Enhancing Malware Detection and Family Identification with Machine Learning Models and Explainable AI

Project Members:

Souptik Dutta (202002021001)
Jyoti Bikash Choudhury (202002022101)
Ritik Raj (202002022097)

Under the Guidance of

Dr. Amitava Nag, Professor, Dept. of CSE

Dept.of Computer Science and Engineering
CENTRAL INSTITUTE OF TECHNOLOGY KOKRAJHAR

Table of Contents

Introduction

Cybersecurity and Cyber Attacks

- Cybersecurity is important in protecting sensitive information and ensuring the integrity of systems
- Cyber attacks are becoming increasingly sophisticated, posing significant threats to individuals, businesses, and governments worldwide.

Malware

- Malware, short for malicious software, refers to any software intentionally designed to cause damage to a computer, server, client, or computer network
- Types of malware include viruses, worms, trojans and ransomware

Introduction

Dangers of Malware

- In 2023, it was reported that over 560,000 new pieces of malware are detected every day
- The global cost of cybercrime, driven significantly by malware, is projected to reach $10.5 trillion annually by 2025
- Malware is a primary cause of data breaches, with over 70% of organizations experiencing at least one malware-related breach in the past year

Research Gap of ML Analysis

- Still a significant gap in leveraging machine learning techniques for accurate and real-time malware detection and analysis, necessitating further research and development in this domain

Objective & Motivation

- To implement suitable feature selection and classification methods for effective malware analysis.
- To analyse model explainability and accountability through XAI methods.
- To effectively cluster malicioius samples and identify their respective families through clustering methods.

Related work

Ref.	Year	Contribution
[4]	2024	Developed a CNN model for ransomware detection
[6]	2024	Malware classification as a graph classification problem approach utilizing function call graphs
[7]	2023	XAI approaches for intrusion detection in IoT network
[1]	2022	Utilization of Boruta FS algorithm to extract best feature set
[2]	2022	XGB based two-stage pipeline approach for intrusion detection
[3]	2021	A unified form of clustering combining K-Means and FCM
[8]	2020	Comparative study of effective cluster size distribution of K-Means and FCM

Table: Related Work Reference Summary

Methodology

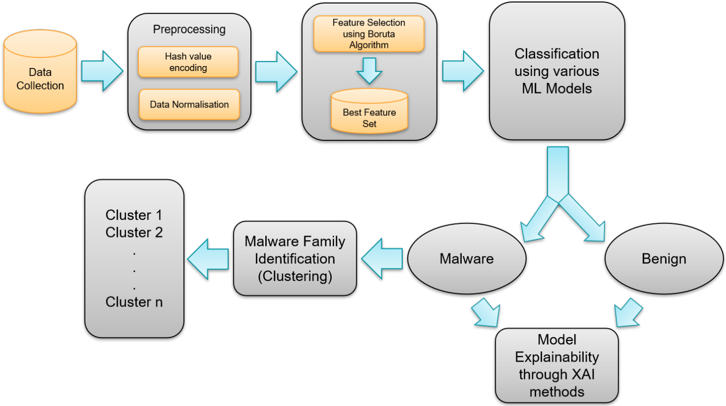


Figure: Proposed Architecture

Dataset

- Dataset has been taken from Practical Security Analytics LLC[5]
- Our dataset has **201549** Windows Portable Executable samples (eg: .exe, .dll,.scr)
- It comprises of **86812 benignware** and **114737 malware** samples
- Contains hash values, which were preprocessed

Field	Description
id	The identifier for the sample that corresponds to the name of the file in the samples directory.
Md5	The MD5 hash of the file.
Sha1	The SHA1 hash of the file.
Sha256	The SHA256 of the file.
total	The total number of antivirus engines that scan this file at the time of the query.
positive	The number of antivirus engines that flag this file malicious at the time of the query.
list	Either blacklist or whitelist indicating whether or not the file is malicious or legitimate respectively.
filetype	This field will always be exe for this data set.
submitted	The data that the sample was entered into my database
User_id	Redacted
Length	The length of the file in bytes.
entropy	The Shannon entropy of the file. The values will range from 0 to 8.

Hash Encoding

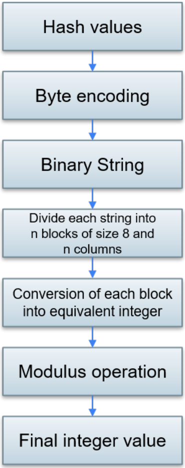


Figure: Proposed Hash-encoding algorithm

Taking a hash value of 'N' bits for instance, say ad27. . . .7ab8

- Converted to binary string by byte encoding
- String divided into 'n' blocks given by $n = N/32$, also 'n' new columns added
- Each string converted to an integer
- Modulus Operation - dividing each integer by the smallest 4-digit prime number, 1009
- Each integer added to the new columns

So, augmented columns for each hash type:

4 for MD5: md5_int1 , ... , md5_int4

5 for SHA-1: sha1_int1, ... , sha1_int5

8 for SHA-256: sha256_int1, ... , sha256_int8

Feature Selection

- **Dimensionality Reduction:** Simplifies models by eliminating irrelevant features
- **Improves Performance:** Enhances accuracy and speed of machine learning algorithms
- **Prevents Overfitting:** Reduces model complexity to improve generalization
- **Interpretability:** Makes models easier to understand and interpret

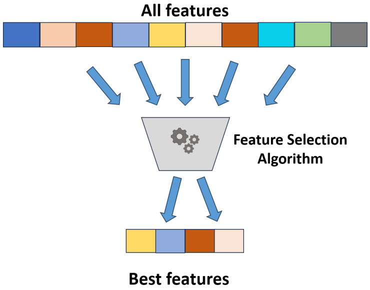
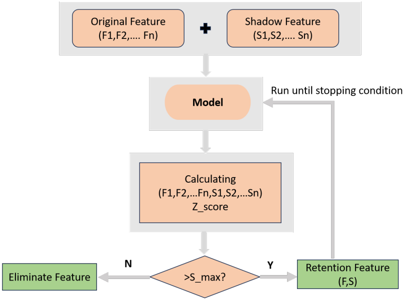


Figure: Feature Selection process

Workflow of Boruta Feature Selection Algorithm

Boruta is wrapper approach of feature selection built around an RF or XGB classifier

- Creates shadow features by creating duplicate features & shuffling the values in each column.
- Model trained on both the original and shadow features.
- Calculates & compares Z_scores of the original features to the shadow features.
- If an original feature is significantly more important than its shadow counterpart, it is considered a relevant feature and retained, or else rejected.



Applying Boruta FS on our dataset gave three best features: **length, entropy and sha1_int3** columns

Machine Learning

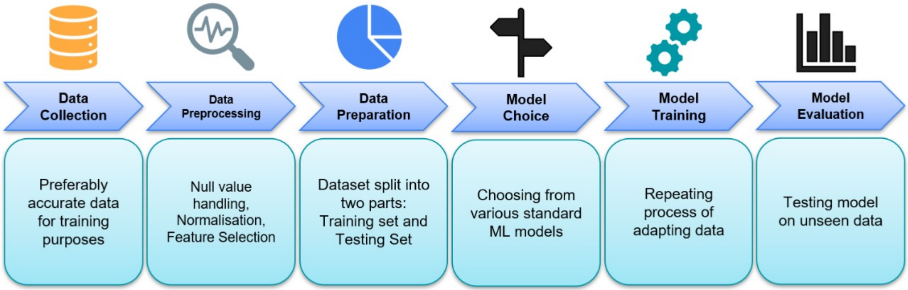
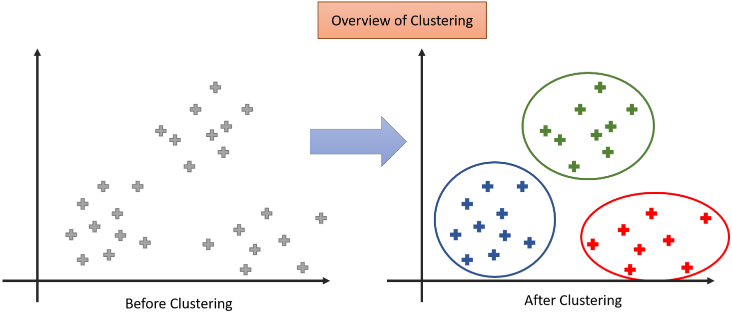


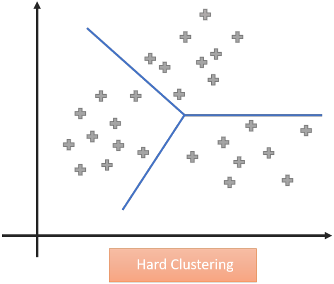
Figure: Workflow of an ML process

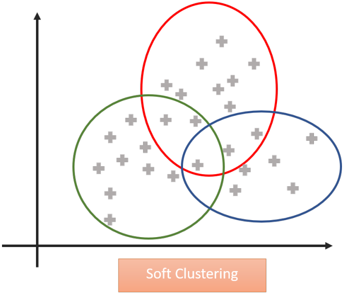
Clustering



- Unsupervised learning method, works with unlabeled data
- Tries to group or "cluster" similar items together
- Used in our study for malware family identification

Clustering Methods





Fuzzy C-Means	**K-Means**
Each data point is assigned a degree of membership to each cluster, indicating the probability or likelihood of the point belonging to each cluster	Each data point is exclusively assigned to one and only one cluster, based on the closest centroid, typically determined using Euclidean distance.
It does not impose any constraints on the shape or variance of clusters. It can handle clusters of different shapes and sizes, making it more flexible	It assumes that clusters are spherical and have equal variance. Thus it may not perform well with clusters of non-spherical shapes or varying sizes.
It is less sensitive to noise and outliers as it allows for soft, probabilistic cluster assignments.	It is sensitive to noise and outliers in the data

Table: Differences between Fuzzy C-Means and K-Means Clustering

Explainable AI & Model Explainability

- The primary goal of XAI is to create AI models whose decisions and predictions can be understood and interpreted by humans
- XAI is essential for enhancing "black box" models, which are frequently opaque and intricate
- Utility of post-hoc approach like SHAP was leveraged in this study

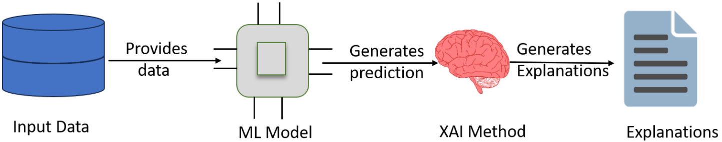
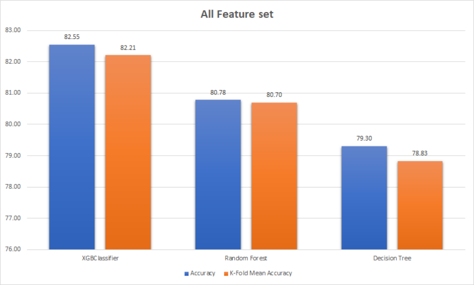
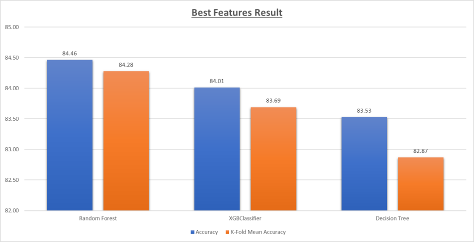


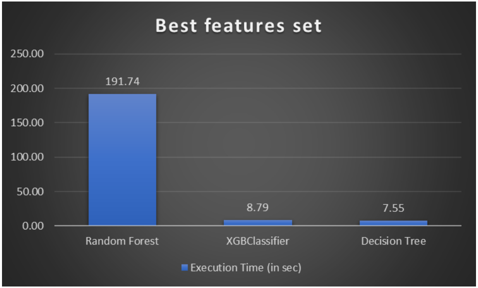
Figure: Overview of XAI

Results and Execution Time of All Features Set

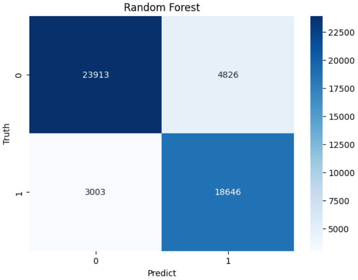


Results and Execution Time of Best Features Set

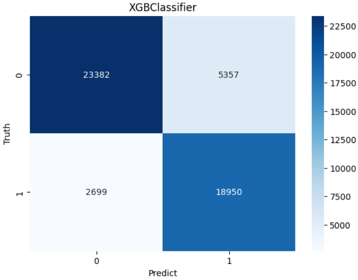




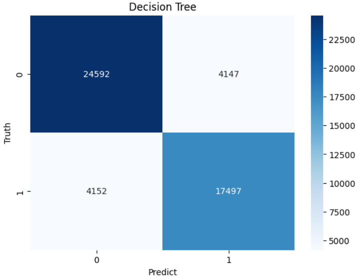
Confusion Matrices of Best Features Set



(a) Confusion Matrix of RF

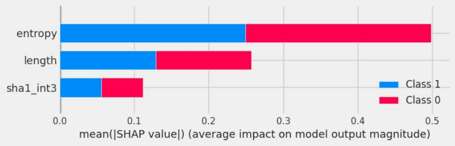


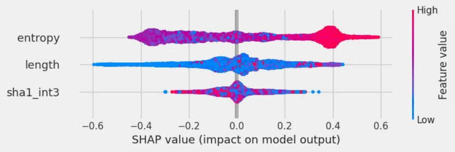
(b) Confusion Matrix of XGB



(c) Confusion Matrix of DT

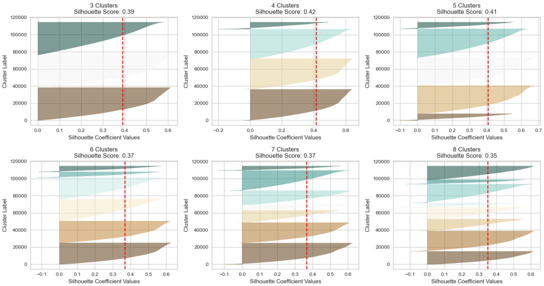
Model Explainability using SHAP

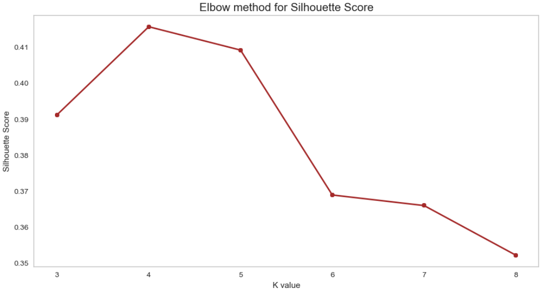




- **Entropy**: Significant impact on Malware class, indicating malware files have higher entropy.
- **Length of file**: Majorly impacts the benign ware class, indicating longer files may be linked with benign files.
- **Sha1_int3**: Has a balanced effect on both the classes.

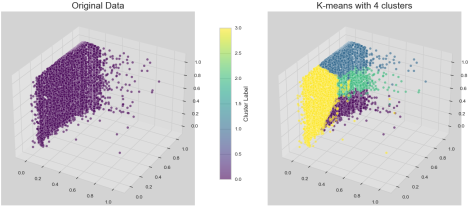
Clustering Analysis



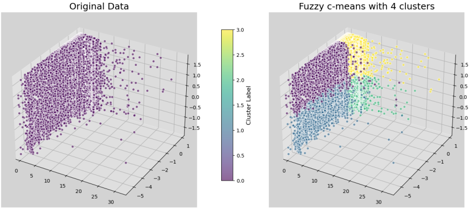


- **Silhouette analysis and Elbow analysis performed on the malware samples**
- **Four families of malware identified**

Clustering Results



(a) For K-Means clustering, silhouette score = 0.42



(b) For Fuzzy C-Means clustering, silhouette score = 0.48

Conclusion & Future scope

- **Real-time Detection Systems:** Developing and deploying real-time malware detection systems
- **Deploy Deep Learning methods:** Investigating the use of deep learning architectures, such as CNNs and RNNs

References I

[1] Hala Ahmed, Hassan Soliman, and Mohammed Elmogy. "Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree". In: *Computers in Biology and Medicine* 146 (2022), p. 105622.

[2] Pieter Barnard, Nicola Marchetti, and Luiz A DaSilva. "Robust network intrusion detection through explainable artificial intelligence (XAI)". In: *IEEE Networking Letters* 4.3 (2022), pp. 167–171.

[3] Ioan-Daniel Borlea et al. "A unified form of fuzzy C-means and K-means algorithms and its partitional implementation". In: *Knowledge-Based Systems* 214 (2021), p. 106731.

[4] Sibel Gulmez, Arzu Gorgulu Kakisim, and Ibrahim Sogukpinar. "XRan: Explainable deep learning-based ransomware detection using dynamic analysis". In: *Computers & Security* 139 (2024), p. 103703.

References II

[5] Michael Lester. "PE Malware Machine Learning Dataset". In: (2021). URL: https://practicalsecurityanalytics.com/pe-malware-machine-learning-dataset.

[6] Vrinda Malhotra, Katerina Potika, and Mark Stamp. "A comparison of graph neural networks for malware classification". In: *Journal of Computer Virology and Hacking Techniques* 20.1 (2024), pp. 53–69.

[7] Nour Moustafa et al. "Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions". In: *IEEE Communications Surveys & Tutorials* (2023).

[8] Kaile Zhou and Shanlin Yang. "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering". In: *Pattern Analysis and Applications* 23.1 (2020), pp. 455–466.

Thank You