e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 9, September 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

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

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209021 |

Android Malware Detection Using Machine Learning Techniques

Mamatha M, Sandhya N

P.G. Student, Department of MCA, The National Institute of Engineering College, Mysore, Karnataka, India

Associate Professor, Department of MCA, The National Institute of Engineering College, Mysore, Karnataka, India

ABSTRACT: Growing mobile device use heightens Android malware risk, imperilling privacy. Our innovative method employs Genetic Algorithm-based extraction and ensemble learning for malware detection. We analyse 429 app features, focusing on permissions, refining to 302 crucial ones via Genetic Algorithm. Our approach deploys varied ML methods: ANN (92%), SVM (88%), Logistic Regression (88%), Decision Trees (88%), and ensembles (90%). We pinpoint essential attributes, bolstering mobile security and advancing anti-malware endeavours.

KEYWORDS: Android malware risk, Genetic Algorithm-based extraction, Malware detection, Mobile security, Anti-malware endeavours.

I. INTRODUCTION

With the increasing popularity and widespread adoption of mobile devices, the Android platform has grown into primary target for cybercriminals seeking to exploit vulnerabilities and compromise user privacy. The proliferation of Android malware presents a significant challenge to the security of users' personal information, financial data, and overall digital well-being. Detecting and mitigating such threats is of utmost importance to ensure a safe and secure mobile environment. In1.1 this research project, we address the critical issue using a fresh strategy that combines Genetic Algorithm-based feature extraction and ensemble learning techniques. The goal of our project is to create robust and efficient system capable of accurately identifying malicious apps among the vast pool of legitimate applications.

The foundation of our approach lies in permission analysis, as Android apps request various permissions to access system resources and perform specific actions. Malicious apps often seek excessive or unnecessary permissions, raising red flags for potential security risks. To harness the wealth of information hidden in the permission data, in this context, we extract an all-encompassing collection of 429 features, which meticulously represent the permissions sought by each application. Our approach employs an array of machine learning algorithms to fabricate a classifier. To refine the feature set, the Genetic Algorithm is harnessed. It employs an iterative evolution of potential solutions, progressively selecting the most enlightening attributes while tossing out those that are unnecessary or inappropriate. Consequently, this process culminates in a distilled feature set containing 302 pivotal traits, thereby ensuring an efficient and effective detection procedure.

In the pursuit of creating a versatile and precise classifier, we embrace a diverse array of machine learning algorithms. These encompass Artificial Neural Networks (ANN), Support Vector Machines (SVM), Logistic Regression, and Decision Trees. Furthermore, we tap into the potential of ensemble learning to amplify classification performance. This is achieved by amalgamating individual classifiers using techniques such as Random Forests (RF) and Stacking with Logistic Regression. This approach harnesses the amalgamated wisdom of distinct models while also mitigating their individual limitations.

II. APPROACH

The rapid proliferation of Android malware has emerged as a substantial hazard to the security and privacy of mobile device users. With malicious applications progressively advancing in their sophistication and elusiveness, the requirement for a resilient and streamlined Android malware detection system has become paramount in safeguarding users against potential cyber risks. Conventional detection techniques frequently encounter challenges in staying abreast of the ceaselessly evolving malware environment. This necessitates the adoption of innovative methodologies that possess the capability to precisely and swiftly distinguish malicious apps amidst the extensive array of genuine applications.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209021 |

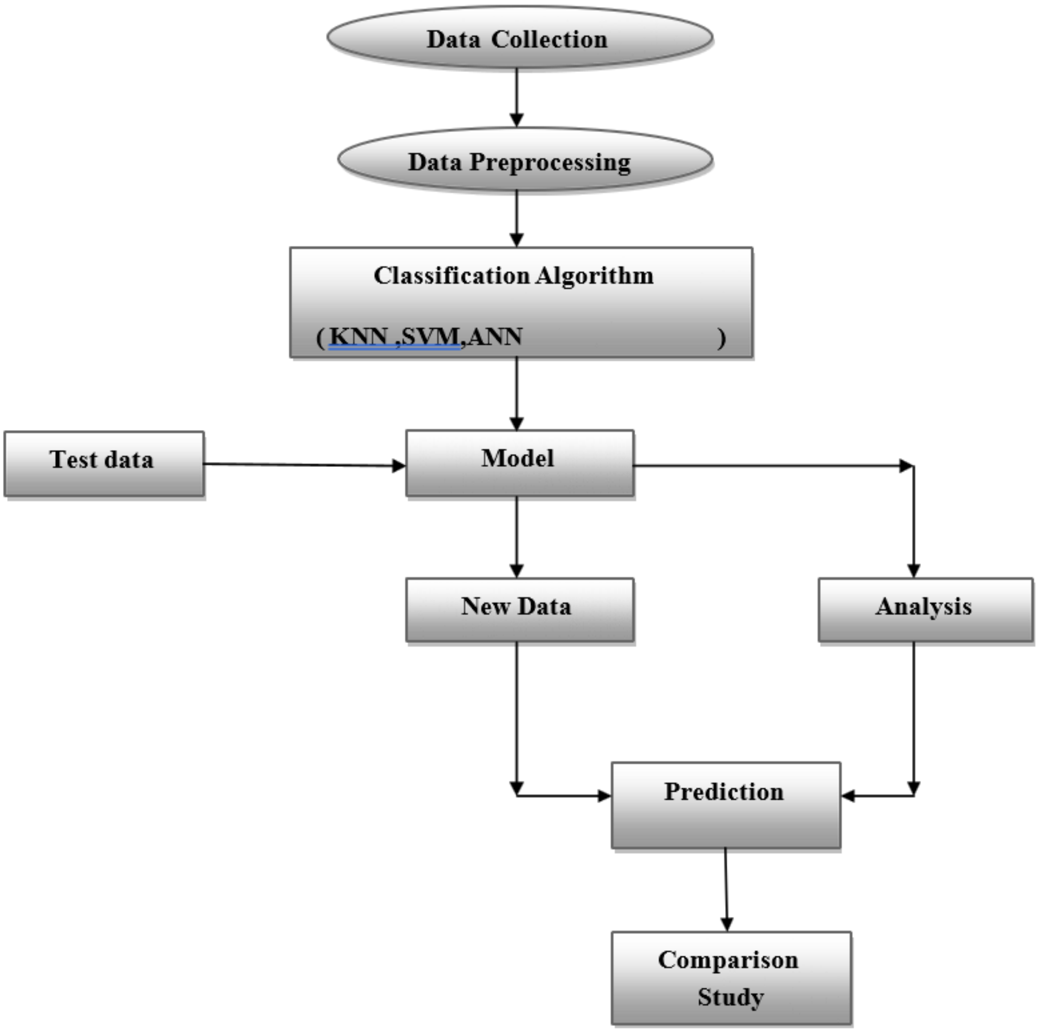
III. AIM

The primary objective of this research project is to create a sophisticated Android ma lware detection system that capitalizes on Genetic Algorithm-driven feature extraction and ensemble learning methodologies. The system's overarching goal is to heighten the precision and effectiveness of malware detection, ultimately furnishing users with an augmented level of security within their mobile interactions.

IV. OBJECTIVES

- To realize a comprehensive system by employing Genetic Algorithm-based feature extraction and ensemble learning techniques. To optimize the feature set using the Genetic Algorithm, reducing computational complexity and improving detection efficiency.
- To design, assess, and validate a varied ensemble that optimally employs machine learning classifiers for enhanced accuracy.
- To conduct rigorous experiments and performance evaluations using real-world Android app datasets.
- To conduct feature importance analysis, identifying the most critical indicators of malicious behaviours.

V. WORKFLOW DIAGRAM



|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209021 |

VI. RELATED WORK

In literature [1], a paper published in 2018, the authors conducted malware detection using 300 malware files and 300 benign APK files. However, only 183 malware samples and 300 benign grayscale images were successfully generated. The remaining 117 malware samples could not be converted into images due to corrupted APK files or the absence of the classes.dex file. The accuracy of the implemented algorithms was lower compared to expectations. The paper explored three classifier techniques: k-nearest neighbour (KNN), random forest (RF), and decision tree (DT).
Another paper [2], published in 2017, evaluated the performance of various machine learning algorithms, such as Naive Bayes, J48, random forest, multiclass classifier, and multilayer perceptron, for Android malware detection. The researchers developed a framework that utilized machine learning techniques to classify Android applications as either malware or normal applications. The study involved a dataset of 3258 Android app samples, and the models were trained and evaluated based on classification accuracy and processing time.

In literature [3], published in 2016, the authors proposed a Robotium program in an Android sandbox for automatic triggering and behaviour monitoring of Android applications. The program incorporated a UI identification automatic trigger program to interact with mobile applications in a meaningful order. They also attempted to build a decision model using behaviour data collected through the random forest algorithm. The model successfully determined whether an unknown application was malware and provided a confidence value, which was stored in their database.
Another paper [4], published in 2018, presented an Android malware detection system that utilized permissions, APIs, and various key app information as features for training and building a classification model. Machine learning techniques were employed to automatically distinguish between malicious Android apps (malware) and legitimate ones.

VII. METHODOLOGY

Data Collection: Gather a large dataset of Android apps, including both benign and malicious samples, along with the permissions they request. Feature Extraction: Extract a set of 429 features representing the permissions from the dataset.
Genetic Algorithm: Apply the Genetic Algorithm to optimize the feature set and reduce it to the most informative features.
Machine Learning Classifiers: Utilize diverse machine learning algorithms, such as Artificial Neural Networks, Support Vector Machines, Logistic Regression, and Decision Trees, to create individual classifiers.
Ensemble Learning: Combine the individual classifiers using ensemble methods like Random Forests and Stacking with Logistic Regression.
Model Evaluation: Evaluate the performance of the proposed system using various evaluation metrics on separate testing datasets.
Feature Importance Analysis: Conduct a thorough analysis of feature importance to identify critical indicators of malicious behaviour.

VIII. EXPERIMENTAL RESULTS

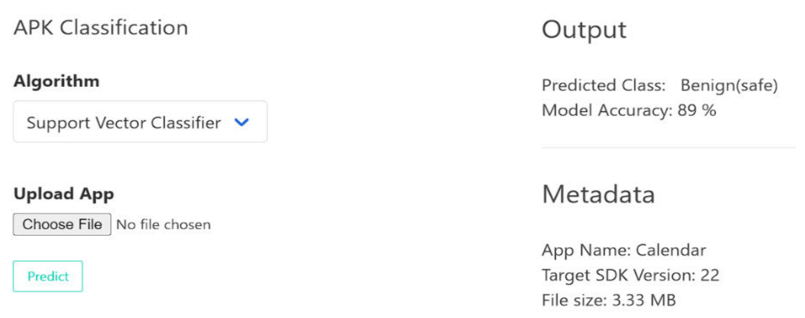


Fig 1

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

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209021 |

The above picture shows that, the apk file that is taken is considered as benign. It doesn't contain any malware features. Here SVM classifier is used for the prediction. Application name is Calendar and its size is 3.33MB. The accuracy level is 89%.

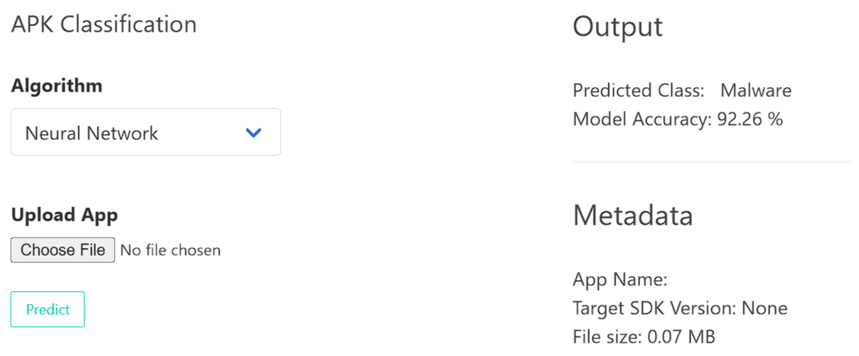


Fig 2

The above picture shows that, the apk file that is taken is considered as Malware. It contains some malware features might affect the system. Here Neural network is used for the prediction. Application name is android service and its size is 0.07 MB. The accuracy level is 92.26%.

IX. ADVANTAGES

- Enhanced Accuracy: The proposed system's ensemble approach ensures higher accuracy by leveraging the strengths of multiple classifiers.
- Efficiency: Genetic Algorithm-based feature extraction reduces the feature set, resulting in a more efficient and faster detection process.
- Adaptability: The ensemble learning approach allows the system to adapt to new and emerging malware threats effectively.
- Improved Security: By accurately detecting Android malware, the system contributes to a safer mobile environment for users.
- Insightful Analysis: The feature importance analysis provides valuable insights into the indicators of malicious behaviour, aiding future research and development in malware detection.

X. CONCLUSION

The Android malware detection system, rooted in Permission Analysis, demonstrated robust performance, notably showcasing strength through the Ensemble model that employed stacking. Employing the feature extraction genetic algorithm yielded positive results by effectively reducing the feature dimensions from an initial 429 down to 302. This reduction significantly enhanced efficiency and concurrently addressed the potential concern of overfitting, all while maintaining the system's accuracy intact. The method of feature extraction was critical in capturing essential patterns and improving the classifiers' ability to differentiate between malware and non-malware instances effectively. Throughout the project, we addressed potential challenges, such as class imbalance and dataset bias. By employing several evaluation metrics, such as recall and precision, and F1-score, we gained a comprehensive understanding of the models' performance and their ability to correctly classify both malware and non-malware samples. The combination of different classifiers and feature extraction using the incorporation of the Genetic Algorithm played a substantial role in elevating both the accuracy and efficiency of the system.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 9, September 2023 ||

| DOI:10.15680/IJIRSET.2023.1209021 |

REFERENCES

1. Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., & Rieck, K. (2014). DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket. Proceedings of the 21st Annual Network and Distributed System Security Symposium (NDSS).
2. Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., &Glezer, C. (2010). Andromaly: A Behavioural Malware Detection Framework for Android Devices. Journal of Intelligent Information Systems, 38(1), 161-190.
3. Liao, Q., & Wang, C. (2019). A Survey of Android Malware Detection and Defence Techniques. IEEE Access, 7, 17492-17514.
4. Dorigo, M., Maniezzo, V., &Colorni, A. (1996). The Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(1), 29-41.
5. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
6. Cortes, C., &Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
8. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
9. Hsu, C. W., & Lin, C. J. (2002). A Comparison of Methods for Multi-Class Support Vector Machines. IEEE Transactions on Neural Networks, 13(2), 415-425.
10. Wolpert, D. H. (1992). Stacked Generalization. Neural Networks, 5(2), 241-259.