PDF Malware Detection: Toward Machine Learning Modeling with Explainability Analysis

Kilaru Pradeepthi[1], Ravi Sri Lakshmi[2], Boddu Vaishnavi [3], Vadugu Narendra Babu[4], Tatineni Vijayasree[5]
[1, 2, 3, 4, 5] *Department of CSE (AI&ML), SRK Institute of Technology, Vijayawada, A.P., India*

Abstract- ***In the digital age, PDF files are widely used for document sharing, but their popularity also makes them a target for malware attacks. This project, titled "PDF Malware Detection: Toward Machine Learning Modeling With Explainability Analysis," aims to develop and evaluate machine learning models for detecting malware in PDF files. Utilizing a dataset from Kaggle, which contains labeled examples of malicious and benign PDFs, various algorithms including Random Forest, C5.0, J48, Support Vector Machine (SVM), AdaBoost, Deep Neural Network (DNN), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN) will be applied. The primary focus is on achieving high detection accuracy while also providing explainability to understand the decision-making process of the models. By leveraging machine learning techniques, this project seeks to enhance cybersecurity measures, offering a robust solution to identify and mitigate potential threats embedded in PDF documents.***

Keywords: PDF malware detection, machine learning, Random Forest, SVM, DNN, explainability, cybersecurity, malicious PDF, classification algorithms, Kaggle dataset.

INTRODUCTION

The increasing reliance on PDF files for document sharing has rendered them a popular medium for both legitimate and malicious activities. While PDFs offer convenience and versatility, their widespread use has made them a prime target for cybercriminals looking to exploit vulnerabilities. Malware embedded within PDF documents poses significant risks to individuals and organizations alike, leading to data breaches, identity theft, and financial loss. As such, effective detection mechanisms are critical to safeguard against these threats.

In response to the growing need for robust security measures, this project explores the application of machine learning techniques to detect malware in PDF files. By utilizing a comprehensive dataset from Kaggle that includes both malicious and benign PDFs, we aim to develop and evaluate various classification algorithms. Techniques such as Random Forest, C5.0, J48, Support Vector Machine (SVM), AdaBoost, Deep Neural Network (DNN), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN) will be employed to identify malware with high accuracy.

A key focus of this research is not only on achieving superior detection performance but also on ensuring that the models provide explainable insights into their decision-making processes. Understanding the rationale behind the models' predictions is essential for cybersecurity

practitioners, as it enables them to trust and verify the outcomes of automated systems. This project aims to bridge the gap between high-performance machine learning models and the need for transparency in their operations, ultimately contributing to enhanced cybersecurity frameworks for PDF document handling. By leveraging advanced algorithms and prioritizing explainability, this study aspires to offer effective solutions for detecting and mitigating potential threats posed by malicious PDF files.

LITERATURE SURVEY

[1] S. S. Alshamrani, "Design and analysis of machine learning based technique for malware identification and classification of portable document format files," Secur. Commun. Netw., vol. 2022, pp. 1–10, Sep. 2022. The paper "Design and Analysis of Machine Learning Based Technique for Malware Identification and Classification of Portable Document Format Files" introduces a novel PDF malware classification system using machine learning. Unlike traditional antivirus software, this system inspects PDF files both statistically and dynamically, enhancing its ability to detect obscure and zero-day malware. Five different classifiers were evaluated, with the Random Forest classifier achieving the best performance, yielding an

F1-measure of 0.986. The system's effectiveness surpasses existing methods, making it a robust solution for identifying malicious code hidden within PDF files.

[2] P. Singh, S. Tapaswi, and S. Gupta, ‘‘Malware detection in PDF and office documents: A survey,’’ Inf. Secur. J., Global Perspective, vol. 29, no. 3, pp. 134–153, May 2020.
This paper surveys the growing threat of malware embedded in PDF and Office documents, which has surged in the past five years alongside increasingly sophisticated attacks.
The flexibility of these document formats, with their numerous exploitable features, makes them prime targets for cybercriminals. Despite ongoing efforts from both industry and academia,
malicious documents remain a significant security challenge. The paper provides a comprehensive classification of document-based attacks, explores the structures of PDF and Office file formats,
and reviews current tools and research in automatic malware detection methods.

[3] N. Livathinos, C. Berrospi, M. Lysak, V. Kuropiatnyk, A. Nassar, A. Carvalho, M. Dolfi, C. Auer, K. Dinkla, and P. Staar, ‘‘Robust PDF document conversion using recurrent neural networks,’’ in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 17, 2021, pp. 15137–15145. The paper "Robust PDF Document Conversion Using Recurrent Neural Networks" presents a novel approach to recovering document structure from PDF files using recurrent neural networks (RNNs). Unlike traditional visual methods, this technique processes low-level PDF printing commands directly, enabling more precise identification of structural components such as titles, sections, and tables. The method is computationally efficient, memory-saving, and able to handle text flow across pages naturally. With a weighted average F1 score of 97% across 17 labels, the model significantly enhances document conversion, particularly in large-scale applications like COVID-19 article analysis.

[4] Q. A. Al-Haija, A. Odeh, and H. Qattous, ‘‘PDF malware detection based on optimizable decision trees,’’ Electronics, vol. 11, no. 19, p. 3142, Sep. 2022. The paper titled "PDF Malware Detection Based on Optimizable Decision Trees" presents a novel system for detecting malicious PDF files by distinguishing them from benign ones. The system employs an AdaBoost decision tree with optimal hyperparameters, trained on the Evasive-PDFMal2022 dataset. It demonstrates high efficiency, achieving a 98.84% prediction accuracy with a quick prediction interval of 2.174 μSec. The model's performance surpasses other state-of-the-art approaches in the field, offering a lightweight and accurate solution for uncovering PDF malware, making it a valuable tool for enhancing cybersecurity with minimal detection overhead.

[5] M. Abdelsalam, M. Gupta, and S. Mittal, ‘‘Artificial intelligence assisted malware analysis,’’ in Proc. ACM Workshop Secure Trustworthy CyberPhys. Syst., Apr. 2021, pp. 75–77. The tutorial by Abdelsalam, Gupta, and Mittal (2021) offers a comprehensive review of current research and applications of Artificial Intelligence (AI) and Machine Learning (ML) in malware analysis. It covers three primary approaches: static, dynamic, and online malware analysis. The tutorial provides background information, key results, and practical insights. It also includes a hands-on session focused on applying ML algorithms for dynamic malware analysis in cloud Infrastructure as a Service (IaaS) environments, making it a valuable resource for those interested in enhancing their understanding and skills in AI-assisted cybersecurity.

[6] W. Wang, Y. Shang, Y. He, Y. Li, and J. Liu, ‘‘BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors,’’ Inf. Sci., vol. 511, pp. 284–296, Feb. 2020. BotMark is an automated model designed to detect botnets by combining flow-based and graph-based traffic analysis. Traditional methods relying solely on one type of analysis often miss sophisticated bots. BotMark addresses this by extracting 15 flow-based and 3 graph-based features, utilizing k-means for flow similarity, and Local Outlier Factor for graph anomalies. The model marks bots through an ensemble of similarity, stability, and anomaly scores. Tested on traffic from five botnets, including Mirai and Zeus, BotMark achieved a detection accuracy of 99.94%, outperforming individual detection methods.

EXISTING SYSTEM

Current systems for PDF malware detection largely rely on PDF files which are commonly used, making them a target for cybercriminals who embed malware to compromise systems. Traditional detection methods struggle to identify all threats due to their reliance on limited feature sets.

To improve detection, they created a dataset of 15,958 PDF samples, including benign, malicious, and evasive files. Using PDFiD, PDFINFO, and PDF-PARSER, they extracted key characteristics and derived new features that enhance malware classification.

They developed an optimized feature selection method, improving detection accuracy using the Random Forest classifier. Additionally, they built a decision tree model to provide explainable rules for better human interpretation.

Disadvantages of Existing System:

Evasive Malware Challenges: Highly sophisticated evasive PDFs (e.g., using deep obfuscation or runtime decryption) may still bypass detection.

Feature Extraction Dependence: The approach relies on PDF-specific tools (PDFiD, PDFINFO, PDF-PARSER) for feature extraction. If new malware variants exploit unknown features, detection accuracy may drop.

Limited Generalization: The dataset (15,958 samples) is comprehensive but may not cover all real-world malware scenarios.

Resource Intensive: Scanning and analyzing files can be resource-heavy, affecting system performance.

Inadequate Explainability: Traditional methods lack transparency in decision-making, making it difficult to understand why a file was flagged.

PROPOSED SYSTEM

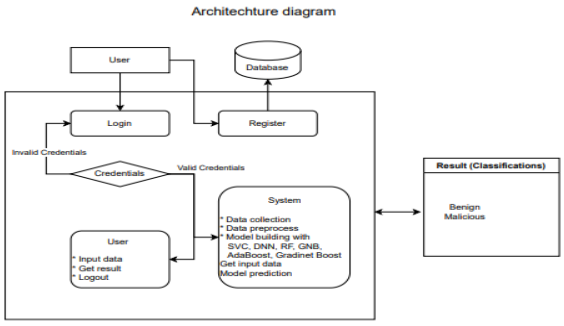

Fig: [1] Architecture

Proposed system algorithms:

KNN: K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for classification and regression. It identifies the 'k' closest data points to a given input and makes predictions based on majority voting or averaging. As a non-parametric and instance-based algorithm, KNN stores all training data and makes decisions at the time of prediction. The performance of KNN depends on the choice of 'k' and the distance metric, such as Euclidean or Manhattan. However, it can be computationally expensive for large datasets due to the need to calculate distances for all points.

SVM: Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression. It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. SVM uses kernel functions to handle non-linearly separable data by transforming it into a higher dimension. It is effective for complex datasets but can be computationally expensive for large datasets. SVM is widely used in applications like image classification, text categorization, and bioinformatics.

Random Forest: Random Forest is a supervised machine learning algorithm used for classification and regression. It operates by constructing multiple decision trees during training and combining their outputs to improve accuracy and reduce overfitting. The final prediction is made through majority voting (for classification) or averaging (for regression). Random Forest is robust to noise, handles missing values well, and works effectively with large datasets. It is widely used in applications like fraud detection, medical diagnosis, and recommendation systems.

Ada Boost: AdaBoost (Adaptive Boosting) is a machine learning ensemble technique that improves weak classifiers by combining multiple iterations to create a strong model. It assigns higher weights to misclassified samples, making the next weak classifiers focus more on difficult cases. The final prediction is based on the weighted sum of all weak learners. AdaBoost is widely used for classification tasks and works well with simple models like decision

stumps. However, it is sensitive to noisy data and outliers, which can impact performance.

Gradient Boosting Machine: Gradient Boosting Machine (GBM) is a powerful ensemble learning technique used for both classification and regression tasks. It builds models sequentially, where each new model corrects the errors of the previous one by minimizing a loss function using gradient descent. GBM combines multiple weak learners (typically decision trees) to create a strong predictive model. It uses boosting to improve performance, making it highly accurate but also prone to overfitting if not tuned properly. GBM is widely used in applications like finance, healthcare, and recommendation systems due to its efficiency and predictive power.

Deep Neural Networks: Deep Neural Networks (DNNs) are advanced machine learning models consisting of multiple layers of artificial neurons. They process data through an input layer, several hidden layers, and an output layer, enabling them to learn complex patterns and representations. DNNs use activation functions (e.g., ReLU, Sigmoid) and backpropagation with gradient descent to optimize weights. They excel in tasks like image recognition, natural language processing, and speech recognition. However, DNNs require large datasets, significant computational power, and careful tuning to avoid overfitting.

METHODOLOGIES

Data Collection: Gather a dataset containing PDF files labeled as benign or malicious. Extract relevant features like page numbers, XREF length, object counts, etc.

Data Preprocessing: Clean the dataset by handling missing values, normalizing feature values, and removing irrelevant data.

Data Splitting: Divide the dataset into training and testing sets for model evaluation.

Model Building: Train various machine learning models: Support Vector Classifier (SVC), Deep Neural Networks (DNN), Random Forest (RF), Gaussian Naive Bayes (GNB), AdaBoost, Gradient Boosting, K-Nearest Neighbors (KNN)

Prediction and Result: The selected ML model predicts whether the PDF is benign or malicious. The system displays results accordingly.

Advantages:

Enhanced Detection Accuracy: Utilizes multiple algorithms to improve detection rates and identify a wider range of malware.

Explainability: Provides transparency in decision-making, allowing users to understand the basis for malware classification.

Adaptability: Capable of detecting novel and evolving threats by leveraging machine learning models trained on diverse datasets.

Reduced False Positives: Advanced algorithms help minimize incorrect identifications of benign files as malicious.

RESULT

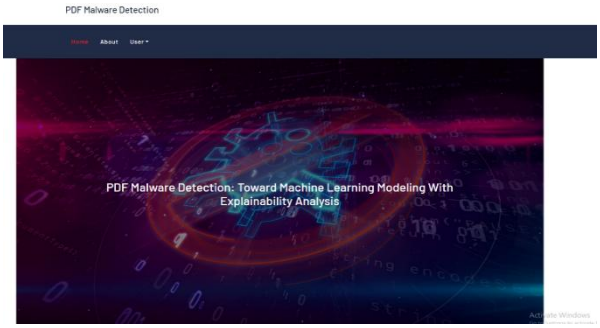

Fig: [2] Home Page

Fig: [3] About Page

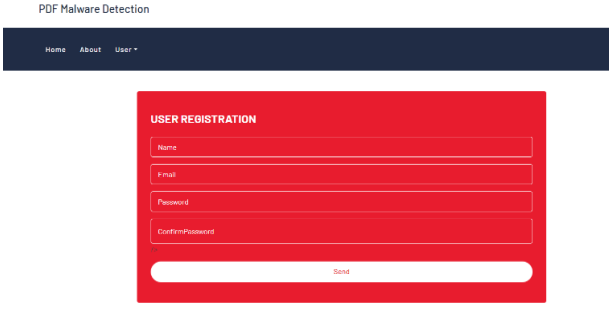

Fig: [4] Registration Page

Fig: [5] User Login Page

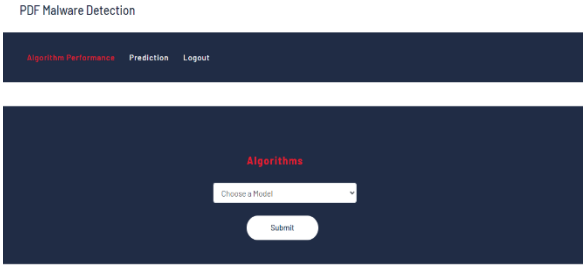

Fig: [6] Model Selection

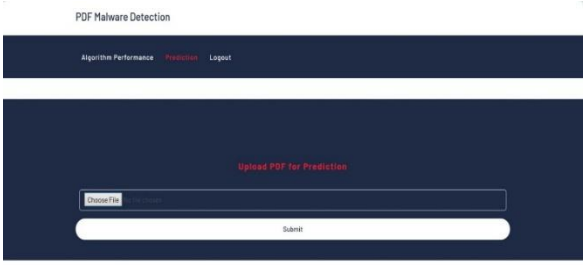

Fig: [7] Prediction Page

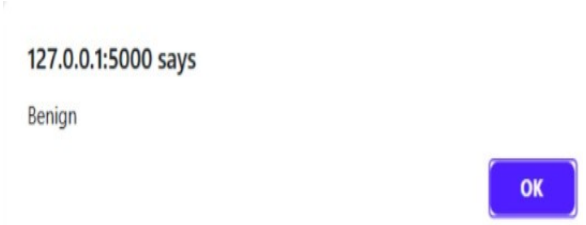

Fig: [8] Result

Fig: [9] Result

CONCLUSION

The project "PDF Malware Detection: Toward Machine Learning Modeling with Explainability Analysis" demonstrates the efficacy of various machine learning algorithms in detecting malicious content within PDF files. By employing models such as Random Forest, SVM, DNN, and others, the project not only achieves high detection accuracy but also emphasizes the importance of model explainability. This dual focus enhances the transparency and trustworthiness of the detection process, making it easier for cybersecurity professionals to understand and respond to potential threats. The successful integration of explainability into the detection models represents a significant advancement in cybersecurity, providing a reliable and interpretable approach to safeguarding against PDF-based malware attacks.

FUTURE SCOPE

Future enhancements for this project could involve integrating advanced explainability tools such as SHAP (SHapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) to further improve the interpretability of complex models like Deep Neural Networks (DNNs). Additionally, expanding the dataset with more diverse and recent PDF samples could increase the model's robustness against evolving malware tactics. Incorporating real-time detection capabilities and automating the update process for the models based on new threats could also enhance the system's effectiveness. Finally, exploring hybrid models that combine machine learning with traditional rule-based methods could provide a more comprehensive approach to PDF malware detection, further strengthening cybersecurity defenses.

REFERENCE

[1] S. S. Alshamrani, "Design and analysis of machine learning based technique for malware identification and classification of portable document format files," Secur. Commun. Netw., vol. 2022, pp. 1–10, Sep. 2022.

[2] P. Singh, S. Tapaswi, and S. Gupta, "Malware detection in PDF and office documents: A survey," Inf. Secur. J., Global Perspective, vol. 29, no. 3, pp. 134–153, May 2020.

[3] N. Livathinos, C. Berrospi, M. Lysak, V. Kuropiatnyk, A. Nassar, A. Carvalho, M. Dolfi, C. Auer, K. Dinkla, and P. Staar, "Robust PDF document conversion using recurrent neural networks," in Proc.

AAAI Conf. Artif. Intell., vol. 35, no. 17, 2021, pp. 15137–15145.
[4] Q. A. Al-Haija, A. Odeh, and H. Qattous, ‘‘PDF malware detection based on optimizable decision trees,’’ Electronics, vol. 11, no. 19, p. 3142, Sep. 2022.
[5] Y. Wiseman, ‘‘Efficient embedded images in portable document format,’’ Int. J., vol. 124, pp. 38–129, Jan. 2019.
[6] M. Ijaz, M. H. Durad, and M. Ismail, ‘‘Static and dynamic malware analysis using machine learning,’’ in Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2019, pp. 687–691.
[7] Y. Alosefer, ‘‘Analysing web-based malware behaviour through client honeypots,’’ Ph.D. dissertation, School Comput. Sci. Inform., Cardiff Univ., Cardiff, Wales, U.K., 2012.
[8] N. Idika and A. P. Mathur, ‘‘A survey of malware detection techniques,’’ Purdue Univ., vol. 48, no. 2, pp. 32–46, 2007.
[9] M. Abdelsalam, M. Gupta, and S. Mittal, ‘‘Artificial intelligence assisted malware analysis,’’ in Proc. ACM Workshop Secure Trustworthy CyberPhys. Syst., Apr. 2021, pp. 75–77.
[10] W. Wang, Y. Shang, Y. He, Y. Li, and J. Liu, ‘‘BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors,’’ Inf. Sci., vol. 511, pp. 284–296, Feb. 2020.