# PROJECT REPORT II : CIC-Evasive-PDFMal2022

**Authors: Group 5**

Himanshu Shekhar (220454)
Lokesh Yadav (220594)
Kamal Kant Tripathi (241110086)

November 7, 2024

# 1 Introduction

CIC-Evasive-PDFMal2022 is a notable variant of malware that specifically targets PDF documents to evade detection by security systems. This type of malware typically utilizes sophisticated techniques to conceal its malicious payload, such as embedding harmful scripts within seemingly benign PDF files. Its evasive capabilities pose significant challenges for cybersecurity professionals, as traditional detection methods often struggle to identify these hidden threats.

In this project, a thorough analysis of the existing dataset and model has been carried out to identify the limitations and finding the innovative ways to overcome the same. The details have been explained in subsequent paragraphs.

The **GitHub repository** can be found **here**

# 2 Dataset Analysis

In addition to the existing dataset from CIC Website, we have added a diverse dataset obtained from VirusShare, GOVDocs, PDFRep, VirusTotal etc. The dataset contains equal proportion of malicious and benign PDF files making a balanced and robust set of data to test and validate our improved version of Model. The Dataset analysis is illustrated below:

## 2.1 Data Collection

In the process of data collection, multiple sources were explored to collect more than 1 lakh benign and approx 60 thousand malicious PDF files whereas the existing dataset were using only 9 and 11 thousand files respectively. **Hence, we have used an enhanced dataset which is a novel idea in this project.**

## 2.2 Data pre-Processing

Feature extraction was carried out from these files using the python script developed by the team. After dropping the duplicate records, a balanced proportion of 30 thousand records each of malicious and benign were selected to incorporate in the final CSV file. The details will be covered during the presentation.

## 2.3 Feature Selection

The existing work had taken into account only 31 features. However, we build upon that judiciously and explored a total of 61 features and finally selected **TO FILL** most important and effective features after due deliberation. **The enhanced feature set has contributed towards a robust model and better accuracy than the existing model and hence is a second novel idea in this project**.

## 2.4 Implementing SHAP (SHapley Additive exPlanations)

**SHAP measures each feature's contribution to the model's output, is the third novel idea in this project** and has been incorporated for feature selection process. The details will be covered during our presentation.

# 3 ML Model

The various independent models were trained and tested, as well as various combinations of different models were tried for stacking model on the dataset prepared by us rigorously. Highlights of this exhaustive process are enumerated in subsequent paragraphs:

## 3.1 The stacking model

The stacking model was dropped due to very high execution time and no significant improvement in overall accuracy.

## 3.2 The XGBoost Classifier

The XGBoost classifier performed exceedingly well for all combination of features and also on the **MalwareBazaar** dataset, which is the most updated set of malicious data available on open source.

## 3.3 Optuna

We have used Optuna, an open source library, for hyperparameter optimization framework which led to some increase in the overall accuracy.

# 4 Comparative Analysis and Results

The comparative analysis between the existing work in the selected research paper and our work was carried out and following are the highlights of the same:

## 4.1 Time taken on Enhanced and Diverse Dataset

Our model has performed exceedingly well in terms of time than the existing model on enhanced and diverse dataset prepared by us.

## 4.2 Accuracy on most Updated Open Source Dataset

Our model has achieved better accuracy than the existing model when tested on most updated open source dataset from MalwareBazaar as well as the original dataset.

# 5 Conclusion

All the things proposed in the first deliverable of the project submitted on 07 Oct 24 have been complied with and successfully implemented in this deliverable. In addition, SHAP has been implemented for feature engineering.