International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Utilizing Random Forest for DDoS Attack Detection

Anupama Mishra
Computer Science & Engineering, Himalayan School of Science & Technology, Swami Rama Himalayan University, Jolly Grant, Dehradun, Uttarakhand, India

Abstract- The "Distributed Denial of Service" (DDoS) attack represents one of the most common forms of cyber assaults. The goal of DDoS is to overwhelm the server machine with an overwhelming number of data packets. This causes the bulk of the network bandwidth and server resources to be used leading to a Distributed denial-of-service problem. In this paper, we employed a random forest classifier for detecting the DDoS attack. This leads to an improvement in accuracy as well as a reduction in the amount of processing overhead required. Utilizing the CICDDOS2019 dataset, our experimental results showcased an impressive accuracy rate of 99.81%.

Keywords**:** DDoS, Random Forest, Machine Learning, Cloud computing.

1. INTRODUCTION

Technology has made sharing storage, network bandwidth, and computing resources easier. Cloud computing helps companies and individuals meet their needs. The Service Level Agreement (SLA) provides detailed information about the services and safety measures individuals who store their data on the cloud often express concerns about the potential for the cloud to undermine their privacy [1]. Cloud servers have security measures, yet an undetected attack is possible. On-premises software applications and static security approach is challenged by cloud platform' dynamic features. The exploitation of a large number of hosts, known as zombies, is used to perform DDOS attacks against a targeted system. A sudden and rapid spike in traffic congests the network, preventing regular data transfer to its intended recipients. The phenomenon disrupts network traffic patterns. The attack has been linked to extortion. These attacks have also grown in size and complexity [2-3]. DDoS attacks can affect services of server. Figure 1 presents the abstract view of DDoS attacks.

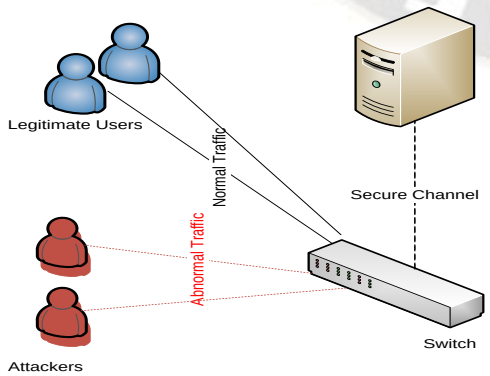

Figure 1: Abstract view of DDoS Attack

Current DDoS mitigation systems classify data packets as malicious or genuine, according to authors signature-based and anomaly-based approaches are the main categories. signature-based detection uses database-stored attack signatures. Signature-based detection cannot discover novel malware variants before their signatures are added to the database. Cybercriminals can avoid detection by altering database specifications between new attacks.

Another attack prevention strategy is anomaly-based detection. Anomaly-based attack detection uses specified criteria to identify attacks in network activity and patterns. These laws are highly adaptive due to their nature. In the event that any rules are violated, the system will audible notifying you. Despite threshold settings, anomaly-based detection strategies require manual development by knowledgeable professionals. The complexity, time, and human intervention required to design anomaly-based detection systems make them difficult to execute [4].

To address challenges in both signature-based and anomaly-based detection, ML techniques have been employed. While ML is a common approach in numerous research endeavors, there is a need to pioneer novel models to enhance accuracy. This research work suggests using Random Forest classifier we can detect DDoS attack and its type. The proposed method strives for high accuracy and low false alarms. Cybersecurity methods increasingly require the ability to automatically recognize and stop malicious network activity.

The paper proceeds with the following structure: Segment 2 provides Related work, while Segment 3 outlines the proposed methodology. Subsequently, Segment 4 delves into

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

the findings and subsequent discussion, while Segment 5 provides the concluding remarks of the research.

2. RELATED WORK

In their study, Gaurav et al. Introduced an innovative strategy to tackle authentication and security challenges within the realm of smart vessels in maritime transportation. The authentication of access for smart vessels and devices is facilitated through the utilization of an identity-based method. However, it should be noted that the methodology employed in this study is constrained by its focus on maritime transportation alone.

In a paper [5], a proposal was put up for the implementation of an auction mechanism featuring many attributes. The objective of this mechanism is to efficiently mitigate DDoS attacks targeting a network. The researchers developed a detection strategy based on reputation, where the user's reputation is determined by their least utility. This proposal suggests the implementation of two distinct payment plans, one for legitimate users and another for fraudulent users, along with the incorporation of an identification mechanism in conjunction with the existing identification technique. This approach employs a greedy resource allocation model to effectively divide resources among authorized users. Differential payment systems are specifically designed to impose penalties on those who engage in manipulative behavior with the intention of maximizing their allocation of limited resources.

In this research paper [6], the author introduces a methodology employing Bayesian game theory to detect instances of DDoS attacks. It is often suggested that both the service provider and authorized users participate in network monitoring to collect probabilistic data to determine whether a certain user is engaging in hostile activities or not with access to probabilistic information, Both the service provider and authorized users possess the capability to adjust their operation and responses in response to malicious activities within the system.

3. PROPOSED WORK

In this study, our primary focus is on datasets, the extraction of key features from those datasets, modeling with a classifier, and, lastly, prediction based on a testing dataset. After then, the effectiveness of the Random Forest classifier is assessed using several performance measures. The necessary actions are broken down into numerous tasks, such as the setting up of experiments, the selection of data sets and features, including the utilization of classifiers and the assessment of confusion metrics and performance metrics Figure 2 depicts "process sequence.

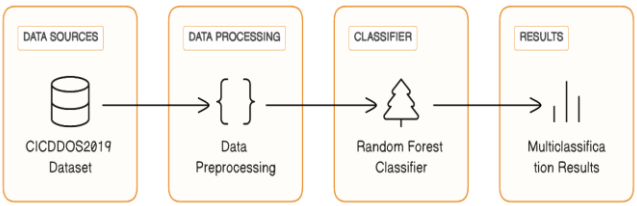

Figure 2: Assessment of confusion metrics and performance metrics process sequence.

Dataset

The CICDoS2019 dataset [7] was developed specifically for the purpose of DDoS attack detection research and is an important resource in the field of cybersecurity. It simulates a wide number of different types of DDoS attacks, in addition to conventional network traffic data and many different types of network traffic features. It makes it easier to create and evaluate DDoS detection algorithms by providing labeled instances of both benign and malicious behavior. This contributes to the advancement of network security measures. We have divided the dataset, allocating 70% of it for training purposes and reserving 30% for testing.

Random Forest

The machine learning technique commonly referred to as "random forest" was developed by Leo Breiman and Adele Cutler. This phrase specifically denotes this approach. In this algorithm, multiple individual decision trees are utilized, and their outcomes are aggregated to make a final decision. This method is versatile and applicable to both classification and regression scenarios, and it is user-friendly and adaptable, which are both factors that have contributed to its widespread deployment [8-10].

Since the random forest model is constructed from a group of decision trees, it is advantageous to commence with a concise elucidation of the algorithm governing decision trees. This would be done to maximize the efficiency of the random forest model. The first step in the decision-making process involves posing a fundamental query and selecting the best feature for root to build a tree. In most cases, the Classification and Regression Tree algorithm is used to train decision trees, which are then used to search for the optimal data split in order to create a subset of the data. To determine how well a split was performed, Measures like Gini impurity, information gain, or mean square error can be utilized.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 12 Issue: 2
Article Received: *25 November 2023* ***Revised:*** *12 December 2023* ***Accepted:*** *30 January 2024*

Though decision trees are a common type of supervised learning algorithm, they can still encounter various challenges, such as bias and overfitting. However, with the random forest algorithm, when numerous decision trees are combined to form an ensemble, they are able to make more accurate predictions, and this is especially true when the individual trees have no correlation with one another.

4. EXPERIMENTAL CONFIGURATION

The experiments were carried out on a laptop utilizing Python 3.10.4 64-bit through Jupyter Notebook within VS code. The laptop was equipped with an 11th Generation Intel(R) Core (TM) i5-1135G7 CPU with a clock speed ranging from 2.40GHz to 3.32GHz. It had 16GB of DDR4 GPU RAM and storage comprising a 1TB HDD along with a 256GB SSD. Furthermore, the experiments utilized several libraries including scikit-learn, Matplotlib, NumPy, and Pandas.

Feature selection

The first thing that was done was an inspection of the data validation process to ensure accuracy and reliability. Because the dataset was so vast, a piece of it had to be removed in order to maintain the imbalance problem and avoid overfitting or underfitting the data, depending on the situation. Following the completion of the whole data set's processing and the use of the tree classifier [11-12], we chose the top 15 features for the prediction. The top 15 distinguishing characteristics are detailed in Figure 3.

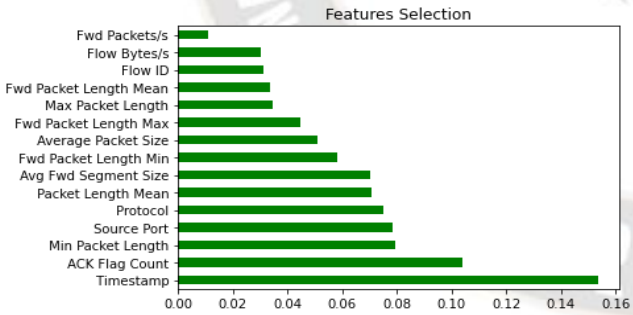

Figure 3: Best 15 features by using Extra Tree Classifier

5. OUTCOME ASSESSMENT

In this research, to analyze our method, we utilized machine learning metrics such as precesion, recall, and F1-score to measure the quality level. The performance metrics mentioned in prior works [13-14] encompass precision, recall, F1-score, and accuracy.

Precision evaluates the ratio of accurately identified instances to the total instances labeled as positive. Recall denotes the ratio of correctly identified instances among all actual positive instances. Accuracy shown in figure 4. The accuracy of the model is 99.81% and the other metrices are shown in figure 5.

ACCURACY=(TP+TN)/(TP+FP+TN+FN)

PRECISION=TP/(TP+FP)

RECALL=TP/(TP+FN)

F1 SCORE=2* ((PRECISION*RECALL))/((PRECISION+RECALL))

Figure 4: Performance Metrics Formulas

	precision	recall	f1-score
BENIGN	0.99	1.00	1.00
LDAP	1.00	1.00	1.00
MSSQL	0.92	0.99	0.96
NetBIOS	1.00	1.00	1.00
Portmap	0.98	1.00	0.99
Syn	1.00	1.00	1.00
UDP	1.00	1.00	1.00
UDPLag	0.92	0.73	0.81

Figure 5: Performance Measurement for Random Forest Classifier

In ML, The Receiver Operating Characteristic curve is employed for assessing binary classification models it plots the sensitivity against 1-specificity across different threshold values, providing an evaluation of a model's ability to distinguish between classes .The model performs better when the Receiver Operating Characteristic curve approaches the plot's upper left corner. The area under the ROC curve (AUC) measures behavior, with 1 indicating perfect classification and 0.5 random chance. The ROC curve helps compare and select models and comprehend classification task sensitivity-specificity trade-offs. Here, Figure 6 shows the effectiveness of the model.

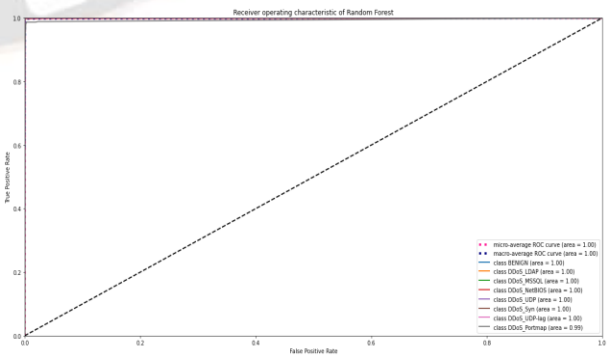

Figure 6: Receiver Operating Characteristic of Random Forest Curve

6. CONCLUSION

In conclusion, achieving a notable accuracy rate of 99.81% using a random classifier on the CICDDoS2019 dataset is an impressive outcome. However, it is important to note that this level of accuracy may not directly transfer into an effective and viable solution for practical DDoS detection. Additionally, the rest of other metrices like precision, recall and f1-score also help to identify and classify the DDoS attacks effectively.

REFERENCES

[1] Alhalabi, W., Gaurav, A., Arya, V., Zamzami, I. F., & Aboalela, R. A. (2023). Machine Learning-Based Distributed Denial of Services (DDoS) Attack Detection in Intelligent Information Systems. International Journal on Semantic Web and Information Systems (IJSWIS), 19(1), 1-17.

[2] Ling, Z., & Hao, Z. J. (2022). Intrusion detection using normalized mutual information feature selection and parallel quantum genetic algorithm. International Journal on Semantic Web and Information Systems (IJSWIS), 18(1), 1-24.

[3] Gaurav, A., Gupta, B. B., & Panigrahi, P. K. (2022). A novel approach for DDoS attacks detection in COVID-19 scenario for small entrepreneurs. Technological Forecasting and Social Change, 177, 121554.

[4] Negi, P., Mishra, A., & Gupta, B. B. (2013). Enhanced CBF packet filtering method to detect DDoS attack in cloud computing environment. arXiv preprint arXiv:1304.7073.

[5] Dahiya, A., & Gupta, B. B. (2020). Multi attribute auction based incentivized solution against DDoS attacks. Computers & Security, 92, 101763.

[6] [6] Govindaraj, L., Sundan, B., & Thangasamy, A. (2021, December). An intrusion detection and prevention system for ddos attacks using a 2-player bayesian game theoretic approach. In 2021 4th International Conference on Computing and Communications Technologies (ICCCT) (pp. 319-324). IEEE.

[7] [7] Akgun, D., Hizal, S., & Cavusoglu, U. (2022). A new DDoS attacks intrusion detection model based on deep learning for cybersecurity. Computers & Security, 118, 102748.

[8] [8] Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

[9] [9] Qi, Y. (2012). Random forest for bioinformatics. Ensemble machine learning: Methods and applications, 307-323.

[10] [10] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[11] [11] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

[12] [12] Baby, D., Devaraj, S. J., & Hemanth, J. (2021). Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach. Turkish Journal of Electrical Engineering and Computer Sciences, 29(8), 2742-2757.

[13] [13] Kleijnen, J. P., & Smits, M. T. (2003). Performance metrics in supply chain management. Journal of the operational research society, 54, 507-514.

[14] [14] Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006.

[15] [15] Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). ROC curve estimation: An overview. REVSTAT-Statistical journal, 12(1), 1-20.

[16] [16] Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean journal of radiology, 5(1), 11-18.