International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

A Machine Learning Based Classification and Prediction Technique used for DDoS Attacks.

Dr. Bhargavi Peddi Reddy, Ravi Teja Kondagorla, Paidi Sai Surya Gupta, Sirangi Sushank
Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, India

Abstract - The field of network security has undergone numerous attacks through DDoS attacks that are designed to affect the access for the intended users. Previous studies that have been conducted regarding intrusion detection with the use of the KDD cup 1999 datasets have failed to account for the current nature of the attack. In our study, we relied on two models which include random forest and XGBoost by applying the UNSW-NB15 dataset. Our work encompasses everything ranging from data cleaning to modeling. On average, the random forest had 89% accuracy whereas XGBoost scored 90%. These findings are much better compared to others using deep learning methods, which scored an average of 79% and 85%.

Keywords -DDoS Attacks, Random Forest, XGboost, UNSW-NB15, Intrusion Detection, Network Security & Machine Learning.

I. INTRODUCTION

Today, one of the major cyber threats is associated with the rise of DDoS attacks. Such an attack implies flooding a targeted network with lots of fake packets. As a result, normal users cannot access related services. Banking systems, shopping sites, governmental portals, as well as Internet of Things (IoT) devices, usually become victims of DDoS attacks. With each new year, the number of interconnected devices increases. Thus, attackers get more targets for attacking companies using botnets.

Nevertheless, despite the actuality of the issue, many researches on the creation of DDoS algorithms work with the same database, namely KDD Cup 1999 Dataset, which is already twenty-five years old. Nowadays, Internet looks rather different from what it used to look like when this set of data had been created. Moreover, there have appeared plenty of new types of both traffic and attacks that are not mentioned in this dataset.

In order to fill this void, this research employs the UNSW-NB15 dataset created by the Australian Centre for Cyber Security that contains real-world network traffic data across nine attack types. The paper examines the classification capabilities of two widely used supervised machine learning algorithms, namely Random Forest and XGBoost, under an eight-step framework. Accuracy, precision, recall, and F1-scores are applied to measure the performance of both models in recognizing DDoS attacks.

A. Types of DDoS Attacks

Several forms of DDoS attacks could be used to attack different parts of the computer network infrastructure. First of all, there is the SYN Flood attack, in which attackers exploit the TCP three-way handshaking procedure to conduct an attack on the server. The problem occurs when there is a massive amount of request packets to create connections with the server without establishing them; therefore, system resources are depleted. Second, the UDP Flood attack involves sending a lot of packets through the User Datagram Protocol, thereby consuming all the computing power of the targeted machine. Third, the HTTP Flood attack implies that attackers utilize their botnet to flood the server with web traffic. Since the attack looks like legitimate website usage from the side of the server, it is extremely hard to differentiate between actual users' requests and the attack traffic, thus becoming one of the hardest types of attacks to detect. The Ping of Death attack implies sending ping packets, which are larger than the maximum packet size, specified by the IP protocol. Upon receiving such packages, the server becomes unable to handle the request and crashes.

Smurf and Fraggle attacks are some of the most popular ways of conducting a DDoS attack, whereby the former is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

done via the ICMP protocol and the latter through the use of the UDP protocol by sending packets to other computers within the network using the attacking computer. Through such a mechanism, the reply packets reach the same IP address of the attacked computer. Finally, the NTP amplification attack can be said to be a form of a reflection attack whereby a smaller-sized packet is sent to the NTP server, making the replied packets reach the attacker's address.

B. Motivation and Contributions

Considering the classical approaches, the main difficulty lies in the lack of an approach that would enable these algorithms to address issues like irrelevant features and missing values in the data. In this regard, the most suitable solution to this problem lies in the application of the Random Forest algorithm, which consists in creating many decision trees based on random subsamples of input data. However, the upgraded algorithm XGBoost performs even better than the traditional one because the construction of decision trees takes place incrementally to reduce mistakes made by earlier created trees. In contrast to artificial neural networks used to detect DDoS attacks, the proposed algorithms are faster, consume fewer computing resources, and are easier to adjust. The major contributions of this research paper include: (i) an approach for preparing input data for multi-classification problems in terms of DDoS attacks step by step; (ii) an evaluation of the effectiveness of Random Forest and XGBoost algorithms for DDoS attack multi-classification; and (iii) comparison with the best approaches found in recent scientific literature.

II. RELATED WORKS

Various model implementations based on machine learning and deep learning models were conducted. According to Karatas, Demir, and Sahingoz [1], various machine learning models were analyzed in terms of their attack detection capabilities, where the model that attained the highest accuracy was the K-Nearest Neighbors. In another work, Su et al. [2] introduced a hybrid model involving the combination of the CNN and LSTM models, which were then applied to the KDD data set with an accuracy of 85%. The implementation of models like CNN, BAT-MC, BAT, and RNN in the UNSW-NB15 provided the highest accuracy of 79%. Similarly, Nagaraja et al. utilized the CNN and LSTM models in the KDD data set with an accuracy rate of 85.14%. One of the researchers' works is by Larriva-Novo et al., who demonstrated how intelligent data preprocessing could attain the highest accuracy rate of 45%. Furthermore, Aamir et al. indicated that models like SVM and isolation forest could provide over 90% accuracy when implementing the CICIDS2017 dataset. Finally, Wanjau et al. designed a model based on convolutional neural networks in detecting SSH Brute force attacks, providing an accuracy rate of 94.3% and F1-Score of 91.8%.

III. System Architecture

Our process utilizes a pipeline of eight stages that facilitates learning and evaluation using the same data by both classifiers. Stages 1 to 4 constitute a typical data processing pipeline performed before presenting either algorithm with the data. After processing through the four stages, the data set undergoes splitting into two parts for use as the train and test sets respectively. Both algorithms are trained using the exact same train and test sets. Thus, any difference in performance must be as a result of differences in the algorithms used.

A. Stage Descriptions

Steps 1 through 3 involve data preparation. The UNSW-NB15 dataset, containing 82,332 rows and 45 features, is imported with the aid of the Python and Jupyter Notebook programming environments. The detection and removal of outliers, handling of missing data, and consistency checking are done through statistical tests. The categorical features are converted to numeric with the help of label encoding, and all the numeric features are scaled using Standard Scaler such that none of the features will influence the algorithm with respect to their magnitude.

The fourth step concerns the splitting of the data set. There will be one split made, with the ratio between the training and testing sets being 80/20. This is uniform across both algorithms from here onwards.

Steps 5A to 7A constitute the Random Forest algorithm. Hyperparameters of the algorithm are tuned to determine the best parameters through the training of the 80% training set. Confusion matrix is generated using the test results from the 20% testing set. Four measures; precision, recall, f-score, and accuracy are calculated, each being about 89%. Similarly, Steps 5B to 7B constitute the XGBoost algorithm. The algorithm yields a score of about 90% for all

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

four measures. Finally, in Step 8, comparison is made between both algorithms, showing that XGBoost outperforms Random Forest.

Table 1: UNSW-NB15 Dataset Summary

Total Rows	Total Columns
82,332	45

IV. RESULTS AND DISCUSSION

A. Random Forest Classifier

The Random Forest algorithm is a type of machine learning algorithm whereby several decision trees are constructed in the training process. The final outcome is obtained through consideration of the votes made by all the decision trees. Thus, the model is stronger than when a single decision tree is used. Additionally, there is reduced risk of overfitting since the machine learning algorithm is derived from several decision trees and not a single one. In this case, the random forest algorithm was applied to all the ten classes of the UNSW-NB15 data set. From Table 2, the random forest algorithm had an accuracy of about 89%.

Table 2: Performance Measure — Random Forest

AC (%)	PR (%)	RE (%)	F1 (%)
89	89	89	89

B. XGBoost Classifier

The working principle behind the XGBoost model is the gradient boosting approach, which involves building a sequence of trees that concentrate on minimizing the residual errors generated by their predecessors. The effectiveness of the algorithm improves at each step. Moreover, XGBoost has integrated regularization and optimization capabilities and, therefore, can process large data volumes without overfitting; hence, it is an optimal choice for network intrusions recognition purposes. For all four measures tested, XGBoost yielded around 90% accuracy, whereas the Random Forest model produced 89% results.

Table 3: Performance Measure — XGBoost

AC (%)	PR (%)	RE (%)	F1 (%)
90	90	90	90

C. Work Comparison

The comparison between our proposed model and several other models already present in the current literature is outlined in Table 4 below. Our proposed model's excellent performance is evident from the accuracy of 90% in terms of attack prediction using the UNSW-NB15 dataset. It is important to note that such good performance is a significant step forward from the previous record-high performance level of 79% in terms of predicting attacks through the use of CNN algorithms, which is discussed in detail in the paper by Jiang et al.
Furthermore, the accuracy of the results attained by our model is also better compared to 85% attained through the use of CNN-LSTM algorithm by Su et al. based on KDD dataset. There are two reasons for our model's good performance. First, the UNSW-NB15 dataset used in our analysis is more recent and therefore more realistic than the outdated KDD dataset used in the previous studies. Hence, the models we created were based on the data similar to what would be observed in reality. Second, ensemble algorithms (i.e., Random Forest and XGBoost) outperform CNN and LSTM deep learning algorithms because of low computational complexity and simple tuning process.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Table 4: Work Comparison — Proposed Models vs. Related Literature

Research Work	Dataset	Algorithm	Avg. Accuracy
Jiang et al. [3]	UNSW-NB15	CNN, BAT-MC, BAT	79%
Su et al. [2]	KDD	CNN + LSTM	85%
This Work	UNSW-NB15	Random Forest	89%
This Work	UNSW-NB15	XGBoost	90%

V. CONCLUSION AND FUTURE WORK

The current research has proposed a comprehensive machine learning process involving eight phases, namely, data acquisition, data preprocessing, training model, validation of the trained model, and evaluation of model performance. The accuracy scores of Random Forest and XGBoost models were estimated to be approximately 89% and 90%, respectively. These results are significantly higher than those that have been reported in previous literature for similar or the same data set using deep learning approaches (79% and 85%). Therefore, XGBoost would be the preferred algorithm for DDoS attack detection in real-time scenarios.

In order to advance, some issues must be addressed in the course of future research. For example, unsupervised learning methods could be applied to detect network traffic in real time, but with no labels attached. Next, transformers must be experimented on with the UNSW-NB15 dataset in order to see if attention-based systems would improve accuracy. Finally, an adaptive threshold real-time detector for addressing concept drift would prove invaluable.

ACKNOWLEDGMENT

In this regard, it is our pleasure to thank the Australian Centre for Cyber Security (ACCS) for providing us with an open-source dataset called “UNSW-NB15”. Moreover, we would also like to acknowledge the open-source Python software development community, especially scikit-learn, XGBoost, and Pandas, who have provided us with the necessary packages for conducting the experiments described in this paper. Specifically, we wish to acknowledge the support provided by the Department of Computer Science & Engineering of Vasavi College of Engineering, Hyderabad.

REFERENCES

[1] G. Karatas, O. Demir, and O. K. Sahingoz, "Improving the effectiveness of machine learning based IDS on imbalanced and updated dataset," IEEE Access, vol. 8, pp. 32150–32162, 2020.
[2] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep learning approaches for network intrusion detection using NSL-KDD dataset," IEEE Access, vol. 8, pp. 29575–29585, 2020.
[3] H. Jiang, Z. He, G. Ye, and H. Zhang, "A network intrusion detection approach based on PSO-XGBoost model," IEEE Access, vol. 8, pp. 58392–58401
[4] Proceedings of the Vasavi Conference on Emerging Technologies (VCET-2025), Vasavi College of Engineering, Hyderabad, 2025.