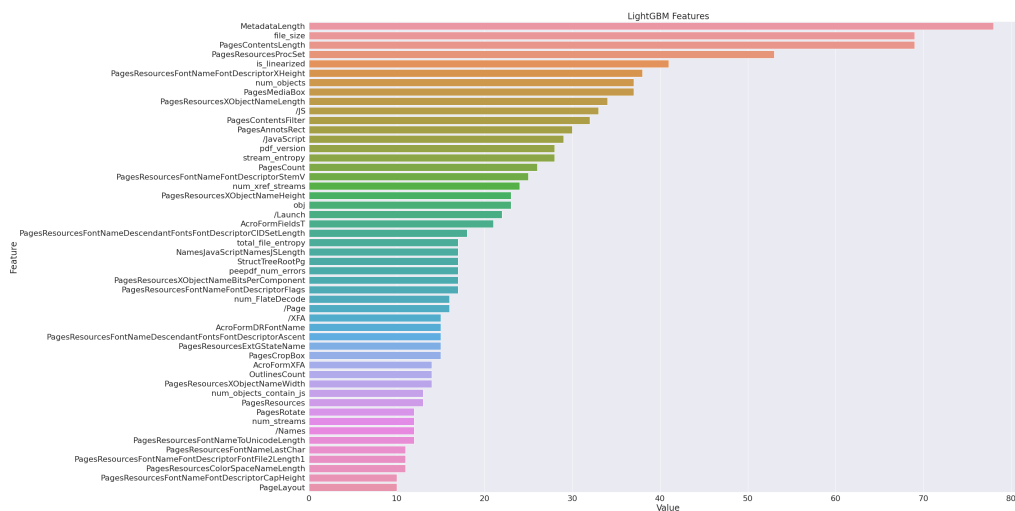
Như đã đề cập ở mục 1.3, một quy trình phát hiện tự động tệp PDF độc hại gồm trích xuất đặc trưng và xây dựng mô hình phân loại. Kết quả của quá trình thử nghiệm trích xuất đặc trưng trong mục 2.2 mang lại một tập các dữ liệu tệp đã được xử lý, là đầu vào cho quá trình phân loại này. Khi thực hiện phân loại, với mục tiêu xây dựng một mô hình học máy có thể mang lại kết quả dự đoán tốt nhất, các đặc trưng đã được xếp hạng, lựa chọn và đánh giá trên tập kiểm định.

Hai thuật toán học máy được cân nhắc đưa vào thử nghiệm gồm Random Forest và LightGBM. Ngoài ra, ở cả hai thuật toán này, tôi sử dụng thêm **optuna** ⁵ - một framework hỗ trợ điều chỉnh tự động các tham số cho mô hình học máy để mang lại kết quả tốt nhất mong muốn. Sau khi thử điều chỉnh tham số với 100 vòng lặp, nhận thấy LightGBM mang lại kết quả tốt hơn trên tập kiểm định với độ chính xác tốt nhất là 99,68%, trong khi RandomForest mang lại độ chính xác tốt nhất là 99.47%. Từ đây, thuật toán LightGBM sẽ được sử dụng trong tất cả các kết quả dưới đây.

2.3.1 Xếp hạng đặc trưng theo độ quan trọng

Quá trình trích xuất sử dụng các công cụ PDFiD, PeePDF và Hidost mang lại tổng 382 đặc trưng, trong đó có 67 đặc trưng thống kê và 315 đặc trưng đường dẫn cấu trúc. Thông qua thuật toán LightGBM, mức độ quan trọng của các đặc trưng được đánh giá như hình 2.4

⁵<https://optuna.org>



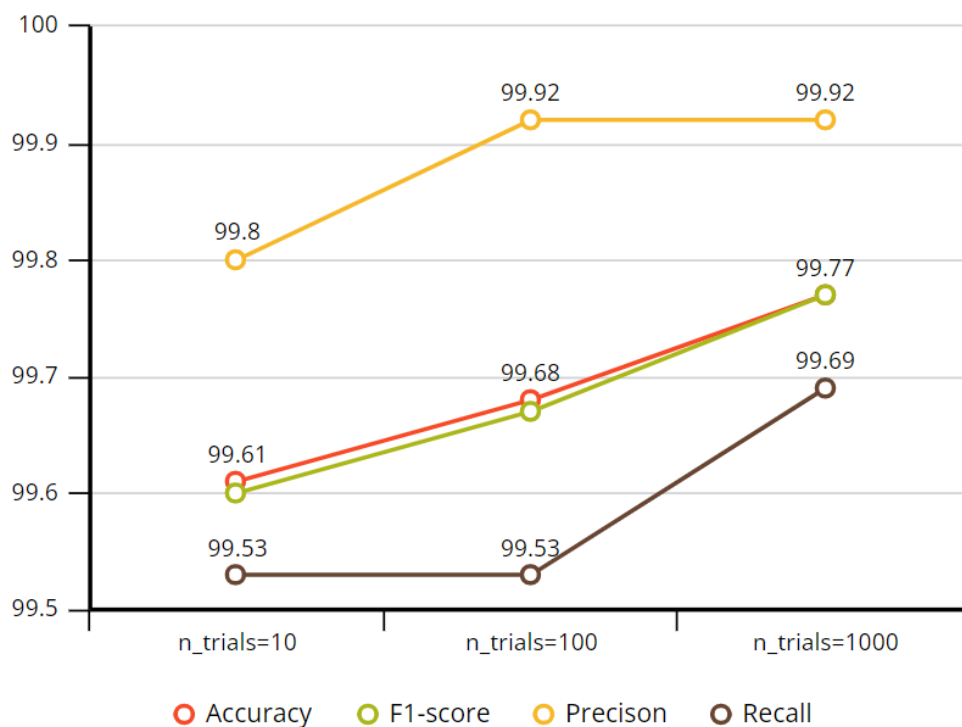
Hình 2.4: Biểu đồ 50 đặc trưng có độ quan trọng cao nhất trong các đặc trưng thống kê và đường dẫn cấu trúc.

Quan sát kết quả xếp hạng độ quan trọng các đặc trưng, có thể thấy các đặc trưng liên quan đến kích thước dữ liệu, nội dung trong tệp PDF mang lại độ quan trọng cao nhất, ví dụ như Metadata/Length, Pages/Contents/Length hay file_size.

Ngoài ra một số đặc trưng liên quan đến đoạn mã javascript hay liên quan đến biểu mẫu tương tác AcroForm trong tài liệu PDF cũng mang lại sự hữu ích trong mô hình phân loại khi xuất hiện nhiều trong top 50 này, ví dụ /JS, /Javascript, AcroFormDRFontName, AcroFormXFA, AcroFormFieldsT, NamesJavaScriptNameJS.Length.

2.3.2 So sánh kết quả khi điều chỉnh tham số và số lượng đặc trưng

Như đã giới thiệu, ở thí nghiệm này, tôi sử dụng optuna để tự động điều chỉnh tham số đầu vào cho thuật toán học máy LightGBM. Qua quá trình lặp lại với nhiều lần điều chỉnh, tôi đã chọn được bộ tham số mang lại kết quả tốt nhất cho thuật toán. Dưới đây là biểu đồ so sánh kết quả kiểm định trên các mô hình LightGBM tốt nhất được ghi lại tại lần thử thứ 10, 100 và 1000. Biểu đồ 2.5 đã cho thấy hiệu quả của việc điều chỉnh tham số trong mô hình học máy.

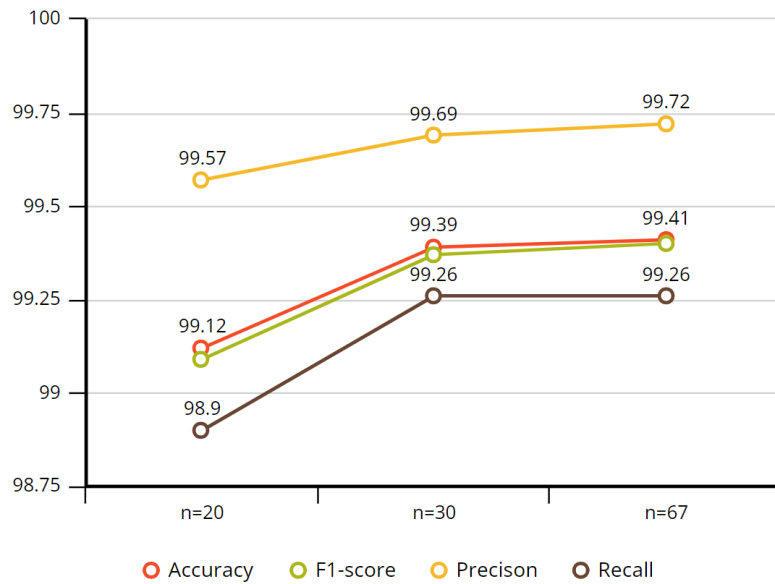


Hình 2.5: Biểu đồ so sánh kết quả điều chỉnh tham số thuật toán LightGBM qua các lần lặp trên tập kiểm chứng

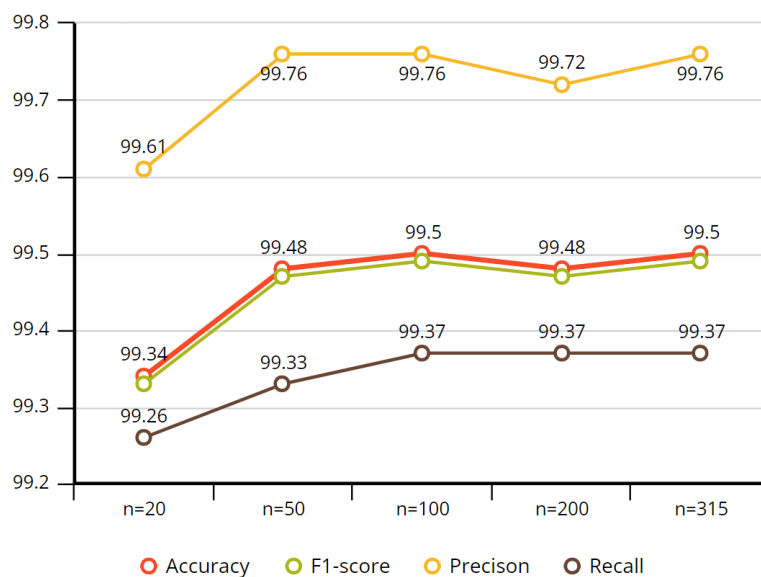
Bên cạnh việc điều chỉnh các tham số, tôi tiếp tục thử nghiệm thuật toán với việc điều chỉnh số lượng đặc trưng từ tập đặc trưng ban đầu:

- Hình 2.6: chỉ sử dụng các đặc trưng thống kê, và thực hiện thay đổi số lượng đặc trưng, kết quả cho thấy việc sử dụng toàn bộ tập đặc trưng thống kê mang lại kết quả tốt nhất với độ chính xác 99.41%
- Hình 2.7: chỉ sử dụng các đặc trưng đường dẫn cấu trúc, và thực hiện thay đổi số lượng đặc trưng, kết quả cho thấy việc sử dụng toàn bộ tập đặc trưng cấu trúc cây mang lại kết quả tốt nhất với độ chính xác 99.5%
- Hình 2.8: thay đổi số lượng đặc trưng trên toàn bộ đặc trưng đã trích xuất được (bao gồm cả đặc trưng thống kê và đặc trưng cấu trúc cây). Kết quả cũng mang lại độ chính xác tốt nhất là 99.77% khi sử dụng toàn bộ đặc trưng.

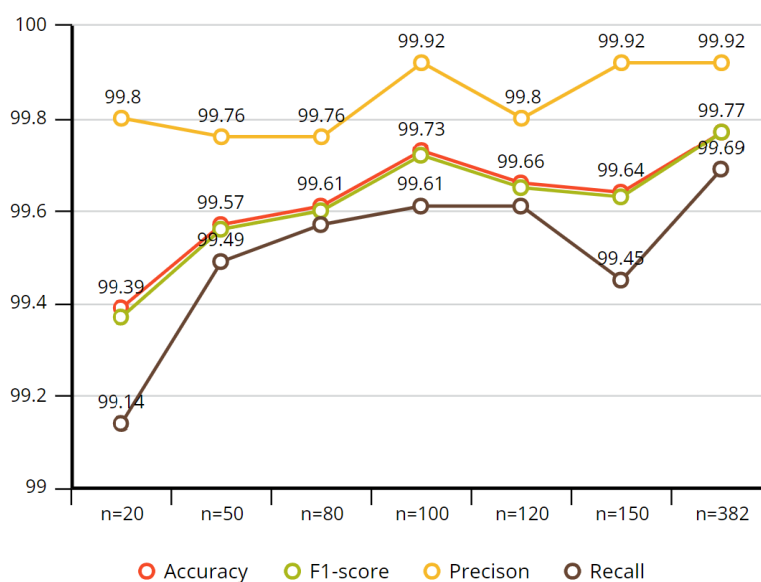
Việc sử dụng nhiều hơn các đặc trưng bằng thực nghiệm trên đã chứng minh là mang lại kết quả tốt hơn trên mô hình học máy LightGBM.



Hình 2.6: Biểu đồ so sánh mô hình khi sử dụng số lượng đặc trưng tăng dần từ tập đặc trưng thống kê.



Hình 2.7: Biểu đồ so sánh kết quả của mô hình khi sử dụng số lượng đặc trưng tăng dần từ tập đặc trưng cấu trúc cây

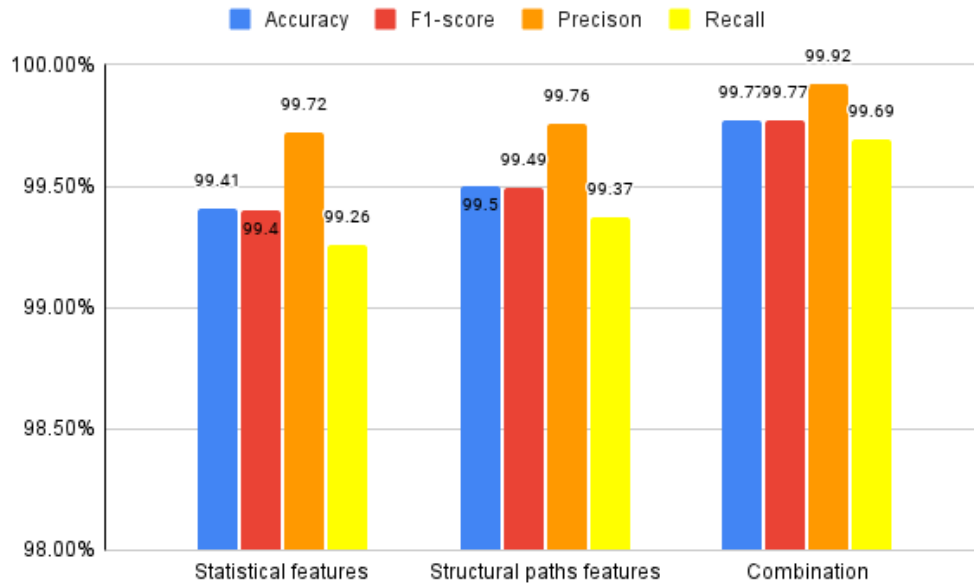


Hình 2.8: Biểu đồ so sánh kết quả của mô hình khi sử dụng số lượng đặc trưng tăng dần từ sự kết hợp các tập đặc trưng thống kê và đặc trưng cấu trúc cây.

2.3.3 So sánh kết quả khi kết hợp các loại đặc trưng

Để đánh giá sự kết hợp của các đặc trưng thống kê và đường dẫn cấu trúc, thực hiện 3 thử nghiệm cho mô hình khi chỉ sử dụng tập đặc trưng thống kê, chỉ sử dụng tập đặc trưng cấu trúc cây và cuối cùng là sử dụng kết hợp hai loại đặc trưng này. Kết quả được biểu diễn trên biểu đồ 2.9, đã cho thấy một kết quả tốt nhất khi kết hợp các loại đặc trưng này với nhau.

Từ đó có thể kết luận, trong quy trình xác định một tài liệu PDF có độc hại hay không, việc phân tích và trích xuất nhiều loại đặc trưng nhất có thể, ở đây là các đặc trưng mang tính thống kê và các đặc trưng cấu trúc sẽ mang lại một mô hình phân loại đáng tin cậy hơn, thay vì chỉ sử dụng một số loại đặc trưng đơn lẻ.



Hình 2.9: Biểu đồ so sánh kết quả của mô hình khi sử dụng hai loại đặc trưng độc lập và khi kết hợp các đặc trưng thống kê và đặc trưng cấu trúc cây

Chương 3

Xây dựng mô hình phân loại Zero False Positive

3.1 Đặt vấn đề

3.2 Bộ phân loại Zero False Positive

3.2.1 Thuật toán phân loại lặp với mục tiêu $FP = 0$

3.2.2 Thực nghiệm và đánh giá kết quả

Tiến hành thực nghiệm để kiểm định mô hình trên tập huấn luyện gồm các tệp PDF đã được trích xuất đặc trưng thống kê và đặc trưng cấu trúc trong chương 2. Kết quả cho thấy số lượng FP qua các vòng lặp đều được đảm bảo bằng 0 và số lượng FN giảm dần. Tuy nhiên, từ vòng lặp thứ 700 - 1000, kết quả FN không giảm, duy trì số lượng 97, với tỉ lệ FNR bằng 0.0095.

Ngoài ra, mô hình được tiến hành thử nghiệm trên một số tập dữ liệu khác. Bộ dữ liệu đặc trưng tệp PDF được trích xuất và công bố bởi đội ngũ phát triển Hidost [ref hidost reputation] với một số lượng lớn các mẫu PDF độc hại được thu thập qua nhiều năm. Tôi đã liên hệ Nedim Srndic, một trong những tác giả của Hidost để xin phép sử dụng bộ dữ liệu đặc trưng trong phạm vi khóa luận này. Bộ dữ liệu được phân thành nhiều tập, và dưới đây là kết quả của hai tập dữ liệu khi thử nghiệm trên mô hình phân loại Zero False Positive.

Với số lượng tập sạch lớn, chênh lệch so với số lượng tập độc, tập dữ liệu của Hidost yêu cầu thời gian chạy lớn khi thực hiện thuật toán phân loại Zero False Positive, và tỉ lệ FNR còn khá cao, 0.043 với tập thứ 1 và 0.041 ở tập thứ 2.

Bảng 3.1: Kết quả của thuật toán phân loại Zero False Positive trên tập dữ liệu đã trích xuất

Iteration	TN	FP	FN	TP
0	7423	0	2993	7220
5	7423	0	361	9852
10	7423	0	255	9988
50	7423	0	133	10080
100	7423	0	107	10106
200	7423	0	105	10108
500	7423	0	100	10113
700	7423	0	97	10116
1000	7423	0	97	10116

Bảng 3.2: Kết quả thuật toán phân loại Zero False Positive trên tập dữ liệu Hidost (1)

Iteration	TN	FP	FN	TP
0	142796	0	1851	4406
10	142796	0	410	5847
50	142796	0	287	5970
100	142796	0	272	5985
120	142796	0	271	5986
160	142796	0	271	5986

Bảng 3.3: Kết quả thuật toán phân loại Zero False Positive trên tập dữ liệu Hidost (2)

iteration	TN	FP	FN	TP
0	132465	0	4919	2484
10	132465	0	539	6864
50	132465	0	317	7086
100	132465	0	305	7098
190	132465	0	305	7098

3.3 Phương hướng xử lý các tệp được gán nhãn sạch

Sau khi lọc các tệp PDF có chứa mã javascript qua mô hình phân loại Zero False Positive, các tệp được gán nhãn sạch sẽ có nguy cơ chứa các tệp độc hại. Từ đây các mẫu này sẽ tiếp tục được xử lý tại giai đoạn hai: thực hiện phân tích và trích xuất những đặc trưng liên quan tới javascript được nhúng trong tài liệu PDF. Trong phần này, một số cách tiếp cận xử lý nâng cao bao gồm cả phân tích tĩnh và phân tích động sẽ được giới thiệu.

Chương 4

Kết luận

4.1 Kết quả đạt được

Khóa luận đã trình bày những phương pháp phân tích, trích xuất đặc trưng và đề xuất các mô hình sử dụng trong phân loại tự động các tài liệu PDF độc hại.

Kết quả thí nghiệm 1 đã chỉ ra rằng, bằng cách trích xuất một tập hợp các đặc trưng trên phạm vi rộng, kết hợp nhiều loại đặc trưng khác nhau có thể tạo ra một công cụ phân loại và phát hiện phần mềm độc hại mạnh mẽ mang lại độ chính xác cao. Ở thí nghiệm này, em đã sử dụng một số kỹ thuật học máy khác nhau, học nhiều lần để điều chỉnh tham số của từng kỹ thuật và cuối cùng đưa ra một mô hình sử dụng thuật toán LightGBM được đánh giá là mang lại hiệu quả nhất.

Tuy nhiên, với các cuộc tấn công né tránh, khi những tài liệu độc hại sử dụng nhiều kỹ thuật nâng cao, tinh vi nhằm che giấu những hành vi độc hại và bắt chước tính bình thường của các tệp sạch trong tập huấn luyện, thì chỉ sử dụng những đặc trưng thống kê và cấu trúc sẽ không thực sự hiệu quả, gây ra một số lượng âm tính giả cao. Từ đó, với việc đề xuất một phương pháp phát hiện tài liệu độc hại với hai giai đoạn, ở giai đoạn một, đã chỉ ra được một mô hình có thể mang lại tỷ lệ dương tính giả (FP) bằng 0, trong khi tỷ lệ âm tính giả (FN) được giữ ở mức tối thiểu trên tập huấn luyện. Mục tiêu này đã được thử nghiệm trên bộ dữ liệu được trích xuất ở thí nghiệm 1 và một số bộ dữ liệu đặc trưng được trích xuất trong một số công bố khác. Mô hình này được ứng dụng trong phạm vi khóa luận như một tường lửa, bước đầu lọc ra những tài liệu đảm bảo độc hại, trong khi những tài liệu được gán nhãn sạch sẽ tiếp tục được xử lý qua giai đoạn hai, thực hiện phân tích sâu hơn về các đoạn mã nhúng trong tài liệu.

Ngoài ra, mô hình còn có tính ứng dụng cao trong những hệ thống cần

lọc và ngăn chặn các xâm nhập bất thường. Ở các hệ thống nhạy cảm với độ trễ, khi mà tính liên tục của hoạt động là rất quan trọng, thì việc ngăn chặn lưu lượng truy cập hợp pháp do nhầm lẫn (dương tính sai) là điều khó có thể chấp nhận được, ví dụ như một số dịch vụ của chính phủ hoặc các hệ thống cung cấp điện, hệ thống phân phối nước.

Bên cạnh đó, mô hình ở thí nghiệm 2 hoàn toàn có thể ứng dụng trong các hệ thống mong muốn phân loại âm tính giả bằng 0, bằng cách đổi lại cách gán nhãn kết quả. Ở những ứng dụng không mong muốn bỏ sót bất kỳ tài liệu độc hại nào như những hệ thống có tính bảo mật cao, những sự xâm phạm có thể gây hậu quả lớn.

4.2 Hạn chế và khó khăn

- Các công cụ hỗ trợ trích xuất đặc trưng khá cũ, yêu cầu cài đặt môi trường và một số gói phụ thuộc (package dependencies) cụ thể, gây khó khăn trong việc tiến hành nhiệm vụ trích xuất đặc trưng. Ngoài ra nhiều tệp PDF có cấu trúc biến đổi, gây khó khăn cho một số trình phân tích cú pháp.
- Các đặc trưng API Javascript đã được trích xuất nhưng số lượng ít và không mang lại kết quả khả quan nên không được đề cập tới trong thí nghiệm của khóa luận này.

4.3 Hướng nghiên cứu tương lai

Phụ lục A

Phụ lục chứng minh

A.1 Zero FP

Cho một tập dữ liệu với 2 nhãn và một bộ phân loại C , sau một số vòng lặp hữu hạn với thuật toán được đề xuất trên, có thể dẫn tới số lượng $FP = 0$.

Chứng minh:

Cho tập dữ liệu $T(x_i; y_i)$ với $i = 1, \dots, m$ và y thuộc $(-1; 1)$. Bộ phân loại C mục tiêu giảm thiểu hàm chi phí J được định nghĩa trước.

Chú ý rằng thuật toán sẽ tạo ra một tập giảm lược từ tập T ban đầu, gồm tất cả các mẫu âm, do đó, nếu với đường biên B_i của bộ phân loại C giúp FP trên tập giảm lược bằng 0 thì FP trên tập T ban đầu cũng sẽ bằng 0. Từ đó, chúng ta sẽ chứng minh FP sẽ có thể bằng 0 trên tập giảm lược, gọi là tập S qua hữu hạn các vòng lặp.

Giả sử vòng lặp đầu tiên, bộ phân loại với đường biên B_0 gây ra một số FN_s và FP_s trên tập S . Hàm chi phí của bộ phân loại là $J(FN_0, FP_0)$. Theo thuật toán Zero FP, những tập dương gây ra FN sẽ được loại bỏ khỏi tập giảm lược S , thành $S(1)$ ($S(i)$ là tập S sau lần cập nhật thứ i). Khi đó, với biên B_0 , tập $S(1)$ sẽ có $FN_0 = 0$ và hàm chi phí trên tập $S(1)$ chỉ phụ thuộc vào FP , tức:

$$J(FP_0) < J(FN_0, FP_0) \text{ hay } J(FP_0) = r_1 J(FN_0, FP_0) (0 \leq r_1 \leq 1)$$

Tập được cập nhật $S(1)$ sau đó sẽ được phân loại lại với C với đường biên B_1 mới, với hàm chi phí $J(FN_1, FP_1)$. Mục tiêu của hàm phân loại là giảm hàm chi phí, do đó, hàm chi phí lúc này sẽ nhỏ hơn hàm chi phí trước khi phân loại.

$$J(FN_1, FP_1) < J(FP_0) \text{ hay } J(FN_1, FP_1) = q_1 J(FP_0) (0 \leq q_1 \leq 1)$$

Hiển nhiên nếu lúc này FP_1 bằng 0, chúng ta được điều phải chứng minh. Ngược lại, với $FP_1 > 0$, nếu $FN_1 > 0$ thì thuật toán tiếp tục thực hiện loại bỏ những mẫu dương gây FN_1 , thành tập $S(2)$ với hàm chi phí trên B_1 là

$$J(FP_1) < J(FN_1, FP_1) \text{ hay } J(FP_1) = r_2 J(FN_1, FP_1)$$

Tương tự như vậy, với thuật toán giảm lược các mẫu dương bị sai (gây FN), tại vòng lặp thứ k , sau khi tìm ra B_k mới dựa trên phân loại $S(k)$ được giảm lược từ vòng $k-1$, hàm chi phí là $J(FN_k, FP_k) = q_k J(FP_{k-1})$

Kết thúc vòng lặp này, nếu FP vẫn khác 0, thì sẽ thực hiện loại bỏ FN , trở thành

$$\begin{aligned} J(FP_k) &= r_k J(FN_k, FP_k) \\ \Rightarrow J(FP_k) &= r_k q_k J(FP_{k-1}) = \dots = T(r_i q_i) J(FP_0) \end{aligned}$$

Vì $T(r_i q_i) \rightarrow 0$. Do đó $\lim_{k \rightarrow \infty} J(FP_k) = 0$ với biên B_k . Từ đó khi với biên B_k , tập S có $FP = 0$, thì tập ban đầu cũng có $FP = 0$

A.2 Tập giảm lược S luôn có thể phân tách được

Với điều kiện khởi tạo ban đầu đảm bảo tập S phân tách được, ta được điều phải chứng minh ở vòng lặp đầu tiên.

Do đó, ta chỉ cần chứng minh sau một hữu hạn vòng lặp, nhờ thuật toán lược bỏ một số mẫu ra khỏi tập dữ liệu, tập S có thể phân tách được. Sử dụng kết quả của định lý chứng minh trong mục trên, ta được:

$$J(FN_{k+1}, FP_{k+1}) = q_{k+1} J(FP_k)$$

$$\text{Mà } J(FP_k) \rightarrow 0 \text{ khi } k \rightarrow \infty.$$

$$\Rightarrow J(FN_{k+1}, FP_{k+1}) \rightarrow 0 \text{ khi } k \rightarrow \infty.$$

Lúc này, hàm chi phí bị ảnh hưởng bởi số lượng FN và FP tiến tới 0, mang lại kết quả phân tách được trên tập S .

A.3 FN giảm dần khi FP bằng 0

Chú ý rằng mục tiêu của thuật toán là $FP = 0$, nên chỉ khi $FP = 0$, chúng ta mới tập trung giảm FN . Do đó phần này sẽ chứng minh trong khi duy trì $FP = 0$, tập S phân tách được thì thuật toán sẽ giảm FN trên tập T ban đầu. Xét tập giảm lược S , với $S(i)$ là lần cập nhật thứ i của tập S sau mỗi vòng lặp.

Từ điều kiện của giả thiết, giả sử $S(k)$ phân tách được với biên B_k , tức $FP = FN = 0$. Biên này trên tập T ban đầu gây ra một lượng $TP(k) > 0$ và $FN(k) > 0$, thuật toán thực hiện thêm tất cả các mẫu dương được phân loại đúng (TP) và một vài mẫu dương trong $FN(k)$ mẫu dương bị phân loại sai (gây FN) vào tập $S(k)$, thành $S(k+1)$. Khi đó ở vòng lặp tiếp theo, tập $S(k+1)$ giữ nguyên số lượng mẫu âm, tăng số lượng mẫu dương. Do đó biên $B(k+1)$ sẽ cố gắng dịch chuyển về phía mẫu âm để hàm chi phí J

được giảm, tức mô hình sẽ chú ý vào các mẫu dương bị phân loại sai (gây FN), cố gắng dịch chuyển để có thể phân loại đúng các mẫu này. Từ đó $FN(k+1) \leq FN(k)$, tức số lượng FN ở trên tập T có thể được giảm.

Tuy nhiên, theo như giả thiết, FN chỉ giảm khi FP được giữ về 0, ngay khi $FP > 0$, thuật toán sẽ thực hiện nhân bản tất cả mẫu âm hoặc lược bỏ các mẫu âm sai (FN), mục tiêu để tăng trọng số cho các mẫu âm, giảm lượng FP. Điều này có thể làm số lượng FN tăng lên. Do đó, giá trị của FN trong mô hình của Best là giá trị nhỏ nhất cục bộ.

Tài liệu tham khảo

- [1] URL: <https://www.iso.org/standard/51502.html>.
- [2] URL: <https://lightgbm.readthedocs.io/en/v3.3.2/>.
- [3] URL: <https://www.unb.ca/cic/datasets/pdfmal-2022.html>.
- [4] Igino Corona et al. “Lux0R: Detection of Malicious PDF-embedded JavaScript code through Discriminant Analysis of API References”. In: vol. 2014. Nov. 2014. DOI: 10.1145/2666652.2666657.
- [5] Michele Elingiusti, Leonardo Aniello, and Leonardo Querzoni. “PDF-Malware Detection: A Survey and Taxonomy of Current Techniques”. In: Jan. 2018, pp. 169–191. ISBN: 978-3-319-73950-2. DOI: 10.1007/978-3-319-73951-9_9.
- [6] Mohammad Sayad Haghighi, Faezeh Farivar, and Alireza Jolfaei. “A Machine Learning-based Approach to Build Zero False-Positive IPSs for Industrial IoT and CPS with a Case Study on Power Grids Security”. In: (2020), pp. 1–1. ISSN: 0093-9994. DOI: 10.1109/tia.2020.3011397.
- [7] Davide Maiorca, Giorgio Giacinto, and Igino Corona. “A Pattern Recognition System for Malicious PDF Files Detection”. In: vol. 7376. July 2012, pp. 510–524. ISBN: 9783642315367. DOI: 10.1007/978-3-642-31537-4_40.
- [8] Davide Maiorca et al. “A Structural and Content-Based Approach for a Precise and Robust Detection of Malicious PDF Files”. In: Feb. 2015. DOI: 10.5220/0005264400270036.
- [9] Himanshu Pareek. “Entropy and n-gram analysis of malicious PDF documents”. In: *International Journal of Engineering Research and Technology* (Feb. 2013).
- [10] Mila Parkour. *16,800 clean and 11,960 malicious files for signature testing and research*. URL: <http://contagiodump.blogspot.com/2013/03/16800-clean-and-11960-malicious-files.html>.

- [11] Florian Schmitt, Jan Gassen, and Elmar Gerhards-Padilla. “PDF Scrutinizer: Detecting JavaScript-based attacks in PDF documents”. In: July 2012, pp. 104–111. ISBN: 978-1-4673-2323-9. DOI: 10.1109/PST.2012.6297926.
- [12] Charles Smutz and Angelos Stavrou. “Malicious PDF detection using metadata and structural features”. In: Dec. 2012, pp. 239–248. DOI: 10.1145/2420950.2420987.
- [13] Nedim Srndic and Pavel Laskov. “Hidost: a static machine-learning-based detector of malicious files”. In: *EURASIP J. Inf. Secur.* 2016 (2016), p. 22. DOI: 10.1186/s13635-016-0045-0.
- [14] Zacharias Tzermias et al. “Combining static and dynamic analysis for the detection of malicious documents”. In: *EUROSEC ’11*. 2011.
- [15] Caglar Ulucenk, Vijay Varadharajan, and Udaya Tupakula. “Techniques for Analysing PDF Malware”. In: Dec. 2011, pp. 41–48. DOI: 10.1109/APSEC.2011.41.
- [16] Cristina Vatamanu, Dragoş Gavriluţ, and Răzvan Benchea. “A practical approach on clustering malicious PDF documents”. In: *Journal in Computer Virology* 8 (Nov. 2012). DOI: 10.1007/s11416-012-0166-z.
- [17] Contributors to Wikimedia projects. *Random forest*. June 2022. URL: https://en.wikipedia.org/wiki/Random_forest.