INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# DDoS Attack Detection Using Machine Learning Techniques

**Rituparna Borah[1], Satyajit Sarmah[1]*, Nitin Choudhury[1], Hriman Mahanta[1], Anjan Chodhury[1]**

**1** Department of Information Technology, Gauhati University, Guwahati, Assam, India

OPEN ACCESS

**Received:** 23-03-2023

**Accepted:** 26-06-2023

**Published:** 02-11-2023

**Editor:** Guest Editor: Dr. Madhuryya Saikia & Dr. Niranjan Bora

**Citation:** Borah R, Sarmah S, Choudhury N, Mahanta H, Chodhury A (2023) DDoS Attack Detection Using Machine Learning Techniques. Indian Journal of Science and Technology 16(SP2): 76-82. https://doi.org/10.17485/IJST/v16iSP2.9526

* **Corresponding author**.

satyajitnov2@gmail.com

**Funding:** None

**Competing Interests:** None



Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

Network Traffic analysis is an important part of network security. With the increase in the usage of internet, new kinds of network security threats are becoming prominent. One of the biggest threats to it is the Distributed Denial of Service (DDoS) attack. **Objective:** The primary objective of our work is to create a DDoS dataset and to classify the attack based on its behavioural analysis. **Methods:** For creating a DDoS dataset, a proper virtual lab environment is set-up. After setting up the environment and virtual network, DDoS attack is performed on the victim machine and the network traffics are captured. Along with the DDoS data, benign network traffics are captured as well. After creating the dataset, different features are extracted from the network traffics and finally used different Machine Learning approach for classifying the features whether the traffics are benign traffics or DDoS traffics. **Findings:** From the experimental result, it is clear that the proposed method can create DDoS traffic and classify different types of DDoS attacks in an efficient manner. From the result analysis, it is seen that the KNN clustering algorithm performs better classifications than the other machine learning algorithms. **Novelty:** The primary novelty in the proposed work is about the dataset that has been created. The DDoS dataset that is used in the proposed work is heterogeneous. The dataset includes DDoS traffics from both the global internet and local network. On this data, among most of the primary machine learning algorithms, Random Forest and K-Nearest Neighbour Classifier performs better with classification accuracy of 99.44% and 99.58%.

## 1 Introduction

One of the most prominent and costliest network security threats is the Distributed Denial of Service (DDoS) attack. A DDoS attack occurs when an attacker compromises a huge number of computers and then uses those computers to flood a server with an overwhelming number of packets. A DDoS attack is an attack on the availability of the system. Therefore, detection of DDoS attack is becoming very important for preventing the system and resources.

Jiahui Chen and their team proposed a flexible and easily configurable machine learning model for network traffic classification [(1)]. They used the statistics of sequences of packets to differentiate between known traffic from unknown traffic. Their method of classification is mainly based on likelihood estimation.

T. P. Fowdur and their team investigated the performances of various network traffic capturing tools for feature extraction and also find the efficiency of various machine learning algorithms [2]. They used six different Internet applications and two different network anomalies. In their experiment, it is found that ColasoftCapsa network capturing tool performs better than the others tools in terms of classification.

Pranita Mane and their team approached Naïve Bayes algorithm and Support Vector Machine for classifying the network traffics as benign or DDoS [3]. They have designed a system which consists of various operations like capturing packets, processing of the data and classifying them.

OnsAouedi and their team proposed methods for finding the most relevant features for network classification. They compared the components of various features in the classification process. Then used tree-based machine learning algorithm for classifying the data [4].

Mohsen Ghasemi and their team addressed the security risks posed of Internet of Things (IoT) [5]. The paper emphasizes the importance of considering social engineering as a significant threat to IoT security, as human trust and interactions are integral parts of the IoT ecosystem.

Y. N. Soe and their team had designed and developed an artificial neural network model for classifying botnet traffic [6]. They used the BoT-IoT and implemented a data re-sampling technique, SMOTE to balance the data. From the result it has been observed that the system was quite effective with a basic configuration of Artificial Neural Network (ANN).

Jiangtao Pei and their team [7] proposed an approach to classify DDoS traffic using random forest machine learning classifier.

From the literature, it is seen that the DDoS classifications are performed either on the global internet or on a specific local network which makes the classifier model more environment specific. In the proposed work, we have used heterogeneous dataset and performed classification on the extracted features of this data which make the classifier model more flexible for the classification task.

Ali T.E. and their team [8] approached a machine learning and deep learning-based classification techniques for detection of distributed denial-of-service (DDoS) attacks in software-defined networks (SDNs). They have focused on the types of detection approaches, methodologies, strengths, weaknesses, benchmark datasets, pre-processing strategies, and performance metrics employed.

Kumari K and their team introduced a mathematical model to address the significant threat posed by Distributed Denial of Service (DDoS) attacks [9]. These attacks impair a server's ability to provide resources to legitimate users, resulting in decreased bandwidth and buffer size. Machine learning algorithms, namely Logistic Regression and Naive Bayes, are employed for detecting both attack and normal scenarios. The CAIDA 2007 Dataset is utilized for experimental analysis, training, testing, and validation of the algorithms. The Weka data mining platform is utilized for implementation, and the study's results are analyzed and compared with other machine learning approaches for detecting denial of service attacks.

This paper is organized as follows. Section 2 provides the methodology used Section 3 provides the result and discussion and Section 4 discusses the Conclusion.

## 2 Methodology

From literature, it is observed that most of the studies are conducted on the data of a single network architecture or a single type of network, i.e., local area network or global network for which the trained model becomes more network specific. Along with that, usage of neural network makes the model more weighted. In order to overcome these problems, in this work, a novel heterogeneous dataset is created and extracting the features out of these data, lightweight machine learning classifiers are used to classify these data. The flow diagram of the methodology is as shown below (Figure 1).

In order to simulate DDoS attacks and capture the network traffic, a virtual environment and a virtual network is set-up. After doing so, from different endpoints of the network, DDoS attack is performed on the victim machine using LOIC (Lower Ion Orbit Cannon), which is a DoS and DDoS traffic generator. LOIC simulates DDoS attack by sending a huge amount of TCP SYN request to the victim machine. In the virtual environment and network, DDoS attack is performed using the tool. To make sure that the packets captured actually simulate packets coming from multiple computers, we have used multiple threads. This is done so that this simulation comes as close as possible to that of a real-world DDoS attack.

The network traffics have been captured in the victim machine using Wireshark. To capture the normal packets, we have used Wireshark. Wireshark is a free and open-source network protocol analyzer [8] through which network traffic can be captured and analyzed. The normal flow of packets in the network is captured for an instant of time and exporting the packets as .pcap files, these packets are labeled as benign traffic. While performing the DDoS attack, the packets are captured as well and exporting as .pcap files, these packets are labelled as DDoS traffic. The data has been recorded for approximately 14 days.

To distinguish DDoS traffic from regular network traffic, it is necessary to apply packet filtering techniques to the captured network traffic. Based on the type of the packets sent by LOIC, the normal and DDoS packets are distinguished. Now, the benign

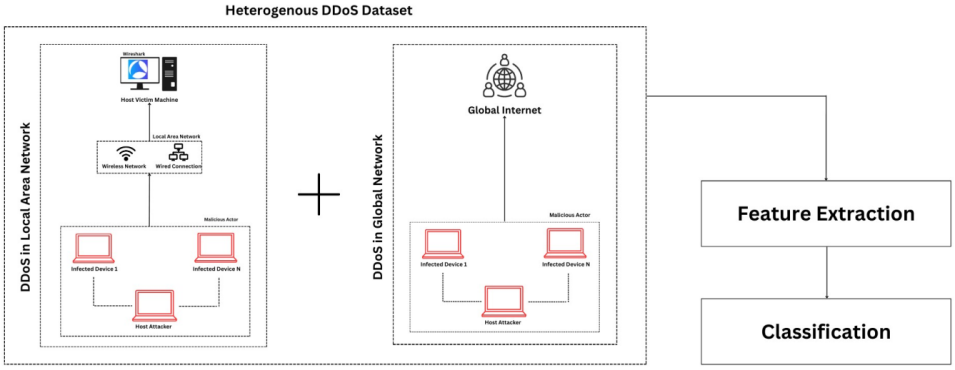


**Fig 1. Methodology Workflow Diagram**

and DDoS traffic of global network is added to the collected data and labelled accordingly.

Now from this heterogeneous dataset, different feature vectors are extracted. Features are extracted by applying a sliding time window of 500 milliseconds on the data. For each time window, a total of 16 features are extracted and accumulated. The extracted features are as shown below.

(i) **Tcp_frame_length:** It is the total length of a TCP packet in the datalink layer.
(ii) **tcp_ip_length:** It is the total length of a TCP packet in the network layer.
(iii) **tcp_length:** It is the total length of a TCP packet in the transport layer.
(iv) **udp_frame_length:** It is the total length of a UDP packet in the datalink layer.
(v) **udp_ip_length:** It is the total length of a UDP packet in the network layer.
(vi) **udp_length:** It is the total length of a UDP packet in the transport layer.
(vii) **num_tls:** It is the total number of TLS connections in a particular time window.
(viii) **num_http:** It is the total number of HTTP packets in a particular time window.
(ix) **num_dhcp:** It is the total number of DHCP packets in a particular time window.
(x) **num_dns:** It is the total number of DNS packets in a particular time window.
(xi) **num_tcp:** It is the total number of TCP packets in a particular time window.
(xii) **num_udp:** It is the total number of UDP packets in a particular time window.
(xiii) **num_igmp:** It is the total number of IGMP packets in a particular time window.
(xiv) **num_connection_pairs:** It is the total number of connected pairs in a particular time window.
(xv) **num_ports:** It is the total number of ports in a particular time window.
(xvi) **num_packets:** It is the total number of packets in a particular time window.

Now, the number of transport layer protocols and number of application layer protocols is calculated. The result of the number of packets containing various transport layer and various application layer protocols are as shown in Table 1.

**Table 1. Distribution of Transport Layer Protocol and Application Layer Protocol**

| Protocols | Layers | No. of packets | Percentage |
|---|---|---|---|
| TCP | Transport Layer | 250117 | 91.58% |
| UDP | Transport Layer | 22981 | 8.42% |
| HTTPS | Application Layer | 261669 | 96.16% |
| DNS | Application Layer | 10257 | 3.77% |
| HTTP | Application Layer | 125 | 0.05% |
| DHCP | Application Layer | 33 | 0.01% |
| BOOTP | Application Layer | 33 | 0.05% |

After extracting the most important features out of the dataset, now different machine learning classifiers such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, Support Vector Machine etc. are trained using the extracted features. For each classifier, performance matrix is evaluated and based on it, classification accuracy and different sensitivity analytic parameters such as precision, recall, f1 scores are calculated.

## 3 Results and Discussion

In the proposed research work, a total of six binary classification algorithms namely Logistic Regression, Decision Tree, Random Forest Classifier, K-Nearest Neighbors Classifier, Naïve Bayes Classifier and Support Vector Machine has been taken into consideration in order to draw comparison by the dataset that has been created. Performance of each classifier algorithm is compared based on its confusion matrices. The derived confusion matrices for each of the classifiers are as shown in Figures 2, 3, 4, 5, 6 and 7.

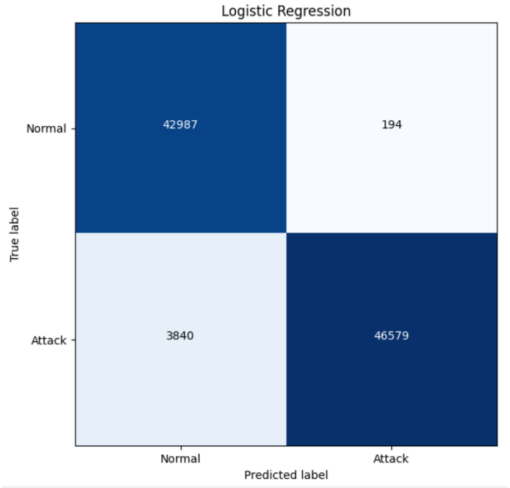


**Fig 2. Confusion matrix for Logistic Regression**

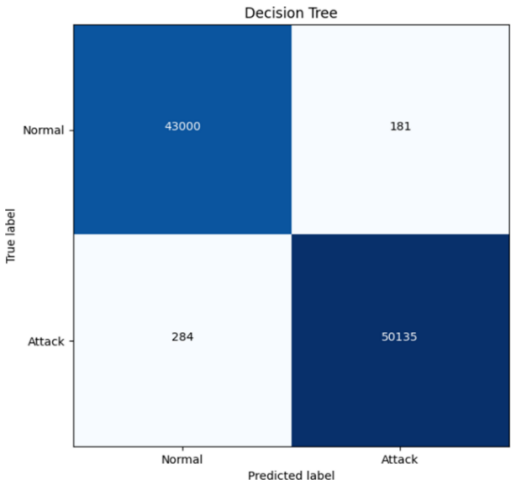


**Fig 3. Confusion Matrix for Decision Tree Classification**

Based on the confusion matrix, different performance calculation parameters such as Accuracy, Precision, Recall are evaluated. Table 2 shows the value of the various parameters of the classifier.

From the performance evaluation of the classifier algorithms, it is found that random forest classifier and k-nearest neighbors classifier perform the best with an accuracy of 99.58% and 99.44% respectively. Ravipati Rama Devi and their team [10], have

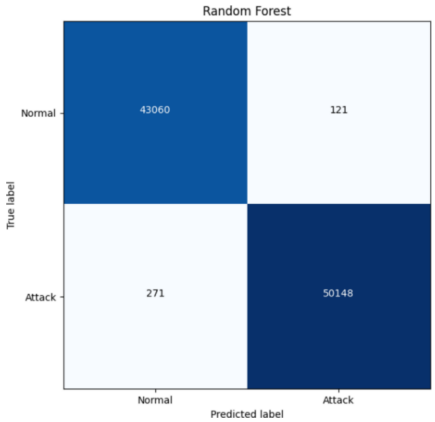


**Fig 4. Confusion Matrix for Random Forest Classifier**

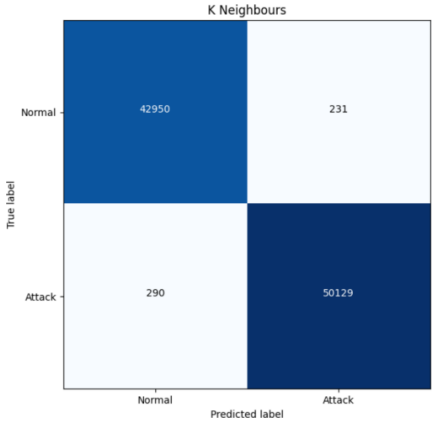


**Fig 5. Confusion Matrix of K-Nearest Neighbors Classifier**

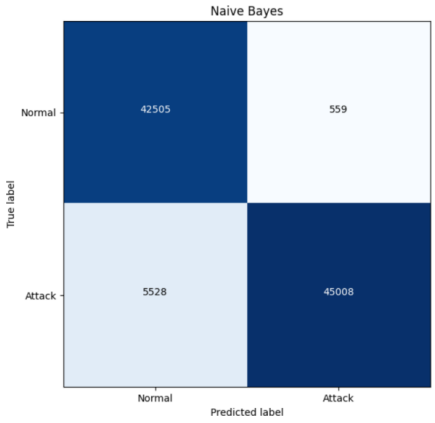


**Fig 6. Confusion Matrix of Naïve Classifier**

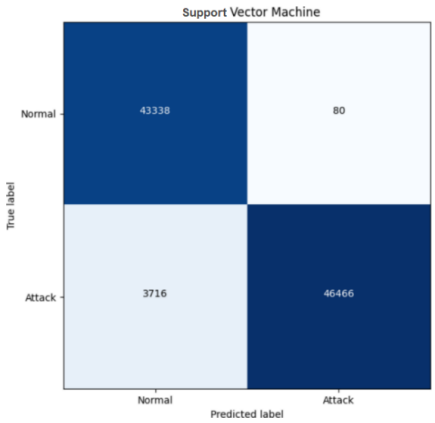


**Fig 7. Confusion Matrix of Support Vector Machine**

**Table 2. Various parameters of the classifiers**

| Classifier | Precision | Recall | F1 Score | Accuracy |
|---|---|---|---|---|
| **Logistic Regression** | 0.9955 | 0.9180 | 0.9552 | 0.9569 |
| **Decision Tree Classifier** | 0.9958 | 0.9934 | 0.9946 | 0.9950 |
| **Random Forest Classifier** | 0.9972 | 0.9937 | 0.9955 | 0.9958 |
| **K-Nearest Neighbor Classifiers** | 0.9947 | 0.9933 | 0.9940 | 0.9944 |
| **Naïve Bayes** | 0.9870 | 0.8849 | 0.9332 | 0.9350 |
| **Support Vector Machine** | 0.9982 | 0.9210 | 0.9580 | 0.9594 |

presented a comparative analysis of the performance of the machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine and Naïve Bayes on the KDD-99 and NSL-KDD dataset. They evaluated the algorithms based on the accuracies of the models. The comparative analysis of the machine learning classifier algorithms in terms of accuracy, using the NSL-KDD and KDD-99 dataset and the proposed heterogeneous dataset are as shown in Table 3 .

**Table 3. Comparitive study of Different Machine Learning Classifier Algorithm in Terms of Accuracy**

| Dataset | Logistic Regression | Decision Tree Classifier | Random Forest Classifier | K-Nearest Neighbor Classifiers | Naïve Bayes | Support Vector Machine |
|---|---|---|---|---|---|---|
| **KDD-99** | 79.7% | 81.05% | 99.0% | 94.17% | 92.4% | 83.09% |
| **NSL-KDD** | 97.4% | - | 99.7% | - | 89.5% | - |
| **Proposed Heterogenous Dataset** | 95.69% | 99.50% | 99.58% | 99.44% | 93.50% | 95.94% |

## 4 Conclusion

As the popularity and demand for IoT devices continue to soar, they have become increasingly attractive targets for malicious actors. Nowadays, a wide range of everyday items, such as light bulbs, outdoor cameras, and watches, are interconnected via Wi-Fi with limited security measures in place. Recent research highlights the vulnerability of these ubiquitous devices, serving as a growing attack surface, especially considering the abundance of sensitive data stored on connected smartphones. This project focuses on combating DDoS attacks independent of network architecture or type.

In this paper, a flexible lightweight machine learning based approach is introduced to mitigate Distributed Denial of Service (DDoS) attack which is independent of network structure. For which a heterogeneous dataset is created that contains DDoS network traffic of local area network as well as global network. Extracting 16 different important features out of the dataset, different lightweight machine learning classifiers are used in order to classify the benign and DDoS traffic. From the experimental result, it is clearly visible that the proposed approach is efficient for classifying DDoS attacks independent of

network architecture.

In the future this model can be used in the real-world server for classifying DDoS attacks and to mitigate it. Another future work that can be done is to use unsupervised learning to classify network traffic.

## 5 Declaration

Presented in Fourth Industrial Revolution and Higher Education (FIRHE 2023) during 23rd-25th Feb 2023, organized by DUIET, Dibrugarh University, India. The Organizers claim the peer review responsibility.

## References

1) Chen J, Breen J, Phillips JM, Van Der Merwe J. Practical and configurable network traffic classification using probabilistic machine learning. *Cluster Computing*. 2022;25:2839–2853. Available from: https://doi.org/10.1007/s10586-021-03393-2.
2) Fowdur TP, Baulum BN, Beeharry Y. Performance analysis of network traffic capture tools and machine learning algorithms for the classification of applications, states and anomalies. *International Journal of Information Technology*. 2020;12:805–824. Available from: https://doi.org/10.1007/s41870-020-00458-0.
3) Mane P, Parkar Y, Patel J, Sanghavi V, Walanje A. Traffic Classification Using Machine Learning. In: 2nd International Conference on Advances in Science & Technology (ICAST-2019), 8th-9th April 2019, Mumbai, India. 2019;p. 1–4. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3372181.
4) Aouedi O, Piamrat K, Parrein B. Performance evaluation of feature selection and tree-based algorithms for traffic classification. In: 2021 IEEE International Conference on Communications Workshops (ICC Workshops), 14-23 June 2021, Montreal, QC, Canada. IEEE. 2021. Available from: https://ieeexplore.ieee.org/document/9473580.
5) Ghasemi M, Saadaat M, Ghollasi O. Threats of social engineering attacks against security of Internet of Things (IoT): the selected papers of the first international conference on fundamental research in electrical engineering. In: Fundamental Research in Electrical Engineering;vol. 480 of Lecture Notes in Electrical Engineering. Singapore. Springer. 2019;p. 957–968. Available from: https://link.springer.com/chapter/10.1007/978-981-10-8672-4_73.
6) Soe YN, Santosa PI, Hartanto R. DDoS Attack Detection Based on Simple ANN with SMOTE for IoT Environment. In: 2019 Fourth International Conference on Informatics and Computing (ICIC), 16-17 October 2019, Semarang, Indonesia. IEEE. 2019. Available from: https://ieeexplore.ieee.org/document/8985853.
7) Pei J, Chen Y, Ji W. A DDoS Attack Detection Method Based on Machine Learning. In: Journal of Physics: Conference Series;vol. 1237. IOP Publishing. 2019;p. 1–6. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1237/3/032040.
8) Ali TE, Chong YW, Manickam S. Machine Learning Techniques to Detect a DDoS Attack in SDN: A Systematic Review. *Applied Sciences*. 2023;13(5):1–27.
9) Kumari K, Mrunalini M. Detecting Denial of Service attacks using machine learning algorithms. *Journal of Big Data*. 2022;9(56):1–17. Available from: https://doi.org/10.1186/s40537-022-00616-0.
10) Devi RR, Abualkibash M. Intrusion Detection System Classification Using Different Machine Learning Algorithms on KDD-99 and NSL-KDD Datasets - A Review Paper. *International Journal of Computer Science and Information Technology*. 2019;11(03):65–80. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3428211.