NTU Journal of Pure Sciences (2024) 3 (3) : 17-29
DOI: https://doi.org/10.56286/ntujps.v3i3

IRAQI Academic Scientific Journals

ISSN: 2789-1089 EISSN: 2789-1097
NTU Journal of Pure Sciences
Available online at: https://journals.ntu.edu.iq/index.php/NTU-JPS/index

Data Mining Techniques used for Evaluation an Efficient DDoS Attack Detection System: A Deep-Learning Model

Ahmed Shihab Ahmed[1], Hussein Ali Salah[2], Safa Bhar Layeb[3]
[1.] Department of Basic Sciences, College of Nursing, University of Baghdad, IRAQ,
[2.] Department of Computer Systems, Technical Institute- Suwaira, Middle Technical University, IRAQ,
[3.] LR-OASIS, National Engineering School of Tunis, University of Tunis El Manar,Tunis, Tunisia.

Article Information

Received: 27-03- 2024,
Accepted: 01-07-2024,
Published online: 30-09-2024

Corresponding author:
Name: Ahmed Shihab Ahmed
Affiliation : Department of Basic Sciences, College of Nursing, University of Baghdad
Email: *ahmedshihabinfo@conursing.uobaghdad.edu.iq*

Key Words:
DDoS attack,
Machine learning,
RapdiMiner,
Deep learning,
CICDDoS2019datasets,
Network security.

A B S T R A C T

Developed recently to solve the shortcomings of traditional networks, software-defined networking is a new paradigm (SDN). Decoupling the control plane from the data plane, which is the SDN's core property, makes network management simpler and promotes effective programmability. This system employs feedforward neural networks and deep learning techniques in the form of autoencoders to identify and detect denial-of-service (DDoS) attacks. Two datasets were analyzed for the training and testing model, initially using a static approach and later using an iterative approach. Every self-encoding model utilizes a concealed layer, and the input layer and concealed layer are vertically arranged to form the auto-encoding model. The new design, on the other hand, is vulnerable to a number of attacks that could exhaust the system's resources and prohibit the SDN controller from offering services to authorized users. One of these threats, the Distributed Denial of Service (DDoS) assault is one that is gaining popularity. A DDoS assault has a severe negative effect on emphasize servers have no ability to access their network facilities as a result. accommodate legitimate users. The resarchers introduce DDoS Net, a DDoS attacks system for intrusion detection for SDN systems. Our techniques is based on data mining software (Rapdiminer) and neural networks working along with auto encoders. As a result, the resarchers have a lot of faith in protecting these networks thanks to our methods. The resarchers assess the CICIDS 2019 dataset, which includes network behavior associated with DDoS threat and benign network activity, is publicly available, and satisfies certain requirements. It evaluates a variety of data mining algorithms' efficacy as well as internet usage characteristics to determine the best attributes for

Introduction

Data mining (DM) is the method of extracting new information from immense amounts of data in terms of patterns or rules. It involves identifying patterns or rules in large amounts of data. High-performance computing, statistics, machine learning, artificial intelligence, information science, and visualization techniques are a few of the fields involved. Data mining techniques, such as association rules, sequential patterns, classification trees, and others, are used to collect various rules and patterns. It must finish data preparation before it may produce useful information. The major goal of data mining is to uncover hidden information in a batch of data. The information gained is beneficial for making choices. Predictive data is currently being effectively found for a variety of applications using a number of well-liked data mining technologies. Results from data mining can be displayed in many different forms, such as a list, graphic outputs, summary tables, and visualizations. [1, 2].

The DM algorithms assess both recent events and previous computations at the input stage. These techniques can make sure that all of the data information is kept with little loss. We use standard DM techniques since we are not interested in understanding long-term temporal dependencies. The typical DM algorithms are simpler and take less time to train than other DM algorithm techniques [3].

A material removal networking technology called SDN (Software-Defined Networking) makes network management and programmability simpler. By centralizing the network and separating the control plane from the data plane, SDN increases network dependability. Additionally, there are numerous security weaknesses in the emerging paradigm that attackers might take advantage of to conduct a wide range of destructive attacks [4]. Furthermore, these threats can impair the entire SDN system, which contains several equipment from various manufacturers, in contrast to traditional networks, if an attack frequently only harms a small number of only a single the maker's networking equipment without impacting the network as a whole [5].

The SDN network can be used in a variety of ways. The Distributed Denial of Service (DDoS) assault, which disables authorized users from using network services, is one of the most frequent and harmful attacks. A DDoS assault sends out packets with a lot of data, which can use up bandwidth on an internet connection or bring down target servers. Additionally, the Internet of Things era has produced a sizable number of internet-connected devices. Many methods based on Machine Learning (ML) techniques for identifying DDoS attacks have been presented in recent years [6–10].

In the majority of those investigations, feature selection was done using machine learning in order to obtain classification model structures that were more effective. However, there are some problems with applying machine learning to enhance feature selection methods. First, network traffic levels are rising quickly as Internet of Things and data warehousing gain popularity. Traditional ML classifiers fail to manage with enormous volumes of data due to their restricted model training capabilities. When looking for similarities in known assaults as opposed to outliers in unknown damaging attacks, conventional machine learning yields better results [11].

Denial of Service (DDoS) assaults include two distinctive traits. One of them is according to a particular aspect, a lot of traffic temporarily changes the allocation of the target host's upstream neighbors [12]. Due to the extensive shielding and the few assault terminals display a variety of attack behaviors on the target [13]. Prior to and following the assault, the computer network architecture is noticeably altered due to these two inherent characteristics of DDoS attacks. As a result, topological changes can also be used to detect DDoS assaults in addition to traffic characteristics [14].

DDoS detection aims to differentiate between legitimate traffic and attack traffic. Statistics, machine learning, and deep learning are the three primary categories of the current mainstream DDoS detection technologies. The statistical method employs metrics like entropy to evaluate the change in the traffic distribution. Additionally, conventional machine learning is shallow learning, which makes it challenging to learn complex relationships. Its accuracy is typically under 90%. With the help of multilayer neural networks, deep learning uses feature extraction that is already built into the neural network structure to become familiar with the fundamental laws of internet traffic. Additionally, multilayer neural networks may mine deep learning and efficiency [15,16].

In this study, a contains and circulation feature-based approach deep learning methodology (GLD-Net) are suggested. It simultaneously gathers fundamental and flow features from duration stream information and uses

structure consideration network (GAT) to find resemblance between would Include-Euclidean features to fuse fundamental and flow properties. To accomplish feature separation and traffic type visualization, the layer that is completely linked works in tandem using the LSTM network that is attached underneath GAT. The detection rate of the GLD-Net method for two primary groups (normal and DDoS flow) and three distinct kinds contains (normal, rapid DDoS flow, and involves DDoS flow), accordingly, obtains 0.993 and 0.942 in testing utilizing the NSL-KDD2009 and CIC-IDS2017 datasets [17].

In this study, introduce DDoSNet, a deep learning method based on DM Algorithms-auto encoder for identifying Threats by DDoS against the SDN. The recommended methodology outperforms various conventional methodologies for purposes of efficiency, recall, and accurateness. The main contrubutions of this paper are :Presents a hybrid method that uses feedforward neural networks and deep learning as autoencoders to detect denial-of-service (DDoS) assaults. For the training and testing model, two datasets were examined, first statically and then iteratively. Each self-encoding model employs a hidden layer, and the input layer and hidden layer are stacked one on top of the other to create the auto-encoding model. Emploied and suggest a deep learning solution based on DM Tools' auto encoder for identifying DDoS assaults against the SDN (DDoSNet). Identifying traffic through networks as malicious or legitimate, integrate DM Tools' auto encoder with a nonlinear activation regression model at the output. An assess our framework with the the recently made available dataset CICDoS2019, which includes a wide completes in information holes and a range of DDoS assaults.Finally,an assess our proposed model's accuracy, recall, and accuracy by comparing it to several cutting-edge machine learning models that are well-known for identifying DDoS attacks. The best performance is achieved by our suggested methodology.

The organization of this manuscript is as the following. Section 2 discussed the relevant related work. In section 3 discuss the DDoS detection using machine learning. In section 4 we present the proposed model. Section 5 explained the evaluation methodology. In section 6 depicts the Data Mining Algorithms. Section 7 describes the experimental analyses. Section 9 illustrates the Limitations .Section 9 illustrates the conclusions.

Related Work

Different DNN algorithms based on methods have been created recently. proposed two self-organizing map-based techniques for detecting DDoS attacks. The suggested techniques and the detection architecture make use of programmable and adaptable SDN technology. We can quickly carry out complex classification and detection algorithms thanks to the SDN controller. We effectively evaluate the precision and computing demands of our proposed approaches in a testbed environment. The results of the experiments show that these algorithms reduce processing time while keeping a respectable degree of precision [18].

Detecting using Support Vector Machine (SVM) technique served as the DDoS assault detection mechanism in the SDN network. The authors employed six packages attributes even during learning stage it is attainable via the SDN takes into consideration. Five virtual hosts were used to simulate the connection SDN while employing the dividend amount and flooding controller to collect samples for the dataset [19].

Three distinct DDoS scenarios, containing ICMP flooding, UDP, and TCP SYN packets, are generated even during the modeling stage. Four different machine learning techniques were utilized by [20] to identify DDoS assaults in the context of SDN. The researchers created both legitimate the Instruction and Validation Dataset was produced using fraudulent traffic scenarios as well. A pair of instances of DDoS are created using the hping3 software (TCP and ICMP floods). The trial's results showed that the J48 is more accurate than the other strategies examined. Abhiroop et al. used SVM, Naive Bayes (NB), and Neural Network (NN) as three different machine learning methods to identify assaults on the SDN network contains flow-tables [21].

Using topological and flow features, the novel deep learning DDoS assault detection system GLD-Net was employed. A graphical model is demonstrated for extracting features. Topological qualities are provided by node attributes, while edge attributes are supplemented by traffic features. By constructing the characteristic table and transferring topographical elements across the period sequence, an adaptive DDoS architecture construction technique is presented. GAT mines complicated structural connections for non-Euclidean inputs, and LSTM recovers sequences association in matrices. [22].

The use of a machine learning-based random forest algorithm model as a new DDOS assault detection technique was suggested. The features of DDoS traffic from attacks are extracted with a high percentage by performing feature extraction and format conversion on The distributed denial of service (DDOS) tool's three protocols assault packages. The extruded characteristics are then used as input characteristics in a machine learning technique for instruction and create the distributed denial of service (DDoS) identification models.

Then, for model testing, the attack data is combined with the regular traffic data. The results of the experiments demonstrate that the proposed identification of DDoS attacks employing machine learning system has an elevated probability of identification for those who conduct frequent DDoS assaults right now [23].

DDos Detection Using Machine Learning

Machine Learning Algorithms

Current machine learning methods are increasingly used, especially at the exception detection stage, to identify and prevent DDoS. Currently used techniques include The decision tree, K-Nearest Neighbor, neural network, the support vector machine, and naive bayes, and more. In the initial phase, applying specified numerical factors or established guidelines, the network's activity is captured and processed to sort and gain insight according with the regulations recorded in the specified database. The retrieved features are normalized in the second phase in order to prepare them for training. Features in the traffic are identified and removed.

Data is transformed into meaningful information using a collection of algorithms called machine learning. It works best when it supplements a topic master's specialized knowledge rather than taking the place of it. As the name suggests, when one value must be predicted using data from other values in the dataset, a predictive model is used. The learning algorithm makes an attempt to comprehend and model the relationships between the objective and other variables. Supervised learning or classification is the process of using a training predictive model. Decision Trees (DT), Nave Bayes (NB), Logistic Regression (LR), Random Forests (RF), and ID3 are examples of supervised learning techniques. In this study, we develop four machine learning models using the ID3, Naive Bayes, SVM, Random Forests, and Decision Trees algorithms. We then compare these models to determine the best. Version 6.1 of the C5.0 decision tree algorithm. His prior C4.5 (j48) algorithm, which was an improvement over his Iterative Measures of Depression 3 technique, is improved by this method (ID3). The C5.0 algorithm has the advantage of having strong opinions about trimming and automatically making many decisions with pretty suitable defaults. The C5.0 algorithm makes use of the notion of information entropy. The output of the procedure is the corresponding class, and it requires a set of input and output training pairs. Both categorical and numerical data are supported, and the output is shown as a tree to make it intelligible by humans. It has several characteristics, such as [24].

- The C5.0 technique may detect noise and incomplete information.
- The C5.0 approach can identify incomplete data and noise. You can think of the extensive decision tree as a collection of basic laws.
- The C5.0 classifier can forecast which characteristics will be important in classification and which ones won't.

Clustering and negligent pruning weren't problems anymore.

Deep Learning Based of DM Approach

DDoS assaults increase demand on the networks they target. to be delayed over time since it does not seem to be malicious traffic. Thus, historical data is required for DDoS detection in order to analyze running traffic. Other historical and statistical patterns must be employed because we are unable to employ identification using a single transmission or its contents for the detection of DDoS attacks in this situation, and it is also insufficient for performance evaluation. The Deep Detection technique draws a distinction between seized and authorized data by using data mining algorithms to evaluate a constant stream of network packets. Data mining algorithms make use of information history to differentiate between legitimate traffic and DDoS assault traffic. Continuous datamining has the benefit of being independent of window size. The window size in earlier DM algorithms varied depending on the task. The algorithms' capacity to recognize the assault is thus constrained. It has long been known that it can be difficult to train standard machine learning algorithms on a continuous stream of data. This limitation has, however, shown to be useful in identifying malicious packets. The performance improves as the dataset's sample count increases [25].

Feature Selection

In order to improve classification performance and conserve memory, choosing characteristics is a process for selecting a subset of important characteristics from a greater number of characteristics and reducing the number of redundant, unnecessary features in a dataset. Feature extraction helps with data comprehension, mitigating the effects of dimensionality, reducing processing needs, increasing learning accuracy, and

identifying the characteristics that can be essential to a specific issue. The classifications of packaging, filtering, and anchoring models include a variety of techniques for supervised selecting features. Informational gain, which assesses the data benefit of each indicator by evaluating the numerical value of a characteristic determined by volatility with respect to the class, is one of the most commonly utilized filter modeling methods for feature selection. The entropy of a characteristic increases with information richness. The information gain approach [26] was used to evaluate the 80 characteristics of the CICIDS2019 dataset [26].

Attributes Selection Measures

Several metrics are used by machine learning and data mining to create and assess models. The Nave Bayes (NB) algorithm, the Logistic regression algorithm, and the decision trees C4.5 approaches have all been created and evaluated on our experimental datasets. These algorithms' accuracy can be evaluated using the confusion matrix they produce (Precision, recall, F-measure, and ROC space were the four performance evaluations utilized [27]). A unique confusion matrix is used to calculate each of the four measures. In the confusion matrix, the categorization outcomes are shown as a matrix. It contains details regarding the present and anticipated classes of a categorization system. The amount of variables categorised as false when they were unmistakably false (FP) and the data set classed as true when they were precisely true (TP) (TN). The other two cells reflect the number of samples that were incorrectly categorised. Additionally, the cells showing the number of tests labeled as false while they were in fact true (FN) and the number of findings produced during the testing that were flagged as true despite being patently false (TF) (FP). It is simple to calculate the precision, recall, and F-measure after the confusion matrices have been generated.

$$Recall = TP / (TP+FN) \quad (1)$$

$$Precision = TP / (TP+FP) \quad (2)$$

$$F\_measure = (2*TP) / (2*TP+FP+FN) \quad (3)$$

$$TPR = TP / (TP+FN) \quad (4)$$

$$FPR = FP / (FP+TN) \quad (5)$$

Proposed Model

The method and application of topological and flow-based deep learning detection. Our suggested DDoS detection system consists of three primary steps, as shown in Figure 1. The extraction module is the initial phase. It is in charge of taking samples and turning them into graph data made up of nodes and edges by extracting features from open datasets or real-world scenarios. The training module is the second stage, where deep-level information gathered from datasets can be extracted using an algorithm for classification. Data samples are used as the inputs, and a label kind is produced, and training allows for parameter optimization. The evaluation module, which is the third stage, evaluates detection effects under various hyper parameters to choose the best configuration. After the aforementioned steps are finished, the trained neural network is saved as a classifier and The characteristic extracted analysis of patterns is stored as an operator. Then, by passing through these processing components just once without retraining, real traffic may be quickly classified.

The structure of the DDoSNet model we propose (see figure 1). DM Algorithms and an auto encoder form the foundation of DDoSNet. A type of data mining tool (Rapidminir) employed in a program is an auto encoder. In this work, the auto encoder was used because, when in contrast to kernels and linear Evaluation by Principal Components, it may significantly increase the accuracy of anomaly identification (PCA). It has the ability to pick up on little anomalies that linear PCA misses. The auto- encoder is also simple to learn and doesn't need complicated computing, in contrast to kernel PCA. Three distinct steps make up the auto-encoder. The entry layer is the first layer to receive the input vector Xi and is responsible for both encoding and decoding it after passing it via several hidden levels (encoder and decoder blocks). The attributes are scaled back in size in comparison to the input data in the encoder phase, and they are rebuilt in the reverse order in the decoder phase to produce the top-layer output. The final feature vector closely resembles the original input data. To enhance the detection model for DDoS attacks, we combined the autoencoder with the conventional DM technique. An issue with traditional feed-forward DM algorithms can be resolved by autoencoding. It may therefore generate models that are substantially more powerful and have good classification accuracy. The usage of DM is widespread in many fields, including speech and language processing. The DM's cyclic connections can be utilized to successfully model sequences, unlike feedforward neural networks [28].

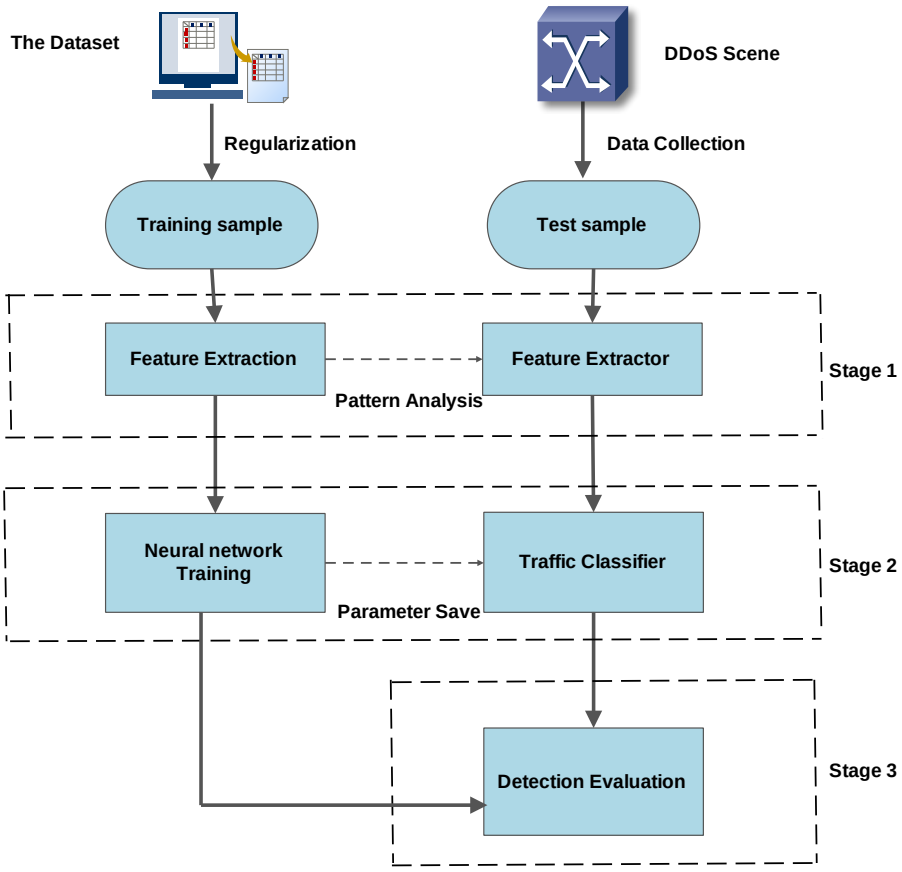

Figure 1. The structure of DDoS attack

The latest features including new sorts of assaults are contained in the Intrusion Detection System (IDS) dataset from the Canadian Institute for Security (CICIDS 2019). A dataset that includes DDoS attacks and that was utilized to create predictions is detailed in the following section. Over eighty network traffic features were gathered and computed for all benign and malicious flows using software called CIC Flow Meter, which is freely available on the Canadian Institute for Cybersecurity website. This dataset has been thoroughly classified. The initial package establishes which way is forwards and which way is backwards of the bidirectional flows it creates. The 80 statistical factors, such as Duration, Package Count, Bytes, and Package Size are established independently for both forward and backward directions. Over 80 network traffic features for each flow are listed in the first six columns: Flow ID, Source IP, Destination IP, Source Port, Destination Port, and Protocol. UDP flows are terminated when a flow timeout happens, but TCP flows are typically terminated when a connection is closed. Each scheme has a flexible way of setting the flow timeout value, 800 seconds, for instance, for TCP and UDP. The program generates a CSV file as its final result.

Evaluation Methodology

Datasets

One of the most important problems for ML/DL intrusion detection algorithms is the availability of datasets. The main reasons why there aren't any datasets in the intrusion detection field are privacy and legal issues. The availability of the network traffic, which contains extremely sensitive information, could reveal corporate and consumer secrets as well as the contents of private conversations. To close the previous gap, several researchers mimic their own data in order to eliminate any sensitive issues. However, the majority of the datasets produced in these situations are unfinished, and the included row samples are insufficient to adequately represent the software operations. In this study, we use the just-released CICDDoS2019 dataset [29] for evaluating our suggested classifiers. The dataset comprises a huge number of potential DDoS attacks that can be

carried out utilizing TCP/UDP-based applications layer protocols. The data set's spread of different assaults is shown in Figure 2. The training data for the testing set's intrinsic evaluation of the detection system did not contain any instances of the PortScan assault. More than 80 flow features were extracted from the dataset using CICFlowMeter software [30]. Both PCAP and flow formats of the CICDDoS2019 dataset are available on the Canadian Institute for Cybersecurity website. Traditional machine learning techniques outperform raw data in feature extraction. The right intrusion system characteristics must be chosen, which is a challenging task that calls for expert guidance. Additionally, it is challenging to pinpoint key characteristics for a certain type of attack because attack scenarios alter on a regular basis. Real-world data are typically non-linear or multivariate in nature. More than three dimensions of the data are challenging to visualize. It is evident that conventional machine learning techniques do not perform well with multidimensional datasets as a result.

Data Preprocessing

Prior to working with actual data, it is essential to pre-process the data. It is employed when the data is ambiguous, challenging to interpret, often heterogeneous (includes errors and outlier values), and hence incomplete. Prior to applying data mining algorithms, preprocessing techniques must be employed to enhance the data's precision and effectiveness, in addition to the effectiveness of the methods used for data mining. Pre-processing activities are regarded as a significant and essential component of data analysis and transportation assessment because of the various circulation patterns in relation to synchronization and size. Techniques for pre-processing data include cleansing, reducing, integrating, discretizing, and transforming data (normalization). The most of methods are employed to normalize data, such z-score, decimal amplification equalization, and min-max. In variables, many calculated values can be discovered. The model will make a classification error if it is only trained on the original data set. The model then requires a lengthy training period, thus the data set is normalized so that the upper limit value is one and the lower bound value is zero [31].

The researchers arrange the data such that it is immediately ready for the learning algorithm. According to figure 2, the CICDDoS2019 dataset is presented in a flow-based fashion, with the CIC Flow Meter extracting over 80 features. The researchers go through a few steps to get the necessary data before the module training.

- Removing connection features: We remove all connection features, including the sources' and targets' IP addresses, ports, timestamps, and stream IDs. Since each network has its own set of these properties, The researchers must use package properties to develop the model.
- Additionally, a regular user and an intruder could have the same IP address. Due to the model's bias toward the connection data, overfitting can emerge when training the DL model with connection data. Processing the data: the original data has a significant amount of incomplete and infinity values, which the researchers entirely erase. After eliminating the unneeded features, we were left with a total of 25 features for the model input.
- Optimize the data input: The values of the characteristics data obtained have different values. The model takes a while to train because using the source data to train it directly can lead to classification errors. The data has been normalized so that zero is the lowest value and one is the greatest value.
- Embedding the classified data: We improved our binary categorizing algorithm to categorize the input data traffic. as legitimate or malicious. As a result, we also count all DDoS classifications as attacks in addition to regular traffic. The binary values of 0 and 1, respectively, are then generated from the string values for the normal and attack labels.

$$W = ((Ni - \min(N))/ (\max(N) - \min(N)) \quad (6)$$

Ni is the data element, N is the minimum and N is the highest value for the entire set of data, and W is a new value. Because part of the values in the CICIDS 2019 dataset are missing, the normalization procedure makes a mistake. Prior to performing the normalization step, the missing value was processed [27].

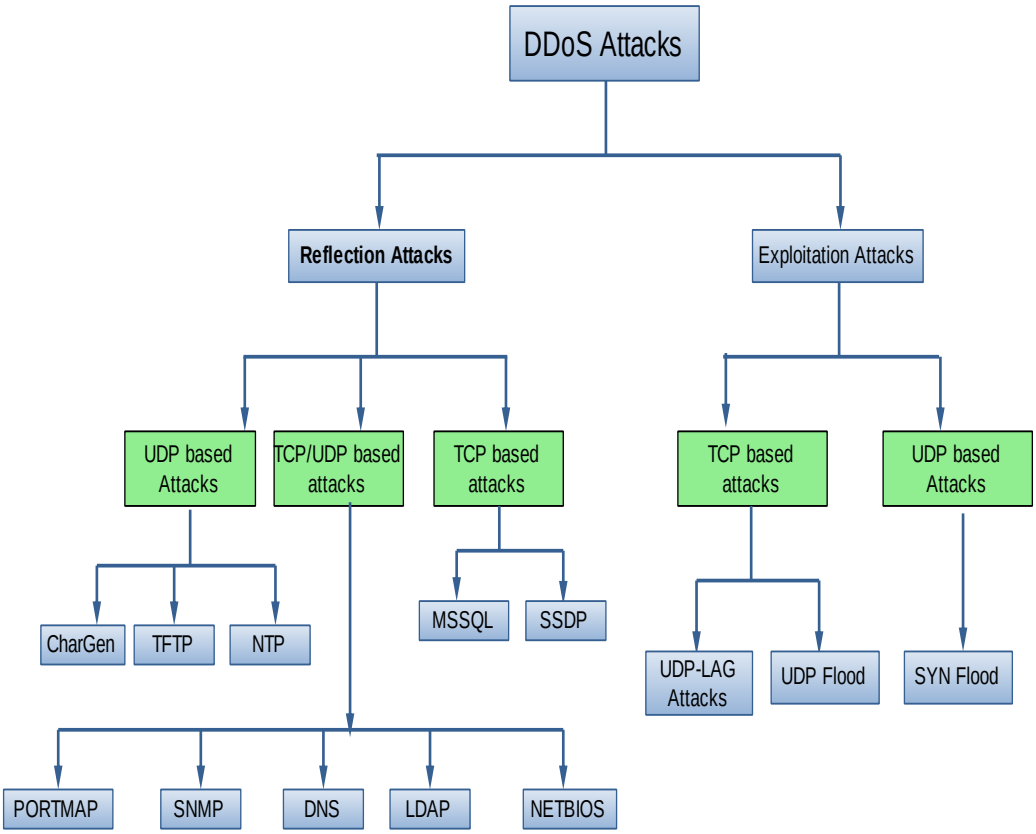

Figure 2. The distribution of DoS and DDoS attack inside the CICDDoS2019 dataset [27]

Data Mining Algorithms

Different algorithms are used in data mining to transform raw data into knowledge that can be applied. A predictive model is used in vocations when it is necessary to forecast a single number utilizing information from the dataset's additional values, as the name suggests. The learning algorithm tries to infer and simulate how the aim and other features relate to one another. Processing a training prediction model is referred to as supervised learning or classification [32].

In this study, four models were developed in order to obtain the findings. The best model was selected after comparing Naive Bayes, Random Forests, Decision Trees, and Logistic Regression. Random Forests is a well-known and effective technique for data exploration, predictive modeling, and analysis. Individual decision trees that are created from a number of independently trained decision trees can deliver results (RF). An ensemble classifier uses learning techniques to generate a group of classifiers, and it then uses a grading system for predictions to classify new data. In the RF technique, which consists of numerous decision trees, the output is accomplished utilizing individual trees [33]. It has several characteristics, such as:

- Offers excellent and efficient services for missing data and methods for dealing with it.
- Because over processing is an issue in some decision trees, this strategy is the best answer [34].

Naïve Bayes (NB)

This method, which is the cornerstone of Bayes Theory, is used when there are a lot of input dimensions. A Bayesian classifier's output can be calculated from its input. You can also upload new data at any time while playing the game and score points for the most accurate probable classifiers. When the category parameter is given, a naive Bayes classifier claims that the availability (or lack) of an attribute assigned to a class is independent to the inclusion or absence of another characteristic [35].

Logistic Regression

It is a method for predictive analysis that operates similarly to other regression analyses. Logistic regression also has the function of describing the data as one of its objectives. Classes and characteristics are connected as a result [27].

Decision tree (DT) C4.5 Algorithm

One of the most important methods is data mining, and machine learning, calculations, and measurements all involve decision trees. A decision tree is used as an insight model to get from a specific idea (shown as branches) to judgments about the object's usage and worth (represented as leaves). Conjunctions of climax refer to the branches that represent the class labels, whereas class labels are represented by the leaves. Decision trees known as regression trees allow the objective variable to acquire permanent properties (often true numbers). Although they are easy to understand and apply, decision trees are considered one of the more popular data mining methods [27].

Experimental Analyses

Before classifying the BENIGN and DDoS attacks, our research divided the 1048560 samples from the CICIDS 2019 dataset into 80 percent training samples and 25 percent testing samples. Several libraries for machine learning techniques are included in the R studio software, which is used to conduct the trials. A confusion matrix is a tool for evaluating the effectiveness of a classification algorithm. It offers true or false categorization findings. Making a prediction model could assist you in identifying the strengths and weaknesses of your categorization model. The choices for categorizing occurrences are as follows, as indicated in Table 1.

Both the True Positive (TP) and True Negative (TN) classifications are correct. When the outcome is incorrectly foreseen as yes (or positives), this is known as a positive prediction that is false (FP) when there isn't any (negative). When a result is projected as negative when it is actually positive, this is known as a false negative (FN) [43]. In order to enhance the outcomes, the performance and accuracy of the chosen characteristics in Table 1 were examined using four machine learning techniques and five folds of cross-validation. Table 1 displays the findings of the confusion matrix for the C5.0 Decision tree, Naive Bayes, Logistic Regression, Random Forest, and ID3 algorithms.

In order to categorize DDoS attack and BENIGN, our analysis used 25 percent of the CICIDS 2019 dataset's 1,059,670 items were used for examination, while 85% were used for learning. Every experiment was run through software. This includes collections of various data mining algorithms. Confusion matrix is a crucial algorithm for analyzing the performance and behavior of classification algorithms. It displays results for categorization into both true and false. One of the most crucial steps in accurately explaining a notion is to calculate confusion. You may do this by creating a two- or matrices with several dimensions that shows the purpose of your categorization model and the kinds of mistakes it makes. The likelihoods of identifying these events as shown in table 1.

Table 1. The confusion matrix for the four algorithms

NB		Predicated class		RF		Predicated class	
		BENGIN	DDoS			BENGIN	DDoS
Actual class	BENGIN	182887	91	Actual class	BENGIN	203053	18
	DDoS	20185	471		DDoS	17	547
DT		Predicated class		LR		Predicated class	
		BENGIN	DDoS			BENGIN	DDoS
Actual class	BENGIN	203061	27	Actual class	BENGIN	203051	34
	DDoS	11	535		DDoS	22	549

Therefore, the classifications of true positivity (TP) and true negativity (TN) are appropriate. When an outcome is incorrectly predicted as a yes (or positive) while it is really a no, the term false positive (FP) is used (negative). False negative (FN) results when an outcome that is truly positive is wrongly forecasted as a negative [33]. With three data mining techniques utilizing 10-fold cross-validation to enhance the outcomes, the efficiency and precision were evaluated based on the attributes chosen as shown in table 3. The performance evaluation of the suggested model using other traditional ML techniques is shown in Table 2 [33].

Table 2. Comparison of the suggested model's performance with various traditional ML techniques. Comparing our suggested method to the other benchmarking techniques, it performs the best

NO	Preprocessing					Algorithm	Result	Ex time
1	sample ALL	Replace missing value	Nominal to numerical Subset Unique integer	Feature correlation 25 feature select	Validation Split>relative Train>0.8 Sampling> shuffled	C5.0 Decision tree Criterion> information-gain depth> 10 Confidence>0.1	accuracy: 98.99%	36 S
2	sample ALL	Replace missing value	Nominal to numerical Subset Unique integer	Feature correlation 25 feature select	Validation Split>relative Train>0.8 Sampling> shuffled	Naïve Bayes	accuracy: 98.09% weighted_mean_recall: 65.17% weighted_mean_precision: 42.18%	25 S
3	sample ALL	Replace missing value	Nominal to numerical Subset Unique integer	Feature correlation 25 feature select	Validation Split>relative Train>0.8 Sampling> shuffled	Logistic regression	accuracy: 98.97% weighted_mean_recall: 95.94% weighted_mean_precision: 97.12%	1.13 m
4	Resampling ALL	Replace missing value	Nominal to numerical Subset Unique integer	Feature correlation 25 feature select	Validation Split>relative Train>0.8 Sampling> shuffled	Random forest	accuracy: 98.98% weighted_mean_recall: 64.39% weighted_mean_precision: 64.67%	27.41 m
5	Resampling 0.3	Replace missing value	Nominal to numerical Subset Unique integer	Feature correlation 25 feature select	Validation Split>relative Train>0.8 Sampling> shuffled	ID3	accuracy: 98.99% weighted_mean_recall: 98.95% weighted_mean_precision: 97.29%	1.04 m

The four most frequently chosen algorithms for data mining are Naive-Bayes (NB), Random Forest (RF), Decision tree (DT), and Logistic regression. Table 3 displays the performance test results of our assessment metrics for these four algorithms. These outcomes depend on the performance measurement equations 2, 3, 4, and 5 as well as the confusion matrices in Table 3.

Table 3. The performance examination results

Model	Accuracy	Recall	Precision	F1 score	Time consumer
RF	0.9898	0.9751	0.9735	0.9742	4.38 m
NB	0.8904	0.8593	0.5214	0.6338	17 S
DT	0.9898	0.9647	0.9800	0.9719	36 S
LR	0.9897	0.9595	0.9712	0.9661	1.61 m

With an accuracy rate of 98.98% and 98.98%, respectively, and a potential success (precision) of 98% for them, Decision tree (DT) and Random Forest (RF) classifiers are superior to the others among the four methods for categorizing data that is numerical in particular that were assessed. The F1scores for DT and RF are 97.19% and 97.42%, respectively, indicating that this experimental approach is preferable. Comparisons with prior studies are shown in Table 4.

Table 4. Comparison with previous studies

Ref	Dataset	Algorithm	Accuracy	
			Previous studies	Our Proposed approach
10	CICDDoS2019	NB	57%	89.04%
		RF	86%	98.98%
		DT	77%	98.98%
		LR	95%	98.97%
19	CICIDS 2017	C4.5	99.96%	98.98%
		LR	92.49%	98.97%

Limitations

Individuals of data mining have numerous obstacles, including dealing with intricate and noisy data, possible reliability concerns, and the significant security and private information consequences.This study is limited to finding a unique way to address this specific class of problems and does not include such issues as distributed systems security and efficacy. Furthermore, the specific detect denial-of-service (DDoS) attacks scenarios described in this study are used for the purpose of this discussion, and are not intended to be an exhaustive generalization of any of the example problems. Other detect denial-of-service (DDoS) attacks problems exist with different seating rules and goals. It is not the intent of this study to account for all methods. Although the Insufficient factual support for a forecast can lead to overestimating its accuracy. Another issue emerges when a database has incomplete data that must be taken into account in order to generate a precise analysis yet to overcome a number of limitations, some of which are, poor flexibility poor maintainability and limited reusability. Also, because the system is still in the implementation and application stage, it is not possible to obtain the related information of the denial-of-service (DDoS) attacks in the application.

Conclusion

Traditional networks do not have the added hazards and vulnerabilities that distributed systems do. Among the latest assault kinds with particularly ferocious tactics, the DDoS assault type wreaks havoc on the entire network infrastructure. In this paper, the researchers suggest DDoSNet, a fresh, DL-based paradigm for identifying DDoS attacks on SDN networks. The researchers used the recently released CICDDoS2019 datasets for instruction and assessment of our suggested technique. The collection contains all of the most current and thorough DDoS assaults. The evaluation of our model showed that DDoSNet offers the most precise assessment measures when compared with currently used, recognized conventional ML approaches. In the future, evaluate how well our suggested model performs using various datasets. In our investigation, the researchers classified using a binary system approach to separate the entrance traffic into categories that were helpful and harmful. However, it is crucial to categorize each attack type separately. This research will be expanded to cover a system of multiple classes. In addition, the researchers will attack traffics and simulate the SDN network in various scenarios to create a heterogeneous dataset that accurately represents modern internet usage. The availability of trustworthy publicly accessible IDS The first of the assessments of the dataset is main problems facing researchers and creators in this sector. Using traditional machine learning algorithms, the researchers discuss the most recent intrusion detection dataset in this study and assess its efficacy. Identification techniques' performance varies depending on the quantity of features and training data samples. The quantity of training data that needs to be gathered increases as the number of features increases.

Acknowledgments. Authors would like to thank Baghdad University, College of Nursing, Department of Basic Sciences , Middle Technical University, Technical Institute-Suwaira, Department of Computer Systems and LR-OASIS, National Engineering School of Tunis, University of Tunis El Manar for the support given during this work.

References

[1] Tedesco, P., Beraldo, P., Massimo, M., Fioravanti, M. L., Volpatti, D., Dirks, R., & Galuppi, R. (2020). Comparative therapeutic effects of natural compounds against Saprolegnia spp.(Oomycota) and Amyloodinium ocellatum (Dinophyceae). Frontiers in veterinary science, 7, 83.

[2] B, V., & Xavier, X. (2024, January 24). A Study on Web Data Mining – Tools and Techniques. International Journal of Research Publication and Reviews, 5(1), 4139–4143.

[3] Pradhan, N., & Dhaka, V. (2020). Comparison-based study of pagerank algorithm using web structure mining and web content mining. Paper presented at the Smart Systems and IoT: Innovations in Computing: Proceeding of SSIC 2019.

[4] Ramani, S., & Jhaveri, R. H. (2022, September 14). ML-Based Delay Attack Detection and Isolation for Fault-Tolerant Software-Defined Industrial Networks. Sensors, 22(18), 6958.

[5] Liu, Y., Wang, Y., & Feng, H. (2023). POAGuard: A Defense Mechanism Against Preemptive Table Overflow Attack in Software-Defined Networks. IEEE Access, 11, 123659–123676.

[6] Santos, R., Souza, D., Santo, W., Ribeiro, A., & Moreno, E. (2020). Machine learning algorithms to detect DDoS attacks in SDN. Concurrency and Computation: Practice and Experience, 32(16), e5402.

[7] Elsayed, M. S., Le-Khac, N.-A., Dev, S., & Jurcut, A. D. (2019). Machine-learning techniques for detecting attacks in SDN. Paper presented at the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT).

[8] Kasongo, S. M. (2023, February). A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework. Computer Communications, 199, 113–125.

[9] Al-Mashadani, A. K. A., & Ilyas, M. (2022, January 20). Distributed Denial of Service Attack Alleviated and Detected by Using Mininet and Software Defined Network. Webology, 19(1), 4129–4144.

[10] Khashab, F., Moubarak, J., Feghali, A., & Bassil, C. (2021). DDoS attack detection and mitigation in SDN using machine learning. Paper presented at the 2021 IEEE 7th International Conference on Network Softwarization (NetSoft).

[11] R L, H. K., K P, B., & A R, D. A. K. (2023, April 30). Mitigation and Detection of DDOS attacks using Software Defined Network (SDN) and Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 11(4), 4821–4829.

[12] Shereen, E., & Dán, G. (2024). Network Topology-aware Mitigation of Undetectable PMU Time Synchronization Attacks. IEEE Transactions on Control of Network Systems, 1–12.

[13] Sharma, K., & Mukhopadhyay, A. (2021). Kernel naïve Bayes classifier-based cyber-risk assessment and mitigation framework for online gaming platforms. Journal of Organizational Computing and Electronic Commerce, 31(4), 343-363.

[14] Shafi, Q., & Basit, A. (2019). DDoS botnet prevention using blockchain in software defined internet of things. Paper presented at the 2019 16th international Bhurban conference on applied sciences and technology (IBCAST).

[15] Alatawi, F. (2021). Defense mechanisms against distributed denial of service attacks: comparative review. Journal of Information Security and Cybercrimes Research, 4(1), 81-94.

[16] Tian, Q., & Miyata, S. (2023, April 12). A DDoS Attack Detection Method Using Conditional Entropy Based on SDN Traffic. IoT, 4(2), 95–111.

[17] Tedesco, P., Beraldo, P., Massimo, M., Fioravanti, M. L., Volpatti, D., Dirks, R., & Galuppi, R. (2020). Comparative therapeutic effects of natural compounds against Saprolegnia spp.(Oomycota) and Amyloodinium ocellatum (Dinophyceae). Frontiers in veterinary science, 7, 83.

[18] Mhamdi, L., McLernon, D., El-Moussa, F., Zaidi, S. A. R., Ghogho, M., & Tang, T. (2020). A deep learning approach combining autoencoder with one-class SVM for DDoS attack detection in SDNs. Paper presented at the 2020 IEEE Eighth International Conference on Communications and Networking (ComNet).

[19] HADI, M. (2022). A Novel Approach to Network Intrusion Detection System Using Deep Learning for Sdn: Futuristic Approach. SSRN Electronic Journal.

[20] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147-167.

[21] Kasim, Ö. (2020). An efficient and robust deep learning based network anomaly detection against distributed denial of service attacks. Computer Networks, 180, 107390.

[22] Hosseini, S., & Azizi, M. (2019). The hybrid technique for DDoS detection with supervised learning algorithms. Computer Networks, 158, 35-45.

[23] Aljuhani, A. (2021). Machine learning approaches for combating distributed denial of service attacks in modern networking environments. IEEE Access, 9, 42236-42264.

[24] Almaraz-Rivera, J. G., Perez-Diaz, J. A., & Cantoral-Ceballos, J. A. (2022). Transport and application layer DDoS attacks detection to IoT devices by using machine learning and deep learning models. Sensors, 22(9), 3367.

[25] Padarian, J., Minasny, B., & McBratney, A. B. (2019). Machine learning and soil sciences: A review aided by machine learning tools.

[26] Hadi, H. J., Hayat, U., Musthaq, N., Hussain, F. B., & Cao, Y. (2022). Developing Realistic Distributed Denial of Service (DDoS) Dataset for Machine Learning-based Intrusion Detection System. Paper presented at the 2022 9th International Conference on Internet of Things: Systems, Management and Security (IOTSMS).

[27] Chen, L., Wang, Z., Huo, R., & Huang, T. (2023, April 5). An Adversarial DBN-LSTM Method for Detecting and Defending against DDoS Attacks in SDN Environments. Algorithms, 16(4), 197.

[28] Alotaibi, F., & Lisitsa, A. (2021). Matrix profile for DDoS attacks detection. Paper presented at the 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS).

[29] Salo, F., Injadat, M., Nassif, A. B., & Essex, A. (2020). Data mining with big data in intrusion detection systems: A systematic literature review. arXiv preprint arXiv:2005.12267.

[30] Saurabh Kumar, & Shwetank. (2023, March 31). Change Detection Analysis of Land Cover Features using Support Vector Machine Classifier. International Journal of Next-Generation Computing.

[31] Chua, T. H., & Salam, I. (2023, June 13). Evaluation of Machine Learning Algorithms in Network-Based Intrusion Detection Using Progressive Dataset. Symmetry, 15(6), 1251.

[32] Mishra, N., Singh, R., & Yadav, S. (2022). Detection of DDoS vulnerability in cloud computing using the perplexed bayes classifier. Computational Intelligence and Neuroscience, 2022.

[33] Mbaabu, O. (2020). Introduction to random forest in machine learning. Engineering Education (EngEd) Program| Section.

[34] Dou, Y., & Meng, W. (2023, April 28). Comparative analysis of weka-based classification algorithms on medical diagnosis datasets. Technology and Health Care, 31, 397–408.

[35] Kwon, H. (2024). Defending Deep Neural Networks against Backdoor Attack by Using De-trigger Autoencoder. IEEE Access, 1–1.