Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comparative Study of Intrusion Detection Models for DDoS and Brute Force Attack Detection Using Public Cybersecurity Datasets

Dr Kazi Abdul Mannan

drkaziabdulmannan@gmail.com

Center for Academic & Professional Career Development and Research (CAPCDR)
https://orcid.org/0000-0002-7123-132X

Md. Admir Hossin

Shanto-Mariam University of Creative Technology https://orcid.org/0009-0005-2657-6053

Md. Soliman Ar Rafi

Shanto-Mariam University of Creative Technology https://orcid.org/0009-0002-4140-9868

Research Article

Keywords: Intrusion Detection System (IDS), Machine Learning, XGBoost, Random Forest, Artificial Neural Network (ANN), CICIDS2017

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10535983/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The rapid growth of internet-based services and connected devices has significantly increased the risk of cyberattacks, making network security a critical concern for organizations and individuals. Traditional Intrusion Detection Systems (IDS) mainly rely on signature-based detection techniques, which are effective for identifying known attacks but often fail to detect new or evolving threats. This study proposes a machine learning-based intrusion detection system to detect Distributed Denial of Service (DDoS) and Brute Force attacks using the CICIDS2017 dataset. The dataset was preprocessed by removing invalid and missing values, encoding attack labels, and preparing the data for supervised learning. Three machine learning algorithms, namely Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), were implemented and evaluated using accuracy, precision, recall, and F1-score. Experimental results demonstrate that all three models achieved excellent classification performance, with XGBoost providing the highest overall performance, achieving approximately 99.9975% accuracy, precision, recall, and F1-score. Random Forest also produced highly competitive results, while ANN achieved slightly lower performance because of its sensitivity to feature scaling. The findings indicate that ensemble tree-based learning methods are highly effective for detecting DDoS and Brute Force attacks and can significantly improve the reliability of modern intrusion detection systems.

1. Introduction

The rapid advancement of information technology and the widespread adoption of the Internet have transformed the way people communicate, conduct business, and access digital ser-vices. Today, billions of users rely on interconnected networks for banking, healthcare, education, ecommerce, cloud computing, and industrial automation. While these technological developments have improved efficiency and accessibility, they have also significantly increased the exposure of computer networks to sophisticated cyber threats [9].

Among the most prevalent cyberattacks are Distributed Denial of Service (DDoS) attacks and Brute Force attacks. A DDoS attack floods a target server or network with an overwhelming volume of malicious traffic, making services unavailable to legitimate users. Brute Force attacks, on the other hand, repeatedly attempt different username and pass-word combinations to gain unauthorized access to systems and sensitive information [3], [7]. These attacks can lead to financial losses, service disruption, data breaches, and reputational damage for organizations.

Intrusion Detection Systems (IDS) are widely deployed to monitor network traffic and detect malicious activities. Traditional IDS mainly rely on signature-based detection, which compares incoming traffic against a database of known attack signatures. Although signature-based IDS perform well against previously identified attacks, they often fail to detect zero-day attacks and newly emerging threats because they cannot recognize attack patterns that are absent from their signature databases [2].

To overcome these limitations, anomaly-based intrusion detection systems have increasingly adopted machine learning techniques. Machine learning enables IDS to automatically learn normal and malicious traffic patterns from historical network data, allowing them to identify both known and previously unseen attacks with high accuracy. Ensemble learning algorithms such as Random Forest and Extreme Gradient Boosting (XGBoost), together with deep learning models such as Artificial Neural Networks (ANN), have demonstrated out-standing performance in modern intrusion detection research [11]-[13].

This study develops a machine learning-based intrusion detection system using the publicly available CICIDS2017 dataset, one of the most widely adopted benchmark datasets for cybersecurity research [10]. Three supervised learning algorithms, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), are implemented and evaluated using Accuracy, Precision, Recall, and F1-score. The primary objective is to identify the most effective model for detecting DDoS and Brute Force attacks while providing a comparative analysis of their classification performance.

The continuous growth of cyberattacks presents significant challenges for existing network security systems. Traditional signature-based intrusion detection systems are effective only for detecting previously known attack patterns and are unable to identify zero-day or modified attacks that do not match existing signatures [2]. As network traffic volume continues to increase, manual monitoring becomes impractical, and conventional security mechanisms frequently generate high false-positive rates while missing sophisticated attack patterns. Although numerous studies have applied machine learning techniques for intrusion detection, variations in datasets, pre-processing methods, feature engineering strategies, and evaluation metrics make direct comparisons difficult. Furthermore, many previous studies evaluate only one or two machine learn-ing algorithms, limiting a comprehensive understanding of their relative performance. Therefore, a systematic comparison of Random Forest, XGBoost, and Artificial Neural Network using identical preprocessing procedures and the CICIDS2017

dataset is required.

The primary objective of this research is to develop and evaluate a machine learning-based intrusion detection system for detecting DDoS and Brute Force attacks using the CI-CIDS2017 dataset. The specific objectives are as follows:

- To preprocess the CICIDS2017 dataset for supervised machine learning.
- To implement Random Forest, XGBoost, and Artificial Neural Network models for intrusion detection.
- To evaluate the performance of each model using Accuracy, Precision, Recall, and F1-score.
- To compare the effectiveness of the three machine learning algorithms.
- To identify the most suitable machine learning model for detecting DDoS and Brute Force attacks.

This research contributes to the field of cybersecurity by providing a comparative evaluation of three widely used supervised machine learning algorithms under identical experimental conditions. The study demonstrates the effectiveness of ensemble learning techniques and deep learning methods for intrusion detection using a realistic benchmark dataset. The findings can assist researchers, cybersecurity professionals, and network administrators in selecting suitable machine learning algorithms for developing intelligent intrusion detection systems. Furthermore, this work provides a foundation for future research involving hybrid models, explainable artificial intelligence, advanced feature selection techniques, and real-time intrusion detection applications [5], [9].

2. Literature Review

The increasing frequency and sophistication of cyberattacks have driven extensive research into intelligent intrusion detection systems (IDS). Traditional signature-based IDS are effective in identifying previously known attacks but often fail to detect zero-day threats and evolving attack patterns. Consequently, researchers have increasingly adopted machine learning techniques to improve detection accuracy, reduce false positives, and enhance the adaptability of intrusion detection systems [2], [9].

Machine learning-based IDS automatically learn network traffic patterns from historical data and classify network flows as either benign or malicious. Compared with conventional rule-based approaches, machine learning models can generalize from training data and identify previously unseen attack behaviors. The availability of benchmark datasets such as CICIDS2017 has further accelerated research by providing realistic network traffic containing multiple attack categories [10].

Machine learning algorithms have become one of the most widely adopted approaches for intrusion detection because of their ability to analyze large-scale network traffic efficiently. Supervised learning techniques, including Decision Trees, Support Vector Machines, Random Forest, Gradient Boosting, and Artificial Neural Networks, have demonstrated excellent performance in identifying malicious network activities [9].

Keskin and Okatan conducted a comparative analysis of several machine learning algorithms for intrusion detection and reported that ensemble learning techniques consistently outperformed conventional classifiers in terms of detection accuracy and robustness. Their study emphasized that appropriate feature selection and data preprocessing significantly improve classifier performance [2].

Distributed Denial of Service (DDoS) attacks and Brute Force attacks remain among the most common threats targeting modern computer networks. Hanif and Ilyas proposed an intrusion detection framework for Internet of Things (IoT) environments that combined feature engineering with ensemble learning techniques to improve the detection of DoS and Brute Force attacks. Their experimental results demonstrated that feature engineering substantially enhanced detection accuracy while reducing computational complexity [3].

Similarly, Neto investigated machine learning techniques for detecting Brute Force attacks and highlighted the importance of supervised learning models in identifying repeated authentication attempts. The study concluded that machine learning provides greater adaptability than traditional signature-based intrusion detection approaches, particularly when detecting previously unseen attack patterns [1].

Random Forest is one of the most widely used ensembles learning algorithms for intrusion detection. Introduced by Breiman, Random Forest constructs multiple decision trees using randomly selected subsets of training data and com-bines their predictions using majority voting. This ensemble mechanism improves classification accuracy while reducing overfitting [11].

Rodr´ıguez et al. evaluated several machine learning algorithms using the CICIDS2017 dataset and reported that Random Forest achieved outstanding detection performance after applying correlation-based feature selection. Their findings demonstrated that Random Forest effectively distinguished malicious traffic from benign network flows while maintaining low false-positive rates [5]. Alduailij et al. proposed a DDoS detection framework that combined Mutual Information feature selection with Random Forest. Their results indicated that selecting the most informative traffic features significantly improved detection accuracy while reducing computational overhead [4].

Extreme Gradient Boosting (XGBoost) is a scalable ensemble learning algorithm based on gradient boosting. Unlike conventional boosting methods, XGBoost incorporates regularization techniques that improve model generalization while reducing overfitting. Chen and Guestrin demonstrated that XGBoost provides high predictive accuracy and computational efficiency across a wide range of classification problems [12]. Recent intrusion detection studies have shown that XGBoost consistently achieves superior classification performance due to its ability to capture complex nonlinear relationships within network traffic. The algorithm has become increasingly popular because it efficiently handles high-dimensional cybersecurity datasets while maintaining excellent accuracy and low prediction error [5].

Artificial Neural Networks (ANN) have also been extensively applied to intrusion detection because of their ability to learn complex nonlinear relationships from network traffic. Deep learning models consist of interconnected layers of artificial neurons that automatically extract meaningful features from large datasets. Goodfellow et al. described deep neural networks as highly effective for solving complex classification tasks involving high-dimensional data [13]. Liu et al. combined machine learning and deep learning techniques for intrusion detection and demonstrated that neural networks achieved high detection accuracy when appropriate preprocessing and feature scaling techniques were applied. However, the study also reported that neural networks generally require greater computational resources and longer training times than ensemble learning algorithms [9].

The performance of intrusion detection models largely de-pends on the quality of the dataset used during training and evaluation. Among publicly available cybersecurity datasets, CICIDS2017 has become one of the most widely adopted benchmarks because it contains realistic network traffic generated under both normal and attack scenarios.

Sharafaldin et al. developed the CICIDS2017 dataset to overcome the limitations of earlier benchmark datasets by including modern attack categories, realistic user behavior, and comprehensive network flow features. The dataset contains multiple attack types, including DDoS, Brute Force, Botnet, Web Attacks, Heartbleed, and Denial of Service attacks, making it suitable for evaluating machine learning-based intrusion detection systems [10].

Although previous studies have demonstrated the effective-ness of machine learning algorithms for intrusion detection, several research gaps remain. Many studies evaluate only a single classifier or compare a limited number of algorithms using different preprocessing techniques and evaluation procedures. Such differences make direct comparisons difficult and may influence reported performance.

Furthermore, relatively few studies perform a comprehensive comparison of Random Forest, XGBoost, and Artificial Neural Network using identical preprocessing methods, training data, testing data, and evaluation metrics. Since each algorithm has different learning characteristics, a systematic comparison under consistent experimental conditions is necessary to determine the most effective model for detecting DDoS and Brute Force attacks.

The reviewed literature demonstrates that machine learning has significantly improved intrusion detection by enabling automated analysis of network traffic and accurate classification of malicious activities. Ensemble learning algorithms, particularly Random Forest and XGBoost, consistently achieve outstanding performance due to their robustness and ability to handle high-dimensional network traffic data. Artificial Neural Networks also provide strong classification capability but generally require more computational resources and careful parameter tuning.

Based on the identified research gap, this study performs a comparative evaluation of Random Forest, XGBoost, and Artificial Neural Network using the CICIDS2017 dataset under identical preprocessing and evaluation conditions. The objective is to identify the most effective machine learning model for detecting DDoS and Brute Force attacks while providing reproducible experimental results for future cybersecurity research.

3. Theoretical And Conceptual Framework

Intrusion Detection Systems (IDS) are designed to monitor network traffic and identify malicious activities that may threaten the confidentiality, integrity, and availability of computer systems. IDS techniques are generally classified into signature-based detection and anomaly-based detection. Signature-based systems compare incoming network traffic with predefined attack signatures and are highly effective for detecting previously known attacks. However, they cannot identify unknown or modified attacks that are not stored in the signature database.

Anomaly-based intrusion detection addresses this limitation by learning the normal behavior of a network and identifying deviations that may indicate malicious activity. Machine learning plays an

important role in anomaly-based IDS because it automatically learns complex patterns from network traffic without requiring manually defined detection rules.

This research adopts a supervised machine learning approach in which labeled network traffic is used for training classification models. Each network flow is assigned a class label representing either benign traffic or a specific attack category. During training, the models learn the relationship be-tween network traffic features and their corresponding labels. After training, the models classify previously unseen network traffic into the appropriate category.

Three supervised learning algorithms are employed in this study.

Random Forest (RF): Random Forest is an ensemble learning algorithm that constructs multiple decision trees using randomly selected subsets of the training data. Each tree independently predicts the class label, and the final prediction is obtained through majority voting. This ensemble strategy improves classification accuracy while reducing overfitting.

Extreme Gradient Boosting (XGBoost): XGBoost is an optimized gradient boosting algorithm that sequentially builds decision trees. Each new tree corrects the prediction errors produced by previous trees. Through boosting and regularization, XGBoost achieves excellent predictive performance while maintaining computational efficiency.

Artificial Neural Network (ANN): Artificial Neural Net-work is a deep learning model inspired by biological neurons. The proposed ANN consists of one input layer, three hidden layers, and one output layer. Hidden layers use the ReLU activation function, while the output layer applies the Softmax activation function for multiclass classification. Dropout layers are incorporated to reduce overfitting during training.

The effectiveness of these algorithms is evaluated using Accuracy, Precision, Recall, and F1-score. These metrics pro-vide a comprehensive assessment of classification performance by measuring prediction correctness, false positives, false negatives, and the balance between Precision and Recall.

The proposed intrusion detection framework follows a sequential machine learning pipeline.

Initially, the CICIDS2017 dataset is collected and preprocessed by removing missing values, replacing invalid values, encoding attack labels, and preparing the data for supervised learning. The processed dataset is then divided into training and testing subsets using an 80:20 ratio.

The training data are used to develop three supervised learning models: Random Forest, XGBoost, and Artificial Neural Network. After training, each model predicts the labels of the testing data. Finally, model performance is evaluated using Accuracy, Precision, Recall, and F1-score. The obtained results are compared to identify the best-performing algorithm for detecting DDoS and Brute Force attacks.

4. Methodology

This research adopts a quantitative approach using supervised machine learning to develop an Intrusion Detection System (IDS) capable of detecting Distributed Denial of Service (DDoS) and Brute Force attacks. The overall frame-work consists of five major stages: dataset collection, data preprocessing, model development, performance evaluation, and comparative analysis.

The CICIDS2017 dataset was selected because it contains realistic network traffic that includes both benign and malicious activities. The dataset provides multiple attack categories, including DDoS, FTP-Patator, SSH-Patator, Web At-tacks, Botnet, Heartbleed, and DoS attacks. This study focuses on detecting benign traffic together with DDoS and Brute Force attacks.

The proposed framework first preprocesses the dataset to remove missing values and invalid entries. The cleaned data are then divided into training and testing subsets using an 80:20 ratio. Three supervised machine learning models, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), are trained using the same dataset. Finally, their performance is evaluated using Accuracy, Precision, Recall, and F1-score to determine the most effective intrusion detection model.

4.1 Dataset Description

The experiments were conducted using the CICIDS2017 dataset developed by the Canadian Institute for Cybersecurity (CIC). This benchmark dataset contains realistic network traffic collected under normal user activities and various cyberattack scenarios. The original dataset consists of multiple CSV files containing network flow records extracted using CICFlowMeter. Each network flow includes 78 numerical features describing packet statistics, flow duration, protocol information, header lengths, packet lengths, byte counts, inter-arrival times, and other traffic characteristics. After preprocessing, the final dataset contained approximately 2,413,177 valid network flow records with 79 at-tributes, including 78 input features and one target class label.

4.2 Data Preprocessing

Data preprocessing was performed to improve data quality before training the machine learning models. Initially, infinite values (+ ∞ and −∞) were replaced with missing values (NaN). Records containing missing values were removed using the dropna() function. The categorical class labels were converted into numerical values using the LabelEncoder technique. The encoded classes were defined as follows:

- BENIGN → 0
- DDoS → 1
- FTP-Patator → 2
- SSH-Patator → 3

Feature scaling was applied only to the Artificial Neural Network because neural networks are sensitive to feature magnitude. StandardScaler was used to standardize all numerical input features before ANN

training. Random Forest and XGBoost did not require feature scaling because they are tree-based algorithms.

4.3 Train-Test Split

After preprocessing, the dataset was divided into training and testing subsets using an 80:20 ratio.

Stratified sampling was employed to preserve the original class distribution in both subsets. Approximately 80% of the dataset was used for model training, while the remaining 20% was reserved for performance evaluation. Using identical training and testing datasets ensured a fair comparison among Random Forest, XGBoost, and Artificial Neural Network.

4.4 Machine Learning Models

Three supervised machine learning algorithms were implemented.

Random Forest: Random Forest is an ensemble learning algorithm that constructs multiple decision trees using randomly selected subsets of the training data. Each tree independently predicts the attack class, while the final prediction is obtained through majority voting. The Random Forest model was trained using 100 decision trees with a random seed of 42.

Extreme Gradient Boosting (XGBoost): XGBoost is an optimized gradient boosting algorithm that sequentially builds decision trees to minimize prediction errors. The model was configured with:

- Number of estimators = 100
- Maximum tree depth = 6
- Learning rate = 0.1
- Subsample ratio = 0.8
- Column sampling ratio = 0.8
- Evaluation metric = mlogloss

Artificial Neural Network: The Artificial Neural Net-work consists of one input layer, three hidden layers, and one output layer. The hidden layers contain:

- 128 neurons
- 64 neurons
- 32 neurons

ReLU activation was applied to all hidden layers, while the output layer used the Softmax activation function for multiclass classification. Dropout layers with a rate of 0.3 were inserted after the first two hidden layers to reduce overfitting. The model was trained using:

- Adam optimizer

- Sparse Categorical Crossentropy loss
- Batch size = 512
- Epochs = 30
- Early Stopping

4.5 Performance Evaluation

The performance of each machine learning model was evaluated using four standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-score

Confusion matrices were generated to analyze True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Among these evaluation metrics, Recall and F1-score received particular attention because undetected cyberattacks (false negatives) may lead to serious security breaches.

4.6 Software Environment

The proposed intrusion detection system was implemented using Python. The following software libraries were employed:

- NumPy
- Pandas
- Scikit-learn
- XGBoost
- TensorFlow
- Keras
- Matplotlib
- Seaborn
- Jupyter Notebook

4.7 Ethical Considerations

This study used the publicly available CICIDS2017 dataset and therefore did not involve human participants, personal information, or confidential organizational data. The dataset was used exclusively for academic research purposes. All referenced studies and software libraries were properly

acknowledged, and the experimental findings are reported objectively to ensure transparency, reproducibility, and academic integrity.

5. Results

The proposed intrusion detection models were evaluated using the CICIDS2017 dataset after preprocessing and label encoding. The cleaned dataset contained approximately 2,413,177 valid network flow records with 78 input features and one target class. The dataset was divided into training and testing subsets using an 80:20 stratified split to ensure consistent class distribution across both datasets.

Three supervised machine learning algorithms, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), were trained using identical training data and evaluated using the same testing dataset. Performance was assessed using Accuracy, Precision, Recall, and F1-score.

Table 5.1 summarizes the overall performance of the three machine learning models. The experimental results indicate that all three models achieved excellent classification performance. XGBoost produced the highest Accuracy, Precision, Recall, and F1-score, followed closely by Random Forest. Although the Artificial Neural Network also achieved high classification accuracy, its performance remained slightly lower than the ensemble tree-based models.

Table 5.1
Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
Random Forest	99.9965%	99.9965%	99.9965%	99.9965%
XGBoost	**99.9975%**	**99.9975%**	**99.9975%**	**99.9975%**
ANN	99.9550%	99.9554%	99.9550%	99.9552%

Confusion matrices were generated for all three classifiers to evaluate their ability to correctly classify benign and malicious network traffic. Random Forest correctly classified almost all benign traffic, DDoS attacks, FTP-Patator attacks, and SSH-Patator attacks with only a very small number of misclassified samples.

XGBoost produced the best confusion matrix among all evaluated models. The classifier correctly identified nearly every network flow while minimizing both false positives and false negatives. This explains its superior Accuracy, Precision, Recall, and F1-score.

The Artificial Neural Network also demonstrated strong classification capability but produced slightly more classification errors compared with Random Forest and XGBoost. The larger number of false negatives slightly reduced its Recall and F1-score.

Feature importance analysis was conducted for both Random Forest and XGBoost to identify the network traffic characteristics that contributed most to intrusion detection. For the Random Forest classifier, the most influential features included:

- Avg Bwd Segment Size
- Bwd Packet Length Max
- Bwd Packet Length Mean
- Average Packet Size
- Packet Length Mean

These features primarily describe backward packet characteristics and were found to be highly effective for distinguishing malicious traffic from benign network activity. For the XGBoost classifier, the highest-ranked features included:

- Bwd Packet Length Std
- Subflow Bwd Bytes
- Avg Bwd Segment Size
- Packet Length Variance
- Bwd Packet Length Mean

The feature importance results demonstrate that backward network flow characteristics contain valuable information for identifying DDoS and Brute Force attacks within the CI-CIDS2017 dataset.

Overall, the experimental evaluation demonstrates that ensemble learning techniques outperform the Artificial Neural Network under the selected experimental configuration. Among the evaluated algorithms, XGBoost consistently achieved the best classification performance while maintaining extremely low false positive and false negative rates. Random Forest also demonstrated outstanding predictive capability with only marginally lower performance. The obtained results confirm that tree-based ensemble learning algorithms provide highly reliable intrusion detection performance when trained using the CICIDS2017 dataset.

6. Discussion

The experimental results demonstrate that all three super-vised machine learning algorithms achieved excellent classification performance for detecting DDoS and Brute Force attacks using the CICIDS2017 dataset. Each model obtained an overall accuracy greater than 99%, indicating that machine learning techniques are highly effective for intrusion detection. Among the evaluated classifiers, XGBoost achieved the highest overall performance with approximately 99.9975% Accuracy, Precision, Recall, and F1-score. Random Forest produced very similar results with only a marginal reduction in performance, whereas the Artificial Neural Network (ANN) achieved comparatively lower evaluation scores.

The superior performance of XGBoost can be attributed to its gradient boosting mechanism, where each successive decision tree learns from the prediction errors of previous trees. This iterative learning process enables the model to capture complex relationships within network traffic while minimizing classification errors. Furthermore, XGBoost incor-porates regularization techniques that improve generalization performance and reduce overfitting.

Random Forest also demonstrated excellent predictive capability because it combines multiple decision trees using majority voting. This ensemble strategy increases model stability while reducing prediction variance. Although Random Forest performed slightly below XGBoost, the difference was very small, confirming that both ensemble learning algorithms are highly suitable for intrusion detection applications.

The Artificial Neural Network achieved strong classification accuracy but produced relatively lower Precision, Recall, and F1-score. Neural networks generally require careful feature scaling and hyperparameter tuning to achieve optimal performance. Under the experimental configuration adopted in this study, the ANN model produced a slightly higher number of false negatives compared with the ensemble learning algorithms.

The findings of this study are consistent with previous research that reported excellent intrusion detection performance using ensemble learning techniques. Rodr´ıguez et al. demonstrated that Random Forest achieved outstanding performance on the CICIDS2017 dataset when combined with appropriate feature selection techniques. Similarly, Hanif and Ilyas reported strong detection capability using ensemble classifiers within IoT environments.

The findings of the present study further demonstrate that XGBoost slightly outperforms Random Forest under identical preprocessing conditions. Compared with the studies reviewed in the literature, the proposed XGBoost model achieved one of the highest reported values for Accuracy, Precision, Recall, and F1-score on the selected attack classes.

The experimental results therefore support previous research indicating that ensemble learning algorithms provide superior performance compared with many conventional machine learn-ing approaches for network intrusion detection.

Feature importance analysis revealed that backward-flow traffic characteristics contributed significantly to attack detection. Random Forest identified Avg Bwd Segment Size, Bwd Packet Length Max, and related backward packet statistics as highly informative features. Similarly, XGBoost assigned the highest importance to Bwd Packet Length Std, Subflow Bwd Bytes, and other backward-flow measurements.

These findings indicate that network behavior associated with backward packet transmission contains valuable information for distinguishing malicious traffic from benign communication. The results also

demonstrate that ensemble learning algorithms effectively exploit these traffic characteristics during classification.

The proposed intrusion detection system has several practical applications.

Organizations can integrate machine learning-based IDS into enterprise networks to improve cyberattack detection while reducing false alarms. Financial institutions, healthcare organizations, educational institutions, cloud service providers, and government agencies may benefit from intelligent intrusion detection systems capable of identifying malicious traffic with very high accuracy.

The experimental findings also demonstrate that tree-based ensemble learning algorithms require relatively modest computational resources compared with many deep learning approaches while maintaining outstanding classification performance. This makes Random Forest and XGBoost practical choices for deployment in operational network security environments.

Although the proposed models achieved excellent performance, several limitations should be acknowledged.

First, the experiments were conducted using the publicly available CICIDS2017 dataset rather than live network traffic. Although the dataset closely represents realistic network behavior, actual production environments may contain additional attack patterns and evolving threats. Second, this study focused only on DDoS and Brute Force attacks. Other attack categories available within the

CICIDS2017 dataset was not included in the experimental evaluation. Third, hyperparameter optimization was limited to experimentally selected configurations. More comprehensive optimization techniques may further improve model performance.

Future research may extend this work by evaluating newer intrusion detection datasets and incorporating additional cyberattack categories. Advanced feature selection techniques, automated hyperparameter optimization, hybrid ensemble learning, explainable artificial intelligence (XAI), and real-time deployment within Software Defined Networking (SDN) environments may fur-ther improve intrusion detection performance. Future studies may also investigate lightweight machine learning models for edge computing and Internet of Things (IoT) applications, where computational efficiency is particularly important.

7. Conclusion

This study presented a machine learning-based Intrusion Detection System (IDS) for detecting Distributed Denial of Service (DDoS) and Brute Force attacks using the CI-CIDS2017 dataset. Three supervised machine learning algorithms, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN), were implemented and evaluated under identical preprocessing and experimental conditions. The dataset was preprocessed by removing invalid values,

encoding class labels, and preparing the network traffic for supervised learning. Model performance was evaluated using Accuracy, Precision, Recall, and F1-score.

The experimental results demonstrated that all three ma-chine learning algorithms achieved outstanding classification performance. Among the evaluated models, XGBoost consistently produced the highest Accuracy, Precision, Recall, and F1-score, achieving approximately 99.9975% across all evaluation metrics. Random Forest achieved nearly identical performance with only a marginal reduction in classification accuracy, while the Artificial Neural Network achieved slightly lower results because of its dependence on feature scaling and parameter optimization.

Feature importance analysis further revealed that backward-flow traffic characteristics played a significant role in distinguishing malicious traffic from benign network activity.

This research contributes to the field of cybersecurity by providing a comprehensive comparison of three widely used supervised machine learning algorithms using the same bench-mark dataset and identical evaluation procedures. The study demonstrates that ensemble learning techniques, particularly XGBoost and Random Forest, provide highly reliable intrusion detection with extremely low false positive and false negative rates. The findings may assist researchers, cybersecurity practitioners, and network administrators in selecting appropriate machine learning models for intelligent intrusion detection systems.

The proposed framework also provides a reproducible methodology that can serve as a reference for future intrusion detection research based on benchmark cybersecurity datasets.

Declarations

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research.

Acknowledgment

The authors would like to express their sincere gratitude to the Department of CSE & CSIT, Shanto-Mariam University of Creative Technology, for providing academic guidance and support throughout this research. The authors also acknowledge the Canadian Institute for Cybersecurity (CIC) for making the CICIDS2017 dataset publicly available, which greatly facilitated the experimental evaluation presented in this study.

Data Availability

The dataset used in this study is the publicly available CICIDS2017 dataset provided by the Canadian Institute for Cybersecurity (CIC). The dataset can be accessed through the official CIC website for academic and research purposes.

References

1. Neto BADS (2024) Brute Force Attacks. Evaluating Machine Learning Solutions for Network Intrusion Detection
2. Keskin S, Okatan E (2023) Machine Learning Methods for Intrusion Detection in Computer Networks: A Comparative Analysis. Interna-tional J Eng Innovative Res 5(3):268–279
3. Hanif AA, Ilyas M (2024) Enhance the Detection of DoS and Brute Force Attacks within the MQTT Environment Through Feature Engineering and Employing an Ensemble Technique. arXiv preprint, arXiv:2408.00480
4. Alduailij M, Khan QW, Tahir M, Sardaraz M, Alduailij M, Malik F Machine Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. Symmetry, 14, 6, Art. 1095, 2022.
5. Rodr´ıguez M, Alesanco A, Mehavilla L, Garc´ıa J (2022) Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection, Sensors, vol. 22, no. 23, Art. no. 9326
6. Abdulkareem S (2026) Employing Functional Machine Learning Principles for Active Detection of Brute Force Attack Relying on the CICIDS2017 Dataset. Dijlah J Eng Sci, 3, 1
7. Vadhii FA, Salihi ML, Nanne MF (2024) Machine Learning-Based Intrusion Detection System for Detecting Web Attacks. IAES Int J Artif Intell 13(1):711–721
8. Hewapathirana IU (2025) A Comparative Study of Two-Stage Intrusion Detection Using Modern Machine Learning Approaches on the CSE-CIC-IDS2018 Dataset, Knowledge, vol. 5, no. 1, Art. no. 6
9. Liu L, Wang P, Lin J, Liu L (2020) Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning. IEEE Access 9:7550–7563
10. Sharafaldin I, Lashkari AH, Ghorbani AA (2018) Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characteriza-tion, in Proc. International Conference on Information Systems Security and Privacy (ICISSP), Porto, Portugal, pp. 108–116
11. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32
12. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System, in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pp. 785–794
13. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, Cambridge, MA, USA
14. Pedregosa F et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
15. Abadi M et al (2016) TensorFlow: Large-Scale Machine Learning on Het-erogeneous Distributed Systems