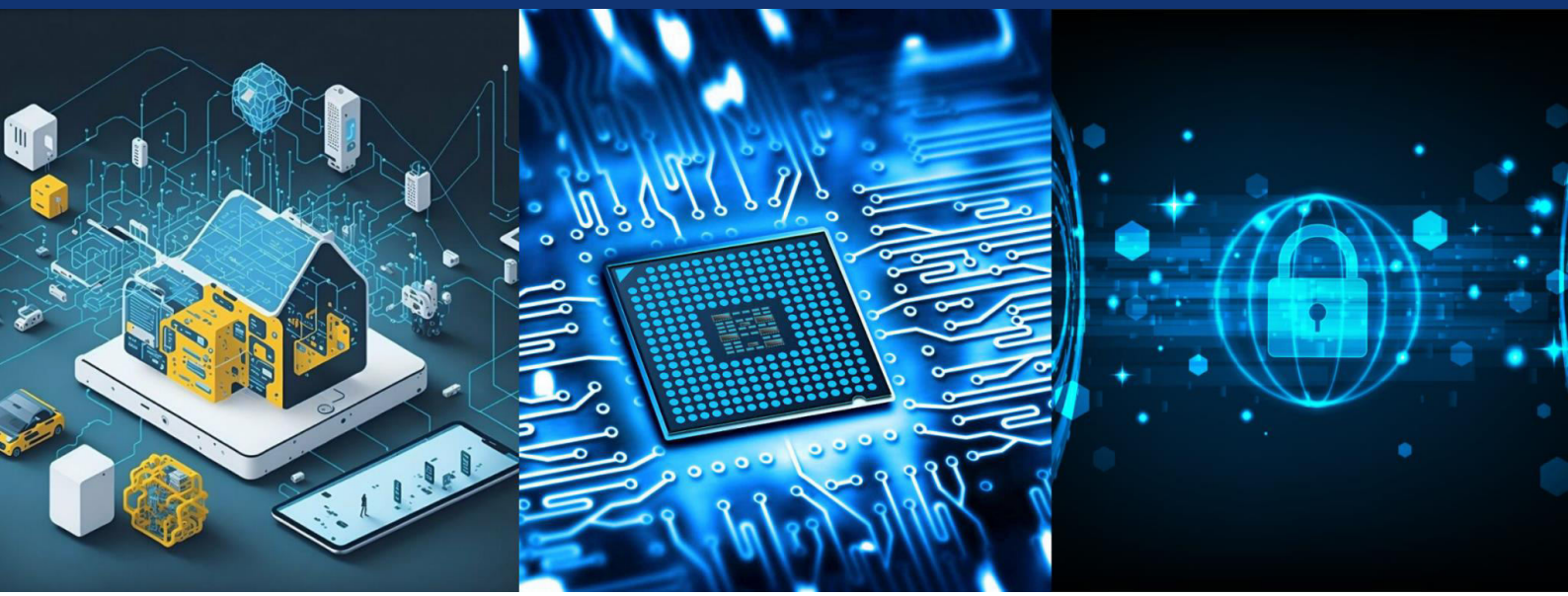
IJIRCCE

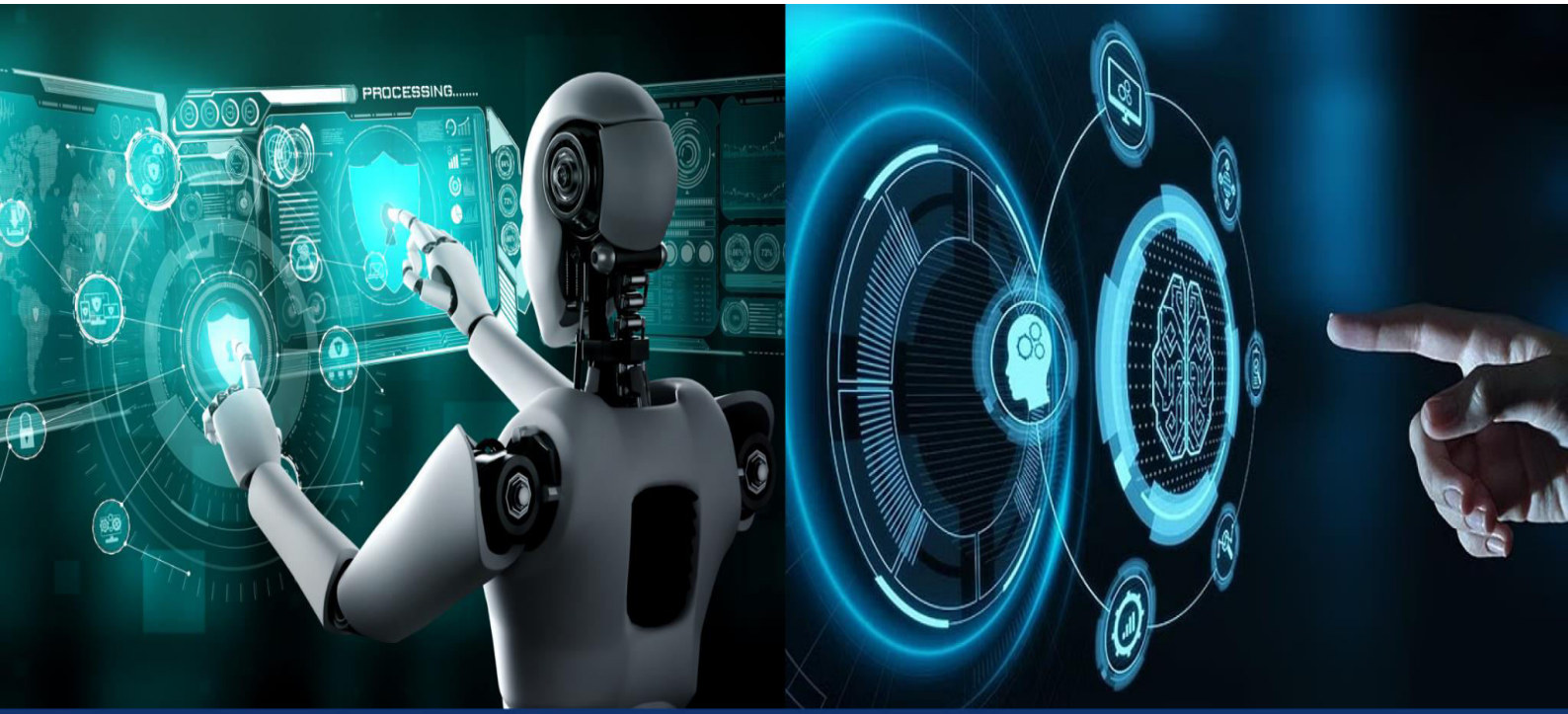
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Ensemble Learning for Enhanced Malware Detection

C Shiva Rama Krishna, Dr. K. Santhi Sree

Post-Graduate Student, Department of Computer Science Engineering, Computer Networks and Information Security, Jawaharlal Nehru Technological University, Hyderabad, India

Professor and Head, Department of Computer Science Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

ABSTRACT: Malware has become increasingly sophisticated, making traditional signature-based detection insufficient for identifying new and obfuscated threats. This paper presents an ensemble learning-based malware detection system using Bagging (Random Forest), Boosting (AdaBoost and Gradient Boosting), and Stacking techniques. The dataset contains 100,000 Android application samples equally divided into malware and benign classes. Data preprocessing includes label encoding, feature scaling, and removal of irrelevant attributes. Experimental results demonstrate that ensemble learning improves accuracy, precision, recall, and F1-score, providing a robust solution for modern cybersecurity applications.

KEYWORDS: Malware Detection, Ensemble Learning, Random Forest, AdaBoost, Gradient Boosting, Stacking

I. INTRODUCTION

Mobile The rapid growth of digital technology and internet usage has significantly increased cybersecurity threats, with malware becoming one of the most critical challenges. Malware is designed to disrupt systems, steal sensitive information, and gain unauthorized access, leading to financial and operational losses. Traditional malware detection methods mainly rely on signature-based techniques, which are effective against known threats but often fail to detect new and sophisticated malware variants such as zero-day, polymorphic, and metamorphic malware.

To overcome these limitations, machine learning techniques have been introduced to identify malicious patterns and behaviors more effectively. However, individual machine learning models may suffer from issues such as overfitting and lower accuracy. In this project, ensemble learning techniques including Bagging (Random Forest), Boosting (AdaBoost and Gradient Boosting), and Stacking are used to develop an efficient malware detection system. The models are trained and evaluated using a preprocessed Android malware dataset, and their performance is measured through accuracy, precision, recall, and F1-score, demonstrating the effectiveness of ensemble learning in modern malware detection.

II. RELATED WORK

A. Advanced Ensemble Learning Techniques for Classification

Ensemble learning has emerged as an effective approach for improving classification performance by combining multiple machine learning models. Singh and Chauhan (2018) discussed various ensemble techniques such as Bagging, Boosting, and Stacking, highlighting their ability to reduce bias, variance, and prediction errors [1]. Their work demonstrated that ensemble models provide better generalization and classification accuracy than individual machine learning algorithms. These findings motivated the use of ensemble learning techniques for malware detection in this project.

B. Deep Learning-Based Ensemble Models for Malware Detection

Ficco (2021) proposed a malware detection framework based on deep learning ensembles that combined multiple neural network models to improve classification performance [2]. The study showed that ensemble approaches could effectively identify complex malware patterns and achieve higher detection accuracy compared to single-model

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

methods. However, increased computational complexity and training time were identified as key challenges. This work highlights the importance of combining multiple models to enhance malware detection reliability.

C. Ensemble-Based Classification Using Machine Learning Models

Damaševičius et al. (2021) introduced an ensemble-based malware detection approach that integrated neural networks with traditional machine learning classifiers [3]. Their study demonstrated that combining models such as Random Forest and Decision Trees significantly improved classification accuracy and robustness when dealing with high-dimensional malware datasets. The research confirmed that ensemble learning can effectively handle complex cybersecurity classification tasks.

D. Stacking-Based Framework for Android Malware Detection

Wang et al. (2022) proposed a stacking ensemble framework called MFDroid for Android malware detection [4]. The framework combined multiple base learners and utilized a meta-classifier to improve prediction accuracy. Experimental results showed that stacking techniques reduced classification errors and enhanced malware detection performance. Their work emphasized the importance of carefully selecting base classifiers and meta-learners for achieving optimal results.

E. Ensemble Learning for Malware Detection in Cloud and IoT Environments

Recent studies have extended ensemble learning techniques to cloud and IoT security applications. Li et al. (2021) combined Bagging and Boosting methods for malware detection in cloud environments, achieving high detection accuracy and scalability [5]. Similarly, Vasan et al. (2020) proposed a stacked deep learning ensemble for IoT malware detection, demonstrating superior performance in identifying sophisticated malware variants [6]. These studies confirm that ensemble learning provides a robust and adaptable solution for modern malware detection systems and inspired the development of the proposed malware detection framework.

III. PROPOSED ALGORITHM

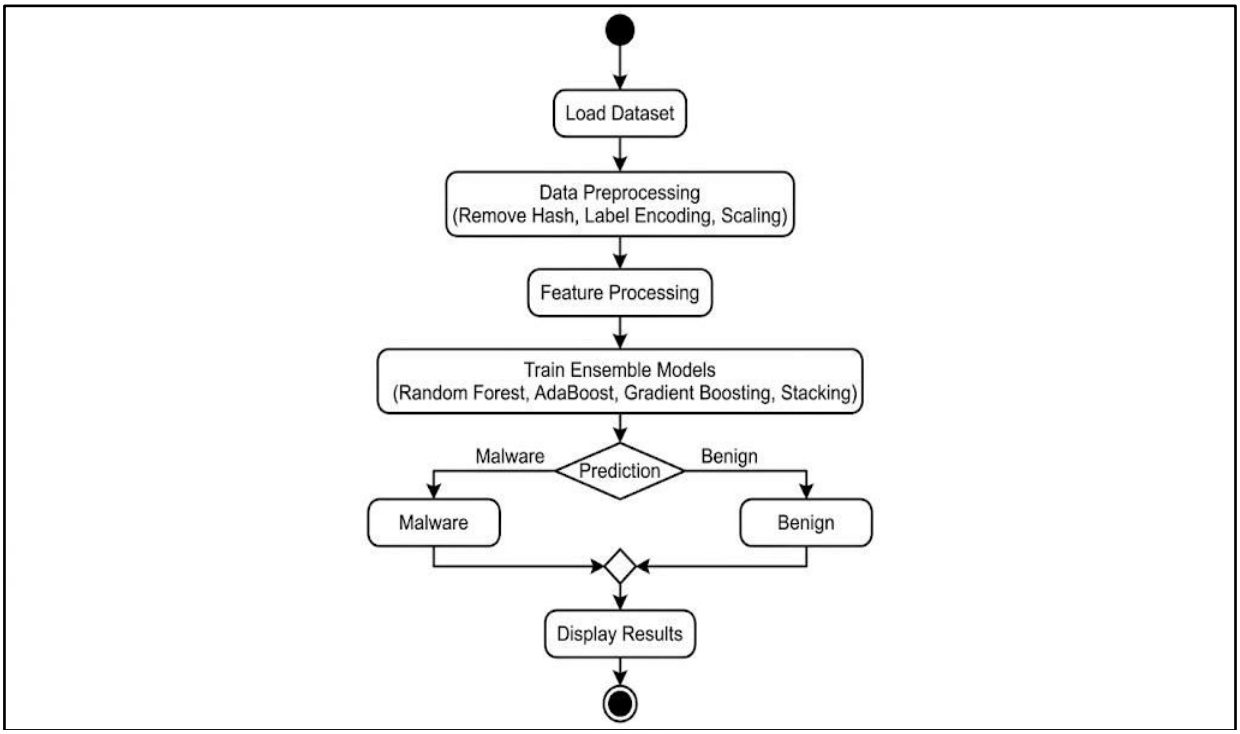


Figure 1: proposed Architecture

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

1) A. Design Considerations

• The malware dataset contains 100,000 Android application samples, equally divided into 50,000 malware and 50,000 benign samples.
• Irrelevant attributes such as hash values are removed during preprocessing.
• Label Encoding is used to convert classification labels into numerical values.
• StandardScaler is applied to normalize feature values and improve model performance.
• The dataset is divided into training and testing sets using a 70:30 ratio.
• Ensemble learning techniques including Bagging, Boosting, and Stacking are used for malware classification.
• Performance is evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

2) B. Description of the Proposed Algorithm

The objective of the proposed algorithm is to accurately classify software samples as Malware or Benign by utilizing ensemble learning techniques. The proposed methodology consists of three major steps.

3) Step 1: Data Preprocessing

The dataset is loaded and cleaned by removing irrelevant features such as hash values. The target class is converted into numerical form using Label Encoding.

Classification = {1, Malware ; 0, Benign} ..... (1)

Feature scaling is then performed using StandardScaler:

$$X_{scaled} = (X - \mu) / \sigma \quad \text{..... (2)}$$

where:
X = Original feature value
μ = Mean of feature values
σ = Standard deviation of feature values

4) Step 2: Ensemble Model Training

The preprocessed dataset is divided into training and testing datasets. Multiple ensemble learning models are trained:
• Bagging – Random Forest Classifier
• Boosting – AdaBoost and Gradient Boosting Classifiers
• Stacking – Decision Tree and Random Forest as base learners with Logistic Regression as the meta-learner.

The prediction function of each model is represented as:

$$\hat{y} = f(X) \quad \text{..... (3)}$$

where:
X = Input feature vector
ŷ = Predicted class label
f(X) = Trained ensemble model

5) Step 3: Performance Evaluation

The trained models are evaluated using standard classification metrics.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad \text{..... (4)}$$

$$Precision = TP / (TP + FP) \quad \text{..... (5)}$$

$$Recall = TP / (TP + FN) \quad \text{..... (6)}$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad \text{..... (7)}$$

where:
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

The model with the highest performance metrics is selected as the final malware detection model. Experimental results indicate that the Stacking ensemble model provides superior classification performance and reliable malware detection.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

IV. PSEUDO CODE

Step 1: Load the malware dataset.

Step 2: Preprocess the dataset.
- a) Remove irrelevant attributes (e.g., hash value).
- b) Apply Label Encoding on class labels.
- c) Normalize features using StandardScaler.

Step 3: Split the dataset into Training Set (70%) and Testing Set (30%).

Step 4: Train the Bagging Model.
- a) Initialize Random Forest classifier.
- b) Train the model using training data.
- c) Predict test samples.

Step 5: Train the Boosting Models.
- a) Initialize AdaBoost classifier.
- b) Train the model and predict test samples.
- c) Initialize Gradient Boosting classifier.
- d) Train the model and predict test samples.

Step 6: Train the Stacking Model.
- a) Define base learners:
  - Decision Tree
  - Random Forest
- b) Define meta learner:
  - Logistic Regression
- c) Train stacking classifier.
- d) Predict test samples.

Step 7: Evaluate all ensemble models.
- a) Calculate Accuracy.
- b) Calculate Precision.
- c) Calculate Recall.
- d) Calculate F1-Score.
- e) Generate Confusion Matrix.

Step 8: Compare performance metrics of all models.

Step 9: Select the model with highest accuracy and lowest classification error.

Step 10: Classify unknown Android applications as Malware or Benign using the selected model.

Step 11: End.

V. SIMULATION RESULTS

The simulation studies were conducted using an Android Malware Dataset containing **100,000 application samples**, equally divided into **50,000 malware** and **50,000 benign** applications. The proposed ensemble learning framework was implemented using Python and Scikit-learn. During preprocessing, irrelevant attributes were removed, class labels were

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

encoded into numerical values, and feature scaling was applied to normalize the dataset. The processed dataset was then divided into training and testing sets using a 70:30 ratio for model development and evaluation.

The proposed framework was evaluated using ensemble learning techniques including **Random Forest, AdaBoost, Gradient Boosting, and Stacking**. Performance was measured using accuracy, precision, recall, F1-score, and confusion matrix analysis. The experimental results show that all ensemble models achieved high classification performance in distinguishing malware from benign applications. Among the evaluated models, the stacking approach provided the most consistent and reliable results by combining the strengths of multiple classifiers. The findings demonstrate that ensemble learning significantly improves malware detection capability, reduces classification errors, and provides a robust solution for identifying both known and previously unseen malware samples.

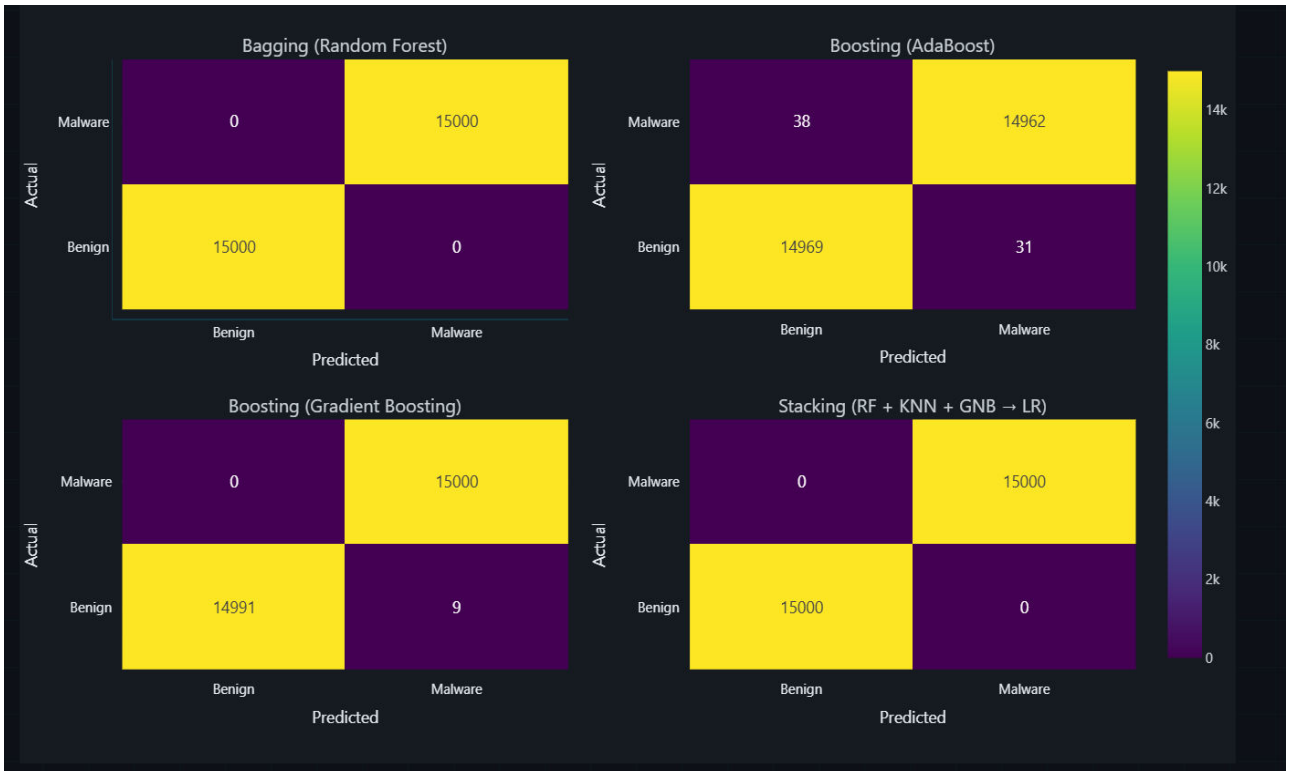


Figure 2: Confusion Matrix Comparison of Ensemble Learning Models (Random Forest, AdaBoost, Gradient Boosting, and Stacking) for Malware Detection.

The confusion matrices illustrate the classification performance of the ensemble learning models, showing the distribution of correctly and incorrectly classified malware and benign samples. The results indicate that the proposed ensemble approaches achieve highly accurate malware detection with very low misclassification rates, demonstrating their effectiveness in distinguishing malicious applications from legitimate ones.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

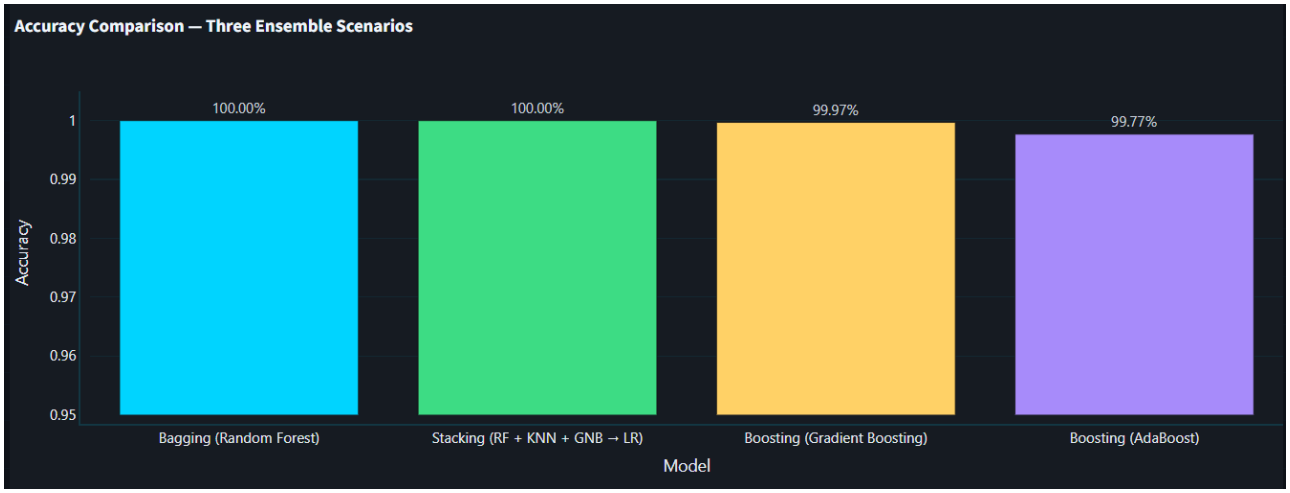


Figure 3: Accuracy Comparison of Ensemble Learning Models for Malware Detection.

The graph compares the accuracy achieved by different ensemble learning techniques, including Random Forest, AdaBoost, Gradient Boosting, and Stacking. The results show that all models achieved exceptionally high accuracy, with Random Forest and Stacking providing the best performance, demonstrating the effectiveness of ensemble learning for malware classification.

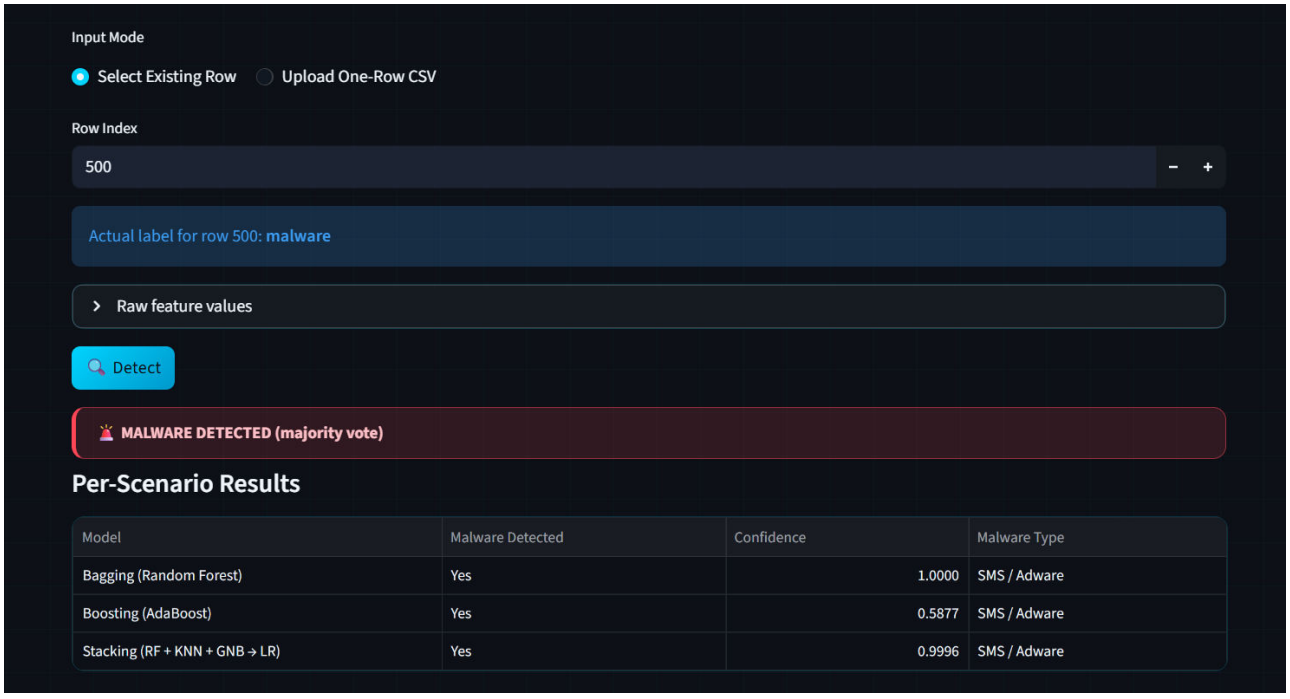


Figure 4: Malware Detection Result Using Ensemble Learning Models.

The figure illustrates the prediction output of the proposed malware detection system for a selected application sample. All ensemble models, including Random Forest, AdaBoost, and Stacking, successfully identified the sample as malware and provided confidence scores, demonstrating the effectiveness of the ensemble-based classification approach.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406042

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

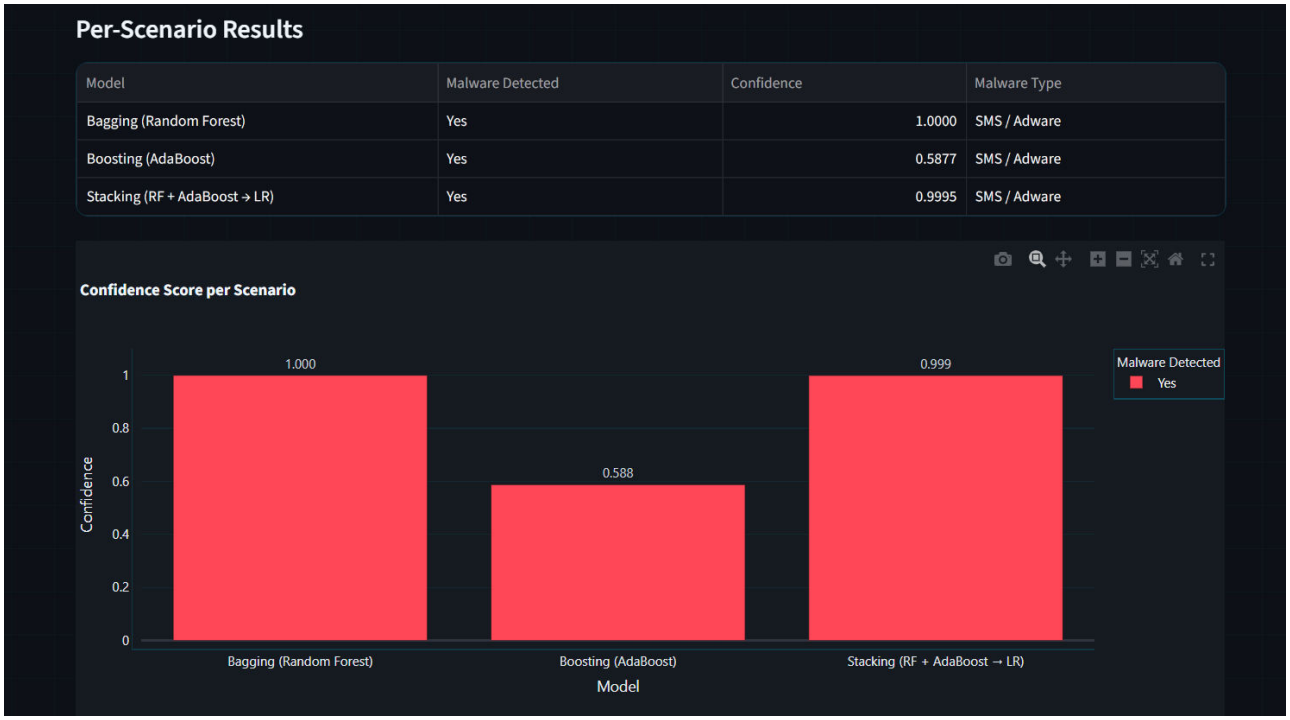


Fig. 5: Per-Scenario Classification Results and Confidence Scores of Ensemble Learning Models for Malware Detection

Model	Accuracy	Precision	recall	F1 score
Stacking	1	1	1	1
RandomForest (Bagging)	1	1	1	1
Gradient Boosting	0.996	0.93	1	0.96
AdaBoost	0.999	0.98	0.92	0.990

Table 1: Performance Comparison of Ensemble Learning Models for Malware Detection

The table presents the performance metrics of the ensemble learning models used for malware detection. The results show that **Stacking** and **Random Forest (Bagging)** achieved perfect classification performance with 100% accuracy, precision, recall, and F1-score, while **Gradient Boosting** and **AdaBoost** also demonstrated excellent detection capability with very high evaluation scores, confirming the effectiveness of ensemble learning in malware classification.

VI. CONCLUSION AND FUTURE WORK

Ensemble Learning for Enhanced Malware Detection provides an effective and intelligent approach for identifying malicious software by combining multiple machine learning models within a unified framework. The proposed system utilizes ensemble learning techniques such as Random Forest (Bagging), AdaBoost, Gradient Boosting, and Stacking to improve malware classification accuracy and reliability. Through data preprocessing, feature scaling, model training, and performance evaluation, the framework efficiently analyzes Android application data and accurately classifies samples as either malware or benign. The use of ensemble techniques enables the system to overcome the limitations of individual classifiers and achieve better generalization across diverse malware patterns.

The experimental results demonstrate the effectiveness of the proposed framework, with Random Forest and Stacking achieving excellent classification performance and very low misclassification rates. High values of accuracy, precision, recall, and F1-score confirm the system's capability to detect both known and previously unseen malware samples.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Furthermore, the modular design of the framework enhances scalability, maintainability, and future expansion. Overall, the proposed ensemble learning-based malware detection system offers a reliable, accurate, and practical solution for modern cybersecurity applications, highlighting the potential of ensemble machine learning techniques in strengthening malware detection and improving digital security.

REFERENCES

1. Loubna Moujoud, Meryeme Ayache and Abdelhamid Belmekki, 'Enhancing Malware Detection through Ensemble Learning Techniques', Cluster Computing, Vol. 28, pp. 477, 2025.
2. Massimo Ficco, 'A Malware Detection Deep Learning Ensemble', Journal of Supercomputing, Vol. 77, pp. 13745–13768, 2021.
3. Robertas Damaševičius, Andrius Venčkauskas, Justinas Toldinas and Saulius Grigaliūnas, 'Ensemble-Based Classification Using Neural Networks and Machine Learning Models for Malware Detection', Electronics, Vol. 10, Issue 4, pp. 485, 2021.
4. Jun Yan, Yibin Qi and Qifan Rao, 'Detecting Malware with an Ensemble Method Based on Deep Neural Network', Security and Communication Networks, 2018.
5. Mohit Dhalaria and Ekta Gandotra, 'Android Malware Detection Using Feature Selection and Ensemble Learning Method', Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, pp. 36–41, 2020.
6. Xin Wang, Lingyu Zhang, Kai Zhao, Xiaojun Ding and Meng Yu, 'MFDroid: A Stacking Ensemble Learning Framework for Android Malware Detection', Sensors, Vol. 22, Issue 7, pp. 2597, 2022.
7. Sheng Li, Yutong Li, Wei Han, Xiaolong Du, Mohsen Guizani and Zhen Tian, 'Malicious Mining Code Detection Based on Ensemble Learning in Cloud Computing Environment', Simulation Modelling Practice and Theory, Vol. 113, pp. 102391, 2021.
8. Dhivya Vasan, Mamoun Alazab, Sitalakshmi Venkatraman, Junaid Akram and Zheng Qin, 'MTHAEL: Cross-Architecture IoT Malware Detection Based on Neural Network Advanced Ensemble Learning', IEEE Transactions on Computers, Vol. 69, Issue 11, pp. 1654–1667, 2020.
9. Shahid H. Khan, Turki J. Alahmadi, Waqar Ullah, Javed Iqbal, Abdul Rahim, Hatem K. Alkahtani, Waleed Alghamdi and Ahmed Q. Almagrabi, 'A New Deep Boosted CNN and Ensemble Learning Based IoT Malware Detection', Computers & Security, Vol. 133, pp. 103385, 2023.
10. Deepak Gupta and Renu Rani, 'Improving Malware Detection Using Big Data and Ensemble Learning', Computers & Electrical Engineering, Vol. 86, pp. 106729, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details