International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05901
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Evaluating Single-Model and Ensemble-Based Intrusion on the CIC-IDS2017 Dataset

Dharavath Sravani
Dept. of CSE
RGUKT
Basar, India
sravanidharavath332@gmail.com

Chirra Sanjana
Dept. of CSE
RGUKT
Basar, India
chirrasanjana24@gmail.com

Myaka Ramya Sri
Dept. of CSE
RGUKT
Basar, India
myakaramyasri@gmail.com

***Abstract*—**The increasing trend of relying on internet-based systems has significantly enhanced the risks of cybersecurity attacks, making network security an essential requirement. Intrusion Detection Systems (IDS) are critical for identifying malicious network behaviors to prevent cyber attacks. Conven- tional methods of using signature-based intrusion detection have limited capabilities to detect unknown attacks. Machine learning algorithms are prone to various problems such as overfitting, high false positive rates, and poor performance while handling high-dimensional data.

In this context, this paper suggests a hybrid intrusion detec- tion system based on machine learning and ensemble learning techniques. The system uses various supervised machine learn- ing algorithms to learn from network traffic data. Ensemble techniques are used to enhance the overall detection capabilities of the system by utilizing the advantages of various machine learning algorithms. The suggested system uses a combination of Random Forest as a bagging technique, Gradient Boosting as a boosting technique to enhance sequential learning, and stacking to integrate the results of various base classifiers using a meta-classifier. Feature selection techniques are used to remove redundant features to enhance the efficiency of the system. Data balancing techniques such as Synthetic Minority Oversampling Technique (SMOTE) and undersampling are used to handle class imbalance problems.

The suggested system is tested using the CICIDS2017 dataset, which consists of realistic benign and malicious network traffic that covers various modern-day cyber attacks. The results show that the suggested system outperforms individual machine learn- ing algorithms by significantly improving the overall accuracy, precision, recall, and F1-score of the system.

In this study, a hybrid intrusion detection system based on machine learning and ensemble learning techniques is suggested. The suggested system uses various supervised machine learning algorithms to learn from network traffic data. Ensemble tech- niques are used to enhance the overall detection capabilities of the system by utilizing the advantages of various machine learning algorithms. Feature selection techniques are used to remove redundant features to enhance the efficiency of the system. Data balancing techniques such as Synthetic Minority Oversampling Technique (SMOTE) and undersampling are used to handle class imbalance problems. The suggested system is tested using the CICIDS2017 dataset, which consists of realistic benign and malicious network traffic that covers various modern-day cyber attacks. The results show that the suggested system outperforms individual machine learning algorithms by significantly improv- ing the overall accuracy, precision, recall, and F1-score of the system.

Index Terms—Intrusion Detection System, Ensemble Learning, Bagging, Boosting, Stacking, CICIDS2017, Machine Learning, Cybersecurity

I. Introduction

The rapid rise of the use of the Internet, cloud computing, and IoT technologies has led to increased threats against the security of contemporary network systems. Intrusion Detection Systems (IDS) have been instrumental in the detection of unauthorized activities, thus enhancing the security of network infrastructure against cyber threats.

In the conventional approach, IDS has been using signature- based detection techniques. Although this approach has been effective in the detection of conventional cyber threats, it has been limited in detecting zero-day attacks. To address the limitations of conventional IDS, machine learning-based IDS has been explored as a potential solution.

Although machine learning-based IDS has been explored as a potential solution, it has been limited in addressing the challenges of high-dimensional spaces and imbalanced datasets. These challenges have been a major contributor to the degradation of performance in IDS.

In addressing the challenges of machine learning-based IDS, ensemble learning has been explored as a potential solution. Ensemble learning has been instrumental in address- ing the challenges of high-dimensional spaces and imbal- anced datasets. Ensemble learning has been instrumental in improving the performance of machine learning algorithms. This paper proposes an ensemble-based intrusion detection system. The main contribution of the proposed system is the application of ensemble learning in improving the performance of IDS.

II. Related Work

Traditionally, intrusion detection systems employed signature-based intrusion detection techniques, where known patterns of attacks were stored in a database. The incoming network traffic is compared with this database to detect any intrusion. Even though this technique is quite effective in detecting known attacks, it is not able to detect unknown attacks.

Machine learning algorithms like Support Vector Machines, Decision Trees, and k-Nearest Neighbors have been employed to improve the intrusion detection capabilities. These machine learning algorithms learn patterns from the network traffic.

Decision Trees are easy to interpret, while Support Vec- tor Machines are good at handling high-dimensional feature spaces. However, machine learning models are often prone to overfitting, which is the lack of generalization.

To overcome this, ensemble learning models have been employed. In bagging, multiple models are trained on randomly sampled datasets, while in boosting, models are trained sequentially, where each model tries to rectify the mistakes made by the previous model.

Stacking is another type of ensemble learning, where the outputs of individual classifiers are combined to improve the overall prediction accuracy.

III. Dataset Description

The CICIDS2017 dataset was developed by the Canadian Institute for Cybersecurity. This dataset offers a realistic benchmark dataset to assess intrusion detection systems. This dataset offers various types of attacks as well as benign network traffic.

The dataset was produced under a controlled environment that replicates realistic enterprise network behavior. This dataset offers detailed network flow features that are extracted by the use of CICFlowMeter.

A. Attack Types

The dataset offers various categories of attacks. These categories are mentioned below.

The CICIDS2017 dataset offers various categories of mod- ern cyber attacks. These attacks are realistic network attacks. Some of the significant categories of attacks that are included in this dataset are mentioned below.

- **Distributed Denial of Service (DDoS):**
  - DDoS attacks send a massive amount of traffic to a targeted server or network.
  - This type of attack tries to exhaust the network bandwidth or the server.
  - These attacks are carried out by utilizing numerous compromised systems.
  - The primary aim of this type of attack is to deny the victims access to the network.
- **Brute Force Attacks:**
  - Brute force attacks attempt different combinations of usernames and passwords.
  - The attackers use tools that try different combina- tions of usernames and passwords.
  - Brute force attacks take advantage of weak passwords.
  - The attacks include FTP-Patator and SSH-Patator attacks.
- **Port Scanning:**
  - Port scanning is a type of reconnaissance method that is carried out by attackers.
  - It helps to detect open ports and services running on the targeted system.
  - Attackers use these open ports to detect vulnerabili- ties.
  - Information is collected from the above process for further attacks.
- **Web Attacks:**
  - Web attacks are carried out to exploit vulnerabilities in web applications.
  - SQL Injection and Cross-Site Scripting are some of the common web attacks.
  - This type of attack helps attackers to manipulate the database or inject malicious scripts.
  - Sensitive information is compromised from the database of the web application.

B. Feature Characteristics

Each network flow consists of approximately 80 statistical features that describe the behavior of the network traffic.

The most important feature categories are:

- Flow duration
- Statistics of packet lengths
- Packet count
- Flow byte rate
- Inter-arrival time
- Protocol information
- TCP flags statistics
- Active and idle time statistics

C. Preprocessing Steps

The data was preprocessed before training any machine learning model.

1) Removal of missing and infinite values
2) Removal of duplicate records
3) Conversion of categorical attributes into numerical val- ues
4) Feature scaling using Min-Max normalization
5) Splitting the data into training and testing sets (80% and 20%)

D. Data Balancing Techniques

The datasets used in intrusion detection contain unbalanced data. Most datasets contain more data points belonging to the ”benign” category than the ”attack” category. This unbalanced data may affect the performance of the model.

To overcome this problem in the datasets, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique creates synthetic data points in the minority category. Also, random undersampling was applied to reduce the number of data points in the majority category.

This problem was overcome to improve the performance of the model in detecting attacks in the minority category.

IV. Mathematical Formulation of Ensemble Learning

Let $h_1(x), h_2(x), ..., h_n(x)$ be individual classifiers.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05901
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

A. Bagging

Bagging involves training multiple classifiers with different samples of the data.

$$H(x) = \frac{1}{n}\sum_{i=1}^{n} h_i(x) \quad (1)$$

B. Boosting

Boosting involves training several classifiers where each classifier attempts to correct the mistakes made by the preceding classifiers.

$$H(x) = \sum_{i=1}^{n} \alpha_i h_i(x) \quad (2)$$

where α_i is the weight of individual classifiers.

C. Stacking

The stacking ensemble combines individual

V. Proposed Methodology

Proposed methodology for the intrusion detection system includes a combination of feature selection, data balancing, and ensemble learning techniques.

A. Base Models

Proposed machine learning algorithms for the base models are as follows:

- Decision Tree
- Random Forest (Bagging)
- Gradient Boosting (Boosting)

B. Feature Selection

Tree-based feature importance techniques have been employed for selecting the important features. The importance score of each feature has been calculated using the random forest algorithm, and the least important features have been removed.

C. Training Process

To balance the dataset, SMOTE and undersampling techniques have been employed. Cross-validation has been used during the training process of the models to avoid overfitting.

D. Stacking Ensemble

To improve the performance of the intrusion detection system, a stacking ensemble has been employed, which combines the predictions of the base models using a meta-classifier.

VI. Machine Learning Methods for Intrusion Detection

In this study, some traditional machine learning methods have been taken into consideration as baseline models for intrusion detection systems. These machine learning methods are widely used due to their simplicity and efficiency.

A. Logistic Regression

Logistic Regression is a supervised machine learning technique that is widely used for binary classification problems. It is a machine learning algorithm that is used to find the probabilities of the class to which a data point belongs. The probabilities are calculated by the logistic function. Logistic Regression is widely used for intrusion detection systems to classify the network traffic into normal and abnormal classes.

However, they are not effective compared to ensemble methods.The logistic function is defined as

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

where,

$$z = w^T x + b$$

w is the weight vector, x is the feature vector, and b is the bias.

Although Logistic Regression is computationally efficient and easy to implement, it only works on linear relationships. This makes Logistic Regression less effective for intrusion detection because intrusion patterns are complex.

B. Decision Tree

Decision Tree is a non-parametric supervised learning al- gorithm that represents decisions in a tree-like structure. Each internal node represents a feature condition, while each leaf node represents a class label.

The model splits the data based on feature values according to criteria such as Gini Index or Information Gain. In IDS systems, decision trees are used to detect patterns of attacks based on conditions such as packet size, flow duration, and protocol type.

Although the model is simple and easy to understand, it faces the issue of overfitting due to its deep levels in the tree structure.

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that aims to find an optimal hyperplane to separate data points with different classes while maintaining maximum distance.

The decision boundary is defined as:

$$w \cdot x + b = 0 \quad (4)$$

where w is the weight vector, and b is the bias.

SVM performs well with high-dimensional data, and it can also handle nonlinear data with the help of a kernel, like the radial basis function (RBF) kernel. SVM has the ability to detect complex patterns of attacks.

However, SVM is computationally expensive, especially when dealing with large data, which makes SVM less scalable for real-time intrusion detection systems.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05901

D. Limitations of Machine Learning Models

Though the above-discussed machine learning models have reasonable performance for intrusion detection, they have some limitations:

- Inability to handle high-dimensional data
- Ineffective performance with imbalanced data
- Overfitting problems, especially with the decision tree algorithm
- Inability to handle complex patterns of attacks

VII. Performance Metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

VIII. Experimental Results

The performance of the proposed intrusion detection system is evaluated using different machine learning and ensemble learning techniques. The evaluation is based on accuracy obtained from both known and unknown datasets.

A. Results of Machine Learning Models

This section presents the performance of individual machine learning models such as Decision Tree (DT), Logistic Regression (LR), and Support Vector Machine (SVM).

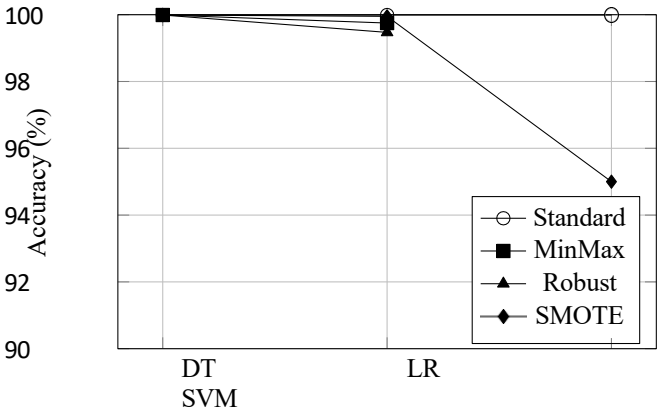

Fig. 1. ML Models Accuracy on Known Dataset (Line Plot)

1) Known Dataset Performance:

2) Unknown Dataset Performance: **Observation:** It is observed that machine learning models achieve very high ac- curacy on known datasets but perform poorly on unknown datasets, indicating poor generalization capability.

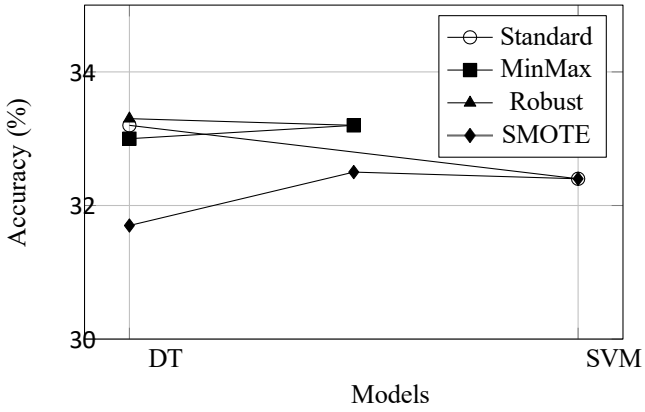

Fig. 2. ML Models Accuracy on Unknown Dataset (Line Plot)

B. Results of Ensemble Learning Models

1) Bagging Technique: **Training Accuracy**

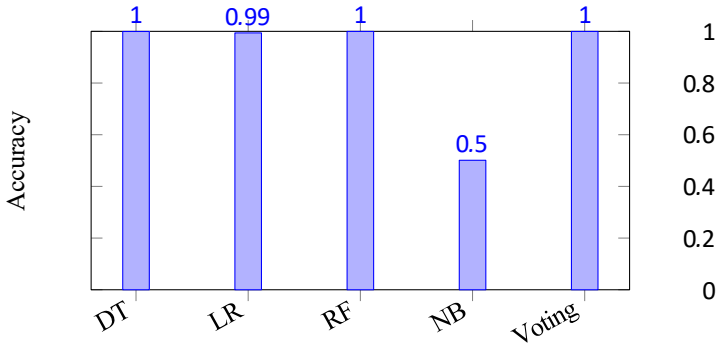

Fig. 3. Bagging Models Training Accuracy

Testing Accuracy

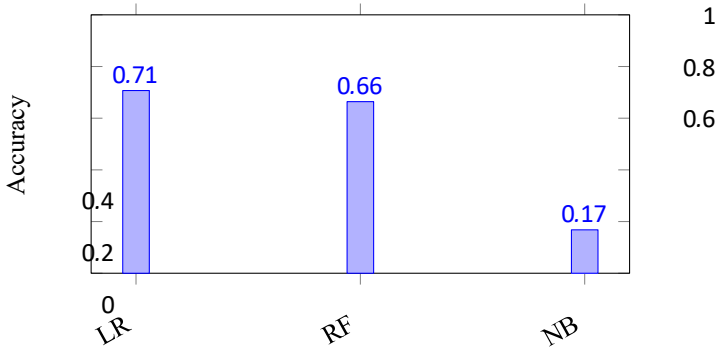

Fig. 4. Bagging Models Testing Accuracy

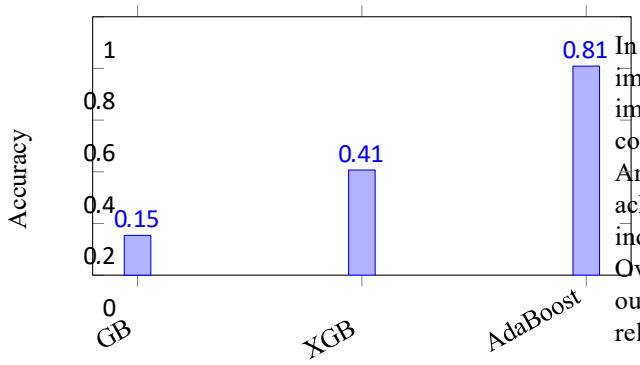

Fig. 5. Boosting Models Performance

2) Boosting Technique:

Fig. 6. Stacking Ensemble Performance

3) Stacking Technique: **Observation:** Stacking achieves very high training accuracy and good testing performance, indicating strong learning capability.

C. Overall Discussion

From the results, it is evident that:

- Individual machine learning models perform well on known data but fail to generalize.
- Ensemble methods significantly improve detection performance.
- Bagging reduces variance, boosting improves learning from errors, and stacking combines strengths of multiple models.

Therefore, ensemble learning provides a more robust and reliable solution for intrusion detection systems.

IX. Discussion

The experimental results highlight the performance differences between individual machine learning models and ensemble learning techniques for intrusion detection.

Machine learning models such as Decision Tree, Logistic Regression, and Support Vector Machine achieve high accuracy on known datasets but show a significant drop in performance on unknown datasets. This indicates poor generalization and difficulty in detecting unseen attack patterns.

In contrast, ensemble learning methods demonstrate better robustness and improved detection capability. Bagging reduces variance, boosting improves learning from misclassified in- stances, and stacking effectively combines multiple models to enhance overall performance.

Among all methods, the stacking ensemble provides the best results, achieving higher accuracy and better generalization compared to individual classifiers.

Overall, the results confirm that ensemble learning tech- niques outperform traditional machine learning models and provide a more reliable solution for intrusion detection sys- tems.

X. Conclusion and Future Work

This study presented a comparative analysis of machine learning and ensemble learning techniques for intrusion detec- tion using the CICIDS2017 dataset. The results demonstrate that individual machine learning models such as Decision Tree, Logistic Regression, and SVM achieve high accuracy on known datasets but fail to generalize effectively on unknown datasets.

In contrast, ensemble learning methods including bagging, boosting, and stacking significantly improve the overall per- formance of the intrusion detection system. Among these, the stacking ensemble model achieved superior results due to its ability to combine multiple classifiers effectively.

The proposed ensemble-based intrusion detection system provides improved accuracy, robustness, and reliability in detecting modern cyber attacks.

Future work will focus on integrating deep learning tech- niques and deploying the model in real-time environments such as cloud computing and IoT-based systems.

Acknowledgment

The authors would like to thank the Department of Com- puter Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Basar, for providing research sup- port.

References

[1] N. Uddamari and P. Sammulal, "Ensemble-Based Network Anomaly Detection Using RFE and Information Gain for Optimized Feature Selec- tion," Informatica, vol. 49, no. 10, 2025, doi: 10.31449/inf.v49i10.8387.
[2] N. Uddamari and P. Sammulal, "CyberAdaptAI: A Dynamic Ensem- ble Learning Framework for Real-Time Cyberattack Detection Using AdaptEnsembleNet," SSRG International Journal of Electronics and Communication Engineering, vol. 12, no. 9, pp. 11–31, 2025.
[3] Z. Ahmad et al., "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," IEEE Trans. Emerging Telecommunications Technologies, 2021.
[4] Y. Alotaibi and M. Ilyas, "Ensemble-Learning Framework for Intrusion Detection to Enhance IoT Security," Sensors, vol. 23, no. 12, 2023.
[5] M. A. Akhtar et al., "Robust genetic machine learning ensemble model for intrusion detection," Scientific Reports, vol. 13, 2023.
[6] C. Lu et al., "Research on Intrusion Detection Based on an Enhanced Random Forest Algorithm," Applied Sciences, 2024.
[7] A. Alsaffar et al., "Enhancing intrusion detection using hybrid feature selection and stacking ensemble learning," Journal of Big Data, 2024.
[8] A. Baluguri et al., "Optimizing Network Security via Ensemble Learn- ing: A Nexus with Intrusion Detection," Journal of Information Security, 2024.

[9] A. Zewdu and H. Kumssa, "An Ensemble Method for Supervised Learning for Intrusion Detection," IntechOpen, 2024.
[10] M. Uppal et al., "Enhancing accuracy through ensemble-based machine learning for intrusion detection," Discover Internet of Things, 2025.
[11] P. A. Doost et al., "A new intrusion detection method using ensemble classification and feature selection," Scientific Reports, vol. 15, 2025.
[12] A. Alabdulatif, "A Novel Ensemble Deep Learning Approach for Intrusion Detection with Explainable AI," Applied Sciences, 2025.
[13] R. Singh et al., "Attentional LSTM-ensemble architecture for intrusion detection in smart grids," Scientific Reports, 2025.
[14] M. Adil et al., "xIDS-EnsembleGuard: Explainable Ensemble Learning- based Intrusion Detection System," arXiv, 2025.
[15] A. Amouri et al., "Enhancing Intrusion Detection in IoT using Hybrid Ensemble Models," arXiv, 2024.
[16] S. Saidane et al., "Optimizing IDS Performance through Advanced Data Processing and Ensemble Learning," arXiv, 2024.
[17] I. Bibers et al., "Comparative Study of ML Models and Ensemble Strategies for IDS," arXiv, 2024.