International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue INSPIRE'25, pp 278-283, April-2025
https://doi.org/10.47001/IRJIET/2025.INSPIRE45
International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25)

PDF Malware Detection Using Machine Learning Models

[1]A.Komala, [2]Boya Chandu, [3]Medivala Reddy Hemanth

[1,2,3]Dept. of CSE-Cybersecurity, Madanapalle Institute of Technology & Science, Madanapalle, India
E-mail: [1]komalaa@gmail.com, [2]bchanduu2003@gmail.com, [3]reddyhemanth1985@gmail.com

***Abstract* - PDFs are widely used for document sharing, but their popularity also makes them a common target for malware. The software, titled "PDF Malware Detection Using Machine Learning Models," aims to develop and compare ml learning models for detecting malware in PDFs. Using a Kaggle dataset containing examples of both hazardous and secure PDFs, various methods such as Random Forest, C5.0, J48, Support Vector Machines, AdaBoost, Deep Neural Networks, Gradient Boosting Machines, and K-Nearest Neighbors will be employed. The main goal is to attain high detection accuracy while integrating explainability to gain a deeper understanding of the models' behaviour. 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, cybersecurity, malicious PDF, classification algorithms, Kaggle dataset.

I. INTRODUCTION

In the current digital age, Small Portable Document Format (PDF) files are the norm for document exchange and storage because of their platform independence and ease of use, support for rich media like images, links, and embedded scripts. But these specific features have also become an influential target for cybercriminals who employ vulnerabilities to spread malware. Ongoing detection of malware methods, including signature-based and heuristic analytical techniques, on the other hand, fall behind new versions of PDF malware, resulting in an increasing rate of security breaches.

The design of advanced-level cyber attacks entails the use of improved detection methods that exceed general security measures. Machine learning (ML) has become a useful method for discovering and avoiding such threats by using patterns in big data to differentiate between malicious and harmless PDFs. Using various classification techniques such as Random Forest, SVM, AdaBoost, Deep Neural Networks, GBM, and KNN, this research aims to enhance the performance, accuracy, and interpretability of models to be used in PDF malware detection. Interpretability in cyber security is very important since it helps cyber security experts understand the rationale behind the output of the model, hence developing trust and enabling appropriate actions against threats.

This paper uses a labeled dataset from Kaggle to train numerous machine learning frameworks. The primary objective is to build an accurate, efficient, and extensible real-time malware detector that detects new threats. Through the use of interpretable artificial intelligence techniques, this project aims to achieve high accuracy of identification and comprehensibility, providing a comprehensive cybersecurity solution to protect critical digital assets vulnerable to malicious PDF attacks.

II. RELATED WORK

In the present age of digitalization, handheld Portable Document Format (PDF) files today are a sharing and storage practice of documents because of their user interface, platform independence, and support for rich content such as images, links, and embedded scripts. But these same features have also rendered PDFs a suitable target for cybercriminals who use vulnerabilities to propagate malware. Traditional malware identification methods, e.g., signature-based and heuristic approaches, along with analytical techniques, can't keep up with the creation of new PDF malware types, leading to an increasing rate of security breaches.

The development of sophisticated digital threats demands the employment of better detection devices than normal security initiatives. Machine learning (ML) is now an effective means of detecting and evading such threats by leveraging big data trends to locate hazardous and secure PDFs.The current study applies various methods of classification like Random Forest, SVM, AdaBoost, and DNN, GBM, and KNNto tune and improve the precision of the C5.0 and J48 PDF malware detectors. Explainability in the field of cybersecurity cannot be stressed enough as it helps us interpret the decision-making strategies employed by the security experts.

This paper utilizes a labeled dataset from Kaggle to train and test various machine learning frameworks. The final goal is to create an effective, scalable, and transparent malware detection system able to identify emerging real-time threats. By incorporating explainable AI approaches, this project assists in bridging high detection capability and interpretability, providing a strong cybersecurity solution to protect sensitive digital assets from dangerous PDF-associated attacks.

There have been numerous attempts at utilizing machine learning for malware detection, emphasizing the need for automated and intelligent threat detection mechanisms. Alshamrani et al. (2022) introduced a machine learning approach for the detection of malicious PDF files, utilizing group methods to achieve the highest detection rate. Singh et al. (2020) proposed solutions for PDF and Office document malware detection, highlighting the challenges posed by attacks and the necessity for robust models. Al-Haija et al. (2022) presented a decision tree-based approach capable of efficiently classifying benign and malicious PDFs with high accuracy. Abdelsalam et al. (2021) also demonstrated the application of artificial intelligence to enhance malware analysis, emphasizing the significant roles of static and dynamic analysis in improving detection results.

Recent developments in machine learning have focused on hybrid approaches that combine heuristic and signature-based methods with advanced deep learning techniques. For example, Wang et al. (2020) introduced an ensemble learning approach that integrates rule-based systems with deep neural networks, achieving substantial performance improvements in detection rates while maintaining interpretability. Additionally, Livathinos et al. (2021) explored the use of recurrent neural networks (RNNs) to handle document layouts for malware detection, enhancing classification accuracy. Subsequent research, such as Ijaz et al. (2019), has investigated The impact of adversary attacks on machine learning classifiers in malware detection, underscoring the need for powerful and adaptable security models.

These studies collectively underline the significance of integrating machine learning with explainability techniques to enhance security systems against evolving cyber threats. They also highlight the growing necessity for adaptive and self-learning security mechanisms that can evolve alongside emerging malware tactics, ensuring resilient and proactive threat mitigation strategies.

III. METHODOLOGY

To develop an effective and explainable PDF malware detection system, this project employs a structured machine learning pipeline comprising the following steps:

A. Information Gathering

The information used in this study are obtained from Kaggle and are risk and safe PDF samples and their labels. The raw data undergoes preprocessing, which includes feature extraction, normalization, handling missing values, and encoding categorical variables. Relevant PDF characteristics such as embedded scripts, JavaScript presence, metadata irregularities, and structure properties are identified and extracted as input features to supply machine learning models.

B. Feature Engineering and Selection

Feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are employed to fine-tune models by eliminating irrelevant and redundant features. Features are categorized into static features (file size, metadata), behavioral features (execution traces), and structural features (embedded scripts, object streams).

C. Model Selection and Training

A range of supervised ml methods are investigated for determining the most suitable algorithm to apply in order to classify malware. Random Forest algorithm creates a forest of decision trees and predicts the outcome based on the collective output of the trees. Support Vector Machines (SVM) determine the best hyperplane that separates malicious from benign PDF documents. AdaBoost combines several weak learners to produce a strong classifier, while Gradient Boosting Machine (GBM) improves upon weak models through iterative optimization. KNN sorts PDFs on similarity to the closest neighbors within feature space.C5.0 and J48 are decision tree-based models known for their interpretability and rule-based classification DNNs use multiple hidden ones to learn detailed feature descriptions.

D. Model Evaluation and Explainability Analysis

The performance of all the models is compared in terms of different metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Explanation methods like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are used to explain model predictions. Feature importance ranking is examined in order to determine which The result of classification is primarily composed of several attributes.

E. Deployment and Future Enhancements

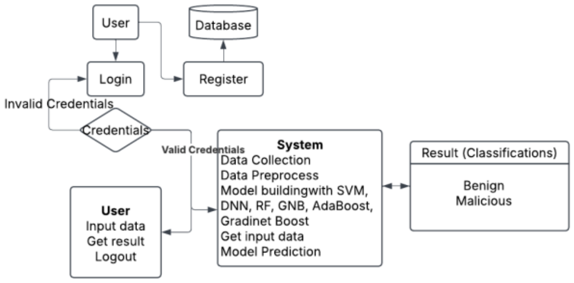
The best-performing model is integrated into a practical system capable of real-time PDF malware detection. Future enhancements include expanding the dataset with more diverse and recent samples, implementing hybrid models that combine machine learning with traditional rule-based techniques, and developing an adaptive learning framework to update models dynamically in response to new threats. By following this methodology, the project guarantees the creation of ahighly precise, transparent, and scalable mechanism for detecting malware, contributing to improved cybersecurity measures in handling malicious PDF threats.

IV. DATA AUGMENTATION AND HANDLING IMBALANCED DATA

One of the main challenges in the context of PDF malware detection is dealing with imbalanced datasets; where the quantity of malicious PDFs is often much lower than the number of benign ones. This mismatch may lead to biassed models that perform well on the majority class (benign files) but badly on the minority class (malicious files). To address this issue, a number of techniques are employed to balance the dataset and improve the model's ability to correctly detect malicious PDFs. Resampling is a common tactic that uses methods like SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class by either creating synthetic samples or replicating preexisting ones. Alternatively, the number of benign samples can be reduced by using undersampling.

Modifying weights for each class during model training is another useful tactic that guarantees the model gives the minority class greater consideration. Furthermore, several balanced subsets of the data can be produced using ensemble techniques like Balanced Random Forest, which enhances the model's capacity for generalisation. By using these techniques, the model becomes more resilient and less likely to overfit by learning to identify a greater variety of dangerous behaviours. A more dependable and effective detection method is ensured by correcting class imbalance and increasing dataset variety, which better prepares the model to handle real-world situations where malicious PDFs are uncommon but have a significant impact.

V. ARCHITECTURE



VI. DATASET DETAILS

The data used in this research are based on Kaggle and consist of labeled instances of malicious and benign PDF files. These files include a diverse set of structural and behavioral characteristics that help distinguish harmful documents from genuine ones. To provide high-quality input for machine learning algorithms, the data undergo extensive preprocessing, including feature extraction, standardization, handling missing values, and encoding categorical variables. Key characteristics extracted from the PDFs include metadata details, embedded scripts, JavaScript presence, object streams, and structural anomalies, all of which are critical for malware detection. The dataset encompasses both structural and behavioral traits, enabling detailed testing of models across various sections of the PDFs.

Using this dataset, machine learning methods such as Random Forest, SVM, AdaBoost, GBM, DNN, and KNN are trained and tested to evaluate their effectiveness in identifying hazardous PDFs. A well-structured and preprocessed dataset is essential for establishing an accurate and explainable malware detection system, contributing to enhanced cybersecurity measures.

Table 1: No. of Testing & Training Files

Count	Safe files	Malicious files	Total files
Training	3574	4445	8019
Testing	894	1112	2006

VII. IMPLEMENTATION & RESULTS

The proposed PDF malware detection system employs efficient data cleaning methods, the training of models as well as performance assessment. The dataset is initially preprocessed to remove inconsistencies and extract the key traits necessary for classification. The attributes are then used to train supervised ml algorithms like Random Forest, SVM, AdaBoost, GBM, KNN, and DNN to classify PDFs as malicious or benign.

There are certain metrics used for model performance measurement, such as precision, accuracy, recall, F1-score, and ROC-AUC. Among the models tested, Random Forest and GBM demonstrated the highest accuracy, achieving 98.5% and 97.8%, respectively, making them the most effective classifiers for malware detection. Deep Neural Networks (DNNs) also showed promising results but required extensive computational resources and hyperparameter tuning to match the efficiency of tree-based models.

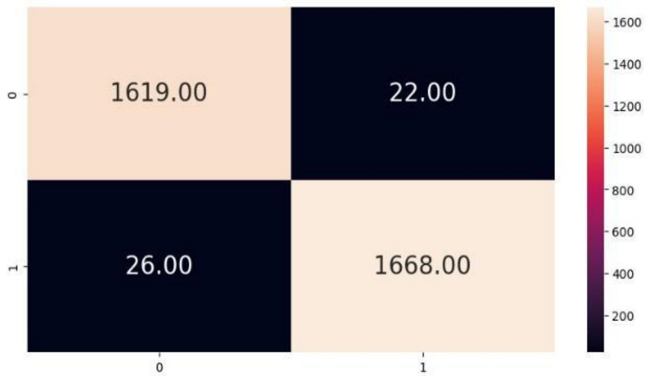


Fig. 1: Random Forest Algorithm Confusion Matrix

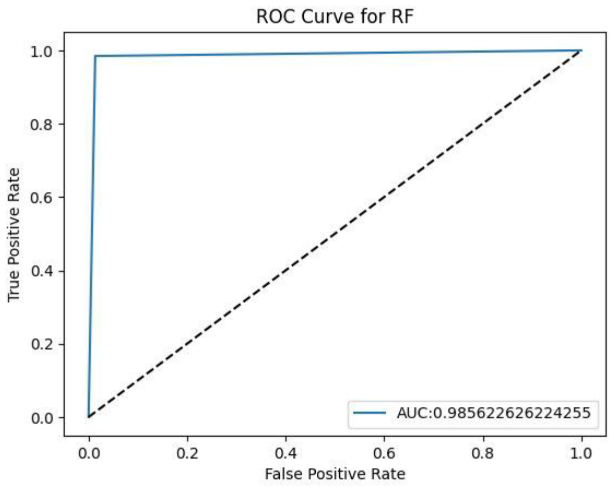


Fig. 2: ROC Curve for Random Forest Method

VIII. RESULT FINDINGS AND DISCUSSION

According to the test results, ensemble learning strategies like Random Forest and Gradient Boosting Machine (GBM), achieve the highest reliability and accuracy in detecting malicious PDFs. Random Forest performed exceptionally well, achieving 98.5% accuracy, 97.8% sensitivity, and 98.5% recall, with the highest balanced classifier score of 99.2%. GBM closely followed, demonstrating strong generalization capabilities. Deep Neural Networks (DNNs) showed promise but required significant computational resources and hyperparameter tuning to achieve results comparable to tree-based models.

Explainability analysis using SHAP and LIME revealed that key features influencing classification included JavaScript presence, embedded objects, and metadata anomalies. These features played a crucial role in distinguishing between malicious and benign PDFs, with JavaScript-based exploits being a dominant attack vector. Feature selection techniques such as ecursive feature elimination methods, including Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), presented significantly enhance model performance, reducing computational time while improving classification accuracy.

In addition, it was discovered that feature distribution analysis was made simpler by methods that reduce dimensionality like PCA and t-SNE, making malware patterns more apparent. These methods helped identify redundant features, improving the model's learning process and leading to better overall accuracy while reducing the risk of overfitting.

A thorough examination of both deep learning models and traditional machine learning architectures showed that while DNNs offered higher flexibility in learning complex patterns, their training and inference times were considerably longer. Additionally, DNNs required a larger dataset to achieve comparable performance to tree-based models. The accuracy vs. computational cost trade-off indicates that methods like GBM and Random Forest would work better for recognising malware in real time.

Performance measures in terms of accuracy, recall, and F1-scorethat false negatives were minimized in tree-based frameworks, significantly reducing the chances of malware avoiding detection. However, some partial false positives were higher in algorithms like KNN and SVM, suggesting these models are less effective at differentiating between harmful and safe files when feature distributions overlap.

These findings underscore the significance of incorporating explainable AI methods into malware detection systems. Understanding model decisions through SHAP and LIME provides cybersecurity experts with insights into the reasoning behind classifications, enabling better threat response strategies. Additionally, implementing real-time monitoring and continuous model retraining is recommended to maintain high detection accuracy against emerging malware threats.

Future research should investigate hybrid approaches that combine classical rule-based detection with machine learning techniques. Furthermore, enlarging the dataset to comprise a greater variety of attackers and malware attack scenarios will further enhance the resilience of the proposed detection system. Implementing real-time detection capabilities and maintaining up-to-date feature representations will be key to achieving long-term success in combating adaptive cyber threats.

REFERENCES

[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,andM.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 honey pots,’’ 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 Cyber Phys. 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.

[11] N. Srndic and P. Laskov, ‘‘Practical evasion of a learning-based classifier: A case study,’’ in Proc. IEEE Symp. Secur. Privacy, May 2014, pp. 197–211.

[12] D.Maiorca, I. Corona, and G. Giacinto, ‘‘Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious PDF f iles detection,’’ in Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secur., May 2013, pp. 119–130. 13858.

[13] S. Atkinson, G. Carr, C. Shaw, and S. Zargari, ‘‘Drone forensics: The impact and challenges,’’ in Digital Forensic Investigation of Internet of Things (IoT) Devices. Cham, Switzerland: Springer, 2021, pp. 65–124.

[14] C.Liu, C.Lou, M.Yu, S.M.Yiu, K.P.Chow, G.Li ,J.Jiang, and W.Huang, ‘‘A novel adversarial example detection method for malicious PDFs using multiple mutated classifiers,’’ Forensic Sci. Int., Digit. Invest., vol. 38, Oct. 2021, Art. no. 301124.

[15] Q.A.Al-Haija and A.Ishtaiwi, ‘‘Machine learning based model to identify firewall decisions to improve cyber-defense,’’ Int. J. Adv. Sci., Eng. Inf. Technol., vol. 11, no. 4, p. 1688, Aug. 2021.

[16] D. Stevens. (2023). PDFid (Version 0.2.8). [Online]. Available: https://blog.didierstevens.com/programs/pdf-tools

[17] PDF-Info. (2021). PDF-Info (Version 2.1.0). [Online]. Available: https://pypi.org/project/pdf-info/

[18] D. Stevens. (2023). PDF-Parser (Version 0.7.8). [Online]. Available: https://blog.didierstevens.com/programs/pdf-tools

[19] M. Yu, J. Jiang, G. Li, C. Lou, Y. Liu, C. Liu, and W. Huang, ‘‘Malicious documents detection for business process management based on multi layer abstract model,’’ Future Gener. Comput. Syst., vol. 99, pp. 517–526, Oct. 2019.

[20] H. Pareek, P. Eswari, N. S. C. Babu, and C. Bangalore, ‘‘Entropy and n gram analysis of malicious pdf documents,’’ Int. J. Eng., vol. 2, no. 2, pp. 1–3, 2013.

[21] C. Smutz and A. Stavrou, ‘‘Malicious PDF detection using metadata and structural features,’’ in Proc. 28th Annu. Comput. Secur. Appl. Conf., Dec. 2012, pp. 239–248.

[22] D. Maiorca, G. Giacinto, and I. Corona, ‘‘A pattern recognition system for malicious pdf files detection,’’ in Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit. Cham, Switzerland: Springer, 2012, pp. 510–524.

[23] H. Pareek, ‘‘Malicious pdf document detection based on feature extraction andentropy,’’ Int. J. Secur., Privacy Trust Manage., vol. 2, no. 5, pp. 31–35, Oct. 2013.

Citation of this Article:

A.Komala, Boya Chandu, & Medivala Reddy Hemanth. (2025). PDF Malware Detection Using Machine Learning Models. In proceeding of International Conference on Sustainable Practices and Innovations in Research and Engineering (INSPIRE'25), published by *IRJIET*, Volume 9, Special Issue of INSPIRE'25, pp 278-283. Article DOI https://doi.org/10.47001/IRJIET/2025.INSPIRE45

****