

RANSOMWARE DETECTION AND PREVENTION



Explainable deep learning-based ransomware detection
using dynamic analysis

NHÓM G5

Hồ Hoàng Diệp - 22520249

Nguyễn Đặng Nguyên Khang - 22520617

Trần Vỹ Khang - 22520628



... TỔNG QUAN



- Ransomware là một loại phần mềm mã độc, mã hóa dữ liệu quan trọng khiến người dùng mất quyền truy cập và yêu cầu tiền chuộc để lấy lại quyền truy cập.
- Tấn công ransomware tăng mạnh từ 2022 và xu hướng phát triển ngày càng mạnh.

Mà :

- **Phân tích tĩnh:** Nhanh, không cần thực thi mã nhưng dễ bị qua mặt (mã hóa, packing)
=> Do đó, sử dụng **phân tích động** theo dõi hành vi thực tế khi chương trình chạy, ghi nhận các chuỗi gọi API, DLL, mutex – các đặc trưng quan trọng của ransomware

Việc áp dụng deep learning/ machine learning tập trung vào real-time monitor để giúp cho việc phân tích dữ liệu thời gian thực. Kết hợp với XAI (Explainable Artificial Intelligence) để hiểu được cách mô hình đưa ra quyết định

MỤC TIÊU NGHIÊN CỨU



Phát hiện ransomware dựa trên dynamic analysis (through Cuckoo Sandbox) kết hợp với mô hình deep learning CNN 2 lớp, sử dụng mô hình XAI (LIME, SHAP) để giải thích quyết định của mô hình deep learning.

PHƯƠNG PHÁP ĐỀ XUẤT

■ Sequences

Tạo ra chuỗi đặc trưng kết hợp ba đặc điểm quan trọng trong việc nhận diện ransomware: API calls, Dlls và Mutexes

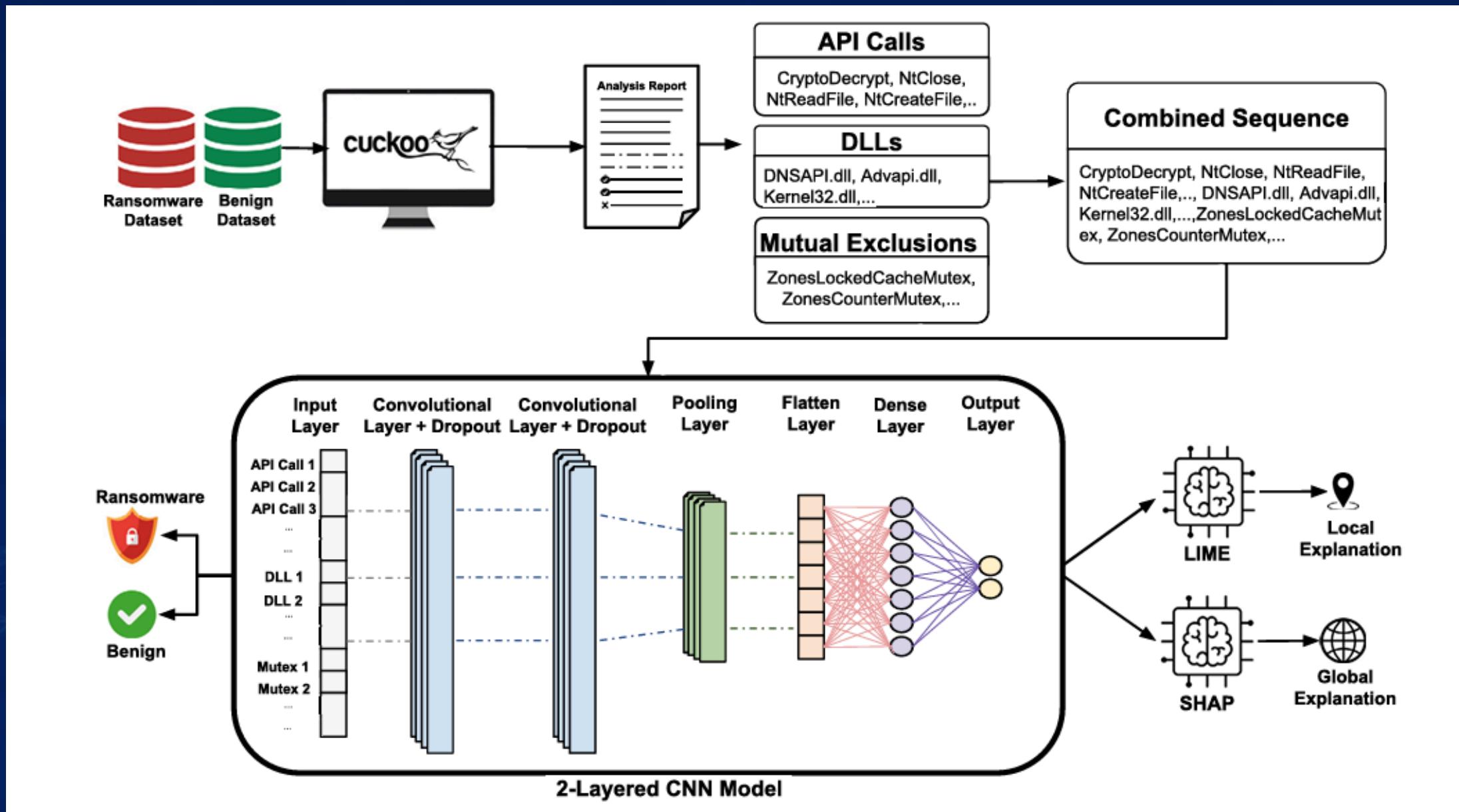
■ 2L-CNN

Triển khai hai lớp cho CNN vừa đảm bảo độ mạnh (đủ khả năng học pattern phức tạp), vừa đảm bảo độ nhẹ (không quá sâu gây overfitting hoặc chậm)

■ LIME, SHAP

Đặc điểm nào của mã độc khiến nó bị phát hiện. Cung cấp giải thích cục bộ và giải thích tổng thể

MÔ HÌNH TRIỂN KHAI



Workflow được áp dụng phân loại cho hai trường hợp:

- Phân loại ransomware với benign
- Phân loại ransomware với malware

Gồm các thành phần chính:

- Cuckoo Sandbox
- Features Extraction
- Combined Sequence
- 2-Layered CNN Model
- LIME - Local Explanation
- SHAP - Global Explanation



SANDBOX CUCKOO & EXTRACT FEATURES

```
1  {
2      "dlls": [
3          "ADVAPI32.dll",
4          "SHLWAPI.dll",
5          "C:\\Windows\\system32\\IMM32.DLL",
6          "gdi32.dll"
7      ],
8      "apis": [
9          "GetSystemTimeAsFileTime",
10         "GetSystemTimeAsFileTime",
11         "NtCreateMutant",
12         "GetSystemTimeAsFileTime",
13         "NtOpenKeyEx",
14         "NtQueryKey",
15         "NtOpenKeyEx",
16         "LdrLoadDll",
17         "LdrGetProcedureAddress",
18         "RegOpenKeyExW"
19     ],
20     "mutexes": []
21 }
```

05

Thực thi mẫu ransomware trong hệ thống máy ảo an toàn (sandbox) của Cuckoo để ghi lại hành vi động như:

- API calls được sử dụng.
- DLLs được nạp.
- Mutexes được tạo.
=> Xuất report dạng JSON để xử lý tiếp.

Extract Features :

- Từ báo cáo JSON, trích xuất 3 chuỗi hành vi có thứ tự là API Calls, DLLs, Mutexes ghép thành vector đặc trưng hành vi.

Ví dụ: API1 || API2 || DLL1 || DLL2 || Mutex1 || Mutex2

=> Tạo Vector được nhúng và đưa vào mô hình CNN để phân loại ransomware.

DEEP LEARNING & XAI (LIME, SHAP)

■ Deep Learning

Model : 2-layer 1D-CNN, phân loại binary classification
CNN gồm 4 lớp chính: Convolutional Layer, Activate Layer (Sigmoid), Pooling Layer (Max Pooling), Full-Connected Layer.

Input : Chuỗi vector đặc trưng từ API Calls + DLL + Mutex
(do Cuckoo trích xuất)

Trainning : sparse categorical cross-entropy, có tập validation để tránh overfitting

Mục tiêu đề ra

1. Ransomware vs. Benign
2. Ransomware vs. Malware



Ý nghĩa: Phân loại tách biệt trong từng ngữ cảnh, hiểu rõ sức mạnh, từ đó nâng cao khả năng áp dụng trong thực tế.

■ XAI

LIME (Local Interpretable Model-agnostic Explanations)

- Giải thích **mỗi dự đoán** cụ thể của mô hình.
- Xác định feature nào quan trọng nhất trong một mẫu.

SHAP (SHapley Additive exPlanations)

- Cung cấp giải thích **toàn cục** cho mô hình.
- Tính mức độ ảnh hưởng của mỗi đặc trưng đến toàn bộ dự đoán.

DATASET



Nhóm đã tổng hợp, **sàn lọc** và sử dụng từ nhiều nguồn dataset làm phong phú kết quả thực nghiệm, gồm :

■ Ransomware: 511 files - 95 families (LockBit, Conti, REvil, Cerber, WannaCry...)

MarauderMap <https://github.com/THUWingTecher/MarauderMap/tree/main/ransomware-samples>

Bazaar <https://bazaar.abuse.ch/>

■ Malware: 591 files

AnyRun <https://any.run/>

theZoo <https://github.com/ytisf/theZoo>

SOREL-20M <https://github.com/sophos/SOREL-20M>

■ Benign: 507 files

Benign-NET

<https://github.com/bormaa/Benign-NET>

07



THU-WingTecher/MarauderMap

[ICSE'24] Latest 7,796 active and unique ransomware samples from 95 families. Code is in another repository. 勒索软件数据破坏攻击分析

MarauderMap/ransomware-samples at main · THU-WingTecher/MarauderMap

[ICSE'24] Latest 7,796 active and unique ransomware samples from 95 families. Code is in another repository. 勒索软件数据破坏攻击分析 - THU-WingTecher/MarauderMap



bormaa/Benign-NET

Benign .NET files

bormaa/Benign-NET: Benign .NET files

Benign .NET files. Contribute to bormaa/Benign-NET development by creating an account on GitHub.

[GitHub](#)



EXPERIMENT & RESULT

- Thực hiện chạy mô hình trên máy 11th Gen Intel(R) Core(TM) i7-11390H, 3.4GHz, 16GB RAM. Triển khai một máy guest VM Windows 7 trên Cuckoo Sandbox để thu thập dữ liệu API calls, Dlls, Mutexes
- Trung bình mỗi report Cuckoo cho ra kết quả 10471.9 API calls, 11.1 DLLs và 10.7 Mutexes.
- Số lượng API calls, Dlls, Mutexes cần thiết để mô hình xử lý tốt là 500 API calls, 10 DLLs và 10 Mutexes
- Số lượng Epoch cần thiết để mô hình deep learning đạt hiệu suất tốt là 10 epoch.
- Thực hiện quá trình training, validation và testing trên bộ dữ liệu đã thu thập cho mô hình 2L-CNN và thử nghiệm trên các mô hình machine truyền thống như Decision Tree, Random Forest và mô hình CNN 1 lớp để kiểm tra tính hiệu quả của mô hình.
- Xây dựng website demo ở phía end-user việc kiểm tra file ransomware





EXPERIMENT & RESULT

Table 1
Comparison with existing methods (ransomware/benign)

| Model | Accuracy | TPR | FPR | F-Score |
|---------------|----------|--------|--------|---------|
| Decision Tree | 0.9820 | 0.9608 | 0.0000 | 0.9800 |
| Random Forest | 0.9900 | 0.9783 | 0.0000 | 0.9890 |
| CNN | 0.9820 | 0.9804 | 0.0167 | 0.9804 |
| 2L-CNN | 0.9910 | 0.9804 | 0.0000 | 0.9901 |

Kết quả nổi bật:

- **2L-CNN** đạt Accuracy cao nhất: 99.10%, thể hiện khả năng phân loại chính xác vượt trội.
- TPR (True Positive Rate) cũng đạt 98.04%, tương đương hoặc cao hơn các mô hình khác.
- FPR (False Positive Rate) = 0.00% – nghĩa là không có mẫu benign nào bị nhầm thành ransomware.
- F-Score của 2L-CNN là 0.9901, cao nhất trong tất cả mô hình, cho thấy sự cân bằng giữa độ chính xác và khả năng phát hiện ransomware.





EXPERIMENT & RESULT

Table 2
Comparison with existing methods (ransomware/malware)

| Model | Accuracy | TPR | FPR | F-Score |
|---------------|----------|--------|--------|---------|
| Decision Tree | 0.9118 | 0.9412 | 0.1176 | 0.9143 |
| Random Forest | 0.9412 | 0.9608 | 0.0784 | 0.9423 |
| CNN | 0.9216 | 0.9412 | 0.0980 | 0.9231 |
| 2L-CNN | 0.9412 | 0.9412 | 0.0588 | 0.9412 |

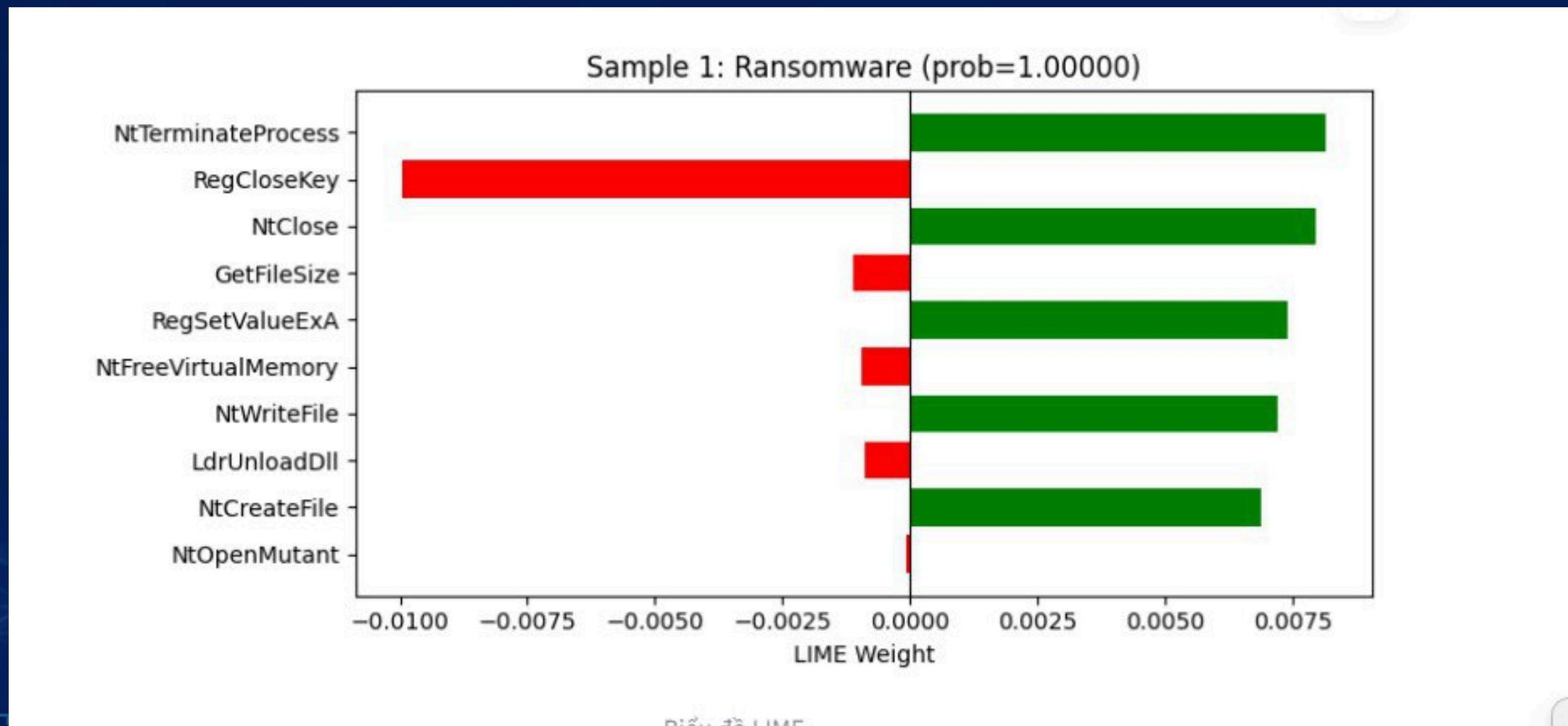
Kết quả nổi bật:

- Accuracy:
2L-CNN và Random Forest cùng đạt 94.12%, cao nhất bảng, thể hiện khả năng phân loại hiệu quả giữa ransomware và các loại malware
- TPR :
Random Forest cao nhất với 96.08% → phát hiện được phần lớn ransomware.
2L-CNN và các mô hình khác đạt 94.12%, vẫn ở mức cao và ổn định.
- FPR
2L-CNN có FPR thấp nhất: 0.0588, tức là ít nhầm malware thông thường thành ransomware nhất, giúp giảm dương tính giả .
- F-Score:
Random Forest dẫn đầu nhẹ (0.9423), nhưng 2L-CNN theo sát (0.9412), cho thấy tính ổn định và hiệu quả trong phân loại.





EXPERIMENT & RESULT - LIME



Các LIME feature đóng góp vào quyết định cho Sample 1

Trong hình là một mẫu được mô hình đánh giá là Ransomware với tỉ lệ là 100%

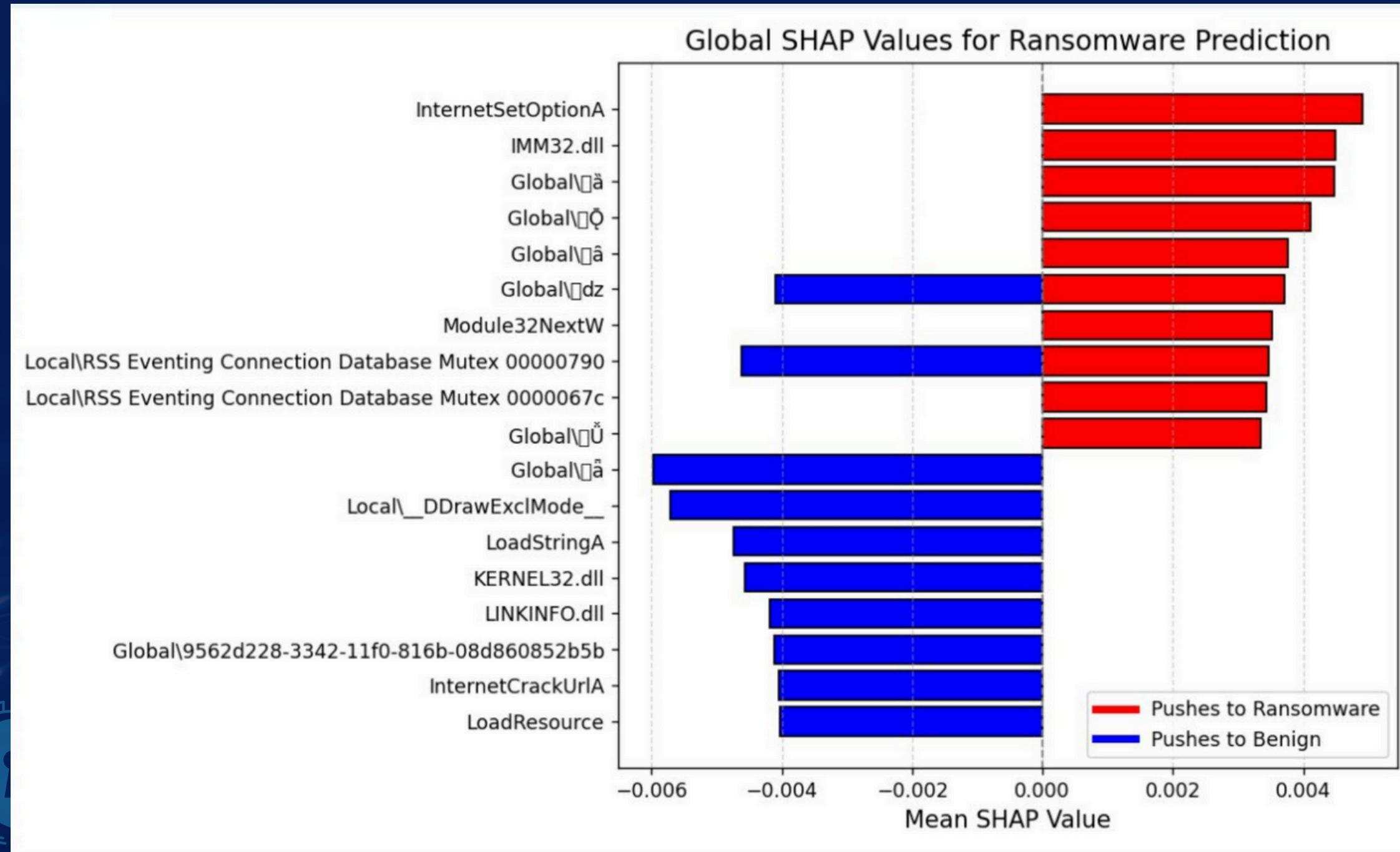
- Các API thuộc các hàng màu xanh như NtAllocateVirtualMemory, NtTerminateProcess, NtClose... hỗ trợ cho quyết định Ransomware của mô hình.
- Các API có giá trị âm là những API màu thuẫn với quyết định của mô hình.

Hình trên chỉ đưa ra top 5 API có tỉ lệ màu thuẫn cao nhất và hỗ trợ cao nhất. Việc quyết định ransomware phải dựa vào tổng lượng của các features, tần suất sử dụng của một thuộc tính...

Một số feature nổi bật khác ủng hộ quyết định ransomware: ZonesCounterMutex, ZonesLockedCacheCounterMutex, NtAllocateVirtualMemory, Advapi.dll



EXPERIMENT & RESULT - SHAP



SHAP feature đánh giá là ransomware theo feature

Ngoài giải thích cục bộ, thì hệ thống sử dụng SHAP để diễn giải tổng thể mô hình dựa trên đặc trưng của các features.

Trong hình bên, các feature màu đỏ sẽ cung cấp quyết định ransomware. Các feature màu xanh thì chống lại quyết định ransomware.
Mô hình quyết định còn dựa trên mối quan hệ của các feature với nhau.

TESTING WEBSITE FOR USER



Malware Analyzer

Chọn loại phân tích:

Ransomware and Benign

Chọn kiểu file đầu vào:

attribute

Tải lên file

Drag and drop file here
Limit 200MB per file • JSON

Browse files

extract_report_24dd41444b7367166ad4cfca3441dd1303fcb5aabcb8f226ac47719... 2.7KB X

Phân tích

Phân tích thành công!

Nhãn dự đoán: Ransomware

Độ tin cậy: 1.00

Website thực hiện chức năng kiểm tra một file có phải ransomware/benign hay ransomware/malware

- Upload File
- Prediction
- Giải thích từ mô hình LIME



DEMO



**THANK YOU FOR
YOUR ATTENTION**