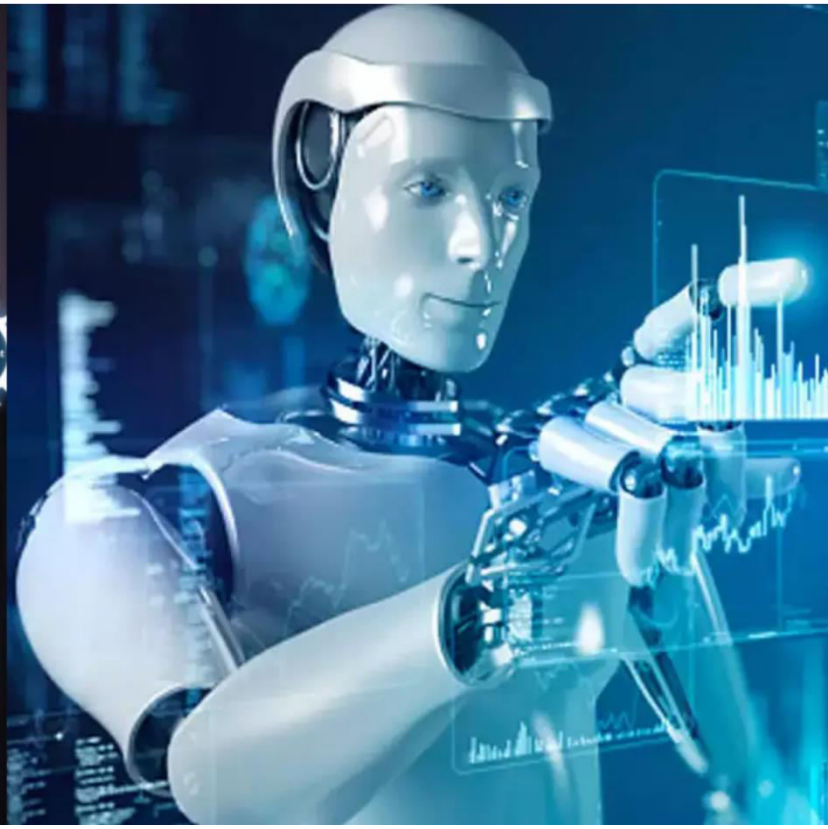
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

Malware Detection using Machine Learning

K. Yaswanth Kumar, J. Lethisiya Nithiya

Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: The Malware Detection System is an intelligent, machine learning-driven solution built to accurately detect and classify files and URLs as either malicious or legitimate. With the increasing number of cyber threats, especially through Portable Executable (PE) files and suspicious URLs, the system aims to enhance the reliability and efficiency of malware detection. It employs two powerful algorithms—K-Nearest Neighbors (KNN) and Random Forest—to learn from patterns in data and make accurate predictions. The system focuses heavily on effective feature extraction and selection, analyzing attributes such as API calls, imported DLLs, section headers, and entropy for PE files, as well as URL structures and domain statistics. These features are preprocessed and normalized to improve the learning ability of the models, and only the most relevant features are selected to boost accuracy and reduce computational load. KNN offers simplicity and effectiveness by classifying based on similarity with neighboring data points, while Random Forest leverages the power of multiple decision trees to improve stability and reduce overfitting. Together, they provide a dual-layered classification system that improves detection confidence and allows result comparison and ensemble predictions. The system is rigorously trained and validated using performance metrics like accuracy, precision, recall, and F1-score to ensure it performs well in real-world scenarios. The main goal is to reduce both false positives and false negatives, which are critical in cybersecurity as even a small error can lead to significant data breaches or system compromise.

I. INTRODUCTION

In today's digital world, almost everything we do - from banking and shopping to socializing and working - relies on computers and the internet. While this connectivity has made life easier, it has also opened the door to a major threat: malware. Malicious software like viruses, worms, Trojans, and backdoors can silently enter our systems, steal sensitive data, damage files, or even take full control of our devices. With millions of these harmful programs hiding inside downloadable files and websites, especially Portable Executable (PE) files and URLs, detecting and stopping them before they cause damage has become more important than ever.

This project focuses on building a smart and efficient Malware Detection System using machine learning. The goal is not just to detect malware but to do it with high accuracy so that even the most cleverly disguised threats can be identified. Unlike traditional antivirus solutions that rely on known signatures, this system learns patterns from both safe and unsafe files using machine learning models - specifically, K-Nearest Neighbors (KNN) and Random Forest algorithms. These models can detect previously unknown malware by understanding how malicious files behave compared to legitimate ones.

A key part of this project is analyzing and extracting useful information from files and URLs - things like what parts of a program it uses, how it's structured, or whether a website link looks suspicious. By training the models on these features, the system becomes capable of recognizing dangerous files more accurately. This means fewer false alarms and better protection for users and organizations.

The ultimate aim of the project is to create a system that not only detects malware but does so with high accuracy, improving cybersecurity defenses and reducing the chances of dangerous files slipping through unnoticed. With cyber threats constantly evolving, this project is a step toward smarter, more proactive protection in the digital space.

To make the system truly effective, we focused on gathering high-quality datasets from reliable sources like Kaggle and security research repositories. These datasets contain thousands of examples of both malicious and legitimate PE files and URLs. For each entry, multiple features were extracted—ranging from structural elements in executable files like

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

section sizes and import tables, to lexical properties in URLs such as domain length and use of special characters. These features provide critical clues that help the system learn what sets a harmful entity apart from a safe one.

In today's threat landscape, malware is no longer just about damaging a system—it's often tied to data theft, financial fraud, corporate espionage, and national security threats. By using a machine learning-based approach, this project goes beyond reactive methods and moves toward proactive, intelligent cybersecurity. The system learns and adapts, offering a strong foundation for future enhancements like integrating live threat feeds, real-time scanning, and cloud-based deployment. As a research-based yet practical solution, it bridges the gap between academic study and real-world application in the fight against evolving cyber threats.

II. LITERATURE SURVEY

1. **Prabath Singh. (2021).** Malware Detection Using Machine Learning.
   • **Objective**: The objective of the proposed malware detection system is to build an intelligent solution that can accurately identify and classify files and URLs as malicious or safe using machine learning. It aims to overcome the limitations of traditional antivirus methods by reducing false positives and false negatives, ensuring better protection for users, enterprises, and sensitive data against the growing range of cyber threats.
   • **Methodology**: The methodology involves extracting meaningful features from Portable Executable (PE) files and URLs, such as section headers, import tables, and structural details, which are then preprocessed and fed into machine learning models. Several algorithms, including Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), are evaluated to find the one that performs best in terms of detection accuracy, precision, and recall. The selected model is then validated using confusion matrix analysis to ensure it effectively distinguishes between malicious and legitimate inputs.
   • **Drawbacks**: t depends heavily on the quality and size of the training dataset, which can limit its effectiveness if the data is outdated or unbalanced. It may also struggle to detect highly obfuscated or new malware strains that do not share patterns with previously seen data. Furthermore, without regular updates and model retraining, its performance may degrade over time as malware techniques evolve.

2. **Nor Zakiah Gorment. (2023).** Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges, and Future Directions.
   • **Objective**: The objective of the paper is to systematically review existing research on malware detection using machine learning and to develop a comprehensive taxonomy that addresses key issues such as detection accuracy, analysis types, and algorithm classification, thereby guiding future work in this domain.
   • **Methodology**: The methodology involves conducting a systematic literature review (SLR) of 77 selected studies related to malware detection. These studies are analyzed to classify different types of malware, detection approaches, and machine learning algorithms. The paper evaluates recent techniques, identifies challenges, and performs an empirical analysis to assess the performance of various algorithms.
   • **Drawbacks**: The drawbacks include the reliance on existing literature, which may not fully cover the latest malware trends or unpublished techniques. There is also a risk of bias in study selection, and the findings may not be generalizable to real-time or large-scale deployment environments.

3. **Irina Baptista. (2019).** A Novel Malware Detection System Based on Machine Learning and Binary Visualization.
   • **Objective**: The objective of this paper is to develop an intelligent malware detection framework that leverages binary visualization and self-organizing incremental neural networks (SOINNs) to effectively detect and classify advanced malware, including zero-day threats, in various file formats.
   • **Methodology**: The methodology involves transforming raw binary files into image-like visual representations using binary visualization, enabling neural networks to analyze malware patterns visually. SOINNs are employed for their continuous learning capabilities, allowing the model to update itself without forgetting previous data. The system is tested on various file types, including ransomware embedded in .pdf and .doc formats, as well as executable files, to validate detection performance.
   • **Drawbacks**: The drawbacks include a reliance on computational resources for real-time image processing and neural training, potential challenges in handling extremely obfuscated or encrypted samples, and limited explainability of deep learning models, which can hinder trust and transparency in critical security applications.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

III. METHODOLOGY

The methodology of this Malware Detection System involves a structured pipeline beginning with the collection and analysis of two primary types of data: Portable Executable (PE) files and URLs, both of which are common vectors for malware distribution. First, the system performs feature extraction by identifying and isolating key attributes that characterize malicious behavior, such as imported DLLs, API calls, section headers, and entropy levels from PE files, along with domain-based statistics and URL structures. These raw features undergo preprocessing techniques like normalization and scaling to ensure consistent data quality and improve the training efficiency of the models. To enhance accuracy and reduce computational overhead, feature selection techniques are employed to retain only the most relevant and informative features. The refined dataset is then used to train two machine learning classifiers—K-Nearest Neighbors (KNN) and Random Forest. KNN is utilized for its simplicity and effectiveness in detecting patterns based on similarity, whereas Random Forest, an ensemble learning method, constructs multiple decision trees to enhance generalization and reduce overfitting. Both models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. The system also supports comparative analysis and blending of results from both classifiers to maximize prediction accuracy. This multi-stage methodology ensures that the system can reliably distinguish between legitimate and malicious inputs and lays the groundwork for future enhancements like integration with deep learning, real-time detection, and cloud-based deployment.

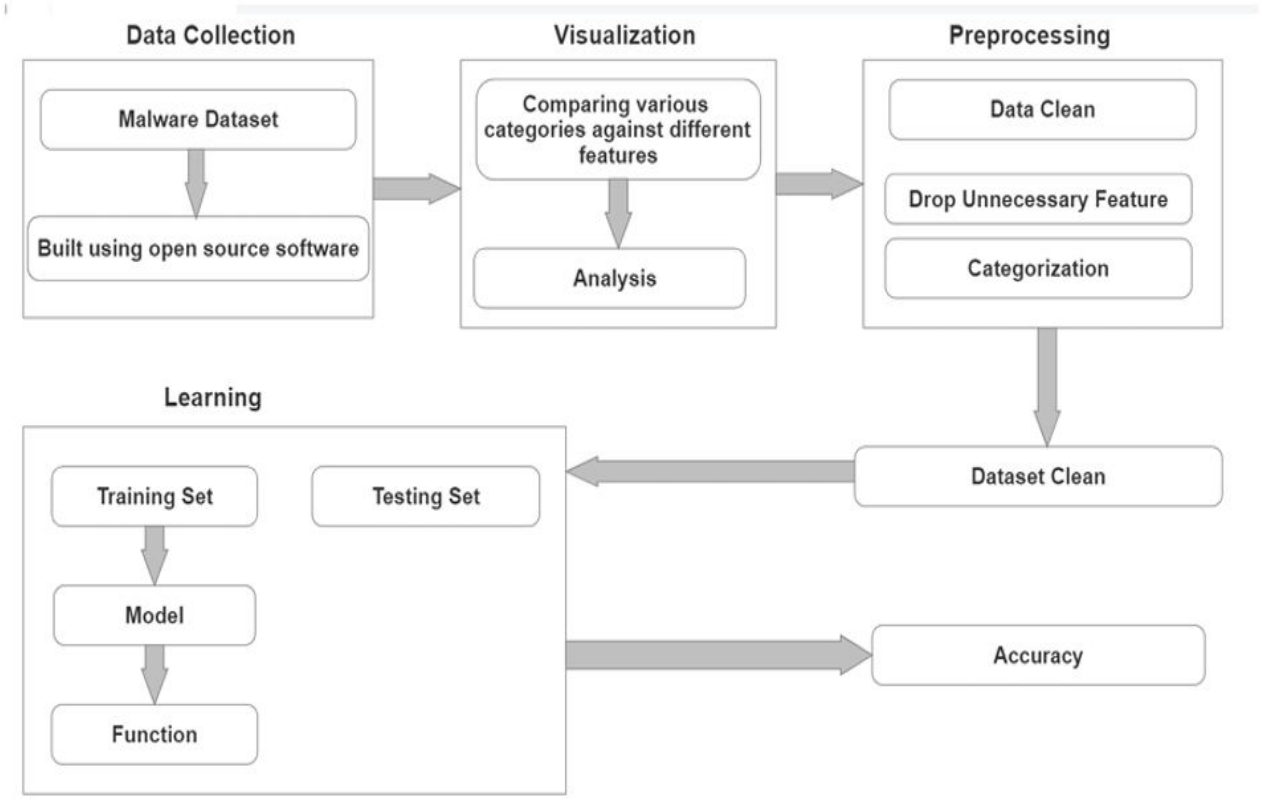


Fig1 System Architecture

1. Dataset Collection and Preparation Module:

- Collect PE file datasets and malicious/benign URL datasets from sources like Kaggle.
- Combine multiple datasets if needed.
- Handle file formatting and ensure consistency across datasets.
- Save preprocessed datasets for model training and testing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

2. **Feature Extraction Module:**

- Use libraries like pefile for PE file structure analysis.
- Extract features such as entropy, section count, imported DLLs, file size, etc.
- For URLs, extract lexical features like length, number of dots, use of symbols, domain type, and WHOIS info.
- Transform raw data into a structured format (CSV or DataFrame).
- Normalize or encode data where needed.

.3. **Data Preprocessing Module:**

- Handle missing or inconsistent data entries.
- Perform label encoding or one-hot encoding on categorical data.
- Normalize numerical values using techniques like MinMaxScaler or StandardScaler.
- Split the dataset into training and testing sets (typically 80:20).

4. **Machine Learning Model Module:**

- Implement K-Nearest Neighbors (KNN) algorithm.
- Implement Random Forest Classifier.
- Evaluate models using metrics: Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

5. **Web Application Interface Module (Flask):**

- Users can: Upload a PE file, Enter a URL for analysis, View the result as "Malicious" or "Legitimate".
- Backend handles file processing, feature extraction, and model prediction.
- Result is shown on the frontend with success/failure messages.

6. **System Testing and Output Module:**

- Conduct unit testing for individual components (e.g., file upload, prediction).
- Test with different file types and URLs to check for edge cases.
  Ensure the model gives correct classifications with minimal false positives.

IV. RESULT

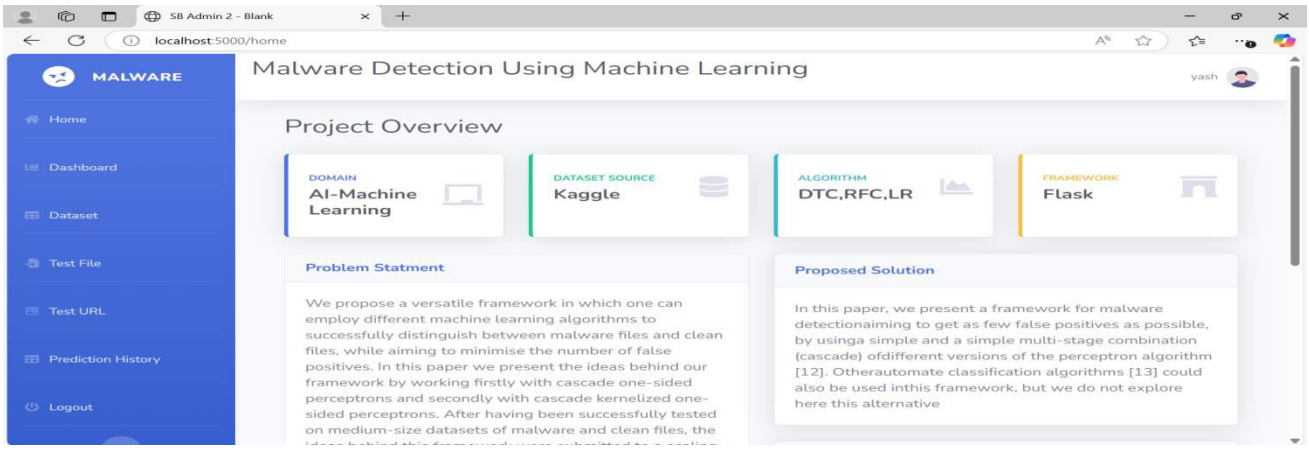


Fig1 Home Page

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

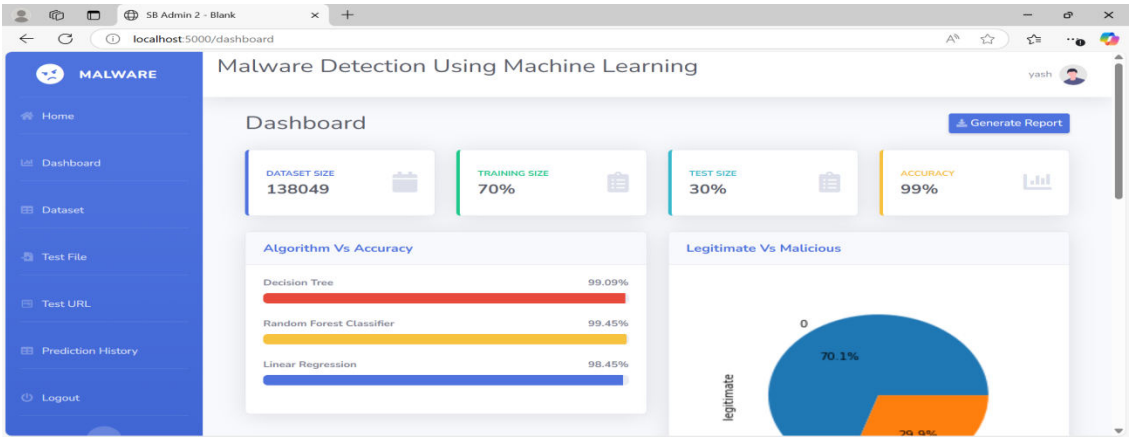


Fig2 Dashboard

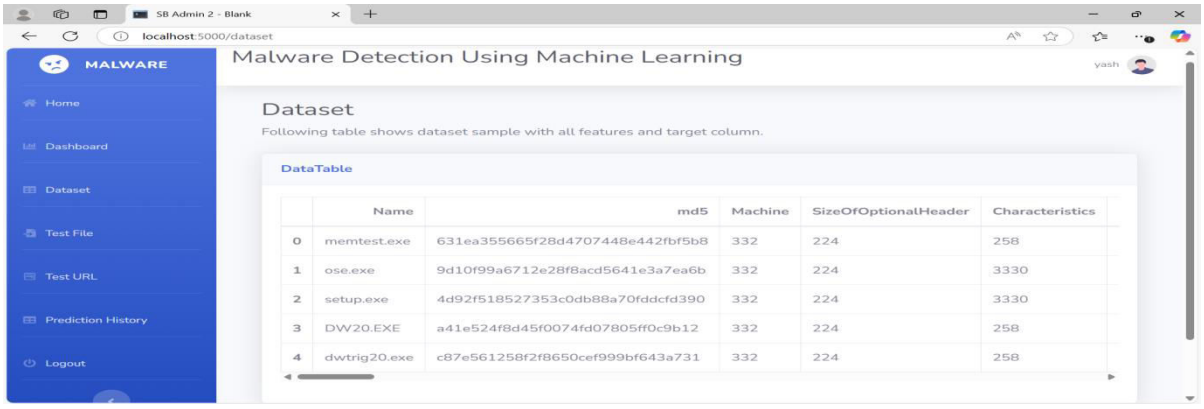


Fig3 Kaggle Dataset

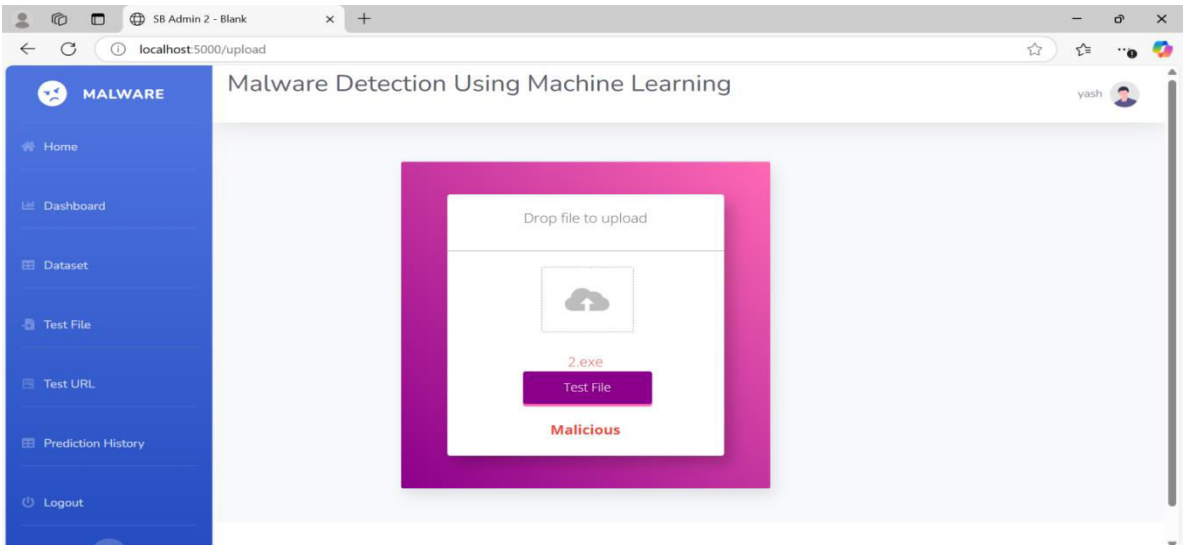


Fig4 PE File Classifier

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

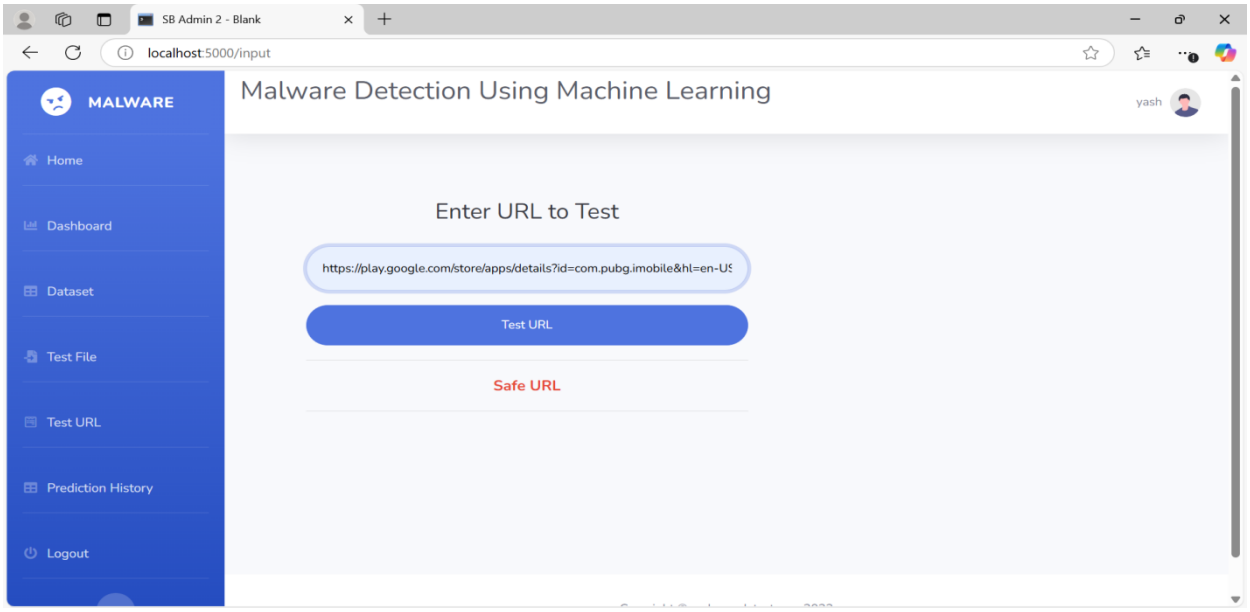


Fig5 URL Classifier

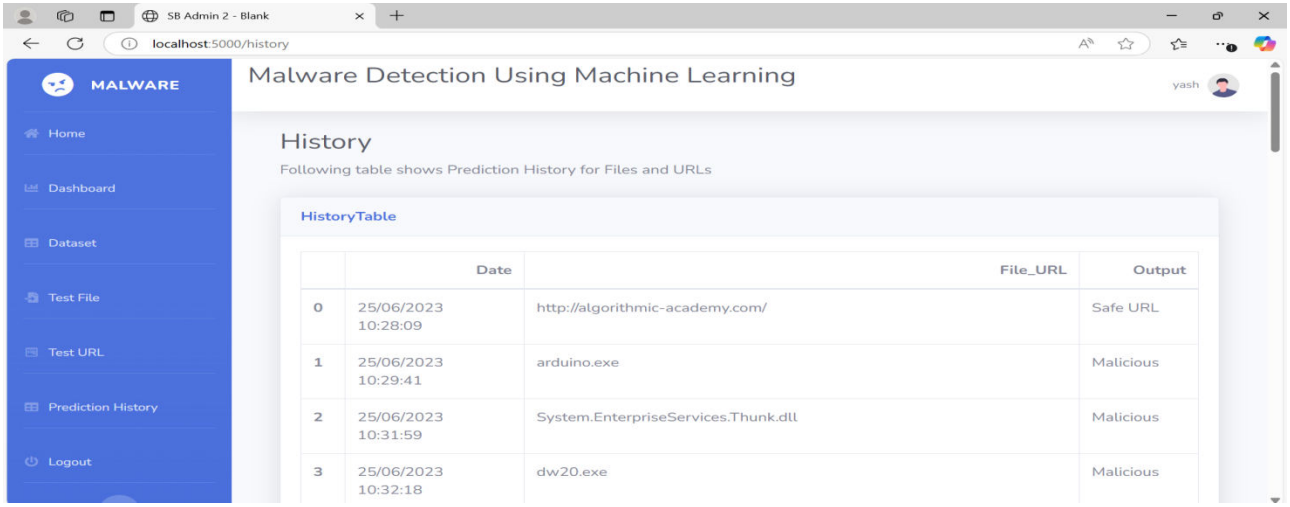


Fig6 Prediction History

V. CONCLUSION

The development of the Malware Detection System using machine learning techniques has demonstrated a highly effective and scalable solution for identifying and classifying malicious files and URLs. By integrating algorithms like K-Nearest Neighbors (KNN) and Random Forest, this system leverages the strengths of both distance-based and ensemble-based learning approaches to provide accurate and reliable predictions. Through comprehensive preprocessing and feature extraction from Portable Executable (PE) files and URLs, the system can detect suspicious characteristics that are typically invisible to the end user. These extracted features are then used to train robust models that can distinguish between malware and legitimate instances with high precision. The use of Random Forest helps improve accuracy through multiple decision trees and reduces overfitting, while KNN provides a simple yet powerful method to classify based on pattern similarity. The project also highlights the importance of having a user-friendly interface. By integrating the backend models with a Flask-based web application, users are given a seamless experience

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404516|

where they can upload a file or enter a URL and immediately receive a result. This practical integration of machine learning into a real-world web application demonstrates the true value of AI in cybersecurity. In conclusion, the system not only contributes to the growing need for intelligent malware detection tools but also proves that with the right dataset, features, and algorithm selection, detection accuracy can be significantly improved. The modular and scalable design of the project ensures that it can be expanded in the future with more sophisticated models, real-time threat analysis, and integration with antivirus platforms—making it a valuable foundation for further innovation in the field of cyber threat prevention.

ACKNOWLEDGMENT

We express our heartfelt gratitude to our esteemed Chairman, **Dr.S. Jagathrakshakan, M.P.,** for his unwavering support and continuous encouragement in all our academic endeavors.
We express our deepest gratitude to our beloved President **Dr. J. Sundeep Aanand** President**,** and Managing Director **Dr. E. Swetha Sundeep Aanand Managing** Director for providing us the necessary facilities to complete our project.
We take great pleasure in expressing sincere thanks to **Dr**. **K. VijayaBaskar Raju** Pro- Chancellor, **Dr. M. Sundararajan** Vice Chancellor (i/c), **Dr. S. Bhuminathan** Registrar and **Dr. R. Hariprakash** Additional Registrar**, Dr. M. Sundararaj** Dean Academics for moldings our thoughts to complete our project.
We thank our **Dr. S. Neduncheliyan** Dean, School of Computing for his encouragement and the valuable guidance.
We record indebtedness to our Head, **Dr. S. Maruthuperumal,** Department of Computer Science and Engineering for his immense care and encouragement towards us throughout the course of this project.
We also take this opportunity to express a deep sense of gratitude to our guide **MS.J. Lethisiya Nithiya** and our Project Co-Ordinator **Dr. K.V.Shiny** for their cordial support, valuable information, and guidance, they helped us in completing this project through various stages.
We thank our department faculty, supporting staff and friends for their help and guidance to complete this project.

REFERENCES

[1] G.D. Penna, L.D. Vita and M.T.Grifa, “MTA-KDD’19: A Dataset for Malware Traffic Detection“ in ITASEC- 2020.
[2] M. Gao, Li Ma, H. Liu, Z. Zhang, Z. Ning and J. Xu, “Malicious Network Traffic Detection Based on Deep Neural Networks and Association Analysis” in Sensors (Basel)- 6 March 2020.
[3] Sudarshan N P.Dass, “Malicious Traffic Detection System using Publicly Available Blacklist’s” in IEEE Conference of International Journal of Engi neering and Advanced Technology (IJEAT)- August 2019.
[4] Paul Prasse, Lukas Machlica, Tomas Pevny, Jiri Havelka and Tobias Scheffer, “Malware Detection by Analysing Network Traffic with Neural Networks” in IEEE Conference of Symposium on Security and Privacy Workshops- May 2017.
[5] NancyAgarwalandSyedZeeshanHussain, “A Closer Look at Intrusion Detection System for Web Applications” in IEEE Conference of Security and Com Communication Networks Volume- 2018.
[6] Gonzalo Marin, Pedro Casas, German Capdehourat, “ DeepMal- Deep Learning Models for Malware Traffic Detection and Classification” on 10 March 2020.
[7] Felipe N. Ducau, Ethan M. Rudd, Tad M. Heppner, Alex Long, Konstantin Berlin “ Automatic Malware Description via Attribute Tagging and Similarity Embedding ” on 15 May 2019.
[8] Luca Demetrio, Scott E. Coull, B. Biggio, G. Lagorio, A. Armando, Fabio Roli “A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection” on 17 Aug 2020.
[9] T. M. Mohammed, L.Nataraj, S. Chikkagoudar, S.Chandrasekaran, B. S. Manjunath “Malware Detection Using Frequency Domain-Based Image Visualization and Deep Learning” on 26 Jan 2021.
[10] Sanjay K. Sahay, C. Rama Krishna and Sanjay Sharmal, "Detection of Advanced Malware by Machine Learning Techniques", 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details