American International University-Bangladesh (AIUB)

DDOS ATTACK DETECTION USING MACHINE LEARNING ALGORITHM

19-40767-2	SHAWON, MOHAMMED IBRAHIM
19-40786-2	JANIN, MD. RAFSAN
19-40426-1	ALI, MD. SAKIB
19-40440-1	MUKTA, MAHMUDA AKHTER

A Thesis submitted for the degree of Bachelor of Science (BSc)

in Computer Science and Engineering (CSE) at

American International University Bangladesh in May 2023

Faculty of Science and Technology (FST)

Abstract

Distributed Denial of Service (DDoS) attacks have grown to be a serious threat to network security in the modern world due to the prevalence of cyberattacks. These attacks clog up the network's bandwidth with a lot of incoming or outgoing traffic, preventing normal business from proceeding. Due to the use of numerous protocols and the disguise of the attacker's identity, identifying and preventing DDoS attacks has become more difficult. Recently, it has become clear that machine learning is a promising method for automating the detection and mitigation of DDoS attacks. Attacks known as Distributed Denial of Service (DDoS) are difficult to identify and counter because they send a lot of traffic through networks, overloading the bandwidth. The application of machine learning algorithms, like Decision Trees, Logistic Regression, Naïve Bayes as well as Random Forests, to automatically detect DDoS attacks is explored in this thesis. The procedure includes gathering data, preprocessing, feature extraction, training a model, evaluating it, and deploying it. Each algorithm has particular benefits and drawbacks, and the choice is made based on the needs of the application. Machine learning can be used to increase network security by quickly identifying and thwarting DDoS attacks.

The purpose of this thesis is to examine how machine learning algorithms like decision trees, logistic regression, and random forest can be used to enhance network security and DDoS attack detection. The research process includes data collection, preprocessing, feature extraction, model training, evaluation, and deployment of the most effective model. Various criteria, including accuracy, interpretability, and computational efficiency, will be used to evaluate each algorithm's effectiveness. This study will advance our knowledge of how Machine learning is employed to identify DDoS attacks and how it might enhance network security.

Declaration

We certify that this thesis is our original work and has not been submitted in any way to another university or other tertiary educational institution for a different degree or diploma. A list of references is provided, and information taken from both published and unpublished works of others is acknowledged in the text.

We acknowledge that the owner(s) of the copyright for all of the material in my thesis are responsible for maintaining that material's integrity. For the purposes of reproducing material in this thesis, we have, when necessary, received permission from the copyright holder and, in the case of any jointly authored works, we have asked the co-authors for their consent.

SHAWON, MOHAMMED IBRAHIM
19-40767-2
Department of Computer Science

ALI, MD. SAKIB
19-40426-1
Department of Computer Science

MUKTA, MAHMUDA AKHTAR
19-40440-1
Department of Computer Science

JANIN, MD. RAFSAN
19-40786-2
Department of Computer Science

Approval

This thesis titled “DDOS Attack Detection Using Machine Learning Algorithm” has been submitted to the following respected member of the board of examiners of the Faculty of Science and Technology impartial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering on dated and has been accepted satisfactory.

SUPTA RICHARD PHILIP
Supervisor
Lecturer, Faculty
Department of Computer Science
American International University-Bangladesh

Md. MASUM BILLAH
External
Assistant Professor, Faculty
Department of Computer Science
American International University-Bangladesh

DR. AKINUL ISLAM JONY
Associate Professor & Head (UG)
Department of Computer Science
American International University-Bangladesh

DR. MD. ABDULLAH - AL - JUBAIR
Assistant Professor & Director
Faculty of Science and Technology
American International University-Bangladesh

PROF. DR. DIP NANDI
Professor & Associate Dean
Faculty of Science and Technology
American International University-Bangladesh

MASHIOUR RAHMAN
Sr. Associate Professor & Dean-in-charge
Faculty of Science and Technology
American International University-Bangladesh

Contributions by authors to the thesis

List the significant and substantial inputs made by different authors to this research, work, and writing represented and/or reported in the thesis. These could include significant contributions to the conception and design of the project; non-routine technical work; analysis and interpretation of researchdata; drafting significant parts of the work or critically revising it so as to contribute to the interpretation.

	Shawon, Md. Ibrahim	**Ali, Md. Sakib**	**Mukta, Mahmuda Akhter**	**Janin, Md. Rafsan**	**Contribution (%)**
	19-40767-2	*19-40426-1*	*19-40440-1*	*19-40786-2*	
Conceptualization	25%	25%	25%	25%	100 %
Data curation	25%	25%	25%	25%	100 %
Formal analysis	25%	25%	25%	25%	100 %
Investigation	25%	25%	25%	25%	100 %
Methodology	25%	25%	25%	25%	100 %
Implementation	25%	25%	25%	25%	100 %
Validation	25%	25%	25%	25%	100 %
Theoretical derivations	25%	25%	25%	25%	100 %
Preparation of Figures	25%	25%	25%	25%	100 %
Writing – original draft	25%	25%	25%	25%	100 %
Writing – review & editing	25%	25%	25%	25%	100 %

Acknowledgments

First and foremost, we want to express our gratitude to Allah, the Almighty, for making the dissertation thesis a huge success. We would like to express our gratitude to the Faculty of Technology for continuing to include thesis credit in the requirements for graduate study and for giving us the chance to experience the academic quests with our care. We are also grateful to the American International University-Bangladesh (AIUB) Faculty of Science and Technology, Office of Placement and Alumni, for setting up a chance for the thesis to be finished.

We also acknowledge the invaluable assistance, counsel, and inspiration provided by our esteemed supervisor SUPTA RICHARD PHILIP, Lecturer of Computer Science, throughout the course of the study. He also inspired us to complete our thesis successfully and smoothly, and for that, we are sincerely appreciative of his dedicated support, supervision, guidance, and advice.

Keywords

DDoS, Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, SDN.

Table of Contents

List of Figures

List of Abbreviations and Symbols

Mention all the abbreviations and the different symbols that is used in this document.

Abbreviations	
DDoS	Distributed Denial of Service
ICMP	Internet Control Message Protocol
UDP	User Datagram Protocol
SYN	Synchronize
HTTP	Hypertext Transfer Protocol
TCP	Transmission Control Protocol
RF	Random Forest
NB	Naïve Bayes
DT	Decision Tree
LR	Logistic Regression
SDN	Software-defined networking

Symbols	
$\hat{\rho}$	Density operator
®	Convolution

Chapter 1

Introduction

In the modern era, the emergence of new threats and security issues has made cyberattacks a research concern that is growing. Distributed Denial of Service (DDoS) attacks are one of the various types of cyberattacks that have grown to be particularly troublesome. These assaults entail flooding a network or website with a lot of incoming or outgoing traffic, which messes up normal operations and renders the website or network inaccessible to authorized users. DDoS attacks come in a variety of forms, including ICMP, UDP, SYN, and HTTP flood attacks. They frequently target well-known websites like social media, web services, and banking sites. DDoS attacks are frequently launched when a website is busiest, making it more challenging to separate legitimate traffic from attack traffic. It is also challenging to identify and stop such attacks because attackers frequently hide their identities by utilizing trustworthy third-party components. The fact that DDoS attacks use both transport layer protocols (TCP, UDP, or both) and application layer protocols makes it difficult to detect them. The fact that various attack types can originate from either TCP or UDP further muddles the detection procedure. To make matters worse, malware can be installed on computers via particular software, and this malware can be used to manage bots that conduct a range of attacks, including DDoS and data theft. Machine learning has become increasingly popular in recent years as a way to identify and stop DDoS attacks. Computers can now behave and react similarly to people thanks to machine learning algorithms. As they gain experience, these algorithms become more adept at identifying patterns and behaviors that are suggestive of an attack. They are trained on sizable datasets of network traffic. To identify DDoS attacks, machine learning algorithms like decision trees, logistic regression, Naïve Bayes and random forests are frequently used. DDoS attacks present a significant threat to network security, and it can be difficult to identify and stop them. A promising solution to this issue is provided by machine learning algorithms, which allow computers to recognize and stop DDoS attacks by anticipating patterns and behaviors that an attack will take place. The use of machine learning offers a promising way forward in the fight against cybercrime, even though the difficulty of detecting and preventing DDoS attacks is likely to remain a significant one.

1.1 About DDOS Attack Detection Using Machine Learning

Cyberattacks are a common problem in this day and age. A hot topic for research is the emergence of new threats and security issues. A Distributed Denial of Service (DDoS) attack is primarily a network attack that overloads network bandwidth by sending a large amount of inbound or outbound traffic over it, disrupting regular operations. DoS attacks come in a variety of forms, but the most prevalent ones are ICMP, UDP, SYN, and HTTP flood. Social media, web services, and banking websites are the most common attack targets. The most popular times for attacks on websites are when they are most busy. There are many different DDoS attack types where the attacker's identity is kept secret by using reliable third-party components. DDoS attacks use application layer protocols and either TCP, UDP, or both of these to use transport layer protocols, making them difficult and difficult to detect. From TCP or UDP, various attacks flow. A specific computer is infected with malware using a particular piece of software. Bots can therefore carry out a variety of attacks, including data theft and DDoS. Recent years have seen a rise in interest in machine learning. It makes it possible for computers or machines to function and respond in a manner that is comparable to that of people. By learning the anticipated behavior, these systems get better with use. Detecting Distributed Denial of Service (DDoS) attacks has become easier thanks to the use of machine learning algorithms like decision trees, logistic regression, Naïve bayes and random forests. This entails gathering and preparing data, extracting features, training and assessing models, and deploying the top-performing model to track network traffic. The logistic regression model represents the association between input features and probability, the decision tree algorithm builds a model based on feature values, and random forest combines multiple decision trees to increase accuracy. Network security can be improved by machine learning applications' automatic detection and mitigation of DDoS attacks.

1.2 Problem Statement

Recent years have seen a rise in interest in machine learning. It makes it possible for computers or machines to function and respond in a manner that is comparable to that of people. By learning the anticipated behavior, these systems get better with use. Cyberattacks are a common problem in this day and age. A hot topic for research is the emergence of new threats and security issues. A Distributed Denial of Service (DDoS) attack is primarily a network attack that overloads network bandwidth by sending a large amount of inbound or outbound traffic over it, disrupting regular operations. DDoS attacks come in many different forms, but the most prevalent ones are ICMP, UDP, SYN, and HTTP flood. The most popular attack platforms include social media, web services, and banking websites. When websites are busiest is when attackers tend to launch attacks. There are numerous DDoS attack types where the attacker's identity is concealed by using reliable third-party components. DDoS are difficult to detect because they are carried out using application layer protocols and either TCP, UDP, or both transport layer protocols (TCP and UDP). Many attacks stem from TCP or UDP. Malware is specifically installed on computers through specific software. As a result, bots can carry out a variety of attacks, including DDoS and data theft. Machine learning, which enables computers to behave and respond like people, is becoming more popular as a result of the emergence of new threats and security issues in recent years. Attackers frequently use Distributed Denial of Service (DDoS) attacks to overwhelm networks with a lot of traffic and stop normal operations. DDoS attacks can be carried out in a variety of ways and on a variety of platforms, with the attackers' identities being hidden by the use of trustworthy third-party components. Because they employ various protocols and frequently result from malware that has been installed on computers, these attacks are challenging to identify.

1.3 Research Question

1. How precise is the Decision Tree, Naive Bayes, Logistic Regression, and Random Forest algorithms in detecting attacks in SDN networks and which algorithm performed the best?

2. What are the key features of the SDN dataset that are most important for detecting attacks using machine learning algorithms?

3. How does the chosen machine learning algorithm perform on attacks of various kinds in the SDN dataset, and which types of attacks are the most challenging to detect?

Chapter 2

Literature review

The Naive Bayes classification method, which relies on the Bayes theorem, makes the predictors' independence a given. It is a straightforward algorithm with high computational efficiency that works for both binary and multiclass classification issues.[1]

A classification and regression algorithm called Random Forest creates a lot of decision trees, combines the output from each tree, and then uses the combined information to predict. The algorithm randomly selects subsets of the data and features for each tree, which helps to reduce overfitting and increase accuracy. It is a popular and powerful algorithm for both classification and regression tasks, and can handle high-dimensional data with complex relationships.[20]

A Decision Tree is a classification and regression algorithm that recursively partitions the data into subsets based on the value of the predictors Each internal node represents a test on a predictor variable, each branch represents the test result, and each leaf node represents a class label or a numerical value, creating a tree-like model of decisions and their potential outcomes. It is a simple and interpretable algorithm that can handle both categorical and numerical predictors, but it is prone to overfitting and can be sensitive to small changes in the data.[1]

Logistic Regression is a classification algorithm that models the probability of the response variable as a function of one or more predictor variables. It uses a logistic function to transform a linear combination of the predictors into a probability value between 0 and 1, which can be interpreted as the likelihood of belonging to a particular class. It is a popular and widely used algorithm that can handle both binary and multiclass classification problems, but It presumes that the predictors have a linear relationship and the log-odds of the response, and can be sensitive to outliers and multicollinearity.[1]

Software-Defined Networking (SDN) is an architecture that separates the control plane from the data plane in a network, allowing the network administrator to manage and control the network centrally and programmatically. The data plane is in charge of carrying out the actual forwarding while the control plane is in charge of making decisions regarding how traffic should be forwarded through the network. SDN allows network operators to define policies and rules in software that can be automatically applied to the network devices, making it easier to manage and optimize the network.[2]

TCP (Transmission Control Protocol) is a reliable, connection-oriented transport protocol that operates at the transport layer of the Internet Protocol (IP) suite TCP enables the reliable, sequential delivery of a stream of bytes over an unreliable network from one program on one computer to another program on another computer. It uses a number of mechanisms such as flow control, congestion control, and error detection and recovery to ensure that the data is delivered correctly and efficiently. TCP is widely used for applications that require reliable data transfer, such as web browsing, email, and file transfer. [3]

UDP (User Datagram Protocol) is a connectionless transport protocol that operates at the transport layer of the Internet Protocol (IP) suite. UDP provides a simple, unreliable, and datagram-oriented service to send and receive messages between applications running on different hosts over an IP network UDP is a simpler and faster protocol for applications that do not require a reliable data transfer, such as streaming media, online gaming, and real-time communication, as it does not provide reliability, flow control, congestion control, or error detection and recovery mechanisms, in contrast to TCP.[3]

The Internet Control Message Protocol (ICMP) is a network-layer protocol that hosts and routers use to exchange data and error messages at the network level. ICMP messages are typically generated by network devices in response to errors or exceptional conditions that occur in the IP network, such as a destination host or network being unreachable, a router forwarding loop, or a time-to-live (TTL) exceeded. ICMP messages are encapsulated within IP datagrams and are typically not meant to be seen by the end-user, but rather by network administrators and troubleshooting tools.[3]

The new data is fed through each tree during the prediction stage, and the class predictions from each tree are combined to produce the final output. Implemented a machine learning algorithm that started pinging an IP address until it died. The model was trained using the random forest algorithm, and it correctly classified 99.76% of instances. They have used Network-based intrusion detection (NSL_KDD) dataset for experiment.[4]

Detected DoS attack effectively using Machine learning (ML) and Neural Network (NN) algorithms.by using the latest DoS attack dataset CIC IDS 2017. The primary goal of the essay is to use Random Forest (RF) and The CIC IDS 2017 dataset was classified using multi-layer perceptron (MLP) algorithms into binary classification. They have detected attack by using RF and MLP algorithm where the RF result was far better.[5]

Have put into practice machine learning techniques like J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network. No single machine learning algorithm among them is capable of effectively defending against every type of attack. The conclusion reached after implementing all of these types of algorithms was that this algorithm could support various attacks.[6]

They used trained algorithms to simulate a DDoS attack on the network while training the dataset with the Naive Bayes and Random Forest algorithms. Application of the Naive Bayes and Random Forest algorithms allowed for the detection of the DDoS attack. Following implementation, the Naive Bayes algorithm produces superior outcomes to the Random Forest algorithm.[7]

A feature selection technique based on a clustering approach has been implemented. Five different ML algorithms were used to compare the algorithms. For training, support vector machines (SVM) and random forests (RF) were employed. The highest accuracy in RF was around 96%.[8]

Designed using a semi-supervised, entropy-based machine learning approach. This implementation uses both supervised and unsupervised compositions, and the unsupervised technique performs well with few false positives. However, supervised techniques help to lower false-positive rates. For this experiment, recent datasets were used. [9]

To optimize accuracy, we can use the NB model using Gaussian classifier. DS03_IG is one of the best datasets. AUC scores of DS03_IG are better than the other datasets among all the ML algorithms. The AUC score of DS03_IG is 93.1832% which is 0.1832% higher than other datasets [10]

The accuracy rate and detection rate of the AdaBoost algorithm is very good, which is more than 95%. IDS based on Machine Learning has some advantages. It does not need any modification of its protocol. It can use any kind of encryption algorithm[11]

To prevent Blackhole attacks and Wormhole attacks, EC-BRTT one of the best methods which use trip time for finding data delay time. If there is a time delay it can automatically detect the attacks and it also reduce the communication overhead[12]

By spotting unusual behaviors or inappropriate uses, intrusion detection systems (IDS) aim to keep the network secure. By spotting attempted and successful attacks at the endpoint or inside the network, intrusion detection systems offer more thorough security functionality than access control barriers.[13]

In SDN, a new technique finds DDoS attacks. The three sections of this method are collector, entropy-based, and classification. The UNB-ISCX, CTU-13, and ISOT datasets were used in the experiments, and the results show that this method outperforms its competitors in terms of accuracy when it comes to detecting DDoS attacks in SDN.[14]

It has been demonstrated that USML (unsupervised learning) performs the best in terms of accuracy, false alarm rate (FAR), sensitivity, specificity, false positive rate (FPR), AUC, and MCC when attempting to distinguish between Botnet and regular network traffic. In the fields of computer security and other related areas, this validation is significant.[15]

The IP addresses and MAC addresses of the malicious devices are sent to the SDN controller when abnormal packets are discovered. These MAC and IP addresses are added to blacklists by the SDN controller, and the controller determines the policies for SDN switches based on the blacklists. The SDN switches immediately block the devices on the blacklists.[16]

This ensemble model for the Friday morning dataset has been seen to achieve a prediction accuracy of 97.86%. Similarly, for the Friday afternoon log file, the prediction accuracy for 16 attributes obtained through feature selection based on information gain and ML model based on regression analysis is 73.79%.[17]

In terms of running time, C4.5 takes 0.58s, and Naive Bayesian takes 1.1s. With a score of 98.8 (%) for accurate classification, C4.5 outperforms Naive Bayesian, which received a score of 91.4 (%). The results show that the C4.5 is the superior method in terms of the selected classification technique.[18]

Attack Enhanced detection and response module integration will be the subject of future research and development. There is currently no detector responder coordination; instead, detectors generate brief recommended rules for responders to impose. The individual packet classification decisions made by the responder in a detection/response system with closer coupling could make use of the rich data structures kept by the detector. This would enable more targeted filtering and rate limiting, as well as lessen the possibility of responses having an adverse effect on legitimate traffic.[19]

This paper evaluates four machine learning algorithms for identifying and categorizing DDoS attacks. The KDDCup99 dataset, which is based on the PCA feature selection method to rank and select features, is used as attack data. A Meta-evaluation of Machine Learning Techniques for Detection of DDoS Attacks reveals that the results of the experiment demonstrate that the Random Forest classification algorithm outperforms with an accuracy of. 99.95% in attack detection, but it is not very effective when it comes to machine learning model training time.[20]

Machine learning algorithms are used in this to efficiently detect DDoS attacks. The attack data set is the CAIDA data set, and relevant features have been chosen based on chi-square and information gain ranking. According to experimental findings, fuzzy c-means clustering provides better classification and is quicker than the other algorithms.[21]

The authors of this paper examined various CAV communication-based cyber security attacks. In order to evaluate the efficacy of CAV cyber-attack detection by the two classification models, the newly created CAV-KDD data set was statistically analyzed using the Naive Bayes and Decision Tree machine learning algorithms. Naive Bayes had better PROBE attack detection accuracy than Decision Tree, while Decision Tree had better DoS attack detection accuracy. When trying to identify U2R and R2L attacks, both models had poor performance.[22]

Chapter 3

Methodology

DDoS attacks are a significant risk to network security, and they can be automatically detected using machine learning algorithms like decision trees, logistic regression, and random forests. Data collection, preprocessing, feature extraction, model training, evaluation, and deployment are steps in the process of detecting DDoS attacks using machine learning techniques. The decision tree algorithm is straightforward but efficient, logistic regression is understandable and yields insights, and random forest lessens overfitting and boosts precision. The choice of algorithm is based on the particular requirements of the application because each method has advantages and disadvantages of its own. In the end, machine learning algorithms can greatly improve DDoS attack detection and network security.

Automated DDoS attack detection is possible thanks to machine learning algorithms, which are a serious threat to network security. There are several options for DDoS attack detection, each with strengths and weaknesses, including decision tree, logistic regression, and random forest algorithms. The selection of an algorithm is based on the application's particular needs, including those for accuracy, interpretability, and computational efficiency.

3.1 Introduction

The detection of DDoS attacks can be greatly enhanced by the use of decision tree, logistic regression, and random forest algorithms. The process entails gathering the data, preprocessing it, separating the features, training the model, assessing it, and deploying it. The choice of algorithm depends on the particular requirements of the application because each method has advantages and disadvantages of its own.

The following justifies the popularity of decision trees, logistic regression, and random forests as machine learning techniques for DDoS attack detection:

1. **Decision tree:** A decision tree can be used to determine the key characteristics that distinguish between legitimate and malicious traffic. It is simple to understand and use. Because they can be quickly constructed, decision trees are a viable solution for handling large datasets.
2. **Logistic regression:** It is a kind of linear model that can simulate the association between the characteristics of the input data and the likelihood that the traffic is malicious or not. For binary classification issues like DDoS attack detection, it is an easy and effective technique.
3. **Random forest:** To increase the model's accuracy, this more advanced machine learning algorithm combines different decision trees. It can manage non-linear relationships between features and lessen overfitting, which is a common issue in machine learning.
4. **Naive Bayes:** The Naive Bayes algorithm, which is based on Bayes' theorem, is a straightforward and efficient probabilistic method for classification tasks. Assuming that the features are independent of one another, which is frequently false in datasets from the real world, is what gives the algorithm its "naive" moniker. Naive Bayes has been demonstrated to be effective in a variety of classification tasks, despite this simplification assumption. Naive Bayes is a simple probabilistic classifier built on the Bayes theorem and the feature independence presumption. It is widely used for classification tasks in machine learning and natural language processing.

Data collection, preprocessing, feature extraction, model training, evaluation, and deployment are steps in the process of using machine learning techniques to detect DDoS attacks. This process makes sure the machine learning model is trained on a variety of datasets and can generalize to new data effectively. To choose the best-performing model for deployment in a production environment, it is essential to complete the evaluation step.

3.2 Dataset

The SDN dataset is specific to software-defined networking and was generated using the Mininet emulator for traffic classification with machine learning and deep learning algorithms. The dataset includes data from ten Mininet topologies with switches connected to a single Ryu controller. Both legitimate traffic types, such as TCP, UDP, and ICMP, and malicious traffic types, such as TCP Syn attack, UDP flood attack, and ICMP assault, were simulated. The dataset includes 23 features, with some features calculated and others retrieved from the switches, such as switch-id, packet count, byte count, duration, source and destination IPs, total duration, tx bytes, rx bytes, date and time, packet rate, number of packet ins messages, total flow entries, tx kbps, rx kbps, port bandwidth, and class name. The class name shows whether the traffic is benign or malicious, with 1 indicating malicious traffic and 0 indicating benign traffic. The dataset contains 1,04,345 rows of data collected during a 250-minute network simulation, and more data can be obtained by repeating the simulation for the same interval. The dataset is particularly useful for developing and testing models to automatically detect DDOS attacks in software-defined networking environments.

We have collected SDN Dataset[23] from **Automated DDOS attack detection in software defined networking.**

(https://data.mendeley.com/datasets/jxpfjc64kr/1)

3.3 Graphs and Chart of experiment

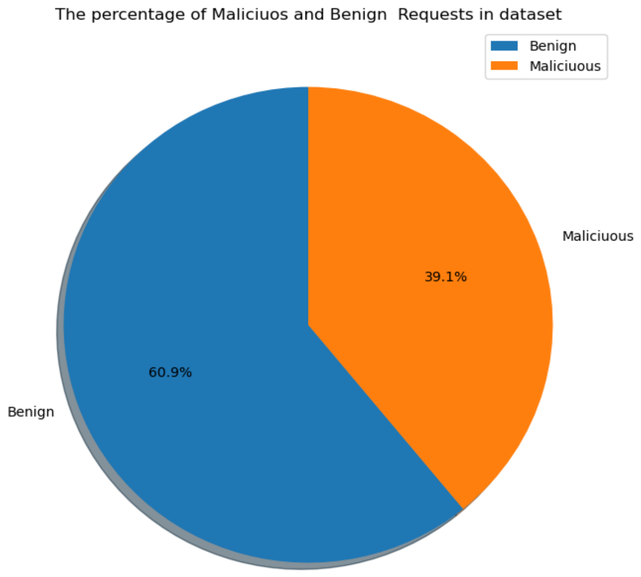


Figure 1:The percentage of Benign and Malicious Request in dataset.

From the Fig-1, we can see the percentage of Benign and Malicious Request in dataset. In the graph it shows that there is more Benign request than Malicious request where Benign with 60.9% and Malicious with 39.1%.

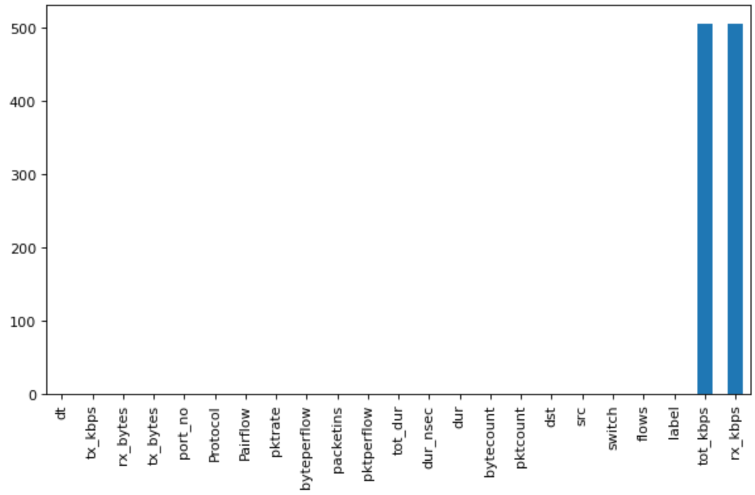


Figure 2:Features which has Null values.

From Fig-2, it shows that most of the features has null values except two features. One is rx_kbps and the other one is tot_kbps.

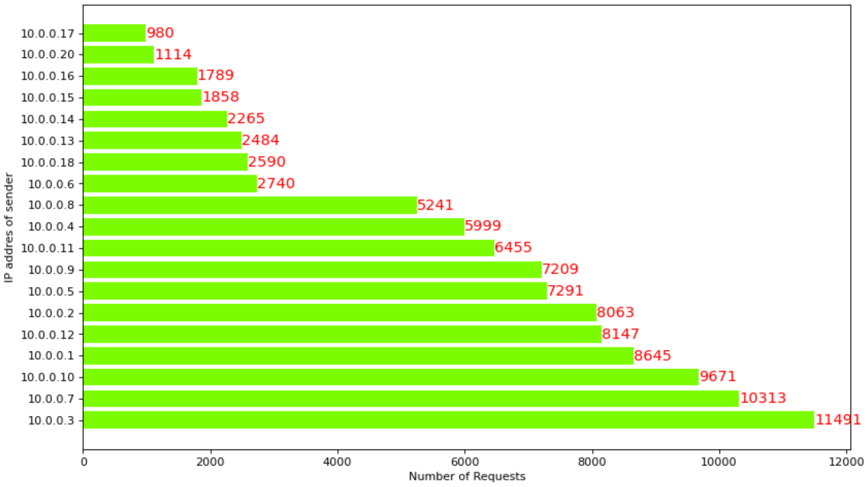


Figure 3:Number of all request.

In Fig-3, it shows the number of requests from different sender IP addresses, where we can know which Ip addresses has more request and which has less. Here we can see IP address 10.0.0.3 has more requests comparing other IP addresses which is 11,491 and IP address 10.0.0.17 has less requests comparing other IP addresses which is 980.

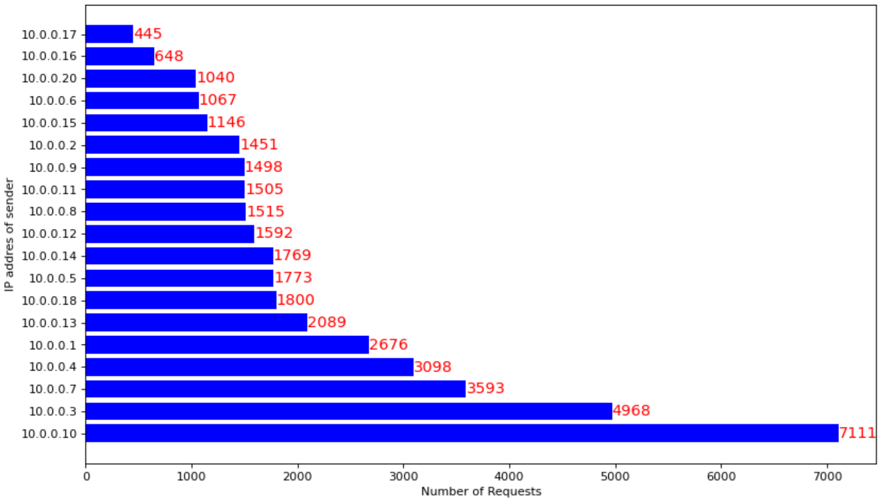


Figure 4:Number of attack requests.

In Fig-4, it shows the number of attack requests from different IP addresses of sender, where we can know which Ip addresses has more attack request and which has less. Here we can see IP address 10.0.0.3 has more attack requests comparing other IP addresses which is 7,111 and IP address 10.0.0.17 has less attack requests comparing other IP addresses which is 445.

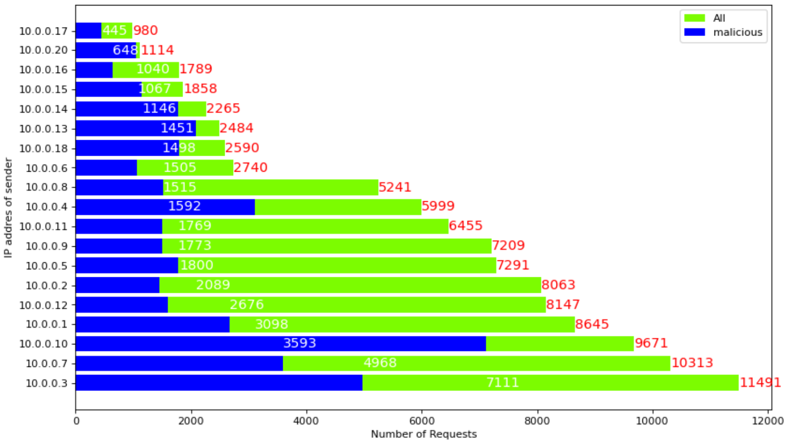


Figure 5:Number of requests from different IP address.

In Fig-5, it shows both the number of requests and the attack requests combinedly from different IP addresses, where we can compare between only requests and attack requests among all the Ip addresses of senders. Here we can see for IP address 10.0.0.10, most of the time it has attack request among normal requests and the difference between normal and attack request is less than any other IP addresses of sender.

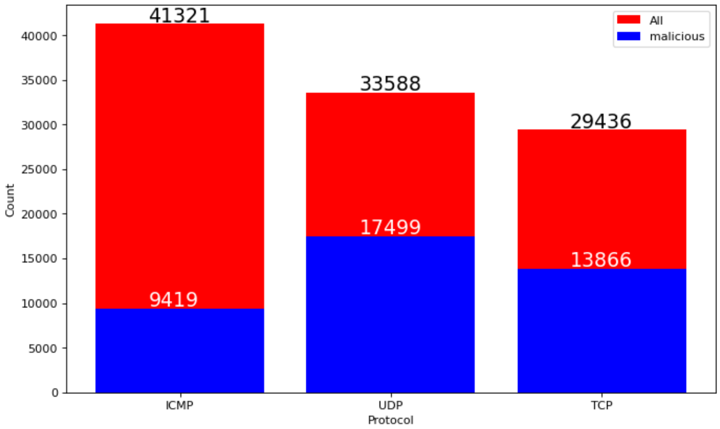


Figure 6: The number of requests from different protocols.

In Fig-6, it shows the number of requests from different protocol such as ICMP, UDP and TCP. Where we can compare malicious requests with all requests. Here for UDP protocol, the difference between normal requests and malicious requests is less than other protocols. So UDP protocol request more malicious than ICMP and TCP protocol by comparing difference of their normal requests and malicious requests.

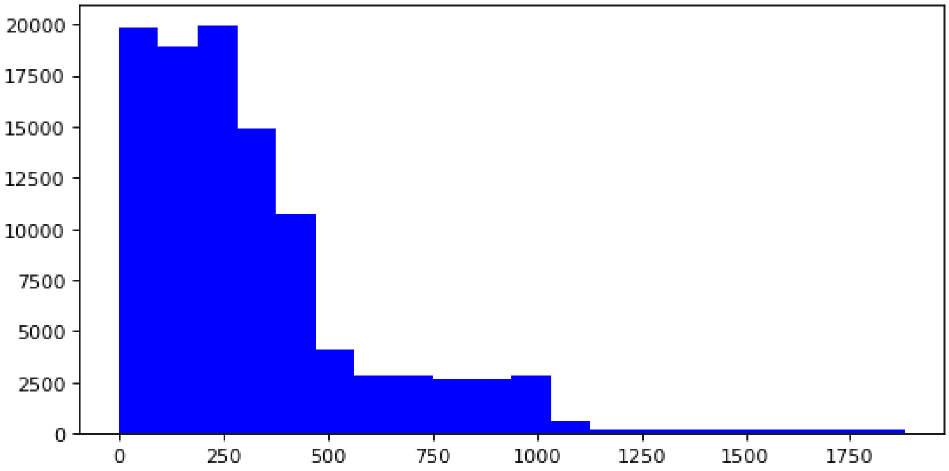


Figure 7:: Duration.

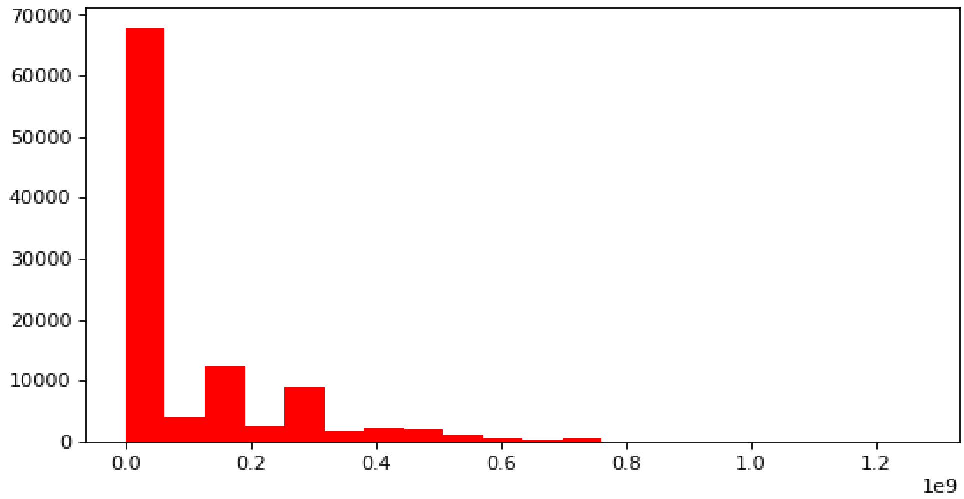


Figure 8:TX_BYTES – Transmitted Bytes.

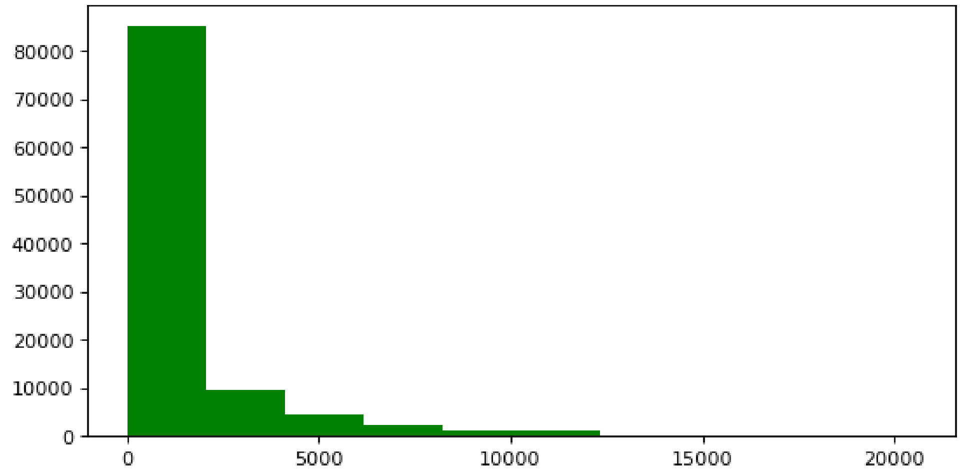


Figure 9: TX_KBPC.

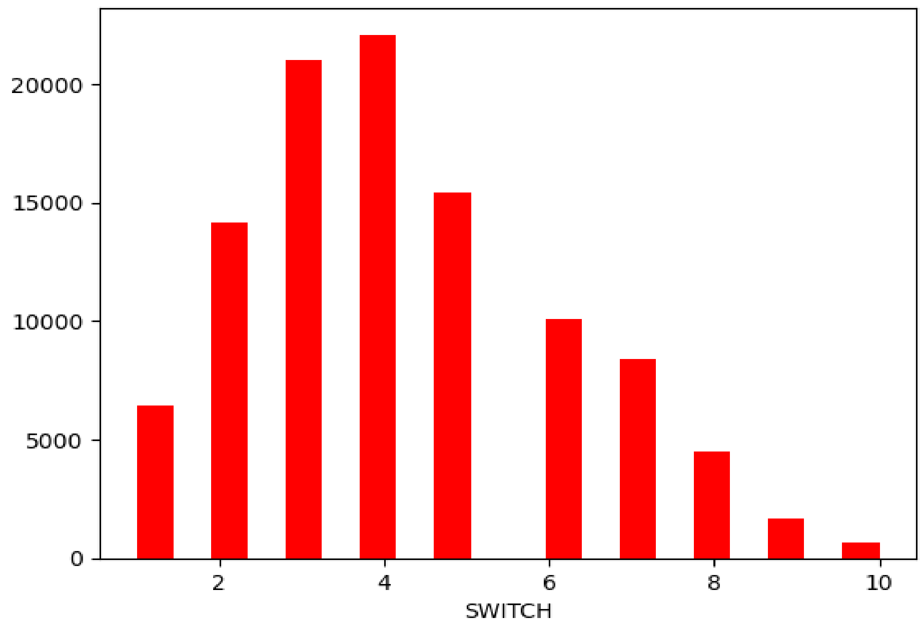


Figure 10: Switch.

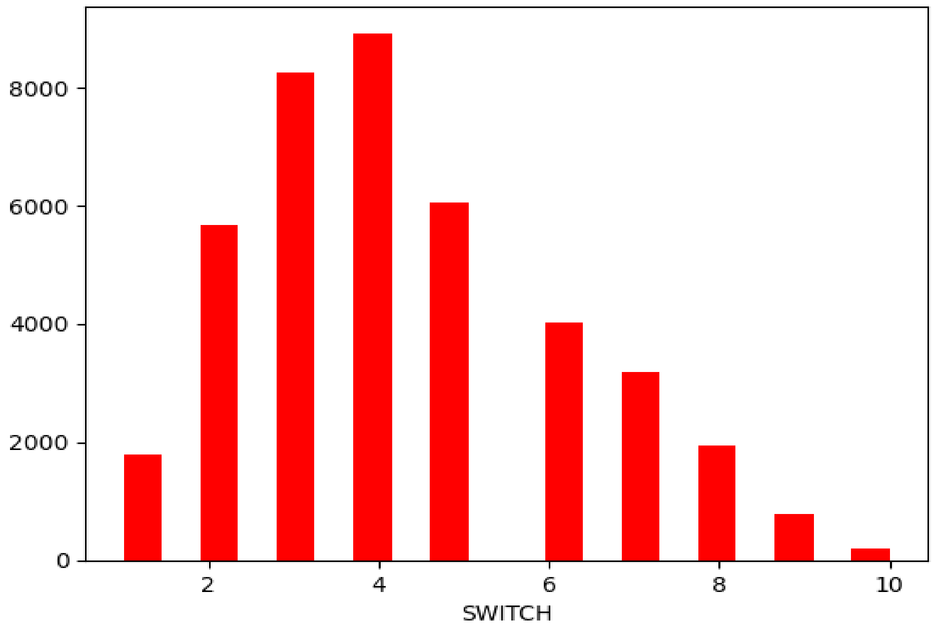


Figure 11: Switch.

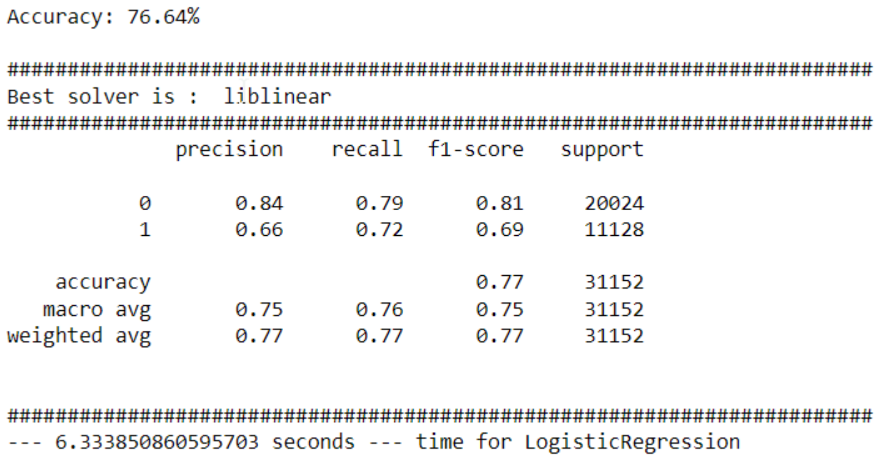
Chapter 4

Results

4.1 Introduction

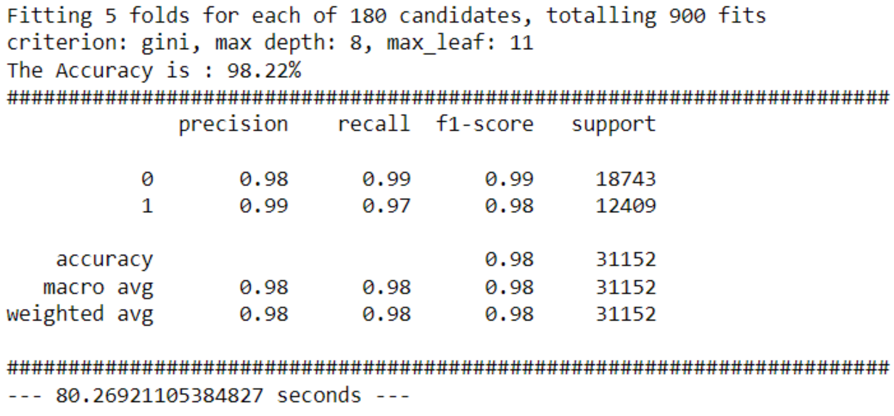
Here, we apply four machine learning models to detect the percentage of attacks. Which is logistic regression, Decision Tree, Naïve Bayes, Random Forest Classification. In this chapter we will discuss the dataset, show the results and also, we will analysis outcome of the operations.

4.2 Result



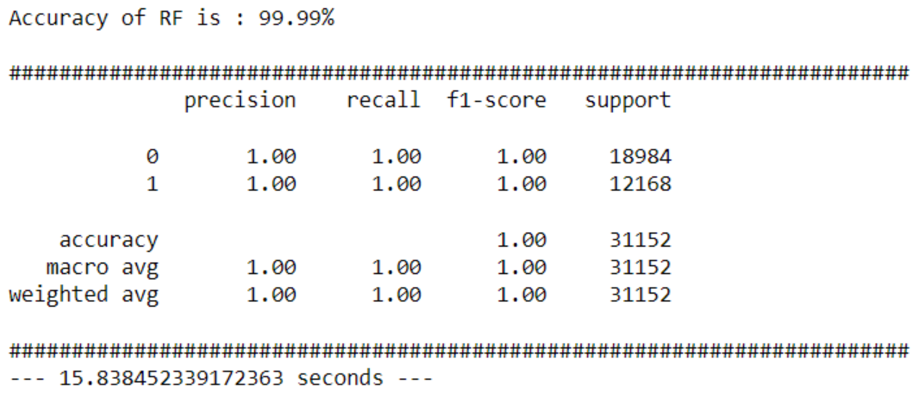
```
Accuracy: 76.64%

##########################################################################
Best solver is :  liblinear
##########################################################################
              precision    recall  f1-score   support

           0       0.84      0.79      0.81     20024
           1       0.66      0.72      0.69     11128

    accuracy                           0.77     31152
   macro avg       0.75      0.76      0.75     31152
weighted avg       0.77      0.77      0.77     31152


##########################################################################
--- 6.333850860595703 seconds --- time for LogisticRegression
```

Figure 12: Result for Logistic Regression.



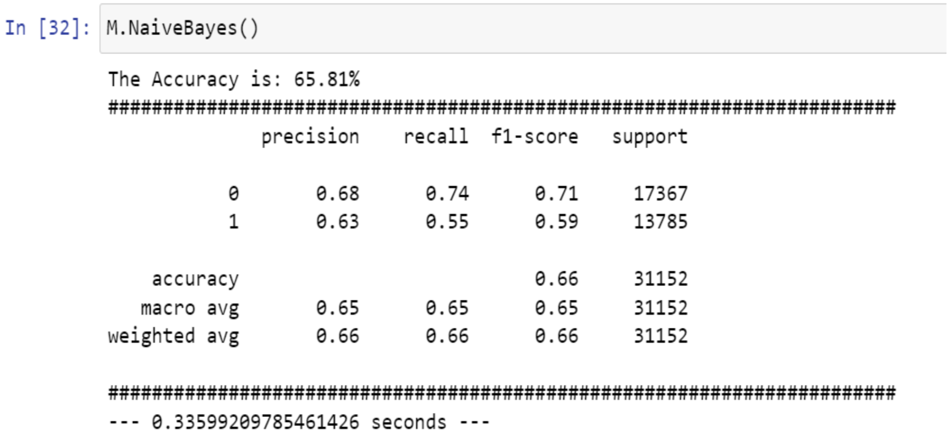
```
Fitting 5 folds for each of 180 candidates, totalling 900 fits
criterion: gini, max depth: 8, max_leaf: 11
The Accuracy is : 98.22%
###########################################################################
              precision    recall  f1-score   support

           0       0.98      0.99      0.99     18743
           1       0.99      0.97      0.98     12409

    accuracy                           0.98     31152
   macro avg       0.98      0.98      0.98     31152
weighted avg       0.98      0.98      0.98     31152

###########################################################################
--- 80.26921105384827 seconds ---
```

Figure 13: Result for Decision Tree.



```
Accuracy of RF is : 99.99%

########################################################################
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     18984
           1       1.00      1.00      1.00     12168

    accuracy                           1.00     31152
   macro avg       1.00      1.00      1.00     31152
weighted avg       1.00      1.00      1.00     31152

########################################################################
--- 15.838452339172363 seconds ---
```

Figure 14: result for Random Forest Classification.



```
In [32]: M.NaiveBayes()

The Accuracy is: 65.81%
########################################################################
              precision    recall  f1-score   support

           0       0.68      0.74      0.71     17367
           1       0.63      0.55      0.59     13785

    accuracy                           0.66     31152
   macro avg       0.65      0.65      0.65     31152
weighted avg       0.66      0.66      0.66     31152

########################################################################
--- 0.33599209785461426 seconds ---
```

Figure 15: result for naive Bayes.

4.3 Outcome Analysis

Based on the findings from the models tested, it appears that the random forest and decision tree models perform better than the other two models in accurately detecting DDoS attacks. The random forest model demonstrates the highest accuracy of 99.99%, indicating its high effectiveness in detecting DDoS attacks. This could be attributed to its ability to handle large amounts of data and detect intricate relationships between variables. The decision tree model also performs well with an accuracy of 98.22%, as decision trees are typically good at identifying anomalies in data. However, the logistic regression and naive Bayes models show lower accuracies of 76.64% and 65.81%, respectively. Logistic regression assumes a linear relationship between input variables and the output, which may not be suitable for detecting DDoS attacks. Naive Bayes may not be as effective in identifying DDoS attacks due to the complex and high-dimensional nature of the data. Overall, the random forest and decision tree models seem to be the most effective in detecting DDoS attacks based on the provided results.

Chapter 5

Discussion and Conclusion

Our thesis on DDOS attack detection using Machine Learning models and comparing the results of four different models, namely Random Forest, Logistic Regression, Decision Tree, and Naive Bayes, is a significant contribution to the field of network security. The results that we have obtained from our analysis are impressive, with Random Forest performing the best with a result of 99.99%, followed by Decision Tree at 98.22%. Logistic Regression and Naive Bayes, while not performing as well, still achieved respectable results of 76.64% and 65.81%, respectively.

The dataset we have used for our analysis, the SDN Dataset, is a valuable resource for researchers and practitioners working in the field of network security. It offers a complete set of features that can be tested and utilized to train machine learning models for DDOS attack detection. Our thesis shows how machine learning models can effectively identify DDOS assaults and emphasizes the significance of choosing the right models to achieve high accuracy rates.

Our findings may have important repercussions for the creation of network security measures that are more effective. The adoption of Machine Learning models could be a crucial tool for network security professionals to detect and prevent such attacks given the rising frequency and sophistication of cyber-attacks. Our study establishes the foundation for further research in this area and could have an impact on the development of a system for more accurate and efficient DDOS attack detection.

In conclusion, the goal of this thesis was to examine how well various machine learning algorithms can identify DDoS attacks. The four algorithms that have been chosen are Logistic Regression, Decision Tree, Random Forest, and Naive Bayes. Based on their accuracy, precision, memory, and F1 score, they were assessed and compared.

The machine learning algorithms have been applied to a training dataset in order to conduct experiments and discover the most important SDN dataset elements that aid in the precise identification of assaults by machine learning algorithms.

The results of the studies demonstrate that Random Forest, followed by Decision Tree, Logistic Regression, and Naive Bayes, attained the highest accuracy. This suggests that the best algorithm for identifying DDoS attacks in a network context is Random Forest.

The study also highlights the importance of selecting appropriate model and engineering to improve the performance of machine learning algorithms in DDoS detection. It also emphasized the need for further research in this area to address the challenges of DDoS attacks. Therefore, further research can be conducted to explore different Models and methods to enhance the accuracy and efficiency of the DDoS attack detection system.

Overall, this study provides valuable insights into the application of machine learning algorithms in DDoS detection, which can be useful for network administrators and security professionals in improving the effectiveness and efficiency of their defense mechanisms against such attacks.

Bibliography

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning with Applications in R Second Edition,” 2021.

[2] D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, “Software-defined networking: A comprehensive survey,” *Proceedings of the IEEE*, vol. 103, no. 1, 2015, doi: 10.1109/JPROC.2014.2371999.

[3] J. F. Kurose and K. W. Ross, *Computer networking : a top-down approach*.

[4] S. Pande, A. Khamparia, D. Gupta, and D. N. H. Thanh, “DDOS Detection Using Machine Learning Technique,” in *Studies in Computational Intelligence*, 2021. doi: 10.1007/978-981-15-8469-5_5.

[5] S. Wankhede and D. Kshirsagar, “DoS Attack Detection Using Machine Learning and Neural Network,” in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 2018. doi: 10.1109/ICCUBEA.2018.8697702.

[6] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, “Evaluation of machine learning algorithms for intrusion detection system,” in *SISY 2017 - IEEE 15th International Symposium on Intelligent Systems and Informatics, Proceedings*, 2017. doi: 10.1109/SISY.2017.8080566.

[7] G. Ajeetha and G. Madhu Priya, “Machine Learning Based DDoS Attack Detection,” in *2019 Innovations in Power and Advanced Computing Technologies, i-PACT 2019*, 2019. doi: 10.1109/i-PACT44901.2019.8959961.

[8] M. Aamir and S. M. Ali Zaidi, “Clustering based semi-supervised machine learning for DDoS attack classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, 2021, doi: 10.1016/j.jksuci.2019.02.003.

[9] M. Idhammad, K. Afdel, and M. Belouch, “Semi-supervised machine learning approach for DDoS detection,” *Applied Intelligence*, vol. 48, no. 10, 2018, doi: 10.1007/s10489-018-1141-2.

[10] M. Aamir and S. M. A. Zaidi, “DDoS attack detection with feature engineering and machine learning: the framework and performance evaluation,” *Int J Inf Secur*, vol. 18, no. 6, 2019, doi: 10.1007/s10207-019-00434-1.

[11] 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,” *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 6, 2016, doi: 10.1007/s13042-014-0309-2.

[12] K. Lakshmi Narayanan, R. Santhana Krishnan, E. Golden Julie, Y. Harold Robinson, and V. Shanmuganathan, “Machine Learning Based Detection and a Novel EC-BRTT Algorithm Based Prevention of DoS Attacks in Wireless Sensor Networks,” *Wirel Pers Commun*, 2021, doi: 10.1007/s11277-021-08277-7.

[13] M. Aditya Vikram, “Anomaly detection in network traffic using unsupervised machine learning approach,” in *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020*, 2020. doi: 10.1109/ICCES48766.2020.09137987.

[14] A. Banitalebi Dehkordi, M. R. Soltanaghaei, and F. Z. Boroujeni, “The DDoS attacks detection through machine learning and statistical methods in SDN,” *Journal of Supercomputing*, vol. 77, no. 3, 2021, doi: 10.1007/s11227-020-03323-w.

[15] T. A. Tuan, H. V. Long, L. H. Son, R. Kumar, I. Priyadarshini, and N. T. K. Son, “Performance evaluation of Botnet DDoS attack detection using machine learning,” *Evol Intell*, vol. 13, no. 2, 2020, doi: 10.1007/s12065-019-00310-w.

[16] Y. W. Chen, J. P. Sheu, Y. C. Kuo, and N. van Cuong, “Design and implementation of IoT DDoS attacks detection system based on machine learning,” in *2020 European Conference on Networks and Communications, EuCNC 2020*, 2020. doi: 10.1109/EuCNC48522.2020.9200909.

[17] S. Sambangi and L. Gondi, “A Machine Learning Approach for DDoS (Distributed Denial of Service)

Attack Detection Using Multiple Linear Regression," 2020. doi: 10.3390/proceedings2020063051.

[18] M. Zekri, S. el Kafhali, N. Aboutabit, and Y. Saadi, "DDoS attack detection using machine learning techniques in cloud computing environments," in *Proceedings of 2017 International Conference of Cloud Computing Technologies and Applications, CloudTech 2017*, 2018. doi: 10.1109/CloudTech.2017.8284731.

[19] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred, "Statistical approaches to DDoS attack detection and response," in *Proceedings - DARPA Information Survivability Conference and Exposition, DISCEX 2003*, 2003. doi: 10.1109/DISCEX.2003.1194894.

[20] N. Jyoti and S. Behal, "A meta-evaluation of machine learning techniques for detection of DDoS attacks," in *Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom 2021*, 2021. doi: 10.1109/INDIACom51348.2021.00093.

[21] M. Suresh and R. Anitha, "Evaluating machine learning algorithms for detecting DDoS attacks," in *Communications in Computer and Information Science*, 2011. doi: 10.1007/978-3-642-22540-6_42.

[22] Q. He, X. Meng, R. Qu, and R. Xi, "Machine learning-based detection for cyber security attacks on connected and autonomous vehicles," *Mathematics*, vol. 8, no. 8, 2020, doi: 10.3390/MATH8081311.

[23] N. Ahuja, G. Singal, D. Mukhopadhyay, and N. Kumar, "Automated DDOS attack detection in software defined networking," *Journal of Network and Computer Applications*, vol. 187, 2021, doi: 10.1016/j.jnca.2021.103108.

Appendix

Appendix A

Random Forest Algorithm: Random Forest is a well-liked machine learning algorithm for classification, regression, and other tasks. It is an ensemble learning technique that, during the training phase, creates a large number of decision trees. The technique then produces a class that represents the average of the classes (classification) or average prediction (regression) made by the individual trees.

Random forests create multiple decision trees from subsets of the training data at random, then combine the outcomes to make predictions. As a result, overfitting is avoided and the variance of the model is reduced. The formula for random forest can be broken down into several steps:

1. Randomly select a subset of the training data.
2. Randomly select a subset of the features.
3. Build a decision tree on the selected subset of data and features.
4. Repeat steps 1-3 a specified number of times to create a forest of decision trees.
5. Make a mean prediction (regression) of the class that corresponds to the mode of the classes in the classification or the classification of the individual trees.

Code: Prediction = mode (predict (tree_i, x))

where 'tree_i' is the i-th decision tree in the forest, 'x' is the input feature vector for the new data point, and 'mode' is the function that returns the most common predicted class among all the decision trees.

Combining the predictions from each decision tree in the forest yields the final prediction. This can be achieved by using the mean prediction for regression problems or the mode of the class for classification problems.

References:

- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

- https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Decision Tree: In decision analysis, a decision tree is a tree-like structure that is used to model decisions and potential outcomes. It is a graphic depiction of a decision-making process that lists all potential outcomes of a choice and their probabilities.

Each node in the decision tree represents a decision or an event, and each branch in the decision tree represents one or more potential outcomes of that decision or event. A decision or an event that results in a new set of potential outcomes is represented by each node after the root node in the decision tree.

The formula for calculating the expected value of a decision tree is:

Expected value = (Probability of outcome 1 * Value of outcome 1) + (Probability of outcome 2 * Value of outcome 2) + ... + (Probability of outcome n * Value of outcome n)

Where:

- Probability of outcome i: The probability of a specific outcome occurring
- Value of outcome i: The value or payoff associated with a specific outcome

The decision tree is a commonly used tool in decision analysis, game theory, and finance. It can be used to analyze complex decisions, evaluate risk, and identify the optimal course of action.

Reference:

- Decision tree (2022). In Wikipedia. Retrieved February 28, 2023, from https://en.wikipedia.org/wiki/Decision_tree

Naive Bayes: The Naive Bayes algorithm, which is based on Bayes' theorem, is a straightforward and efficient probabilistic method for classification tasks. Assuming that the features are independent of one another, which is frequently false in datasets from the real world, is what gives the algorithm its "naive" moniker. Naive Bayes has been demonstrated to be effective in a variety of classification tasks, despite this simplification assumption.

A straightforward probabilistic classifier based on the Bayes theorem and the assumption of feature independence is called naive Bayes. It is widely used in machine learning and natural language processing for classification tasks.

The formula for Naive Bayes classification is:

P(y|x1, x2, ..., xn) = P(y) * P(x1|y) * P(x2|y) * ... * P(xn|y) / P(x1, x2, ..., xn)

Where:

- P(y|x1, x2, ..., xn) is the probability of class y given the feature vector (x1, x2, ..., xn)
- P(y) is the prior probability of class y

- P(xi|y) is the probability of feature xi given class y
- P(x1, x2, ..., xn) is the probability of the feature vector (x1, x2, ..., xn)

The Naive Bayes classifier assumes that the features are conditionally independent given the class, i.e., P(x1, x2, ..., xn|y) = P(x1|y) * P(x2|y) * ... * P(xn|y). This assumption simplifies the calculation of the probabilities and makes the classifier computationally efficient.

Reference:

- Friedman, N., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Logistic Regression: A statistical technique known as "logistic regression" is used to simulate the relationship between a categorical dependent variable (commonly referred to as the "outcome" or "response" variable) and one or more independent variables (often called "predictors" or "covariates"). In order to forecast the likelihood of an event occurring given a set of predictor variables, logistic regression is used.

The formula for logistic regression is as follows:

$p = 1 / (1 + e^{(-z)})$

where:

- p is the probability of the event occurring
- e is the base of the natural logarithm (approximately 2.71828)
- z is the linear combination of the predictor variables and their coefficients, calculated as follows:

$z = b0 + b1x1 + b2x2 + ... + bnxn$

where:

- b0 is the intercept or constant term
- b1, b2, ..., bn are the coefficients for the predictor variables x1, x2, ..., xn

The logistic regression model uses maximum likelihood estimation or another appropriate method to infer the values of the coefficients b0, b1, b2,..., bn from the data. Once the coefficients are estimated, the model can be used to forecast the likelihood that an event will occur for brand-new observations with various values for the predictor variables.

Reference:

- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Son.

Code

G.1 Code Link:

https://github.com/mohammedibrahimshawon/DDOS_Attack_models

======================================

Approval

This thesis titled "DDOS Attack Detection Using Machine Learning Algorithm" has been submitted to the following respected member of the board of examiners of the Faculty of Science and Technology impartial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering on dated and has been accepted satisfactory.

SUPTA RICHARD PHILIP
Supervisor
Lecturer, Faculty
Department of Computer Science
American International University-Bangladesh

Md. MASUM BILLAH
External
Assistant Professor, Faculty
Department of Computer Science
American International University-Bangladesh

DR. AKINUL ISLAM JONY
Associate Professor & Head (UG)
Department of Computer Science
American International University-Bangladesh

DR. MD. ABDULLAH - AL - JUBAIR
Assistant Professor & Director
Faculty of Science and Technology
American International University-Bangladesh

PROF. DR. DIP NANDI
Professor & Associate Dean
Faculty of Science and Technology
American International University-Bangladesh

MASHIOUR RAHMAN
Sr. Associate Professor & Dean-in-charge
Faculty of Science and Technology
American International University-Bangladesh

Declaration

We certify that this thesis is our original work and has not been submitted in any way to another university or other tertiary educational institution for a different degree or diploma. A list of references is provided, and information taken from both published and unpublished works of others is acknowledged in the text.

We acknowledge that the owner(s) of the copyright for all of the material in my thesis are responsible for maintaining that material's integrity. For the purposes of reproducing material in this thesis, we have, when necessary, received permission from the copyright holder and, in the case of any jointly authored works, we have asked the co-authors for their consent.

Ibrahim Shawon

SHAWON, MOHAMMED IBRAHIM
19-40767-2
Department of Computer Science

Sakib

ALI, MD. SAKIB
19-40426-1
Department of Computer Science

Mukta

MUKTA, MAHMUDA AKHTAR
19-40440-1
Department of Computer Science

Rafsan

JANIN, MD. RAFSAN
19-40786-2
Department of Computer Science