International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Study on Detection of DDoS Attacks using Machine Learning Techniques

Velisi Nirmala Priya

Student, Rajam, Vijayanagaram, 532127, India.

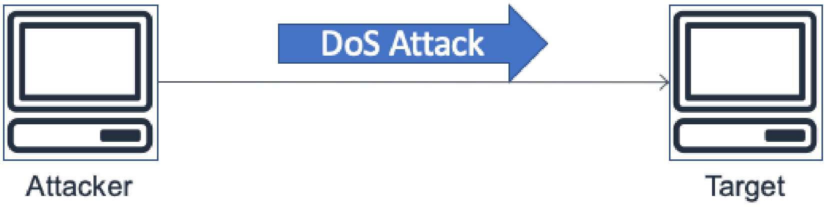
ABSTRACT

Cyber security is one of the crucial domains that helps protect information or data from cyber threats connected via the Internet. Even though there is security, cybercrimes are present in a plethora in our present society. Many cybercrimes, such as website spoofing, IP address spoofing, ransomware, etc., are increasing day by day. DDoS is a type of cybercrime that affects an individual or an enterprise with heavy traffic from multiple botnets. A DDoS (Distributed Denial of Service) attack targets a victim's server or network by overwhelming their network, which results in a denial of service. The detection of these attacks will result in a solution to stop such malicious activities. Some of the networking paradigms like SDN (Software-Defined Networking), which provide flexibility to identify malware attacks, are also vulnerable to DDoS attacks. There were numerous ways to detect these DDoS attacks. Some of them include machine learning techniques like Random Forest, a Support Vector Classifier, and many more that train the dataset and detect heavy traffic with decent accuracy.

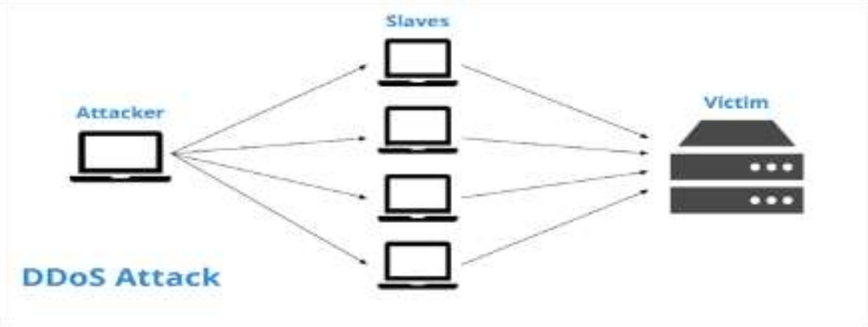
Keywords: Cyber security, Ransomware, DDoS (Distributed Denial of Service), SDN (Software-Defined Networking), Machine Learning, Malware, Vulnerable

1. Introduction

A DoS (Denial of Service) is a cyberattack that stops the service to a benign user by flooding it via a single system and is also known as a system-to-system attack. Specifically, the DDoS(Distributed Denial of Service) attack is similar to the DoS(Denial of Service) attack but occurs from multiple sources to a single victim to disrupt their services temporarily or for an indefinite period. A DoS attack can be easily identified and can be stopped since the flooding is via a single system and can stop the attacker from flooding their server.



Whereas, in these DDoS attacks, an attacker creates a botnet by sending malicious email attachments or some malware software to vulnerable systems. When these systems click on those email attachments, the malicious software is automatically downloaded to their systems, and thus they will be in the control of the attacker. The attacker can create multiple bots worldwide and control all of these bots. This collection or group of such bots creates a botnet that attacks a benign user to stop their service. A DDoS attack is a cyber threat done by a botnet to make the victim degrade his financial status, his/her reputation, or for many other reasons. There are many networking paradigms in which this problem appears. DoS is a cyber-attack in which the hacker attempts to make network resources or services disturb its legitimate users by flooding a large volume of malicious traffic that targets network resources.



DDoS is a large-scale cyber-attack in which the hacker uses multiple attack vectors to generate a large amount of malicious traffic from different sources. The presence of more than one attack vector from different sources causes a challenge to security administrators in the intrusion detection mechanism.

A breach of the Confidentiality, Integrity, and Availability (CIA) security principles in computer networks are referred to as an intrusion, according to the National Institute of Standards and Technology (NIST) [8].

- The two procedures of intrusion detection (ID) are monitoring intrusions and analyzing network events to look for any malicious packets or sources in the computer network or computer system.
- When a distinctly unrelated occurrence takes place in a legal or permitted event, the intrusion detection system (IDS) recognizes it as an intrusion [9].

One networking paradigm that enables software-based programming of networking devices is SDN (Software-Defined Networking). It recognizes malicious traffic and links issues fast. Despite having numerous functions, they are also susceptible to DDoS attacks [6]. To stop this malware attack, research is ongoing and has already begun. Some of the current research develops methods to anticipate DDoS attack flow using machine learning approaches. A key factor in the field of security is the quick development of artificial intelligence [9] using artificial neural networks (ANN), machine learning (ML), and natural language processing (NLP). With the use of feature selection algorithms, classifiers, and machine learning, it is possible to create intelligent IDS that can recognize unidentified attacks [10,16]. This DDoS attack, which was detected by a hybrid IDS, also targeted the most advanced area of information technology, the Internet of Things [15]. This technique has tremendous importance since DDoS assaults can cause major problems and are difficult to identify and neutralize [12]. The elapsed time during any form of real-time detection of DDoS attacks is also a challenge. The current methods for detecting DDoS attacks, however, have significant shortcomings, including high processing costs during detection and an inability to handle heavy network traffic heading toward the server. To distinguish DDoS from regular packets, classification algorithms classify the packets [13].

II. Literature Survey

The DDoS attack is one of the cyber threats that cause availability loss among the CIA triad of cyber security. The potential loss among any of these three would make a system/process be attacked by threats or vulnerabilities. Detection of such attacks can be possible by using many cyber tools as well as using many methodologies. But the classification of benign as well as normal traffic in the network can be well classified by using machine learning techniques. Many authors have experimentally done and detected DDoS attacks using machine learning techniques. Some of them are:

Nisha Ahuja et al. [1] and the other authors differentiated benign and DDoS traffic using Machine Learning techniques like Random forest, Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, Ensemble Classifier, Artificial Neural Network (ANN), Hybrid Machine Learning Model (SVC-RF) on SDN (Software-Defined Networking) dataset. Among all these SVC-RF achieved the highest accuracy of 98.8% and was a significant technique to predict many of the class labels correctly in the majority of the cases.

Mazhar Javed Awan et al. [2] and the other authors identified datasets that are large in size and consist of malware attacks that can be identified via many approaches like the big data approach. Small-size datasets can be done using traditional intrusion detection techniques. These large-sized datasets are classified as per machine learning techniques namely Random Forest, and Multilayer Perceptron with and without using the big data approach. The big data approach resulted in minimum average training and testing in minutes of 14.08 and 0.04 respectively whereas in non-big data approach resulted in maximum average training and testing of 34.11 and 0.46 respectively.

Deepak Kshirsagar et al. [3] and the other authors used the feature reduction method to detect the DDoS attack on a dataset known as CICDDoS 2019 with the J48 classifier. A feature reduction method by the union of information gain and correlation is used to get more relevant features. This feature reduction method obtains a maximum and minimum of 56 and 82.92% respectively of original features. The final dataset is verified with the well-known dataset named KDD Cup 1999. An analysis shows that both datasets show significant results in the detection of the attacks.

Swathi Sambangi et al. [4] and the other authors explained that discrimination between normal traffic and attack traffic is the most difficult task in these attacks where bots also act as legitimate users. Because of this, there may be a chance of an attack on Cloud resources and further, there may be a chance of financial revenues. Therefore, many experts are trying to minimize and detect these attacks by using machine learning techniques. The analysis is done on the CICIDS 2017 dataset using some feature selection techniques which resulted in an accuracy of 97.86%.

Vimal Gaur et al. [5] and the other authors used some of the Feature selection methods namely chi-square, Extra tree, and ANOVA are applied to four classifiers of machine learning. They are Random Forest, Decision Tree, k-Nearest Neighbors, and XG-Boost. This paper results in the early detection of DDoS attacks on IoT Devices. The research work is done on a CICDDoS 2019 dataset in which XG Boost showed an accuracy of 96.677%. The selection methods are continuously checked and the combination of ANOVA and XG Boost resulted in an accuracy of 98.374% with 15 features.

III. Detection of DDoS cyber attacks :

Many detection techniques were proposed by the authors and the detection processes are as follows:

ARTP (Anomaly-based Real-Time Prevention): By tracking system activity and categorizing it as either normal or unusual, an anomaly-based intrusion detection system can detect network and computer intrusions as well as misuse.

3.1 Machine learning techniques used:

Logistic Regression [1]– A classification algorithm used in Machine Learning and widely used in the cyber security domain to classify cases that is malicious or benign. It tries to determine whether a new sample best fits into a category.

Support Vector Classifier [1,18]- A classification as well as regression algorithm in Machine Learning which visualizes a decision boundary among various data points from a class of data points. This algorithm is mostly used to detect malware, intrusions, and spam filtering.

K-Nearest Neighbor [1,5]– It is an unsupervised machine learning algorithm that works on the principle of similar things working together. This algorithm stores the available data and classifies the data based on similar cases. This was used to identify the program's behavior as benign or malicious.

Random Forest [1,2,5,18] – A supervised machine learning algorithm that performs both classifications as well as regression. It considers a group of decision trees and considers the majority classes in the case of classification and the average in the case of regression. It uses bagging (reducing variance using noisy data) and features randomness to create a forest of trees that shows decent accuracy than that of an individual tree[14].

Ensemble Classifier[1]- It is a method used to provide better and more accurate than the rest of the models. It combines all the basic models to generate a new model, that provides more accuracy than the accuracy provided in individuals. Bagging, Stacking, Boosting, etc. are some methods considered ensemble classifiers.

Artificial Neural Network [1,17]– A derived biological neural network model which develops the structure of the human brain. These neural network mimics the human brain and helps to take decisions like humans. This network consists of three layers namely a) the Input layer b) the Hidden layer c) the Output layer where the input layer takes the input and provides an output from the output layer where the hidden layers act as weighted nodes to the input layer to give out the output.

Hybrid Machine Learning Model [1] – It is the combination of one or more machine learning algorithms that helps to reduce the disadvantage of one another to provide better results. No single ML approach is suitable for all problems, just as no single cap fits all heads.

Multi-Layer Perceptron [2,11]- It is one neural network that consists of three layers namely the input layer, hidden layer, and output layer among which the hidden layer plays a major role in computational operations and forward the data from the input to the output layer. These were majorly used in pattern classification, recognition, prediction, and approximation [11].

Multiple Linear Regression [4] - An extension of linear regression which in turn gives out the result by a response variable from several explanatory variables. This model helps to establish a linear relationship between independent and dependent variables.

3.2 Feature Reduction Strategies

Information Gain [2] – The most popularly used synonym of Information Gain is "mutual information". It is one of the most frequently used neural learning algorithms. The majorly used brain network called "vanilla " is sometimes referred to as Multi-Layer Perceptron. This algorithm shows a path for sophisticated Convolutional Neural Networks (CNN).

Correlation [2] – It is a feature reduction machine learning algorithm that, explains the relationship among the variables and derives a target variable by considering them as input data. This technique helps to identify the most important variables which are dependent. It was one of the most significant techniques during data analysis and data mining.

3.3 Classifiers in Machine Learning

J48 Classifier [3]- J48 is a decision tree classification technique for machine learning that is based on Iterative Dichotomiser 3. A continuous and categorical examination of the data is highly beneficial and intends to improve detections and predictions.

Decision Tree [1,5] - A supervised learning method called a decision tree may be used to solve classification and regression issues, but it is often favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's characteristics, branches for the decision-making process, and each leaf node for the classification result.

XGBoost [5]- Gradient Boosted decision trees are implemented using XGBoost technology. Many Kaggle Competitions are dominated by XGBoost models. Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given weight before being put into the decision tree that forecasts outcomes. Variables that the tree incorrectly anticipated are given more weight before being placed into the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It may be used to solve issues including regression, classification, ranking, and custom prediction.

ANOVA [5] - Analysis of variance (ANOVA) is a statistical method for examining differences in means. It consists of several statistical models and the accompanying estimating techniques (such as the "variation" within and across groups). Ronald Fisher, a statistician, created ANOVA. The rule of the total variance, on which the ANOVA is based, divides the observed variance in a given variable into components owing to various causes of variation. ANOVA generalizes the t-test beyond two means by offering a statistical test to determine if two or more population means are equal. In other words, the ANOVA is employed to determine if two or more means differ from one another.

Chi-square [5]– A feature selection machine learning classifier that helps in, whether the chosen predictive works or predicts the necessary operation. This works fine with categorical data in which the value of one variable doesn't matter /change the probability distribution of the other.

Extra Tree [5]- A similar random forest classifier that differs only among the construction of decision trees in the forest. Each decision tree is constructed from the training sample which selects the best feature to split the data as per some mathematical criteria like Gini Index.

3.4 Datasets used in papers:

***CICIDS DDoS 2017* [15]–** The benign and most recent common assaults are included in the CICIDS2017 dataset, which closely reflects data from the real world. It also contains the outcomes of a network traffic analysis performed using CICFlowMeter, with flows categorized according to the time stamp, source and destination IP addresses, source and destination ports, protocols, and attack (CSV files).

***KDD Cup 1999*[16]-** The Fifth International Conference on Knowledge Discovery and Data Mining, which was conducted in connection with KDD-99, The Third International Knowledge Discovery and Data Mining Tools Competition, utilized this data set. Building a network intrusion detector—a predictive model that can distinguish between bad connections, also known as intrusions or attacks—and good, regular connections—was the task for the competition. A standard collection of auditable data, including a wide range of simulated intrusions into a military network environment, is contained in this database.

***CICDDoS2019* [7]-** Despite the availability of IDS data sets, there are no recent datasets in the public domain that are dedicated to DDoS. To provide additional variation, DDOS data is taken from several IDS datasets that were developed in various years and various experimental DDoS traffic production technologies. These datasets include the CSE-CIC-IDS2018-AWS, CICIDS2017[7], and CIC DoS datasets (2016). The collected DDOS flows are coupled with "Benign " flows which are retrieved independently from the same base dataset and turned into a single greatest dataset.

***CAIDA 2007*[20]** - Dataset 2007 from CAIDA (Center for Applied Internet Data Analysis). This dataset's primary goals are to gather and disseminate information for use in scientific or research analyses of internet traffic, topology, routing, performance, and security-related events. The dataset includes information such as the server's IP address, Timestamp, Time Zone, Object ID/URL of the web page, Response code/status, and Number of bytes delivered.

IV. Analysis :

To obtain better results, the combination of both feature engineering and machine learning methods provides a huge contribution along with some specific criteria of frameworks [17]. This would result in our detection of DDoS attacks from the dataset provided.

Reference	**Approach**	**Accuracy**
[1]	Seven machine learning algorithms Logistic Regression, Support Vector Classifier K-Nearest Neighbor, Random Forest, Ensemble Classifier, Artificial Neural Network, Hybrid Machine Learning Model are applied on a created CSV file	Achieved the highest accuracy with SVC-RF (98.8%)
[2]	Two machine learning algorithms Random Forest, Multi-Layer Perceptron with and without a big data approach.	Both show decent accuracy whereas MLP takes less time.
[3]	Feature reduction and J48 classifier is used on two datasets.	Both show significant results 98.9% and 98.8%.
[4]	Multiple Linear Regression is applied	Providing an accuracy of 73.79% With 16 features and 71.6% with 6
[5]	4 classifiers Decision Tree, Random Forest, KNN, XG BOOST with 3 feature selection algorithms Chi-square, Extra Tree, ANOVA	XG-Boost with ANOVA achieved 98.34% accuracy
[7]	Feature selection approach on CICIDS DDoS 2019	87.66% accuracy is achieved
[11]	Feature selection with MLP classifier	97.66 % accuracy is achieved

[14]	Some feature extraction techniques are used and the pre-processed data is classified by using Classifiers.	99.84% accuracy is achieved using Gradient Boosting.
[15]	Feature selection techniques are used on CICIDS2017 dataset.	99.6% accuracy is achieved using Random Forest.
[16]	Detection analysis is made on KDD Cup 1999 dataset	90.2% accuracy is achieved
[17]	Detection of various Machine Learning algorithms Among these ANN provides the highest accuracy.	93.48% accuracy is achieved
[18]	Detection takes place on the dataset using SVM and RF.	RF is superior to SVM in all Cases of detection.
[20]	CAIDA2007 dataset is used to detect HTTP flood attacks	ARTP has achieved the highest accuracy.

V. Conclusion:

Traditional IDS systems are not suitable for all the data. It is useful to detect any intrusions from a small amount of data. To identify the anomalies from a large amount of data, many techniques were discovered. Among them, machine learning techniques were a boon to information security. The entire procedure depends on some simple steps like pre-processing, classification, and detection. Various machine learning techniques were emerging day by day and the detection would be possible in a piece-of-cake manner. The possibility of detection makes us prevent such kinds of cyber threats and protect us from a major loss. Among all the techniques used, feature reduction called Information Gain and Correlation with J48 classifier provides the highest accuracy of 99.9% [3].

VI. Abbreviations

DDoS – Distributed Denial of Service

IP – Internet Protocol

SDN – Software Defined Networking

DoS – Denial of Service

NIST – National Institute of Standards and Technology

ID – Intrusion Detection

IDS – Intrusion Detection System

ANN – Artificial Neural Networks

ML- Machine Learning

NLP- Natural Language Processing

SVC – Support Vector Classifier

RF- Random Forest

MLP – Multi-Layer Perceptron

XG-Boost – Extreme Gradient Boosting

ANOVA – Analysis of Variance

IoT – Internet of Things

ARTP – Anamoly-based Real Time Prevention

CNN – Convolutional Neural Networks

CSV – Comma Separated Values

HTTP- Hyper Text Transfer Protocol

References

1. Nisha Ahuja, Gaurav Singal, Debajyoti Mukhopadhyay, Neeraj Kumar, Automated DDOS attack detection in Software-Defined Networking, ELSEVIER,2021.

2. Mazhar Javed Awan, Umar Farooq, Hafiz Muhammad Aqeel Babar, Awais Yasin, Haitham Nobanee, Muzammil Hussain, Owais Hakeem, and Azlan Mohd Zain, Real-Time DDoS Attack Detection System Using Big Data Approach, Sustainability,2021.

3. Deepak Kshirsagar, Sandeep Kumar, A feature reduction based refected and exploited DDoS attacks detection system, Journal of Ambient Intelligence and Humanized Computing,2021.

4. Swathi Sambangi, Lakshmeeswari Gondi, A Machine Learning approach for DDoS attack detection using multiple Linear Regression, Proceedings, MDPI, 2020.

5. Vimal Gaur, Rajneesh Kumar, Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices, Arabian Journal for Science and Engineering,2021.

6. Ismael Amezcua Valdovinos, Jesús Arturo Pérez-Díaz, Kim-Kwang Raymond Choo, Juan Felipe Botero, Emerging DDoS attack detection and mitigation strategies in software-defined networks: Taxonomy, challenges and future directions, Journal of Network and Computer Applications,2021.

7. Prasad M, Tripathi S, Dahal K, An efficient feature selection based Bayesian and rough set approach for intrusion detection. , Appl Soft Comput 87:105980, 2020.

8. Scarfone, K., Mell, P., Guide to intrusion detection and prevention systems (IDPs)., NIST Spec. Publ. 2007, 800, 94.

9. Mukherjee, B.; Heberlein, L.T.; Levitt, K.N. Network intrusion detection. IEEE Netw. 1994, 8, 26–41.

10. Zhao F, Zhao J, Niu X, Luo S, Xin Y, A filter feature selection algorithm based on mutual information for intrusion detection. Appl Sci 8(9):1535, 2018

11. Wang M, Lu Y, Qin J, A dynamic MLP-based DDoS attack detection method using feature selection and feedback, Comput Secure 88:101645, 2020

12. Dos Anjos, J.C.S.; Gross, J.L.G.; Matteucci, K.J.; González, G.V.; Leithardt, V.R.Q.; Geyer, C.F.R. An Algorithm to Minimize Energy Consumption and Elapsed Time for IoT Workloads in a Hybrid Architecture. Sensors 2021, 21, 2914.

13. Ganguly, S.; Garofalakis, M.; Rastogi, R.; Sabnani, K. Streaming algorithms for robust, real-time detection of DDoS attacks. In Proceedings of the 27th International Conference on Distributed Computing Systems (ICDCS’07), Toronto, ON, Canada, 25–27 June 2007; p. 4.

14. Maranhao, J.P.A.; Costa, J.P.C.L.D.; Freitas, E.P.D.; Javidi, E.; Junior, R.T.D.S.: Error-robust distributed denial of service attack detection based on an average common feature extraction technique. Sensors 20(20), 5845–5866 (2020).

15. Silveria, F.A.F.; Junior, A.D.M.B.; Vargas-Solar, G.; Silveria, L.F.: Smart Detection: an online approach for DoS/DDoS attack detection using machine learning. Secure. Commun. Netw. (2019).

16. M. Revathi, V. V. Ramalingam, B. Amutha, A Machine Learning Based Detection and Mitigation of the DDOS Attack by Using SDN Controller Framework, Springer,2021.

17. Muhammad Aamir, Syed Mustafa Ali Zaidi, DDoS attack detection with feature engineering and machine learning: the framework and performance evaluation, International Journal of Information Security (2019).

18. Jiangtao Pei, Yunli Chen, Wei Ji, A DDoS Attack Detection Method Based on Machine Learning, Journal of Physics, 2019.

19. Arshi M, Nasreen MD, and Karanam Madhavi, A Survey of DDOS Attacks Using Machine Learning Techniques, E3S Web of Conferences, ICMED,2020.

20. Indraneel Sreeram, Venkata Praveen Kumar Vuppala, HTTP flood attack detection in application layer using machine learning metrics and bio-inspired bat algorithm, Applied Computing, and Informatics (2019).