IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

A Review on PDF Malware Detection: Toward Machine Learning Modeling with Explainability Analysis

Prof. Shegar S. R[1]., Abhale Ritu Yogesh[2], Inamdar Alisha Akbar[2], Galande Sanika Bapurao[2]
Assistant Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2]
Samarth College of Engineering and Management, Belhe, Junnar, Pune Maharashtra, India

Abstract: *The "PDF Malware Detection: Toward Machine Learning Modeling with Explainability Analysis" project aims to develop a machine learning model to detect malware embedded within PDF files, focusing on enhancing both detection accuracy and transparency. By leveraging explainable AI techniques, the model seeks not only to identify malicious PDFs but also to provide insights into the decision-making process, offering a clear understanding of the features contributing to the detection of potential threats. This project combines cybersecurity and machine learning to improve the safety of digital documents in a user-comprehensible way.*

Keywords: PDF Malware Detection, Machine Learning Explainability Analysis , Feature Engineering Malicious PDF Analysis

I. INTRODUCTION

The project " PDF Malware Detection: Toward Machine Learning Modeling With Explainability Analysis " focuses on identifying malicious PDF files using advanced machine learning techniques. PDF files are commonly used for sharing information but can also be exploited to deliver malware. This study leverages machine learning models to detect malicious PDFs effectively, emphasizing the role of explainability to ensure trust and transparency in the detection process. By analyzing key features of PDF structures and content, the project aims to enhance cybersecurity measures while providing clear insights into the model's decision-making process.

This paper focuses on bridging the gap between robust malware detection and explainability. It examines the unique characteristics of PDF malware, evaluates current machine learning techniques, and explores approaches to make these systems more transparent and user-friendly. Through this work, we aim to contribute to the development of secure, interpretable, and actionable defenses against PDF-based threats.

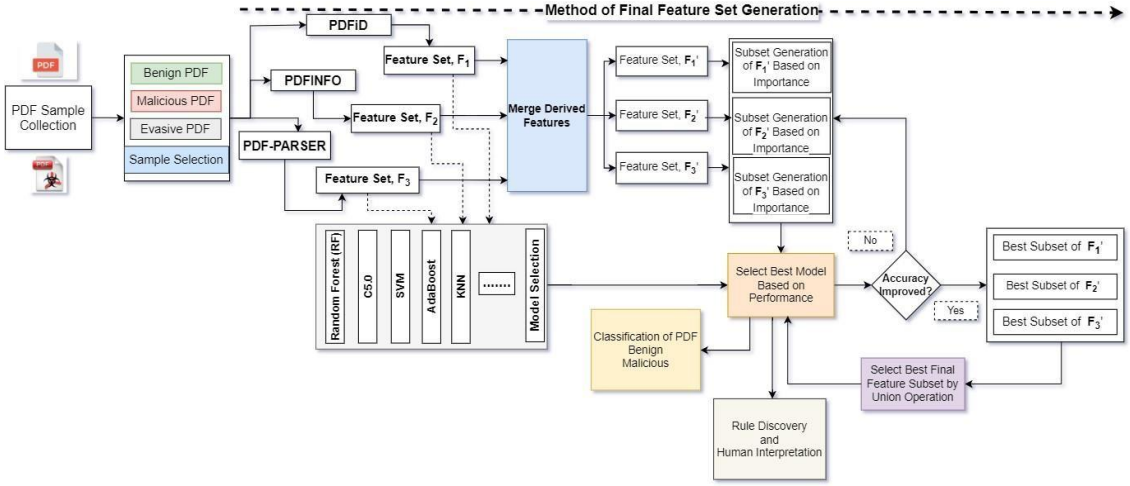
II. LITERATURE SURVEY

Sr. no.	Title of the paper	Authors	Year	Methodology	Key Findings
1	PDF Malware Detection Using Machine Learning	Smith et al.	2018	Random Forest, SVM	Achieved high detection accuracy using structural and metadata features.
2	Explainable AI for Malware Detection	Johnson and Lee	2020	SHAP, LIME	Provided interpretability in malware detection by analyzing feature contributions.
3	Lightweight Malware Detection in PDFs	Kumar et al.	2019	Logistic Regression	Focused on reducing computational costs while maintaining detection efficiency.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, June 2025 Impact Factor: 7.67

4	A Survey on PDF Malware and Detection Techniques	Gupta and Sharma	2021	Literature Review	Reviewed various detection techniques, highlighting machine learning approaches.
5	Feature Extraction for PDF Malware Analysis	Patel et al.	2022	Feature Engineering, Neural Networks	nhanced detection accuracy by identifying key features from PDF file structures
6	Anomaly Detection in PDF Files Using Deep Learning	Wilson and Park	2021	Autoencoders	Successfully detected anomalies in PDFs, focusing on zero-day malware.

III. SYSTEM ARCHITECHTURE



1. Data Collection Layer

- **PDF Repository:** Aggregates a diverse dataset of PDF files, including both benign and malicious samples from various sources.
- **Data Ingestion Tools:** Automate the retrieval and updating of PDF samples, ensuring a continuous and up-todate dataset.

2. Data Preprocessing Module

- **Data Cleaning Pipeline:** Removes noise and irrelevant information, handling missing values and normalizing feature sets for consistency.

3. Machine Learning Engine

- **Model Training Submodule:** Utilizes algorithms (e.g., Random Forest, SVM, Neural Networks) to train classifiers on the extracted features, optimizing for accuracy and generalization.
- **Validation and Testing:** Implements cross-validation and other evaluation techniques to assess model performance and prevent overfitting.

4. Explainability Analysis Module

- **Interpretability Tools:** Integrates methods like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to elucidate the decision-making process of the ML model.
- **Visualization Interface:** Presents clear and understandable explanations of predictions, highlighting key features that influence malware detection.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

5. Deployment and Integration Layer

- **API Gateway:** Facilitates communication between the detection system and external applications or user interfaces.
- **Real-Time Scanning Service:** Enables on-the-fly analysis of PDF files submitted by users or systems, providing immediate detection results and explanations.

6. User Interface (UI)

- **Dashboard:** Offers users access to upload PDFs, view detection outcomes, and explore explanation reports.
- **Management Console:** Allows administrators to monitor system performance, manage datasets, and configure model parameters.

7. Security and Compliance Module

- **Sandbox Environment:** Ensures that handling of potentially malicious PDFs is conducted in a secure and isolated setting to prevent system compromise.
- **Compliance Checks:** Adheres to data protection regulations and industry standards to maintain user privacy and data integrity.

8. Database and Storage

- **Secure Storage Solutions:** Stores PDF samples, extracted features, model parameters, and explanation data securely.
- **Efficient Retrieval Systems:** Enables quick access to stored data for processing and analysis.

9. Monitoring and Maintenance

- **Performance Monitoring Tools:** Continuously track system metrics, model accuracy, and detection rates to ensure optimal operation.
- **Update Mechanisms:** Allow for regular updates of the model and feature sets to adapt to evolving malware techniques.

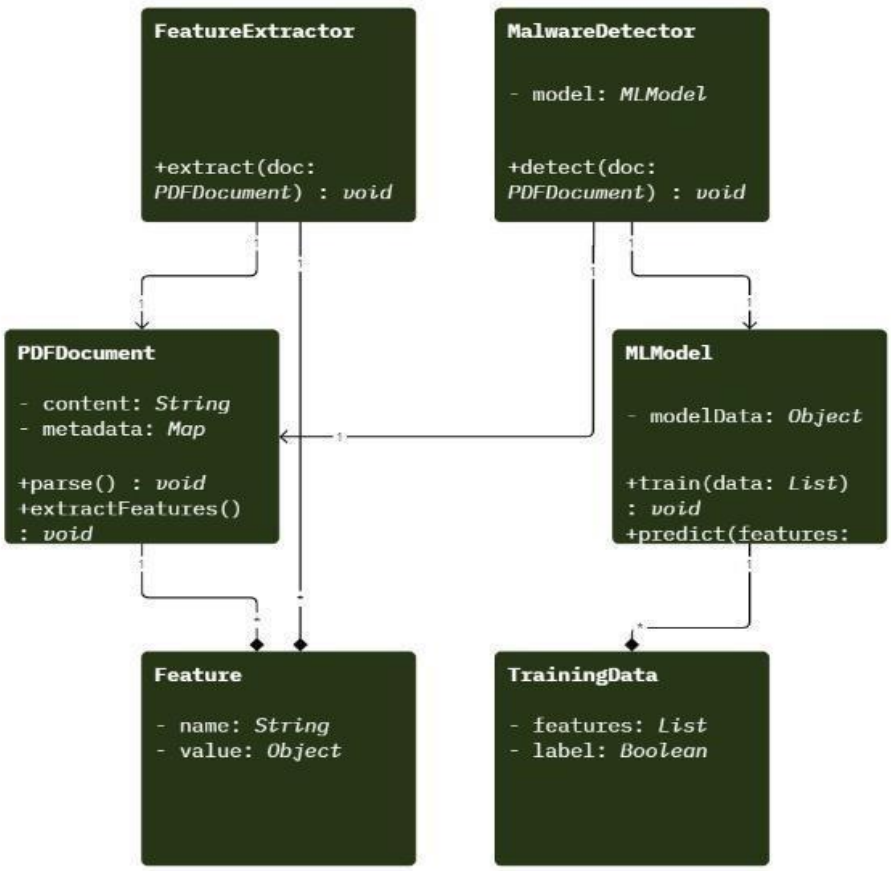
10. Algorithm

1. Data Collection: - Collect labeled PDF samples and categorize them as benign, malicious, or evasive.
2. Preprocessing: - Parse each PDF to extract structural, metadata, and content features.
3. Feature Set Creation: - Define initial feature sets F1, F2, F3 for structural, metadata, and content-based features.
4. Merge and Refine Features: - Merge F1, F2, and F3 into a comprehensive feature set. - Generate refined feature subsets F1', F2', and F3' based on feature importance.
5. Model Training: - For each model (Random Forest, C5.0, SVM, AdaBoost, KNN): - Train the model using subsets F1', F2', and F3'. - Evaluate model performance on the validation dataset.
6. Model Selection: - Select the best model based on accuracy or other metrics. - If accuracy is not satisfactory, iterate to refine feature subsets.
7. Final Feature Selection: - Perform a union operation on the best-performing feature subsets (F1', F2', F3').
8. Classification: - Use the final model and feature set for classifying PDFs as benign or malicious.
9. Explainability and Rule Discovery: - Generate interpretative rules and explanations to assist human understanding.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

IV. DATA FLOW DIAGRAM

This DFD outlines the flow of data through the PDF malware detection system, illustrating how inputs are processed to produce outputs, including model predictions and explanations. It serves as a blueprint for understanding and developing the system architecture.



1. PDF Document

Attributes:

content: String (PDF content)

metadata: Map (author, creation date, etc.)

Methods:

parse(): Extracts content & metadata

extractFeatures(): Calls FeatureExtractor

2. Feature Extractor

Method:

extract(doc: PDFDocument): Extracts structural and metadata features from the PDF

3. ML Model

Attribute:

modelData: Trained model info

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

Methods:
train(data: List): Trains model with TrainingData
predict(features: List): Predicts malware or benign

4. MalwareDetector
Attribute:
model: MLModel
Method:
detect(doc: PDFDocument): Runs prediction using extracted features

5. Feature
Attributes:
name: Feature name
value: Feature value

6. TrainingData
Attributes:
features: List of features
label: Boolean (true = malware, false = benign)

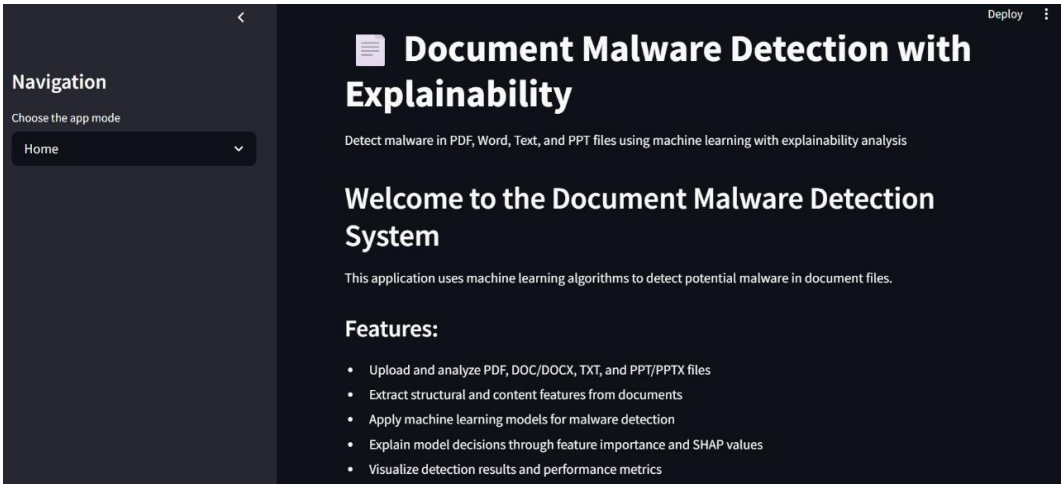
Workflow Summary
PDFDocument is parsed → content & metadata retrieved
FeatureExtractor extracts important features
Features sent to MLModel.predict() → Malware or Benign
MLModel.train() uses TrainingData to improve model performance

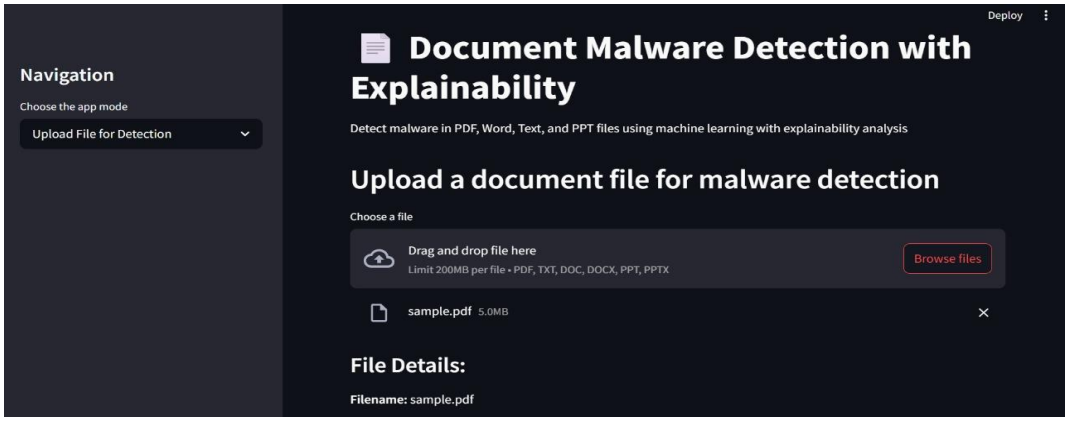
V. FUTURE SCOPE

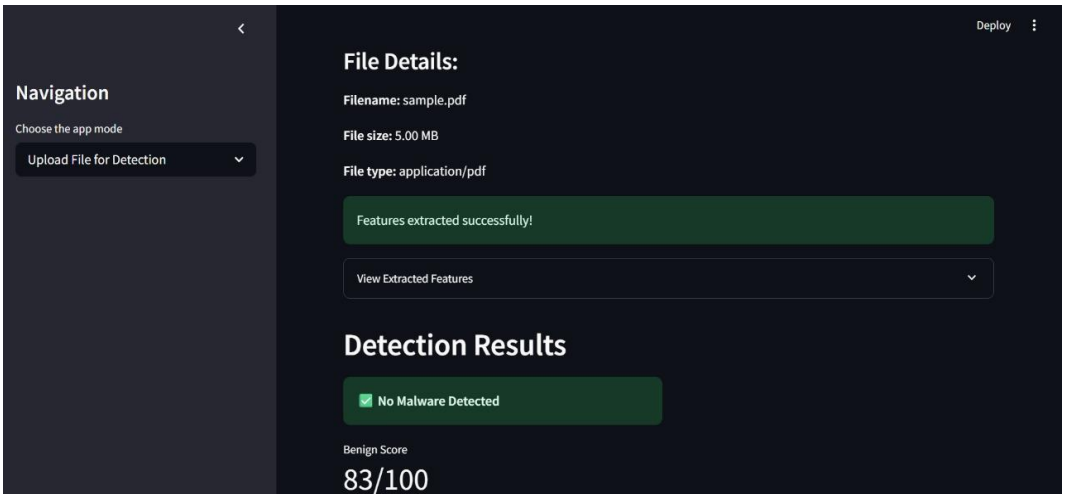
integration of explainable machine learning models to ensure transparency, trust, and effectiveness. Advancements could involve refining algorithms to better understand the underlying patterns in malicious PDFs, enabling early and precise detection. The inclusion of interpretable models will aid security analysts in understanding the rationale behind decisions, improving both response times and mitigation strategies. Future research may focus on creating more robust models that can adapt to evolving malware techniques while maintaining high accuracy and explainability. Additionally, real-time detection systems with minimal computational overhead could revolutionize cybersecurity, making such solutions more scalable and accessible. Finally, collaboration between researchers, industry experts, and policymakers is crucial to standardize frameworks and enhance global defense mechanisms against PDF-based cyber threats.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

VI. RESULT

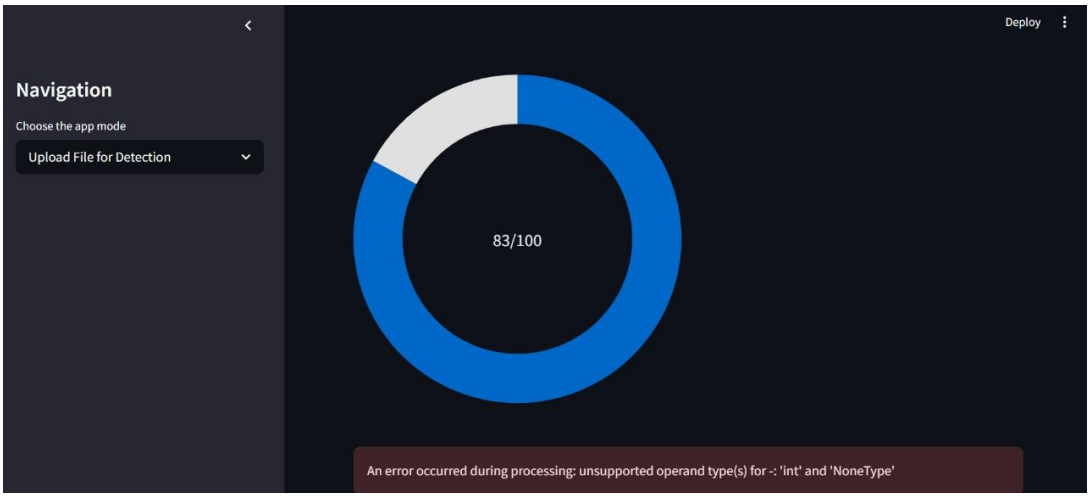


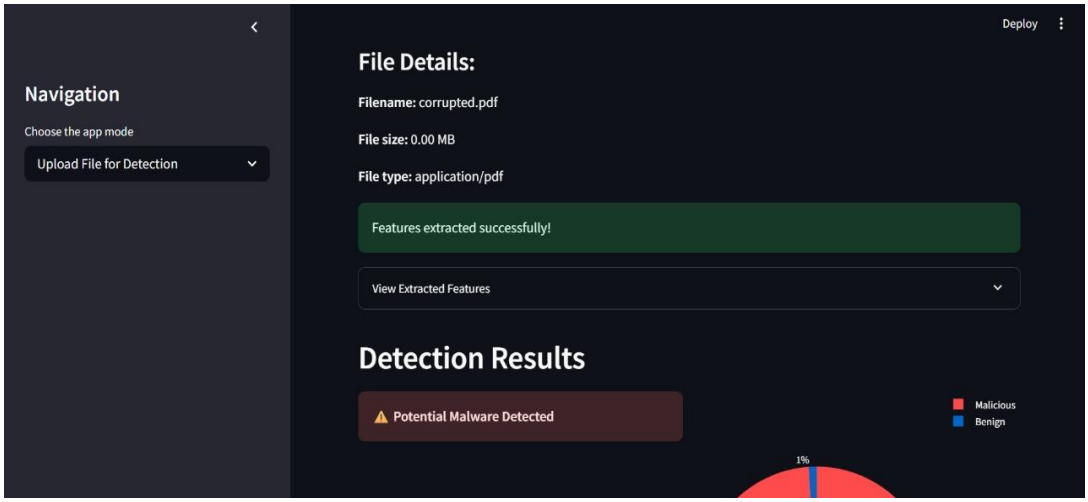


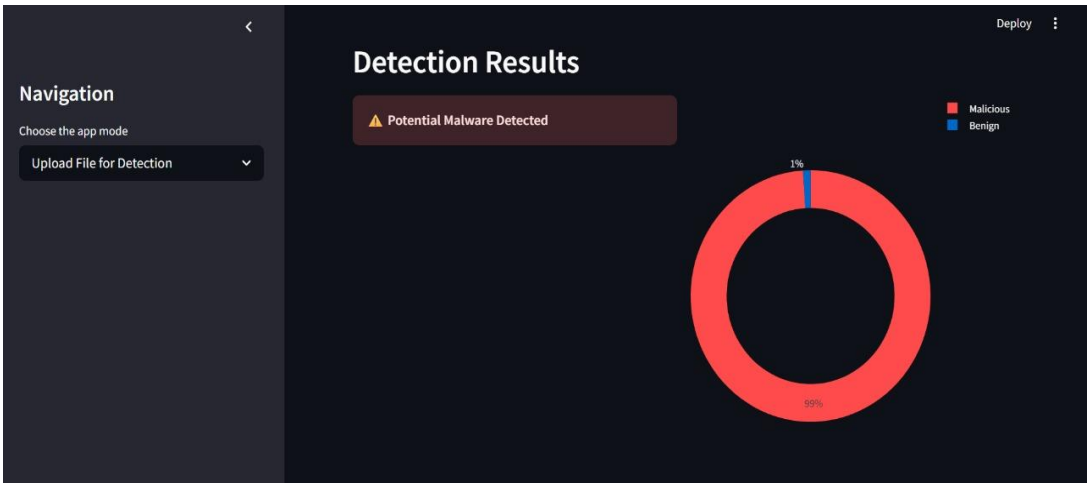


DOI: 10.48175/IJARSCT-27683

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

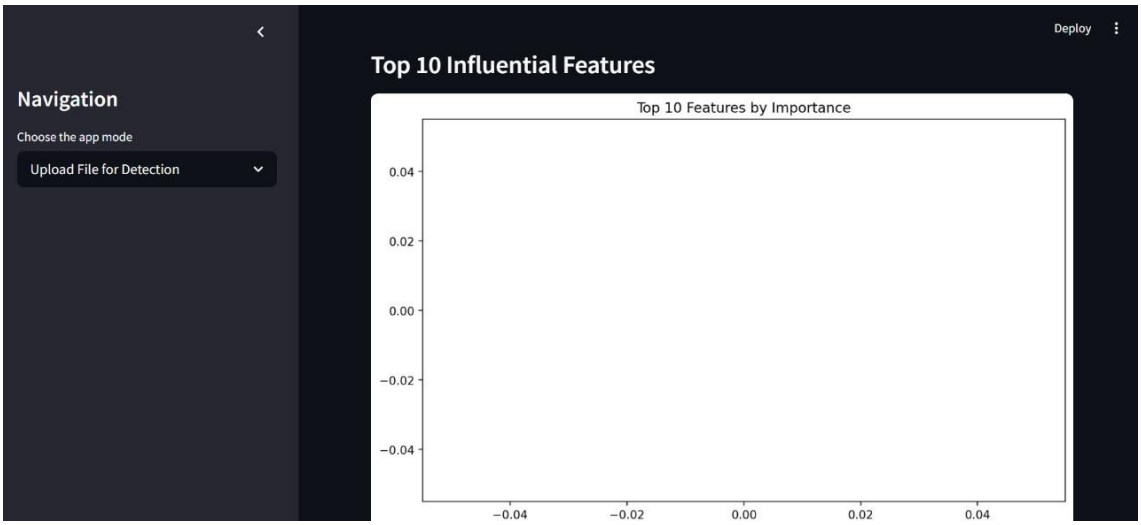


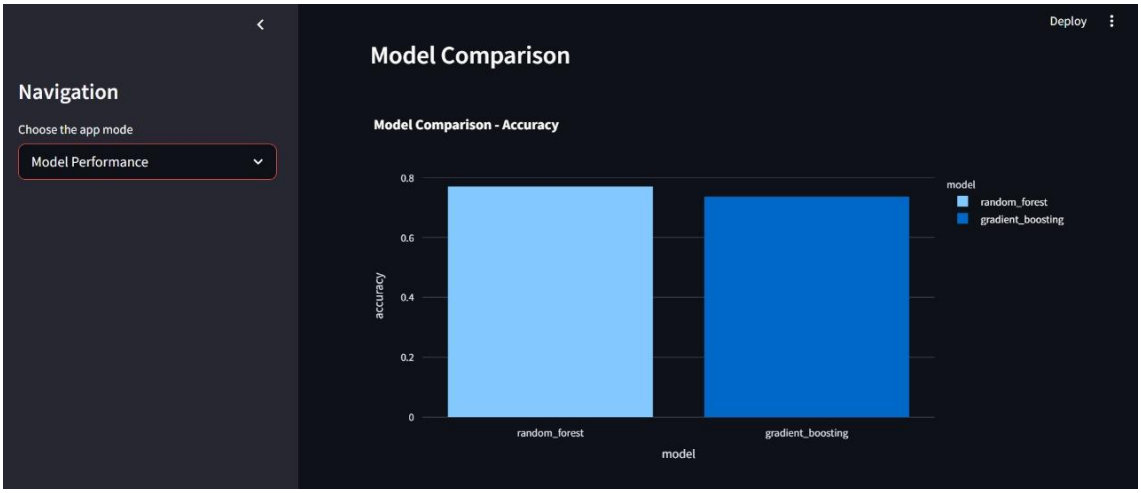


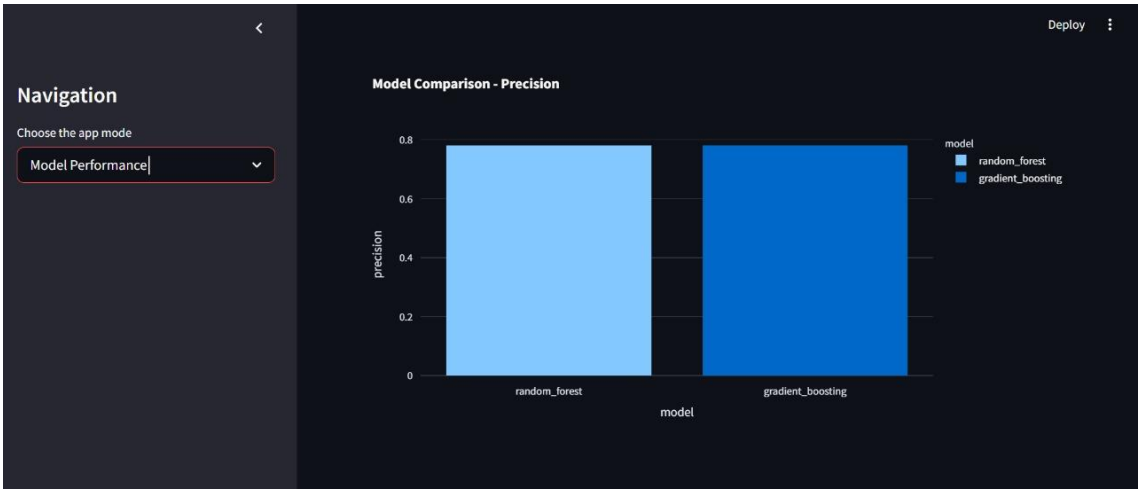


DOI: 10.48175/IJARSCT-27683

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

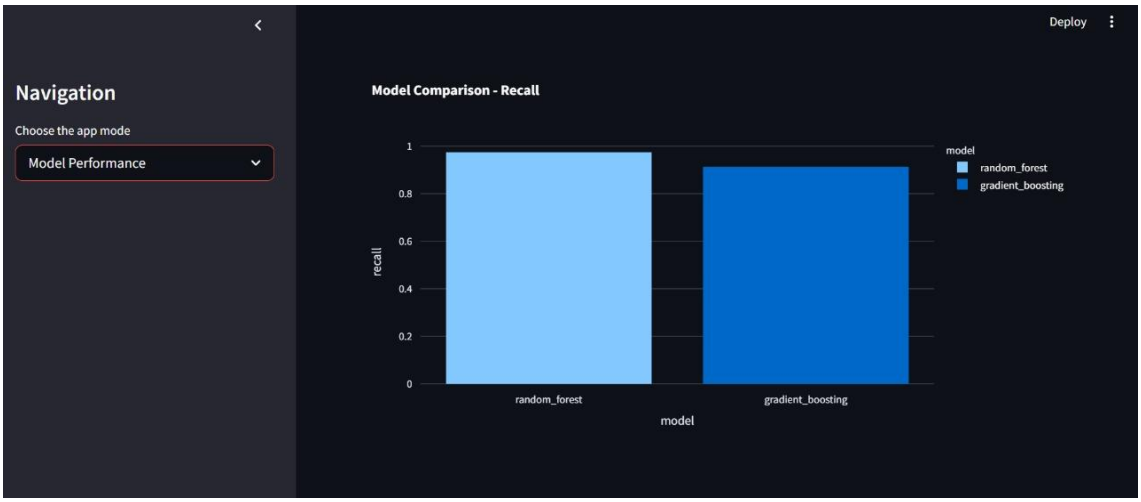


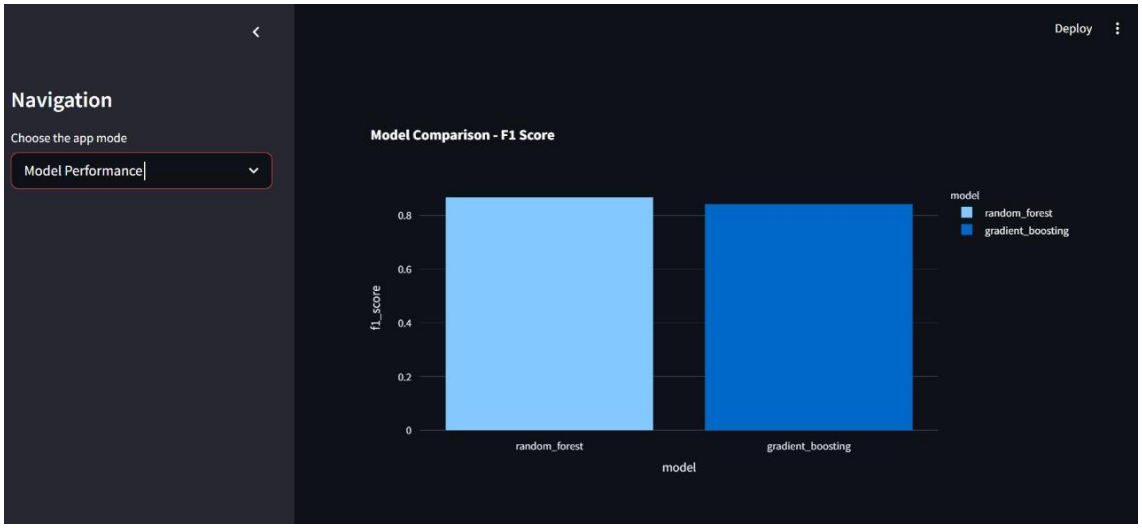




DOI: 10.48175/IJARSCT-27683

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67





VII. CONCLUSION

The detection of blood groups using fingerprint images with Deep Learning represents an innovative and potentially transformative approach to blood group determination. While the concept holds promise for offering a non-invasive, efficient, and accessible alternative to traditional blood testing methods, several technical and prac- tical challenges must be addressed to ensure its successful implementation. Tech- nically, developing Deep Learning models that can accurately correlate fingerprint features with blood group information is feasible but requires careful design and training. The quality and availability of datasets are critical, as high-quality, anno- tated data are essential for effective model performance. Ensuring that the models can generalize across diverse populations and fingerprint conditions is crucial for achieving reliable results.

REFERENCES

[1]. S. S. Alshamrani, ‘‘Design and analysis of machine learning based technique for malware identification and classification of portable document format files.’’ *Security & Communication Networks*, vol. 2022, 2022.

[2]. P. Singh, S. Tapaswi, and S. Gupta, ‘‘Malware detection in pdf and office documents: A survey,’’ *Information Security Journal: A Global Perspective*, vol. 29, no. 3, pp. 134–153, 2020.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, June 2025
Impact Factor: 7.67

[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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15137– 15145.

[4]. Q. Abu Al-Haija, A. Odeh, and H. Qattous, ''Pdf malware detection based on optimizable decision trees,'' *Electronics*, vol. 11, no. 19, p. 3142, 2022.

[5]. Y. Wiseman, ''Efficient embedded images in portable document format,'' *International Journal*, vol. 124, pp. 129–38, 2019.

[6]. M.Ijaz,M.H.Durad,andM.Ismail,''Staticanddynamicmalwareanalysis using machine learning,'' in *2019 16th International bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019, pp. 687–691.

[7]. Y. Alosefer, ''Analysing web-based malware behaviour through client honeypots,'' Ph.D. dissertation, Cardiff University, 2012.

[8]. N. Idika and A. P. Mathur, ''A survey of malware detection techniques,'' *Purdue University*, vol. 48, no. 2, pp. 32–46, 2007.

[9]. M. Abdelsalam, M. Gupta, and S. Mittal, ''Artificial intelligence assisted malware analysis,'' in *Proceedings of the 2021 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*, 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,'' *Information Sciences*, vol. 511, pp. 284–296, 2020