Malware Detection through Memory Dump Analysis by Enhanced Machine Learning Techniques

Leki chom Thungon
Department of Computer Science
Vellore Institute Of Technology
Chennai, India
lekichom.thungon@vit.ac.in

Mandava Nidhish
Department of Computer Science
Vellore Institute Of Technology
Chennai, India
mandva.nidhish2021@vitstudent.ac.in

Dudekula Nahid Shameem
Department of Computer Science
Vellore Institute Of Technology
Chennai, India
dudekulanahid.shameem2021@vitstudent.ac.in

Kilaparthi Ravi Teja
Department of Computer Science
Vellore Institute Of Technology
Chennai, India
kilaparthiravi.teja2021@vitstudent.ac.in

Nagavarapu Divya Charitha
Department of Computer Science
Vellore Institute Of Technology
Chennai, India
nagavarapudivya.charitha2021@vitstudent.ac.in

***Abstract*—In this new era of the digital world, cybersecurity remains a critical concern, demanding frequent updates to counter sophisticated malware such as Trojan horses, Spyware and Ransomware. These threats have reached to high levels, outpacing traditional detection methods. To address this challenge effectively, we propose an innovative approach that combines Gradient Boosting and Random Forest classifiers in a ensemble method. Our study focuses mainly on analyzing the system's memory to identify malware, which often uses conventional detection mechanisms. The hybrid algorithm examines the patterns and behaviors within the system's volatile memory, significantly enhancing the accuracy and efficiency of malware detection. This approach involves collecting and analyzing memory information from diverse digital environments, combining both known and unknown malware scenarios. Initially, we explored the accuracy of four machine learning algorithms: Naive Bayes, Random Forest Classifier, Decision Tree, and Extreme Gradient Boosting. Among these, Random Forest and Extreme Gradient Boosting achieved an accuracy of 83.87% and 86.55% which was highest among the four machine learning models. Building on these results, we developed an ensemble method that blends Random Forest and Extreme Gradient Boosting classifiers. Remarkably, this blended ensemble method achieved the highest accuracy of 87.88%. In conclusion, our proposed approach not only achieves a balanced accuracy but also computational efficiency in malware detection. These findings represent a significant advancement in memory-based malware detection techniques against the evolving cyber threats.**

Keywords—Naïve Bayes, Extreme Gradient Boost, Random Forest, Decision Trees, Ensemble methods.

I. INTRODUCTION

In the rapidly evolving realm of cybersecurity, the evolution of malware presents an ever-growing threat to the integrity and security of digital systems worldwide. Among these, Trojan horses, Ransomware and spyware stand out for their cautious infiltration capabilities, capable of causing extensive damage to sensitive data and critical network infrastructure. Traditional approaches to malware detection often prove inadequate against these sophisticated threats, which makes necessity of innovative methodologies that can effectively counter their problems. Recognizing the critical need for enhanced malware detection techniques, this study sets out to develop an adaptive system that is capable of identifying and mitigating the risks posed by Trojan horses, Ransomware and spyware. Our primary hypothesis revolves around the efficacy of leveraging memory analysis and machine learning algorithms to strengthen the accuracy and efficiency of malware detection. The aim of this research is to design, implement, and validate a hybrid algorithm that harnesses the combined strengths of gradient boosting and random forest classifiers. These two powerful techniques have demonstrated exceptional features in handling complex datasets. By integrating these algorithms into our detection, we aim to enhance its resilience against a diverse spectrum of malware threats, with a particular emphasis on Trojan horses, Ransomware and spyware.

To achieve our objectives, we have outlined a comprehensive experimental design that encompasses several key phases. Firstly, we will acquire and pre-process real-world memory dump datasets containing instances of various malware types, ensuring a diverse and representative sample set. Next, we will implement the four machine learning algorithms that mentioned above and then design and implement the hybrid algorithm with random forest and extreme gradient boosting as they have achieved the highest accuracy, carefully fine-tuning its parameters to optimize performance and accuracy. Throughout the experimental process, we will employ with different ensemble models like voting classifier, stacked classifier and Blended models. The significance of this study lies in its potential to revolutionize existing malware detection methodologies, equipping organizations and security professionals with the tools and insights needed to combat evolving cyber threats effectively. By pushing the boundaries of traditional detection techniques and embracing the working togetherness of memory analysis and machine learning, we want to create stronger and smarter ways to protect against cyber threats as they keep changing.

II. LITERATURE REVIEW

The literature encompasses a series of research endeavors aimed at tackling the rising threat of malware and ransomware attacks in the digital landscape. One key focus revolves around leveraging machine learning techniques,

notably XGBoost models, to detect and combat ransomware variants such as Revil, Lockbit, and BlackCat. By constructing comprehensive datasets and delving into memory-based analysis, researchers have achieved significant milestones, boasting accuracies upwards of 97.85% with minimal false positive rates [1]. These studies highlight the paramount importance of memory traces in ransomware detection, while also underscoring challenges posed by malware obfuscation and the scarcity of adequate analysis environments.

In parallel, efforts have been directed towards addressing the intricate issue of obfuscated malware detection. Through the development of hybrid classification models like MalHyStack, which amalgamate conventional machine learning algorithms with deep learning layers, researchers have made considerable strides in enhancing accuracy rates [2]. By employing meticulous feature engineering techniques and conducting rigorous feature selection analyses, these models have demonstrated exceptional accuracies, often exceeding 99.98% for binary classification tasks and showcasing commendable performances in categorizing malware into various classes [4].

Furthermore, a significant emphasis has been placed on behaviour-based malware detection methodologies, pivotal in identifying and mitigating dynamic threats. Leveraging dynamic analysis environments and runtime features extracted from sandboxing platforms like Cuckoo, researchers have achieved remarkable accuracies using ensemble machine learning algorithms. The integration of diverse feature sets, including printable strings, Shannon entropy, and behavioral attributes, has led to accuracies surpassing 99.54% [3]. Additionally, investigations into machine learning and deep learning techniques have yielded promising results, with accuracies reaching 96% in classifying malware into distinct families [5]. These endeavors collectively underscore the importance of innovative approaches and advanced algorithms in confronting the evolving landscape of malware threats, fostering resilience and robustness in cybersecurity protocols.

The literature surveyed encompasses a variety of approaches to address the pervasive threat of malware across different domains, including computer systems, IoT devices, and mobile platforms. Researchers have recognized the evolving nature of malware, which encompasses viruses, spyware, bots, and ransomware, each posing unique challenges to system security [6]. Traditional signature-based methods, while effective to a certain extent, fall short in accurately detecting zero-day attacks and polymorphic viruses. In response, machine learning-based detection methods have gained prominence, offering modern and adaptive approach to malware detection. These methods, categorized into static, dynamic, or hybrid analysis techniques, have been explored extensively to enhance accuracy and effectiveness in detecting malware across various platforms [7].

Specific research efforts have focused on improving malware detection in specific environments, such as IoT devices running the Android operating system. With the proliferation of Android malware variants employing sophisticated detection avoidance techniques, including obfuscation, there's a growing need to enhance detection mechanisms. Novel approaches, such as leveraging convolutional neural networks (CNNs) trained on Markov images of app executables, demonstrate promising results in accurately detecting obfuscated malware and identifying obfuscation types [8]. The use of CNN models trained from Markov images generated using application byte code showcases a sustainable and cost-effective method for obfuscated malware detection, which surpasses traditional feature engineering-based approaches.

Moreover, research has explored machine learning techniques to detect malware in specific contexts, such as spear-phishing attacks utilizing macro malware written in VBA. Addressing the challenge of imbalanced datasets and practical performance evaluation, methods utilizing language models like Doc2vec and Latent Semantic Indexing (LSI) alongside popular classifiers have been proposed [9]. These methods mitigate class imbalance issues and exhibit robustness in detecting various types of macro malware, irrespective of the family type. Additionally, systematic literature reviews have been conducted to provide a comprehensive taxonomy of machine learning methods for malware detection. These reviews analyses a wide range of research works, categorize machine learning algorithms, and evaluate their performance to address existing challenges and guide future research directions in malware detection [10].

The literature review presented in the provided abstracts offers a comprehensive exploration of malware detection techniques using machine learning (ML) algorithms, particularly focusing on the Android platform and hardware-assisted approaches. The first set of abstracts discusses the effectiveness of various ML algorithms in detecting malware, emphasizing the importance of accurately identifying polymorphic malware that evolves to evade traditional signature-based detection [11]. Through experimental evaluations, the authors compare the performance of algorithms like Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM), highlighting their high detection accuracy rates and low false positive rates, which are crucial for effective malware detection.

Moving on to the context of Android mobile malware detection, the second set of abstracts delves into systematic reviews of ML-based techniques specifically tailored for Android devices [12]. Given the significant market share of Android smartphones, the authors underscore the vulnerabilities inherent in the open-source nature of the platform. They emphasize the evolution of malware threats targeting Android systems and the necessity of employing ML algorithms to effectively detect and mitigate these threats. Additionally, the abstracts highlight the importance of considering contextual features alongside API calls and permissions for improved detection accuracy [13].

Lastly, the literature includes abstracts focusing on hardware-assisted malware detection, acknowledging the

limitations of software-based solutions, such as antivirus programs, in combating sophisticated malware attacks. The authors explore the integration of machine learning with hardware performance counters, trace buffers, and on-chip network traffic analysis to enhance the speed and accuracy of malware detection [14]. This approach aims to address the evolving nature of malware threats and the need for real-time detection in safety-critical systems. Additionally, the authors indicate future directions for research, including the exploration of explainable machine learning for precise classification of benign and malicious programs and interpretation of detection results for identifying malicious behaviors [15].

The literature review provided offers a comprehensive overview of recent advancements in malware detection utilizing machine learning (ML) algorithms, particularly focusing on techniques applied to the Android platform and leveraging Hardware Performance Counters (HPCs) [16]. The first paper highlights the increasing utilization of HPC events by ML algorithms for malware detection, demonstrating the efficacy of Neural Network (NN) algorithms like Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Full Order Radial Basis Function (RBF) in achieving high accuracy rates. Furthermore, it identifies key features of HPCs exploited for recognizing abnormal activities and outlines future research directions, including the evaluation of proposed algorithms with real HPCs datasets [17].

Moving forward, the second paper introduces a novel approach for detecting different categories of Android malware by utilizing a Huffman encoding-based feature vector generation technique. This method, leveraging system call frequencies as features, significantly improves the efficiency of the detection model. Results demonstrate the superiority of the proposed model, particularly with the Random Forest classifier, in achieving a detection accuracy of 98.70% [18]. The paper also underscores the importance of incorporating dynamic features like system calls and suggests integrating static features such as permissions and API calls for future research.

Lastly, the third paper focuses on the evaluation of machine learning algorithms for Android malware detection using features extracted from Android manifest file permissions. It evaluates the effectiveness of various algorithms, with Random Forest emerging as the top performer in terms of precision, accuracy, and F1-score [19]. The paper emphasizes the inadequacy of commercial anti-virus tools and highlights the potential of machine learning algorithms in improving malware detection on Android platforms. Future research directions include the development of novel algorithms tailored to malware detection and the exploration of ensemble approaches and additional ML algorithms for static detection using manifest permission features [20].

III. PROPOSED METHODOLOGY

A. Dataset Description

The dataset used in this study is CIC-MalMem-2022. This is an academic dataset and it is published by Canadian Institute of Cybersecurity. The main purpose of this dataset is to detect the different types of malware using memory analysis. CIC-MalMem-2022 is a balanced dataset with a total of 58,596 records where half of them are benign and half of them are malicious. This dataset contains three different types of malwares: Trojan horse, Ransomware and Spyware. The extra information regarding the different types of malware families are mentioned in Table-1.

Table-1:

Malware Category	***Malware families***	***Count***
Benign	-	29,298
Trojan horse	Reconyc	1570
	Zeus	1950
	Emotet	1967
	Refroso	2000
	Scar	2000
Ransomware	Pysa	1717
	Maze	1958
	Conti	1988
	Ako	2000
	Shade	2128
Spyware	TIBS	1410
	180 Solutions	2000
	Cool web search	2000
	Gator	2200
	Transponder	2410
Total		58,596

B. Data Preprocessing

In the data preprocessing phase, we cleaned real-world memory dump dataset, ensuring the integrity and quality of the data. This involved removing the irrelevant or duplicate entries, handling missing values, and standardizing the data. Additionally, we conducted exploratory data analysis to gain insights into the underlying distribution and characteristics of the dataset, enabling us to make decisions regarding feature selection and model training. Furthermore, in the data processing stage, we implemented one-hot encoding to transform categorical variables such as malware types (Trojan horse, ransomware, spyware, and benign) into numerical representations, facilitating their incorporation into machine learning algorithms. We also employed feature scaling techniques to normalize the data, ensuring that all features contribute proportionally to the model's learning process. This meticulous data preprocessing and processing

steps are essential for optimizing the performance and accuracy of our hybrid algorithm in detecting and clearing malware threats effectively.

C. Machine Learning Algorithms

In the realm of malware detection via memory analysis, the employed methodology encompasses a diverse array of machine learning algorithms tailored to discern patterns indicative of malicious software behaviour. Naive Bayes, a probabilistic classifier, serves to model the likelihood of a sample's classification based on feature distributions. Decision trees, leveraging a tree-like structure, recursively partition feature space to facilitate classification decisions. Random Forest, an ensemble of decision trees, enhances robustness and generalization by aggregating predictions from multiple trees. Further fortification is achieved through ensemble techniques such as Voting Classifier, which amalgamates predictions from various models, and Stacking Classifier, which employs a meta-classifier to blend predictions from diverse base learners. The Blended Model integrates predictions from random forest and XG Boost models with distinct weightings, while the Stacked Classifier augments predictive power by feeding base learners' outputs into a higher-level model.

Naive Bayes: Naive Bayes offers a straightforward yet effective approach to malware detection through memory analysis. By assuming independence between features, it efficiently calculates the probability of a sample belonging to a particular class based on its feature distribution. Its simplicity and computational efficiency make it a viable option for real-time detection tasks, particularly in scenarios where resource constraints are a concern.

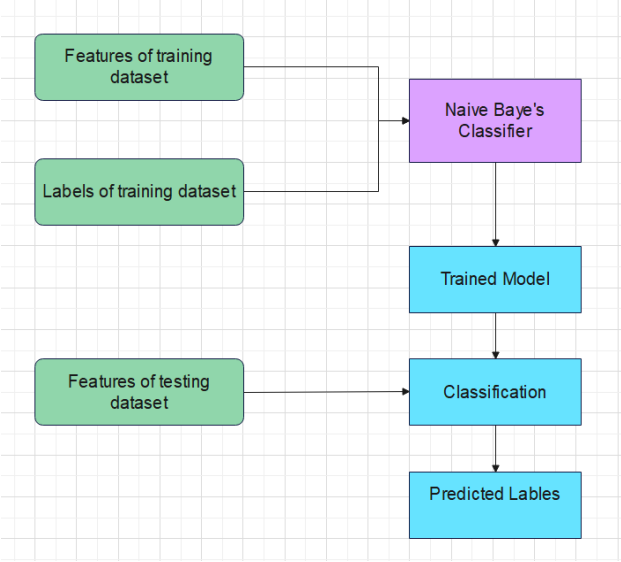

Figure-1: *Naïve Baye's Architecture*

Decision Trees: Decision trees serve as a versatile tool in the arsenal of malware detection methodologies. Through recursive partitioning of feature space, they delineate decision boundaries that separate benign from malicious behavior. Their interpretability aids in understanding the underlying logic of classification, facilitating insights into malware characteristics and aiding in feature selection. However, they may suffer from overfitting and lack robustness when dealing with complex, high-dimensional data.

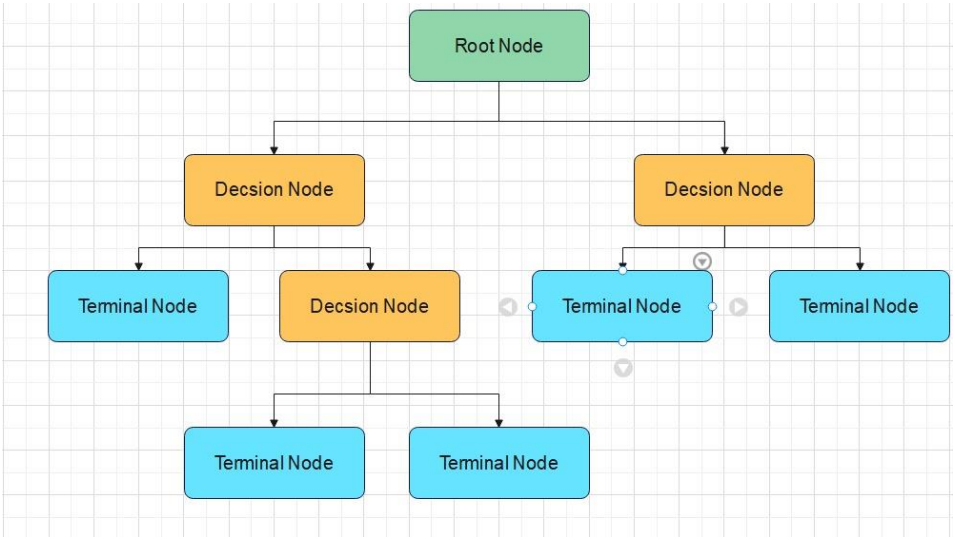

Figure-2: *Decision Tree Architecture*

Random Forest: Random Forest emerges as a powerful ensemble technique for malware detection through memory analysis. By aggregating predictions from a multitude of decision trees, it mitigates overfitting and enhances generalization capabilities. Its inherent robustness to noise and outliers makes it well-suited for handling the intricate and dynamic nature of malware behaviour. Furthermore, its ability to handle large datasets and parallelize computations expedites the detection process.

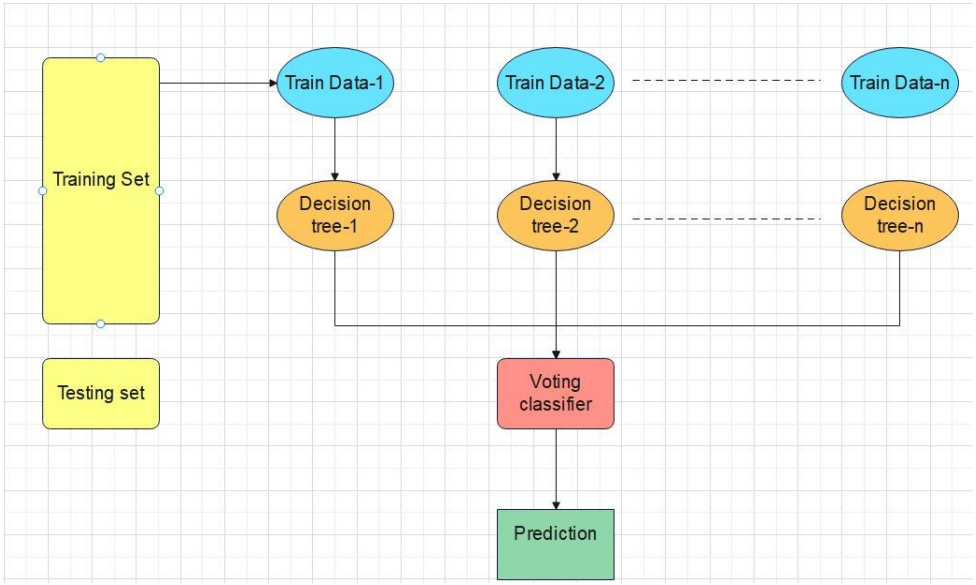

Figure-3: *Random Forest Architecture*

Extreme Gradient Boosting: Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm renowned for its efficiency and accuracy. It employs a boosting technique, combining weak learners to form a strong model. XGBoost minimizes loss functions using gradient descent optimization, enhancing predictive performance. Its parallelization capabilities make it highly scalable, fitting well with large datasets. With its regularization techniques, XGBoost effectively prevents overfitting, ensuring robust model generalization.

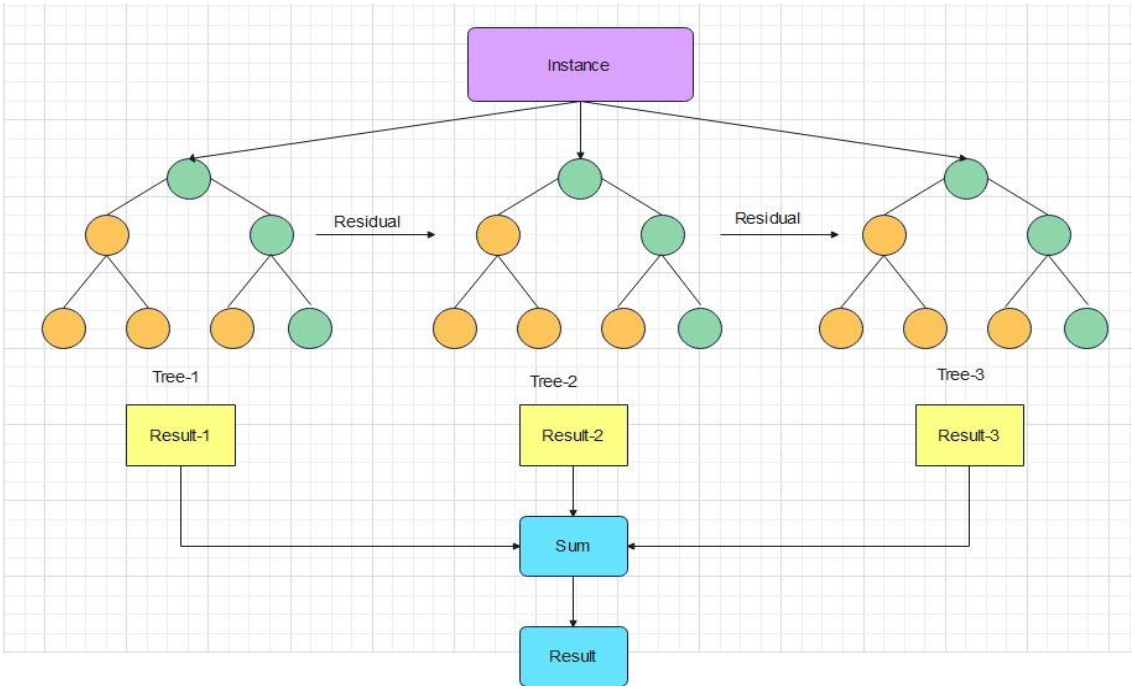

Figure-4: *Extreme Gradient Boost Architecture*

D. Ensemble method of RF and XGBoost

Ensemble Algorithms (Voting Classifier): Ensemble algorithms, such as the Voting Classifier, harness the collective wisdom of multiple base models to bolster malware detection efficacy. By combining predictions from diverse classifiers, it mitigates individual model biases and variance, resulting in more reliable classifications. This democratic approach ensures that each model's strengths contribute to the final decision, enhancing overall accuracy and robustness in detecting malicious software.

Ensemble Algorithms (Stacking Classifier): Stacking Classifier represents a sophisticated ensemble technique that orchestrates a hierarchical fusion of diverse base learners for malware detection. By training a meta-classifier on the predictions of multiple base models, it learns to capture higher-order relationships among their outputs, thus refining the decision-making process. This meta-learning paradigm enables the exploitation of complementary strengths across different models, culminating in enhanced detection performance and adaptability to evolving malware threats.

Blended Model: The Blended Model embodies a strategic amalgamation of diverse machine learning techniques tailored for malware detection via memory analysis. By blending predictions from multiple base models with varying strengths and characteristics, it harnesses the collective intelligence of each component to yield a more robust and accurate detection framework. Through careful weighting and combination strategies, it leverages the complementary nature of different algorithms, synergistically enhancing detection efficacy while mitigating individual model biases.

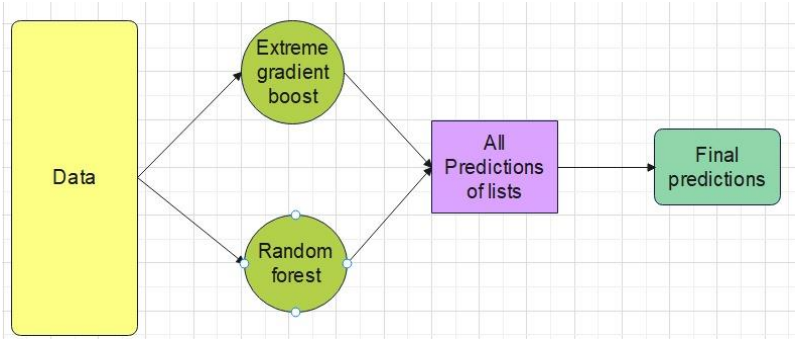

Figure-5: *Ensembled Algorithms Architecture*

E. Architecture of Project

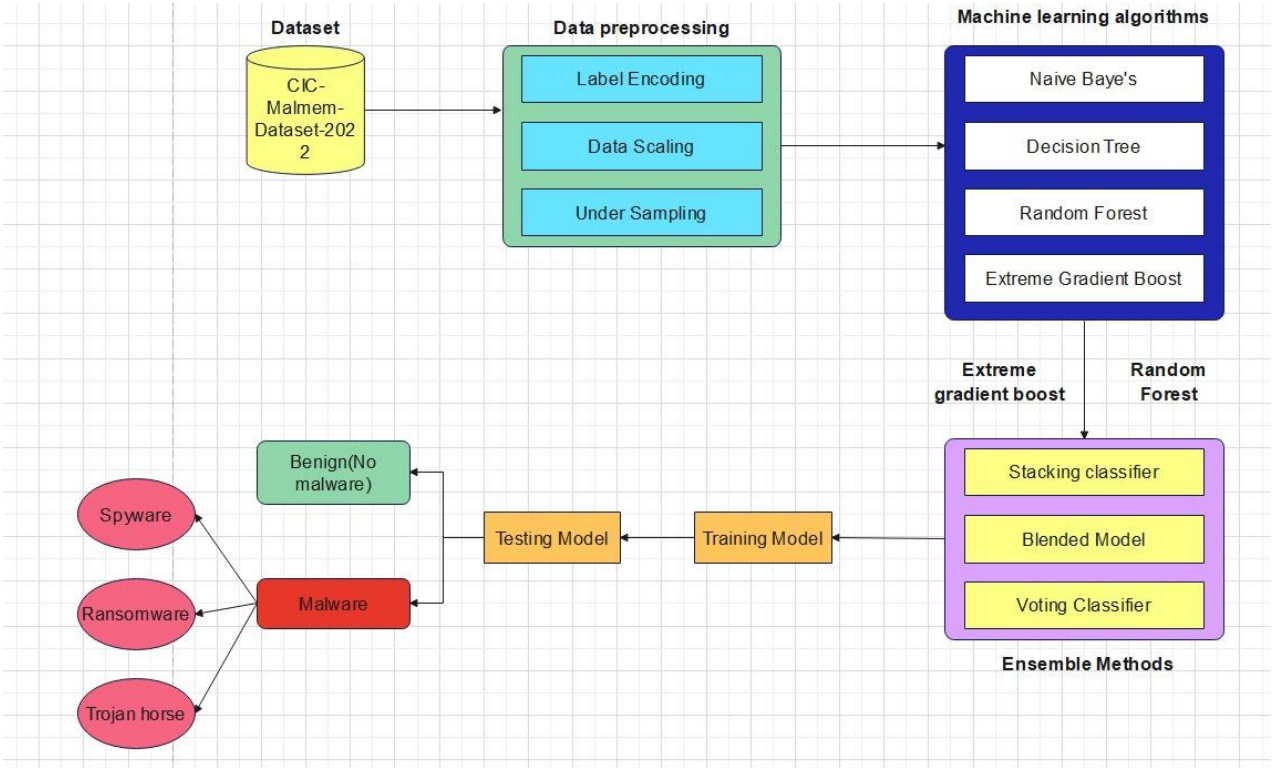

Figure-5: *Complete Architecture of the project*

IV. RESULTS AND DISCUSSION

A. For Individual Models

S. No	Model	Accuracy	Precision	Recall	F1-Score
1.	Gaussian Naïve Baye's	68.12	73.65	68.12	63.66
2.	Random Forest	83.87	82.38	81.87	81.93
3.	Decision Tree	81.64	81.88	81.64	81.61
4.	Extreme gradient boost	86.55	86.52	86.55	86.53

B. Ensemble Models

S. No	Ensemble method	Accuracy
1.	Voting classifier	87.29
2.	Stacking classifier (Logistic Regression)	87.73
3.	Blend Model	87.88
4.	Stacking classifier (Random Forest classifier)	85.67

V. CONCLUSION

In conclusion, our project represents a significant advancement in the field of malware detection using memory analysis, demonstrating the efficacy of a hybrid algorithm combining Random Forest and Extreme Gradient Boosting classifiers. Through precise experimentation and validation, we have found that our blended ensemble model achieves the highest accuracy of 87.88%, outperforming individual classifiers like naive bayes, decision trees, random forest and extreme gradient boost and alternative ensemble methods. The Voting Classifier and Stacking Classifier yielded commendable accuracies of 87.29% and 87.73%, respectively, further highlighting the effectiveness of ensemble techniques in enhancing prediction capabilities. In our research, Gaussian Naïve Bayes exhibited a lower accuracy of 68.12%, by this I got to know importance of employing ensemble algorithms for finding the complexities of malware detection. Our findings underscore the significance of leveraging machine learning and ensemble methods to fight against the evolving landscape of cyber threats effectively. Moving forward, our research also makes way for the development of more resilient and adaptive defense mechanisms, equipping organizations and security professionals with the tools needed to safeguard against malicious malware effectively.

REFERENCES

[1] Aljabri, M.S., Alhaidari, F.A., Albuainain, A., Alrashidi, S., Alansari, J., Alqahtani, W., & Alshaya, J. (2024). Ransomware detection based on machine learning using memory features. Egyptian Informatics Journal.

[2] Roy, K.S., Ahmed, T., Udas, P.B., Karim, M.E., & Majumdar, S. (2023). MalHyStack: A hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Intell. Syst. Appl., 20, 200283.

[3] Jia, L., Yang, Y., Tang, B., & Jiang, Z. (2023). ERMDS: A obfuscation dataset for evaluating robustness of learning-based malware detection system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations.

[4] Singh, J., & Singh, J. (2020). Detection of malicious software by analyzing the behavioral artifacts using machine learning algorithms. Inf. Softw. Technol., 121, 106273.

[5] Bokolo, B.G., Jinad, R., & Liu, Q. (2023). A Comparison Study to Detect
Malware using Deep Learning and Machine learning Techniques. 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI), 1-6.

[6] Al-Janabi, M., & Altamimi, A.M. (2020). A Comparative Analysis of Machine Learning Techniques for Classification and Detection of Malware. 2020 21st International Arab Conference on Information Technology (ACIT), 1-9.

[7] A, D.K., P, V., Yerima, S.Y., Bashar, A., David, A., T, A.N., Antony, A., Shavanas, A.K., & T., G.K. (2023). Obfuscated Malware Detection in IoT Android Applications Using Markov Images and CNN. IEEE Systems Journal, 17, 2756-2766.

[8] Mimura, M. (2020). An Improved Method of Detecting Macro Malware on an Imbalanced Dataset. IEEE Access, 8, 204709-204717.

[9] Gorment, N.Z., Selamat, A., Cheng, L.K., & Krejcar, O. (2023). Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges, and Future Directions. IEEE Access, 11, 141045-141089.

[10] Chowdhury, M.N., Haque, A., Soliman, H.S., Hossen, M.S., Fatima, T., & Ahmed, I. (2023). Android Malware Detection using Machine learning: A Review. ArXiv, abs/2307.02412.

[11] Akhtar, M.S., & Feng, T. (2022). Malware Analysis and Detection Using Machine Learning Algorithms. Symmetry, 14, 2304.

[12] Senanayake, J.M., Kalutarage, H.K., & Al-Kadri, M.O. (2021). Android Mobile Malware Detection Using Machine Learning: A Systematic Review. Electronics.

[13] AlJarrah, M.N., Yaseen, Q.M., & Mustafa, A.M. (2022). A Context-Aware Android Malware Detection Approach Using Machine Learning. Inf., 13, 563.

[14] Lee, J., Jang, H., Ha, S., & Yoon, Y. (2021). Android Malware Detection Using Machine Learning with Feature Selection Based on the Genetic Algorithm. Mathematics.

[15] Pan, Z., Sheldon, J., Sudusinghe, C., Charles, S., & Mishra, P. (2021). Hardware-Assisted Malware Detection using Machine Learning. 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 1775-1780.

[16] Bawazeer, O.M., Helmy, T., & Al-Hadhrami, S. (2021). Malware Detection Using Machine Learning Algorithms Based on Hardware Performance Counters: Analysis and Simulation. Journal of Physics: Conference Series, 1962.

[17] Manzil, H.H., & Manohar Naik, S. (2023). Android malware category detection using a novel feature vector-based machine learning model. Cybersecurity, 6, 1-11.

[18] McDonald, J.T., Herron, N., Glisson, W.B., & Benton, R. (2021). Machine Learning-Based Android Malware Detection Using Manifest Permissions. Hawaii International Conference on System Sciences.

[19] Agrawal, R., Shah, V., Chavan, S.S., Gourshete, G., & Shaikh, N. (2020). Android Malware Detection Using Machine Learning. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE),1-4.

[20] Harshalatha, P., & Mohanasundaram, R. (2020). Classification of Malware Detection Using Machine Learning Algorithms: A Survey. International Journal of Scientific & Technology Research, 9, 1796-1802.