https://doi.org/10.48047/AFJBS.6.Si4.2024.3248-3263

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

DoS Attack Detection using Machine Learning Techniques

Sudhir Kumar Pandey[1], Shambhu Shankar Bharti[2], Zafar Ayub Ansari[3], Ajeet Kumar[4], Bhawana Singh[5], Moazzam Haidari[6]

[1][2][3]Assistant Professor, LokNayak Jai Prakash Institute of Technology, Chapra

[4][5] Assistant Professor, Bakhtiyarpur College of Engineering, Bakhtiyarpur

[6] Assistant Professor, Saharsa College of Engineering, Saharsa

[1] Sudhirp.phd19.cs@nitp.ac.in, [2]shambhu4u08@gmail.com ,[3]zafar786ayub@gmail.com

[4]azit.bce@gmail.com, [5]bhawana.singh.25@gmail.com, [6] moazzam53211@gmail.com

Article History
Volume 6, Issue Si4, 2024
Received : 04 June 2024
Accepted : 25 June 2024
doi:
10.48047/AFJBS.6.Si4.2024.3248-3263

Abstract: Numerous conventional articles have been written focusing on the topic of DoS/DDoS attacks. The majority of these present studies on DoS attacks from the attacker's perspective. Although these research works present assessments of vulnerabilities and then defenses of possible targets, they do not address the attack-enabling 'attack tools'. Despite this, DoS attacks have been extensively evaluated through other various types of security mechanisms which are pre-deployed to avoid DoS attacks. There is no descriptive overview or survey which mainly presents the present techniques, tools, and best practices to avoid DoS attacks. Therefore, this survey is both novel and informative to the researchers and readers in this field.A denial-of-service attack (DoS) is one of the most feared threats of the cyber world. DoS attack is easy to execute, but it is difficult to defend. In recent years, the impact of DoS attacks is widely addressed due to the rising number of DoS attacks. In this paper, we provide an extensive summary and analysis of the present techniques, tools, and best practices which are used to detect and defend against DoS attacks. We also applied Some Machine learning Algorithms with NSL KDD dataset in order to detect DoS attack. We have achieved best result with accuracy of 98.36 in Logistic Regression Algorithm.

Keywords: DoS attack, Machine Learning, Gray wolf optimization, NSLKDD Dataset, Intrusion Detection.

1.Introduction

Attacks known as denial of service (DoS) present serious risks to a number of facets of our existence, including the Internet of Things, smart cities, healthcare, information technology, and

business sectors. In these assaults, an enormous amount of traffic is directed at a system or network, preventing it from responding to valid requests. The complexity, volume, and frequency of DoS attacks have been on the rise, making them a persistent threat to network security across all business sectors. As technology advances, attackers are finding new ways to launch more sophisticated and damaging DoS attacks. DoS attacks share similarities with non-malicious availability issues, such as system maintenance or technical problems. This similarity can make it challenging for system administrators to differentiate between genuine network issues and malicious attacks. Recognizing and effectively defending against DoS attacks can be extremely difficult due to the evolving tactics used by attackers. Symptoms of a DoS attack may include slow network speeds, inaccessible websites, or unresponsive services, making it crucial to have robust detection mechanisms in place.Attackers are continuously enhancing their computing capabilities to carry out more potent DoS attacks. This calls for the creation of intelligent security solutions to prevent susceptible IoT devices from being compromised and utilized in these kinds of attacks. In this project, different machine learning (ML) algorithms that are included in the WEKA tool have been utilized to analyze the detection performance for DoS attacks using the NSL_KDD-master dataset. By utilising ML algorithms, researchers hope to improve the precision and effectiveness of identifying and mitigating DoS attacks. Three classifiers—BayesNet, K-Nearest Neighbors (KNN), and J48—have been evaluated on the dataset in order to predict six different types of attacks and distinguish benign activities. The evaluation's goal is to identify the best classifier for effectively identifying and responding to DoS attacks. The denial of service (DOS) attack is an attack on a computer system or network that causes a loss of service to users, typically the loss of network connectivity and services by consuming the bandwidth of the victim network or overloading the computational resources of the victim system. Many tools are available that can consume either sufficient bandwidth or CPU and memory to cause a denial-of-service condition. Publicly available tools that can be used for good or evil are often referred to as dual-use tools. When used as dual-use tools without authorization, they are considered unethical and their use is illegal. Most of the anti-DOS products available on the market can be circumvented, and false alerts detected. Yet, each one makes the job of the cracker a lot more difficult. With the evolution of communication, automation, and digitization of sensors, a common small-scale DOS attack can become what has been defined as a large-scale DOS attack, one generated through many different locations of the globe (and of course, many different IP addresses).

2. Motivation

In comparison with regular network flows, the detrimental number of dimmed flows is quite little. The DOS and DOS's swarm appear in packets as a mixture of victim and attacker. But the detrimental flows often have the same features, including limited resource usage, high quantity and consistency, and nearly equivalent time-to-live. This uniformity makes it quite challenging to enumerate DOS and substantial traffic. The recent study has underscored the difficulty of detecting DOS and DOS threats, and various investigators have aimed to erect the classifier scheme using either abrupt handcrafted resolutions or automatic method actioniers. In several cases, space-based resolution requests may seem to be currently unavailable. Still, the enactment features of the institution rather must enumerate the unknown and peculiar behaviors of the network traffic in such a manner that it fails to carry out with the network performance.

The concept of Denial of Service (DOS) is to confuse or disrupt information system services. The advent of DOS assaults has become a menace to the flourishing accessibility of online services. With the prevalence of the internet and the rise in system connectivity, the technique to

distribute service assaults has been perpetually enhanced and enlarged. During a DOS assault, the assailant typically develops a great quantity of assault traffic to suffocate the bandwidth of the attack, or accesses great numbers of resources of the attack service, which distinctively falls below the regular processing capacity. Given that a multitude of prevalent pointing devices have restricted bandwidth, it is exceedingly unproblematic to launch a DOS assault. Many generally utilized targeted devices, such as web servers, DNS servers, and other network devices, can be disrupted, severely compromised, or even obstruct the performance of those systems with a fewer packets of synthetic flows.

3.Technique for DOS Attack Detection

In comparison with regular network flows, the detrimental number of dimmed flows is quite little. The DOS and DOS's swarm appear in packets as a mixture of victim and attacker. But the detrimental flows often have the same features, including limited resource usage, high quantity and consistency, and nearly equivalent time-to-live. This uniformity makes it quite challenging to enumerate DOS and substantial traffic. The recent study has underscored the difficulty of detecting DOS and DOS threats, and various investigators have aimed to erect the classifier scheme using either abrupt handcrafted resolutions or automatic method actioniers. In several cases, space-based resolution requests may seem to be currently unavailable. Still, the enactment features of the institution rather must enumerate the unknown and peculiar behaviors of the network traffic in such a manner that it fails to carry out with the network performance.

The concept of Denial of Service (DOS) is to confuse or disrupt information system services. The advent of DOS assaults has become a menace to the flourishing accessibility of online services. With the prevalence of the internet and the rise in system connectivity, the technique to distribute service assaults has been perpetually enhanced and enlarged. During a DOS assault, the assailant typically develops a great quantity of assault traffic to suffocate the bandwidth of the attack, or accesses great numbers of resources of the attack service, which distinctively falls below the regular processing capacity. Given that a multitude of prevalent pointing devices have restricted bandwidth, it is exceedingly unproblematic to launch a DOS assault. Many generally utilized targeted devices, such as web servers, DNS servers, and other network devices, can be disrupted, severely compromised, or even obstruct the performance of those systems with a fewer packets of synthetic flows.

3.1Behavioral Analysis:

Monitoring and analyzing network traffic patterns to detect anomalies and signs of an attack. Behavioral analysis involves monitoring and analyzing network traffic patterns to detect anomalies and signs of an attack, aiming to enhance network security by identifying suspicious activities based on deviations from normal behavior. By observing network traffic patterns, behavioral analysis can establish a baseline of usual behavior within the network environment. Any deviations from this baseline are flagged as anomalies, potentially indicating a security threat or attack. This approach is particularly effective in detecting new and evolving threats that may not be captured by signature-based detection techniques. Unlike signature-based methods that rely on known attack patterns, behavioral analysis focuses on identifying unusual behaviors that may signify a novel attack vector.

3.2Rate-Based Detection:

Detecting attacks based on abnormal traffic rates and volume, initiating protective measures. Rate-based detection involves detecting attacks based on abnormal traffic rates and volume, enabling the system to initiate protective measures when such anomalies are identified. This

method focuses onmonitoring the flow of network traffic and analyzing the rate at which data packets are transmitted. By establishing thresholds for normal traffic rates, any deviations above or below these thresholds can indicate a potential attack scenario. When the system detects abnormal traffic rates or volumes that surpass predefined thresholds, it triggers protective measures such as blocking suspicious traffic, alerting administrators, or implementing traffic filtering rules to mitigate the impact of the attack. Rate-based detection complements other intrusion detection techniques by providing a proactive approach to identifying attacks based on the flow and volume of network traffic, enhancing the overall security posture of the network environment.

3.3Signature-Based Detection:

Identifying known attack patterns and leveraging threat intelligence to recognize and stop attacks.Signature-based detection involves identifying known attack patterns by leveraging threat intelligence to recognize and stop attacks based on predefined signatures stored in a database. This method compares network traffic against a database of attack signatures, raising a detection warning when a match is found with a known attack pattern. While effective in detecting known attacks, it may not identify new or zero-day attacks that do not have recorded signatures. Signature-based detection is particularly useful for detecting well-known attack types for which signatures are available. It is commonly used in intrusion detection systems to match observed network traffic against a library of known attack signatures. By continuously updating the signature database with new threat intelligence, organizations can enhance their ability to detect and prevent known attacks, contributing to a more robust cybersecurity defense strategy.

4.Literature Review

In academic research, a literature review is a critical analysis of the existing body of knowledge on a specific topic. It involves examining and evaluating previous studies, research papers, and publications related to the subject of interest.The literature review chapter in a research project serves to provide a comprehensive overview of the existing research and studies that are relevant to the current work being undertaken. It helps in establishing the context, significance, and gaps in the research area. The highlighted text mentions that the literature review chapter includes various works and studies that directly or indirectly support the present research work on DOS (Denial of Service) attack detection techniques. It highlights that many research studies in the literature have focused on different methods to detect DOS attacks. The text discusses the classification of DOS attack detection into low-rate detection and high-rate detection. This classification helps in understanding the different approaches and strategies used in detecting DOS attacks based on the intensity or frequency of the attacks. The literature review chapter aims to provide a detailed description of previous research studies that have proposed various solutions for detecting DOS attacks. These studies offer insights into different techniques, mechanisms, and algorithms used in the detection of DOS attacks. By providing detailed information about different DOS detection techniques and mechanisms in recent times, the literature review chapter helps in building a strong foundation for the current research work. It allows the researcher to understand the evolution of DOS detection methods and identify areas for further research and improvement.

4.1 Low-rate Distributed Denial-of-Service (DDOS) AttacksDetection: Attackers that use low-rate denial-of-service (DDOS) attacks deliver a lot of attack packets that mimic normal traffic in an attempt to overwhelm the target system. Congestion participation (CPR) is a statistic that Zhang et al. [4] presented, along with a method to identify and filter low-rate DDOS attacks based on this metric. According to their research, regular TCP flows help to reduce network

congestion, but low-rate DDOS flows actively cause it. Zhang et al.'s suggested approach focuses on identifying attacks and legitimate flows by examining how traffic affects network congestion. This differentiation is crucial for effectively identifying and mitigating DDOS attacks.While the CPR-based approach shows promise in detecting low-rate DDOS attacks, the effectiveness of the method needs to be validated through additional experiments and analysis using real-world datasets. This highlights the importance of practical testing to assess the reliability and accuracy of the detection technique.

4.2 Entropy-Based Detection Methods for DDoS Attacks: In order to distinguish between short-term high-rate DDOS attacks and long-term low-rate DDOS attacks, Du and Abe [6] developed an entropy metric based on IP packet size. Their method is predicated on the idea that variations in the distribution of packet sizes during attacks may reveal the existence of malicious activity. However, there are drawbacks to the Du and Abe method, particularly when dealing with low-rate DDOS attacks. These include scalability issues and the need for a lengthy detection period in order to attain high detection accuracy. This indicates the need for more efficient and faster detection mechanisms.Jadhav and Patil suggested an enhanced entropy-based detection method for low-level DDOS attacks, which represents a significant advancement over traditional entropy metrics. Their approach aims to reduce false positives in DDOS detection by improving the differentiation between regular traffic and attack traffic.Despite the improvements in detection accuracy, the distance value between normal and attack traffic remains minimal, leading to an increase in the false positive rate. This highlights the ongoing challenge of balancing detection sensitivity with minimizing false alarms in DDOS detection systems.

4.3 High Rate DDOS Attack DetectionSpoofed DDOS Attacks: In order to initiate a DDoS attack, the attacker modifies the TCP/IP header fields in these attacks. The Time-To-Live (TTL) information cannot be faked, in contrast to the majority of header fields. Thus, to detect fake IP packets, the TTL value is employed. The fact that the header only includes the final TTL value presents the greatest hurdle when utilising TTL for detection. Initial TTL values for an IP address can vary between operating systems and can also fluctuate over time. This may result in false negatives if the attacker correctly estimates the hop-count between the source and the victim, or false positives if a valid packet originates from an unlisted OS. A route fingerprint approach is proposed to separate legitimate traffic from attack traffic. Every packet has an encoded, unique path fingerprint that shows the path the packet took to get to its destination. The packet is considered spoofing if the route fingerprint is not right. Due to the need for calculations at intermediate nodes when the packet enters the same subnet, this approach might not be able to identify subnet spoofing. An alternative method involves determining the operating system of a packet by examining the values of TCP/IP header fields, including TTL, IP Don't Fragment (DF), window size, and total length. These values are used to construct a fingerprint, which is then appended to the packet at the source side for additional examination.

The long-term nature of assault has been one of the major difficulties with the dependability and accessibility of internet network services in preventing the efficacy of DOS attacks. It brings researchers to worry about the vulnerability of today's Internet to avoid DOS studies different strategies linked to the layer, network layer, and higher application identification. This report contributes to the advancements in DOS detection mechanisms, which use machine learning and soft robotics approaches in different ways, to the growing range of detection techniques and presents a deeper understanding to encourage vigilant analysis.

As a denial-of-service (DOS) attack leverages networks with insufficient resources and strict network protocol definition, the potential to subdue the emanation of large volumes of

processing and storage resources, and costs are low, more criminals would induce DOS assaults. Relentlessly, editions of the DOS attack are involved in different security threats that rise as potentially severe risks or potential massive data breaches. Because of its gains, DOS assaults have become one of the earliest and growing online perils helping to sway many criminals.

5.Methodology

The research reported in this thesis is motivated by the advanced implementation techniques for DOS detection and protection. In an attempt to achieve our objectives, we have aimed to develop connect in multiple contexts with three main research questions which are embodied through the following three main stages of the research, the involved processes and methodologies are introduced. We conclude with some insight into the structure of the research and its first chapter, the Review of Related Literature. Additionally, in the first chapter, we expound on the motivation and context for our research, our aim, objectives and approach and the contribution that we seek to make.

This section is structured so as to provide an outline of how the research is structured. The first part of this chapter provides an introductory section outlining what would be discussed in the rest of this journal. The second section highlights a conceptual analysis need for our research and the third part of our methodology justifies why our choice of theoretical underpinning would be in the form of the Review of Related Literature. Lastly, the chapter then concludes with a summation of what has been covered within the section. The number of various kinds of Internet threats is continually rising. Denial of Service (DOS) is becoming one of the most dangerous Internet threats. Current DOS defense systems, in the best practice, cope only with the symptoms but ignore the roots of the problem. This fact results in huge traffic overheads and makes the effectiveness of DOS protection questionable.

5.1Data Selection: Using the NSL_KDD-master data set, the study trained and tested the algorithm. The NSL_KDD-master dataset contains frequent, contemporary, and benign attacks. It is designed to emulate real-world data (PCAPs). It also contains the outcomes of a network traffic analysis performed using a CIC flow metre, in addition to the time stamp, source and destination IP addresses, source and destination ports, protocols, and attack-based flows. The "NSL_KDD-master dataset" refers to a specific dataset that was used to test and train the algorithm during the inquiry. Research on network security and intrusion detection systems frequently uses the NSL_KDD-master dataset. The NSL_KDD-master dataset contains a mix of benign (normal) network traffic data and common attacks.

To differentiate between malicious and benign network activity, algorithms must be trained on a mixture of benign and harmful data. The purpose of the dataset is to replicate packet capture files, or PCAPs, which are actual instances of network traffic data. This suggests that the dataset has network traffic properties and patterns that are commonly observed in actual network environments. The dataset contains the analysis of network traffic performed using the CIC Flow Metre. Among the specific information provided by our study are time stamps, source and destination IP addresses, source and destination ports, used protocols, and flows associated with different types of attacks. The NSL_KDD-master dataset contains more than 80 network flow features that were extracted from the network traffic data. These characteristics are essential for teaching machine learning algorithms to identify patterns and anomalies in network data that may indicate potential security threats.

5.2 Data Preprocessing: In data analysis and machine learning projects, data pre-processing is an essential stage that converts raw data into a clear, structured format. The data must be cleaned, transformed, and organized during this phase in order to be ready for additional

analysis. Making sure the data is in a format that machine learning algorithms can use efficiently is the major goal of data pre-processing. Before being fed into a machine learning model, errors, missing values, and inconsistencies in the raw data gathered from diverse sources must be fixed. The model's performance in machine learning projects is greatly impacted by the quality of the input data.Through the preparation of the data in a consistent and practical manner, data pre-processing aids in increasing the precision and effectiveness of machine learning algorithms. In data pre-processing, managing null or missing values in the dataset is a typical task. Regarding missing data, several machine-learning algorithms have different specifications. Before using the algorithm on the data, these issues must be resolved because methods such as Random Forest do not accept null values. Preparing the data beforehand guarantees that the format of the dataset is suitable for a variety of deep learning and machine learning methods. It is simpler to apply several algorithms to the same dataset and compare their performance to choose the optimal model for the job at hand when the data format is standardised.

5.3 Feature Selection:A key idea in machine learning is feature selection, which is selecting the most pertinent characteristics or features from the dataset to increase the effectiveness of the model. The machine learning model's performance is greatly influenced by these particular features. The model's ability to predict outcomes is strongly influenced by the properties of the training data. It is possible to improve the model's efficacy and accuracy by choosing the features carefully. In the process of developing a model, feature selection is as important as data cleaning and gathering. It assists in determining which features are most important in producing the intended result or predictive variable. Experts in the field can choose features manually or automatically using algorithms. On the one hand, automatic approaches find key aspects through statistical techniques, whereas manual methods require professional knowledge and understanding of the data. Irrelevant features in the dataset can have a detrimental effect on the accuracy and functionality of the model. Unnecessary complexity and noise can be introduced by irrelevant features, which could cause the model to learn from irrelevant patterns. The model can concentrate on learning from the crucial information in the data, increasing its accuracy and predictive capacity by only choosing the most pertinent elements. This procedure aids in the development of a machine-learning model that is more effective and efficient. The Grey Wolf Optimization (GWO) algorithm, which finds the best solution by locating the wolves' alpha, beta, and delta at the best areas, is used in this research study for feature selection. The algorithm begins with a random community of grey wolves.In this updated approach, feature selection is carried out using Stochastic Gradient Descent (SGD) in lieu of the fitness function of the traditional GWO algorithm. This modification aims to enhance the feature selection process by leveraging the optimization capabilities of SGD, which is commonly used in machine learning for optimization tasks. By integrating SGD into the GWO algorithm, the feature selection process benefits from the stochastic nature of SGD, which allows for efficient optimization of the selected features based on the gradient of the loss function. This integration can potentially lead to improved accuracy and performance in feature selection tasks**.**

5.4Machine Learning Algorithms:A branch of artificial intelligence (AI) called "machine learning" focuses on creating models and algorithms that let computers learn from data and make decisions without explicit programming. Put more simply, it allows machines to gain experience and become more proficient at a task. To extract valuable insights from data, machine learning makes use of a variety of approaches, including statistics, pattern recognition, information discovery, and data mining. Machine learning algorithms are capable of making predictions and judgements on their own by examining patterns and trends in data. One of the main types of

machine learning is this one, in which a labelled dataset is used to train the algorithm. The dataset in supervised learning comprises labels for the corresponding input data as well as output data. With the help of the labelled examples supplied during training, the algorithm gains the ability to map inputs to outputs or generalize. Unsupervised learning is different from supervised learning in that it uses unlabeled data to train the algorithm. Without explicit instructions, the algorithm must identify structures and patterns in the data. Tasks like dimensionality reduction, anomaly detection, and grouping can benefit from unsupervised learning. Numerous industries, including healthcare, banking, marketing, and autonomous cars, heavily rely on machine learning algorithms. Applications like recommendation systems, natural language processing, picture recognition, and predictive analytics can be made of them. A machine learning algorithm processes a training dataset in order to identify underlying patterns and relationships in the data during the training phase. In order to reduce errors and enhance forecast accuracy, the algorithm iteratively modifies its parameters. A machine learning method creates an inferred function that may be used to forecast fresh, unknown data after training on a dataset. This function can be used to make decisions in real-world situations by capturing the patterns that were learned from the training data. Supervised learning is suggested as a method to identify distributed denial of service (DDoS) attack in the context of network security. Using supervised learning techniques like Naïve Bayes Machine and Support Vector Machine, the goal is to efficiently classify data according to its effectiveness and performance for network security initiatives. Naïve Bayes and Support Vector Machines are the preferred methods for DDoS attack detection because of their track record of success in classification tasks and their applicability to network security applications. These algorithms are useful tools for spotting and reducing security risks because they can recognise trends and abnormalities in network traffic data.

5.4.1Naive Bayes: The Bayes Theorem, an essential theorem of probability theory, serves as the foundation for the naive Bayes statistical categorization method. It is a fundamental probabilistic classifier that bases its assumptions on feature independence. The foundation of Naïve Bayes is the formula P(A|B) = P(B|A) P(A) / P(B). It determines the likelihood that event A will occur in light of event B's occurrence. The probability that event A will occur when event B is known to have happened is represented by the symbol P (A|B). Because it represents the likelihood after accounting for new information, it is often referred to as the posterior probability. The premise of the Naive Bayes Hypothesis is that the existence of one characteristic in a class is independent of the existence of any other feature. This reduces complexity and improves the computational efficiency of the calculation.The probabilities that occurrences A and B will occur separately are denoted by P(A) and P(B). The conditional probability that event B will occur given that event A has already occurred is denoted by P(B|A). Because Naïve Bayes is easy to use and effective at managing massive data sets with many of features, it is frequently used in text categorization, spam filtering, and medical diagnosis.

For example, given a class y and a dependent feature vector [x1, x2, …….,xn], we obtain

$$P(y\,|x_{1\,,}\,\ldots,x_n) = \frac{P(y)P(x_{1\,,\ldots,}x_n|\;y)}{P(x_{1\,,\ldots,}x_n)} \qquad (1)$$

And using the assumption that all the x_i 's are independent,

$$P(x_i|\;y,x_{i,}\,\ldots,x_{i-1,}x_{i+1},\ldots,x_n) = P(x_i|\;y)$$

On simplifying, following result is obtained

$$P(y\,|x_{1\,,}\,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1\ldots,x_n)} \qquad (2)$$

which is the result of Bayes' theorem.

5.4.2Support Vector Machine (SVM): A supervised learning technique called Support Vector Machine (SVM) is applied to both regression and classification problems. The way it operates is by identifying the hyperplane in the dataset that best divides the various classes. Given a dataset containing n features, each data point in SVM is represented as a point in an n-dimensional space. Next, in this high-dimensional space, the algorithm seeks to identify the hyperplane that maximally divides the classes. The data points that are closest to the decision boundary (hyperplane) separating the classes are known as support vectors. These vectors establish the boundary between several classes; hence, they are essential in identifying the ideal hyperplane. Support vectors serve as the fundamental components of support vector machines (SVM) that facilitate the efficient separation of two classes by delineating the decision boundary. They play a key role in figuring out which hyperplane maximises the margin between classes. SVM is regarded as one of the top techniques for challenges involving pattern and image classification. It works especially well in situations where a large collection of training images needs to be divided into two different classes according to certain attributes. With each data point represented by a vector in a d-dimensional feature space, SVM is built to handle datasets with many dimensions. Based on the feature vectors of the data points, the method is able to efficiently categorise them into multiple classes represented by {−1, +1}.

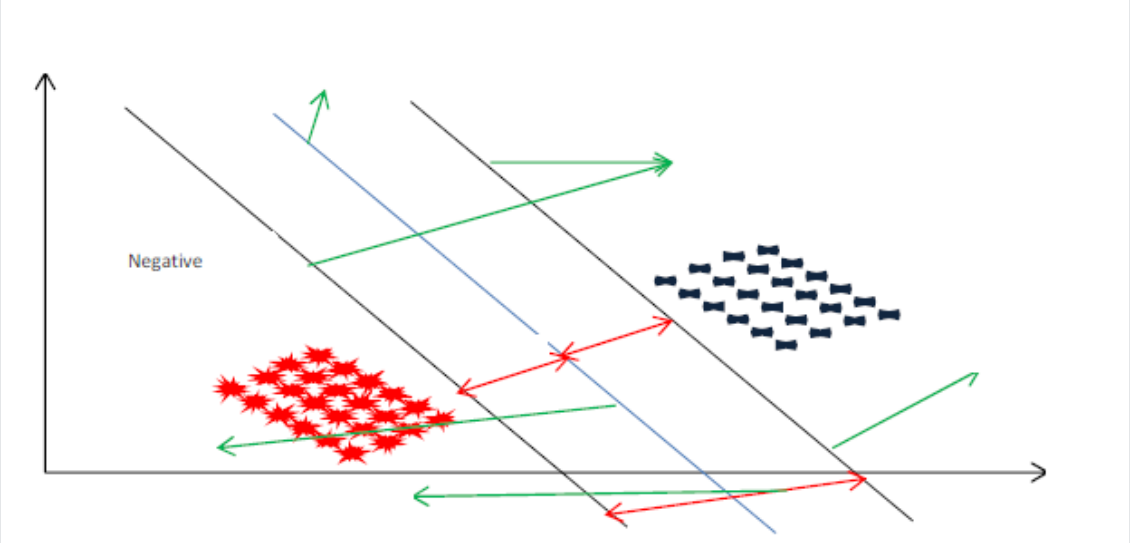

Figure 1. Support Vector Machine

5.4.3 K-Nearest Neighbors (KNN) Algorithm:A machine learning algorithm that is classified as supervised learning is K-Nearest Neighbors (KNN). This indicates that in order to forecast or classify data that hasn't been seen yet, the algorithm learns from labelled training data. KNN functions based on the presumption that recent and old data points are comparable. A new data point is classified by comparing it to the existing data and assigning it to the category that most closely resembles the current categories. Every data point that is accessible is stored by KNN, which then uses the degree of similarity between the new and old data to classify it. In order to identify the K nearest neighbors, this technique involves finding the distance between the new data point and every existing data point. When it comes to classification, KNN is useful in giving a new data point a class label based on the majority of its closest neighbors. For example, KNN can be used to classify a new data item x1 into one of two classes, Category A or Category B, if there are two categories. Based on the majority class among its neighbors,the algorithm may identify which category x1 belongs to by taking into account its Knearestneighbors.

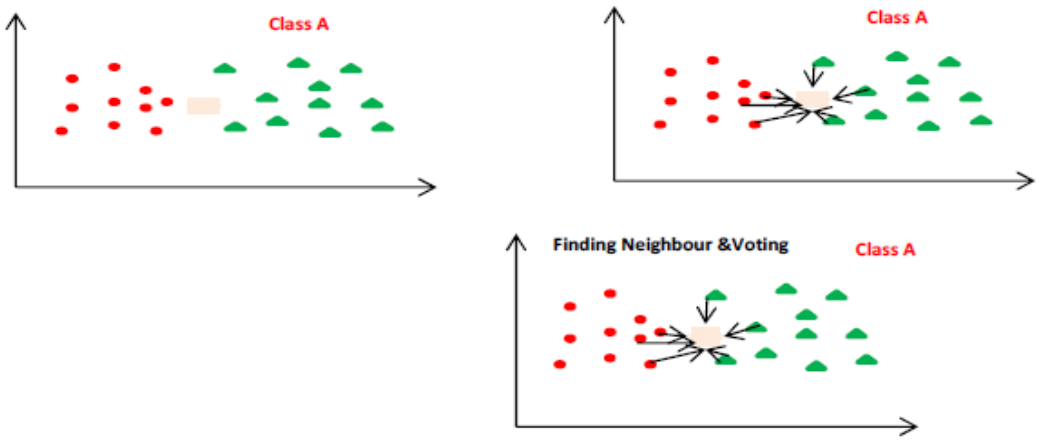

Figure 2. K-Nearest Neighbors (KNN) Algorithm

5.4.4 Random Forest: One example of a supervised learning machine learning algorithm is Random Forest. This means that in order for the algorithm to understand how to map input data to the proper output during the training phase, labelled data is needed. The idea of group learning is the foundation of Random Forest. In ensemble learning, several independent models (classifiers) are combined to produce a more robust and accurate model. These distinct models in the random forest scenario are decision trees. A group of decision trees, each trained on a distinct subset of the original dataset, make up a random forest. To generate forecasts, these decision trees operate on their own. The Random Forest method aggregates the individual predictions made by each decision tree by means of a voting process. The majority vote among all the decision trees determines the Random Forest's final output. By lessening the influence of individual biases and errors, this method produces predictions that are more reliable. Overfitting, a prevalent problem in machine learning when a model performs well on training data but poorly on unknown data, is one of Random Forest's benefits. Random Forest achieves improved accuracy by employing numerous trees and combining their predictions, which improves generalization to new data. The performance of a random forest is dependent on the number of trees in it.A random forest that has more trees in it will typically be more accurate. On the other hand, an excessive number of trees might also result in higher computational complexity. Achieving the ideal performance requires striking the correct balance when it comes to tree count.

6. DATASET:

We used in our work the NSL-KDD Dataset to train the different ML Algorithms.

S.No.	Name of File	Number of Features	Number of Samples	Labels
1.	KDDTrain+.csv	43	125972	5 [DoS, U2R, R2L, Probing Normal]
2.	KDDTest+.csv	43	22542	5 [DoS,U2R, R2L, Probing Normal]

Table 1.Illustration of NSL KDD Dataset

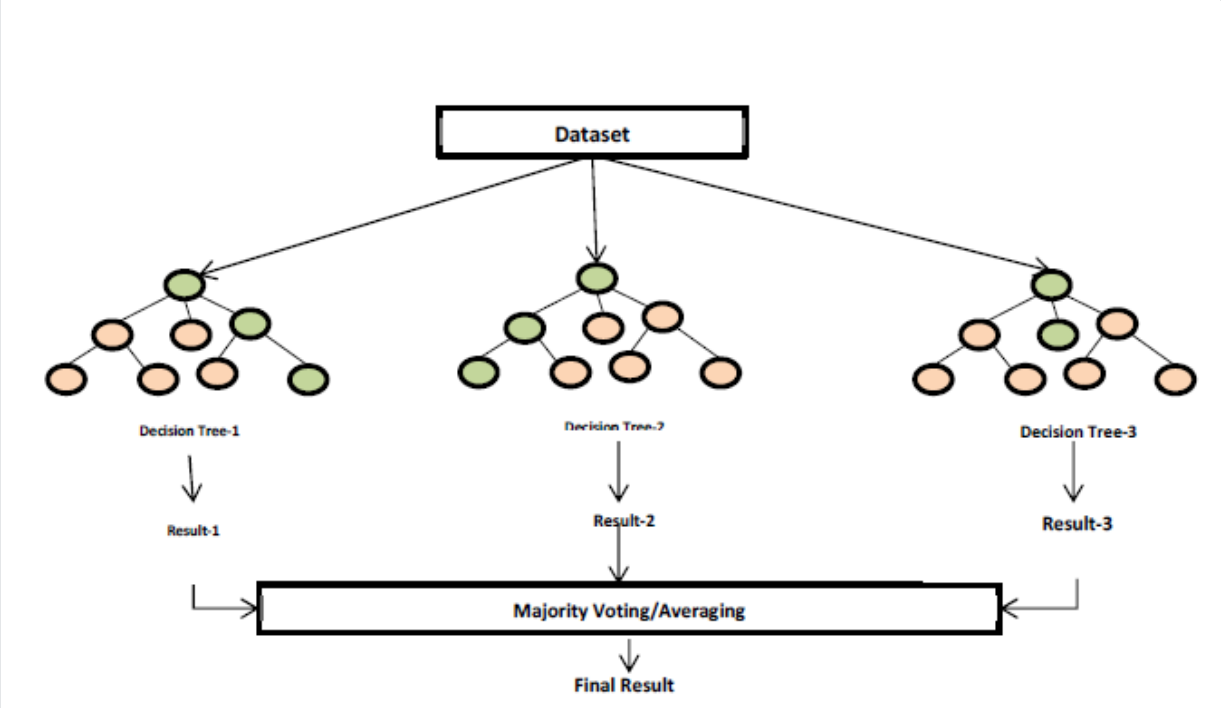

Figure 3.Random Forest

7.PROPOSED WORK

One particular dataset that was used in this work for model testing and training was the NSL_KDD-master dataset. In the fields of network intrusion detection and cybersecurity, it is a well-known dataset. This dataset includes a variety of network traffic data that is frequently used to assess how well intrusion detection systems are working. Researchers can evaluate how well various machine learning techniques identify and categorise network intrusions by using the NSL_KDD-master dataset.

Support Vector Machine (SVM) and Naïve Bayes algorithms:In machine learning, two well-liked classification algorithms are Support Vector Machine (SVM) and Naïve Bayes. SVM is an efficient supervised learning technique that finds the best hyperplane in a high-dimensional space to divide several classes from data points. Based on the Bayes theorem, Naïve Bayes is a probabilistic classifier that presumes the existence of a specific characteristic in a class is independent of the presence of other features. The model used in this study is based on a number of features, including packet flows, port numbers, and traffic analysis, that were taken from the network traffic data. In order to detect any intrusions or assaults, these qualities are critical for spotting trends and anomalies in the network data.The model may be trained to differentiate between typical network behaviour and unusual activity that can point to a security risk by examining these characteristics. The dataset is split into three subsets: training data, test data, and validation data, in order to assess the model's performance. 70% of the dataset in this study is used to train the model, which entails providing labelled data to the algorithm so it can identify underlying trends. Twenty percent of the dataset will be used to test the model's performance on untested data so that researchers may evaluate its generalization and accuracy. The detection model is implemented in two distinct phases: the training stage and the detection stage. The classifiers are trained in the training phase using support vector machine (SVM) machine

learningcomputations. The SVM method is used to identify patterns and traits from the input data—packet
flows, ports, etc.—in order to differentiate between legitimate and Distributed Denial of Service (DDoS)traffic.Choosing the machine learning algorithm that offers the best overall classification accuracy for the proposed approach.

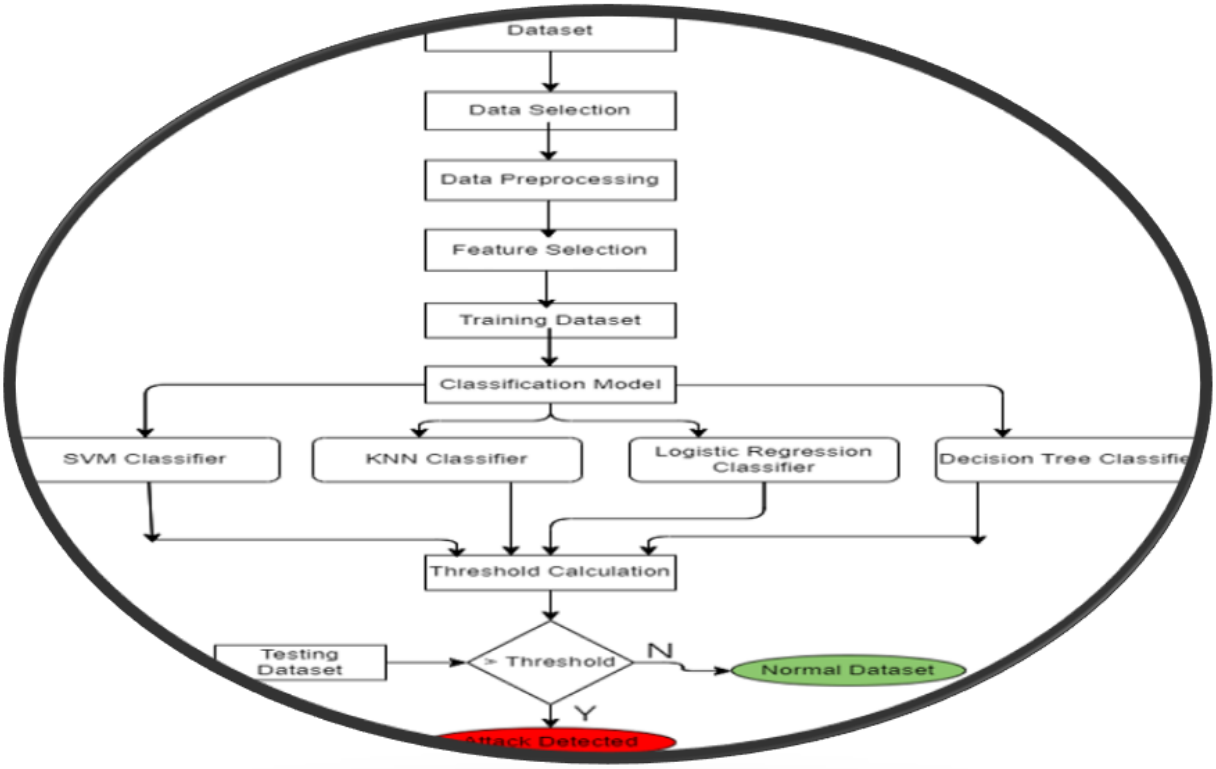

Figure 4: Proposed Approach

During the training phase, support vector machine (SVM) machine learning computations are used to train the classifiers. Using packet flows, ports, and other input data, the SVM approach finds patterns and characteristics that help distinguish between legal and distributed denial of service (DDoS) activity. The primary objective of the training phase is to select the machine learning algorithm that provides the best overall classification accuracy for the recommended detection model. During the evaluation stage, several machine learning methods are tested, and the approach that performs the best in terms of accuracy is chosen to be incorporated into the detection model.

Support Vector Machines, or SVMs, are a well-liked supervised machine learning technique for classification applications. In this instance, the model uses Support Vector Machines (SVM) to create an imaginary line that effectively separates several kinds of data points, allowing it to classify incoming traffic as either legitimate or possibly malicious (DDOS). By training the SVM model on labelled data during the training stage, the approach gains the ability to generalise and generate accurate predictions on unknown data during the detection phase. The Support Vector Machine (SVM) technique aims to increase the model's accuracy in categorising new samples by

optimising the margin between the classes. Using the classifiers that were taught during the training phase, ports, incoming packet flows, and other relevant features are monitored and analysed during the detection phase of the model.
The model compares the features of the incoming traffic to the patterns shown in order to evaluate whether the traffic exhibits symptoms of a denial-of-service attack. Based on a comparison of the trained classifiers with the observed traffic patterns, the detection model makes decisions in real-time about whether to identify traffic as possibly harmful or benign. In the course of real-time observation and classification, the model can swiftly identify and neutralise potential DoS/DDoS attacks, enhancing network security and speed.

8. EVALUATION METRICES:

The following performance measurements were all used, and they are all expressed as percentages.

a. **Precision:**The ratio of genuine positives to the total of false positives and true positives is known as precision.

$$\text{Precision}=TP/TP+FP$$

where the number of DDoS instances that are accurately classified as true positives (TP) and the number of benign instances that are incorrectly labelled as false positives (FP) are, respectively.

b. **Recall:**Remember that this is the ratio of real positives to the total of false negatives and true positives.

$$\text{Recall}=TP/TP+FN$$

where the inaccurate designation of an attack as a benign incident is known as a false negative (FN).

c. **Accuracy:**The number of occurrences in the dataset that are accurately classified as either benign or DDoS attack instances is known as accuracy.

$$\text{Accuracy}=TN+TP/TN+TP+FN+FP$$

where the accurate designation of benign occurrences as benign is known as true negative (TN).

d. **F1-Score:**It is known as the Precision and Recall harmonic mean.

$$\text{F1 score}=(2\times\text{Precision}\times\text{Recall})/ (\text{Precision}+\text{Recall})$$

10. RESULTSAND DISCUSSION

Machine Learnin Algorithm	Accuracy(%)	F1-Score(%)	Precision (%)	Recall(%)
Naïve Bayes	0.9694	98	99	98
SVM	0.7288	76	90	73
K-Neighbor	0.9022	91	94	90
Logistic	0.9836	98	99	98

CONCLUSION

The research suggests a technique for DDoS attack detection based on packet analysis and machine learning algorithms. The work uses the NSL_KDD-master dataset to identify DDoS attacks. 20 of the dataset's 80 features were chosen. Subsequently, the extracted features are employed as machine learning input features, and various algorithms are employed to train and build the DDoS attack detection model. The support vector machine, KNN, logistic regression, and naïve bayes algorithms were used on the data set. The outcomes of the experiment demonstrated that the logistic regression method outperforms SVM, K-Neighbour, and Naïve Bayes in terms of accuracy. When all the findings were compared, it was found that the logistic regression had the highest effectiveness (98.36%).

REFERENCES

[1] Net losses: Estimating the global cost of cybercrime. Technical report, Intel Security Group (previously McAfee, Inc.), 2014.

[2] Zargar ST, JoshiJ, Tipper D “A surveyof defense mechanismsagainst distributed denial ofservice (DDoS)flooding attacks” [J]. IEEE, Corn— munications Survey & Tutorials 2013, 15(4) : 2046—2069

[3] WangBing, ZhengYao, LouWenjing, et a1. DDoSattack protection inthe era of cloud computing and software—defined networking[J]. Computer Networks , 2015, 81(4) : 308—319.

[4] C. Zhang, Z. Cai, W. Chen, X. Luo, and J. Yin, “Flow level detection and filtering of low-rate DDoS,” Computer Networks, vol. 56, no. 15, pp. 3417–3431, 2012.

[5] P.N. Jadhav and B. M. Patil, “Low-rate DDOS Attack Detection using Optimal Objective Entropy Method,” International Journal of Computer Applications, vol. 78, no. 3, pp. 33–38, 2013.

[6] P. Du and S.Abe, “IP packet size entropy-based scheme for detection ofDoS/DDoS attacks,” IEICE Transaction on Information and Systems, vol. E91-D, no. 5, pp. 1274–1281, 2008.

[7] H. Wang, C. Jin, and K.G. Shin, "Defense Against Spoofed IP Traffic Using Hop-Count Filtering." IEEE/ACM Transactions on Networking, 15(1), 40-53, Feb. 2007.

[8] F.Y. Lee and S. Shieh, "Defending against spoofed DDoS attacks with path fingerprint."

International Journal of Computers and Security, Elsevier, 24(7), 571-586, March 2005.
[9] S. K. Pandey and B. M. Mehtre, "Performance of malware detection tools: A comparison," 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 2014, pp. 1811-1817, doi: 10.1109/ICACCCT.2014.7019422.
[10] NSL-KDD data set, https://www.kaggle.com/NSL-KDDdataset.
[11] S. Wankhede and D. Kshirsagar, "DoS Attack Detection Using Machine Learning and Neural Network," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5.
[12] D. Moon, H. Im, I. Kim, and J. H. Park. Dtb-ids: an intrusion detection system based on decision tree using behavior analysis for preventing apt attacks," The Journal of supercomputing, vol. 73, no. 7, (2017): pp. 2881-2895.
[13] C. Yin, Y. Zhu, J. Fei, and X. He. A deep learning approach for intrusion detection using recurrent neural networks," Ieee Access, vol. 5, (2017): pp. 21954{21961.
[14] S.-Y. Ji, B.-K. Jeong, S. Choi, and D. H. Jeong. A multi-level intrusion detection method for abnormal network behaviors," Journal of Network and Computer Applications, vol. 62, (2016): pp. 9-17
[15] I. A. Khan, D. Pi, Z. U. Khan, Y. Hussain, and A. Nawaz. Hml-ids: A hybrid-multilevel anomaly prediction approach for intrusion detection in Scada systems," IEEE Access, vol. 7, (2019): pp. 89507-89521.
[16] N. Shone, T. N. Ngoc, V.D. Phai and Q. Shi, “A deep learning approach to network intrusion detection,” IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no., pp.41-50, 2018
[17] R. Chakraborty and N. R. Pal, “Feature selection using a neural framework with controlled redundancy,” IEEE Trans. Neural Netw. Lear. Syt., vol. 26, no. 1, pp.35-50, 2015.
[18] C. Yin, Y. Zhu, J. Fei and X. He, “A deep learning approach for intrusion detection using recurrent neural networks,” IEEE Access, vol. 5, pp.21954- 21961, 2017.
[19] M. Agarwal, D. Pasumarthi, S. Biswas and S. Nandi, “Machine learning approach for detection of flooding DOS attacks in 802.11 networks and attacker localization,” Int. J. Mach. Learn. & Cyber, vol. 7, no. 6, pp. 1035- 1051., 2016.
[20] A. Shabtai, R. Moskovitch, Y. Elovici, C.Glezer, Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey, Information security technical report 14, 2009.
[21] 1Tripathi, Nikhil &Hubballi, Neminath& Singh, Yogendra. (2016). How Secure are Web Servers? An Empirical Study of Slow HTTP DoS Attacks and Detection. 10.1109/ARES.2016.20.
[22] Gao N, Gao L, Gao Q. An intrusion detection model based on deep belief networks. In: Proceedings of the Second International Conference on Advanced Cloud and Big Data (CBD), Huangshan, China, 2014; pp. 247–252.
[23] Anand, C., Vasuki, N. Trust Based DoS Attack Detection in Wireless Sensor Networks for Reliable Data Transmission. Wireless PersCommun 121, 2911–2926 (2021).
[24] N. Nishanthand A. Mujeeb, "Modeling and Detection of Flooding-Based Denial-of-Service Attack in Wireless Ad Hoc Network Using Bayesian Inference," in IEEE Systems Journal, vol. 15, no. 1, pp. 17-26, March 2021.
[25] S. K. Pandey and B. M. Mehtre, "A Lifecycle Based Approach for Malware Analysis,"

2014 Fourth International Conference on Communication Systems and Network Technologies, Bhopal, India, 2014, pp. 767-771, doi: 10.1109/CSNT.2014.161.

[26] Meng Wang, Yiqin Lu, Jiancheng Qin, (2020) A dynamic MLP-based DDoS attack detection method using feature selection and feedback, Computers & Security, Volume 88, 101645, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2019.101645.

[27] Pandey, Sudhir Kumar, et al. "Gan-based data generation approach for ids: Evaluation on decision tree." Advanced Computing and Systems for Security: Volume 14 (2021): 43-57.

[28] Sharma, Srishti, YogitaGigras, Rita Chhikara, and AnuradhaDhull. "Analysis of NSL KDD Dataset Using Classification Algorithms for Intrusion Detection System." Recent Patents on Engineering 13, no. 2 (2019): 142-147.