

# MALICIOUS PDF DETECTION BASED ON MACHINE LEARNING WITH OPTIMAL FEATURES USING AN UP-TO-DATE DATASET

**Group Number: 5**

## **Group Members**

- Himanshu Shekhar (220454)
- Lokesh Yadav (220594)
- Prabhat Mishra (220775)
- Kamal Kant Tripathi (241110086)

## **Problem Statement**

PDF is one of the most popular document file formats due to its flexibility, platform independence, and ability to embed different types of content. Over the years, PDF has become a popular attack vector for spreading malware and compromising computer systems. Existing signature-based defense systems have extremely high recall rates but quickly become obsolete and ineffective against zero-day attacks, making them easy to circumvent by malicious PDF files.

Recently, Machine Learning (ML) has emerged as a viable tool to improve the discovery of previously unseen attacks. Hence, in this project, we present enhanced ML-based models for the detection of malicious PDF documents. We develop an approach for ML-based detection with static features derived from PDF documents leveraging existing tools and optimize the feature set, which may include new, previously unused features to enhance the performance of the ML-based classifiers.





Our investigative study will be conducted on the published **CIC-Evasive-PDFMal2022** dataset. We will also endeavor to update the **CIC-Evasive-PDFMal2022** dataset by including data published post-2022 till date. We also intend on using new types of attacks that have been introduced and testing this extended dataset on existing models and try to create a new model that would be up-to-date with this extended dataset. We also plan to test out the existing feature encoding schemes on this extended dataset and explore the possibility of any improvement in the encoding scheme.

## **Dataset**

For the dataset, we intend on using the **CIC-Evasive-PDFMal2022** dataset along with other malicious PDFs found from MalwareBazaar and datasets from other external sources (like the reverse mimicry attack dataset). The **CIC-Evasive-PDFMal2022** dataset consists of 4468 benign samples and 5557 malicious PDF samples. From the other sources we will use malicious data with the tag of pdf, js/exe embeddings etc.

# Methodology

## Reference Papers

- **Explainable Ensemble Learning Based Detection of Evasive Malicious PDF Documents** : One area where we can improve upon this paper is by extending the system with content-based features. We can also work on improving the resilience of other types of existing PDF detection systems to incorporate more resilience against reverse mimicry attacks.
- **A Novel Feature Encoding Scheme for Machine Learning Based Malware Detection Systems** : We plan to validate the proposed entropy-based encoding scheme on additional datasets and test it on more complex machine learning models, including deep learning architectures. We also aim to test the feature encoding scheme on the extended dataset and explore the possibility of any improvement.
- **Malware Detection in Portable Document Format (PDF) Files with Byte Frequency Distribution (BFD) and Support Vector Machine (SVM)** : We will test this method on our updated dataset and then explore the feasibility of improving the Sequential Forward Selection method for feature selection and will try to optimize the existing accuracy achieved by Decision Tree, Naïve Bayes, Support Vector Machines and Random Forest.
- **PDF Malware Detection: Toward Machine Learning Modeling With Explainability Analysis** : We will use the noise elimination techniques from this paper and apply the SHAP technique mentioned in this on our new dataset trying to increase its robustness.

## Proposed ML Techniques

Our first approach would be to create an ensemble or a stacking architecture based on common classifiers as the base models. For further analysis, we will also build a Neural Network and compare the results from both our models.

## Proposed ML Libraries

- We will use **scikit-learn** for creating the ensemble or stacking architecture and the base model.
- We will also implement a Neural Network for classification and compare it with the previous model; for this, we will use **TensorFlow** or **PyTorch**.

## Team Members ML Experience

- **Himanshu Shekhar** - Has done the course “Intro to Machine Learning” and many projects involving ML, therefore has decent hands-on experience with ML.
- **Lokesh Yadav** - Has done the course “Machine Learning Specialization” on Coursera, therefore has some hands-on experience with ML.
- **Prabhat Mishra** - Has Done a Project on “Stock Prediction” where I have used LSTM Model and TensorFlow library, therefore has some hands-on experience with ML.
- **Kamal Kant Tripathi** - The member is a beginner in the ML field with no prior hands-on experience in ML. However, he has undertaken the courses on “Intro to Machine Learning” and is looking forward to availing this opportunity to gain hands-on experience in ML.

## Deliverables

- **October 7, 2024:** The extended dataset using data post-2022 till data and an ML model that can classify benign and malicious PDFs. Improvement in the feature set for this new data will be explored and reported. Code, accuracy and other metrics of the model and a write up will be given to the mentor.
- **November 7, 2024:** Test our extended dataset on the already existing models and compare our models accuracy with theirs. A report on our attempt to improve upon the already existing feature encoding scheme and other detection/pre-processing methods using the new dataset.
- **Final Result:** A robust ML model that is trained on an up-to-date dataset using innovative features and techniques, capable of detection all kinds of malicious PDFs.