REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

DATA MINING AND MACHINE LEANING METHODS FOR CYBER SECURITY

Mohammad Saeed Yasein BAKAR

Master's Thesis

Supervisor

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, 2022

DATA MINING AND MACHINE LEANING METHODS FOR CYBER SECURITY

Mohammad Saeed Yasein BAKAR

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2022

The thesis titled DATA MINING AND MACHINE LEANING METHODS FOR CYBER SECURITY prepared by MOHAMMED SAEED YASEIN BAKAR and submitted on 20/12/2022 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

Asst. Prof. Dr. Sefer KURNAZ

the Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Sefer KURNAZ	Department of Computer Engineering, Altınbaş University	_____
Asst. Prof. Dr. Oğuz KARAN	Department of Software Engineering, Altınbaş University	_____
Assoc. Prof. Dr. Adil Deniz DURU	Department of Trainer Education Marmara University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.
Submission data of the thesis to Institute of Graduate Studies: ____/____/____

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Mohammad Saeed Yasein BAKAR

Signature

ABSTRACT

DATA MINING AND MACHINE LEANING METHODS FOR CYBER SECURITY

BAKAR, Mohammad Saeed Yasein

M.Sc., Information Technologies, Altınbaş University

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: 12/2022

Pages: 61

There is drastic increase in needs of networking and data sharing in today's world. Such globalization of increased information technology and development there exists need of network security. Firewalls may provide some level of security but they never alert administrator for upcoming attacks. In order to find such abnormal behavior of network packets there is need of reliable detection system for improvement of efficiency and accuracy. Many research works focused on machine learning approach for enhancing the efficiency of the intrusion detection system and to detect malicious network activity automatically on the basis of network packet behaviors. The proposed model is designed using machine learning approach for detection of malicious activities of the network packets. For that CICDDoS2019 dataset is used. In this research, we proposed statistical methods are used obtain Z-scores, mean, median and mode for determining the significant features, training was performed for 80% of the dataset. The remaining 20% dataset was tested and validation using decision tree, K-NN and Gradient boosting algorithm.

Keywords: Machine Learning, IDS, Cyber Attacks.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

IDS : Intrusion detection systems

KNN : K-Nearest Neighbors

SVM : Support Vector Machines

IPS : Intrusion Prevention Systems

HIDS : Host-based Intrusion Detection Systems

YUV : Luma and Chroma Components

1. INTRODUCTION

1.1 OVERVIEW

Forestalling unapproved admittance to PC frameworks is a main issue in PC framework security. To forestall such unapproved admittance to the framework, a recognition and counteraction framework is required. Such dubious and unapproved clients are ordinarily alluded to as "Interlopers." They are banned from getting to any piece of the PC framework. The acknowledgment cycle is utilized to decide whether a couple of individuals will endeavor to invade the objective framework, on the off chance that they are effective, and to find the movement logs. [1]. Outsiders are probably not going to understand messages, use PC frameworks to assault or upset different frameworks, send counterfeit messages or messages from a PC framework, or check individual data on your PC framework that contains data, for example, fiscal reports, account subtleties, etc. Intruders may likewise be alluded to as aggressors, saltines, or programmers. You probably won't be keen on realizing who claims the objective framework. They have consistently assumed command over the PC framework to send off assaults on different PCs. Normally, aggressors deal with target frameworks, like government or monetary frameworks, permitting them to disguise their actual area and in this way effectively send off assaults.

When a PC framework is associated with the Web, it has no effect whether it plays out a couple of mystery exercises or just messes around while talking with companions; the framework is likewise focused on. Gatecrashers might have the option to deal with each of our exercises on our framework. It is feasible to think twice about data, reformat the hard plate, and cause other harm. Sadly, interlopers continually find new weaknesses, otherwise called "escape clauses" [2, 3], making it challenging to safeguard the framework. These blemishes should be taken advantage of on the PC framework or framework programming. It is challenging to painstakingly test the security of PC frameworks because of the trouble in programming. It is the client's liability not exclusively to acquire and introduce record fixes, yet in addition to arrange the product for more secure activity. There are likewise programming applications with predefined custom settings that permit different clients to get to the PC framework except if the settings are changed to make the framework safer. Visit programs, for instance, permit outside clients to execute orders on their PC

frameworks, permitting a couple of individuals to carry out disastrous projects that run when the client is clicked. This would undoubtedly keep an outsider from inspecting delicate reports [4].

Essentially, it very well might be suitable to keep PC framework assignments classified, whether they are observing our reports or running different applications. Moreover, clients should guarantee that the data entered on the PC framework is flawless and available when required. Purposeful maltreatment of our PC framework by Web gatecrashers could bring about security breaks [5]. Regardless of whether clients are not associated with the Web, they might experience extra dangers like hard circle mistakes, robbery, blackouts, etc. The awful news is that it probably won't occur as expected. Fortunately there are a couple of normal estimates that can be taken to lessen the probability of being impacted by the most well-known dangers. Not many of these means help in the administration of both deliberate and unexpected dangers. Before we realize how we might safeguard our PC framework or home organization, we should check out at a couple of these dangers.

1.2 INTRUSION

Gatecrashers are exercises that endeavor to keep away from the security parts of information systems in an uncertain way. This is a progression of measures that risk information decency, openness, as well as characterization. Security infers that information ought not be revealed to outcasts who are not approved to get to it. Straightforwardness guarantees that the message has not been modified during the trade. An unapproved client changes the substance of the message when a client establishes a connection with another client however before it arrives at the planned recipient. This is known as loss of decency, and it happens because of changes [6]. The openness capacity decides if resources ought to continuously be available to approved clients. Attacks, for example, hindrances lessen resource availability. Table 1.1 portrays the different kinds of assaults that happen on the association. Intruders are habitually brought about by an interloper accessing the system through the Internet, the local association, or the functioning game plan of the sullied machine, or by taking advantage of the shortcoming of an outcast application (middleware), or by aggressors endeavoring to deny explicit endorsed clients. benefit and wellbeing system respects and abuse [7, 8].

Table 1.1: Attack Kinds with Description [6]

Attacks Category	Description	TCP/IP Layer
DoS	Denial-of-service (fake address generate)	Application Layer
DoS	Denial-of-service (fake address generate)	Transport Layer
U2R	Unauthorized admittance to local super user (root) privileges	Application Layer
R2L	Unauthorized admittance from a remote machine	Application Layer
R2L	Unauthorized admittance from a remote machine	Transport Layer
Probe	Surveillance and other probing	Application Layer
Probe	Surveillance and other probing	Transport Layer

1.3 INTRUSION DETECTION

After an assault, IDS recognizes pernicious movement in PC frameworks and performs criminology. Inspect network assets for interruptions and assaults that poor person been come by preventive measures (firewall, switch bundle sifting, intermediary server). An interference is an endeavor to reevaluate the mystery, dependability, or openness of the structure. As a rule, interference area structures are equivalent to real intruder finders. Misuse based interference recognition structures (IDS) (as displayed in Figure 1.1) are expected to distinguish infringement of predefined security plans. Regardless, things become more muddled with the introduction of possibly dangerous ways of behaving that can't be anticipated [9-11]. A model is a fashioner a lot of in an association data in a brief timeframe. This could be an issue with data spillage, however the authorization methodology may not remember it since record moves are allowed [12]. Hence, verifiable characteristic acknowledgment has been introduced, which includes making a profile of a client or system and representing deviations from the profile. While the two sorts of structures are important all alone, joining them can lighten however not take out the shortcomings of each [13-15]. The wellspring of survey data is a significant component that recognizes the sort of

execution acquired from IDS. The two essential sources are have-based shows utilized by have-based IDSs and network data packs utilized by network IDSs. Bit logs, application logs, and gadget explicit logs are instances of host shows [11].

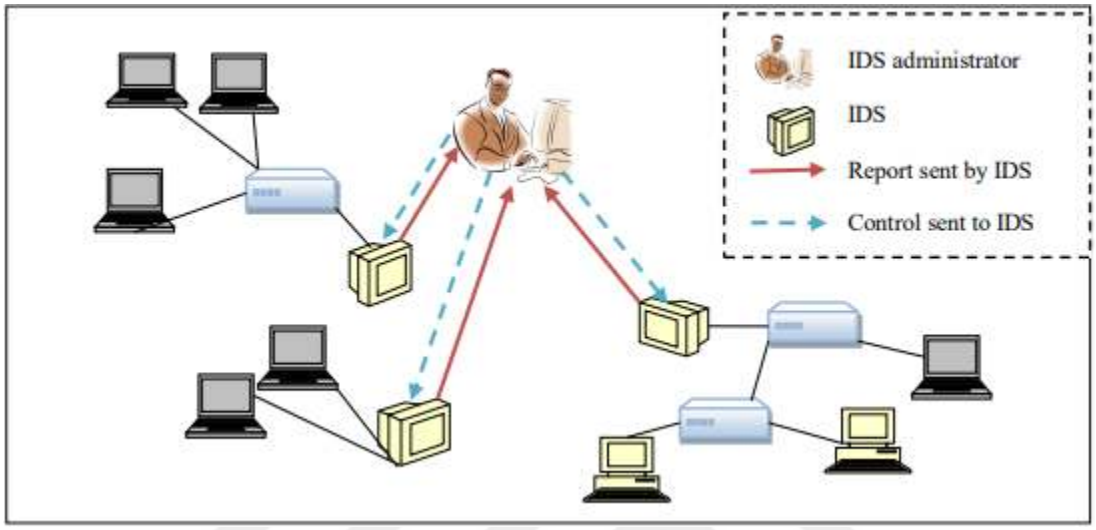

Figure 1.1: Intrusion Detection System

There are several problems with IDS based on host and network IDS. They include:

i. The count of framework explicit discovery boundaries for any framework is very lengthy because of heterogeneous working frameworks..

ii. Increasing the number of critical nodes in the network increases performance.

iii. Diminished have framework execution because of extra security exercises, like B. Enrollment.

iv. Difficulty in detecting attacks at the network level.

v. Deficient host handling ability to give a full host-based IDS. Network-based interference area structures, then again, can utilize a focal system with an association relationship to inactively screen network traffic. When presented at the association's edge, they altogether affect structure execution and can undoubtedly recognize network-level attacks. The execution of association based ID is extremely straightforward [16-19]. Have based IDs ought to be meticulously picked in an essential execution sensitive host association with the goal that each structure's show isn't unduly hampered..

Intrusion Detection Systems are designed to monitor the network at run time and to observe any abnormal or suspicious activities and to generate an alarm as well as creating prevention method [20-23]. Many intrusion detection systems are configured for known attacks or activities. So, there is requirement of daily updation in the detection system to deal with future attacks [7].

Intrusion detection system is categorized into several kinds which are discussed below (as shown in Figure 1.2):

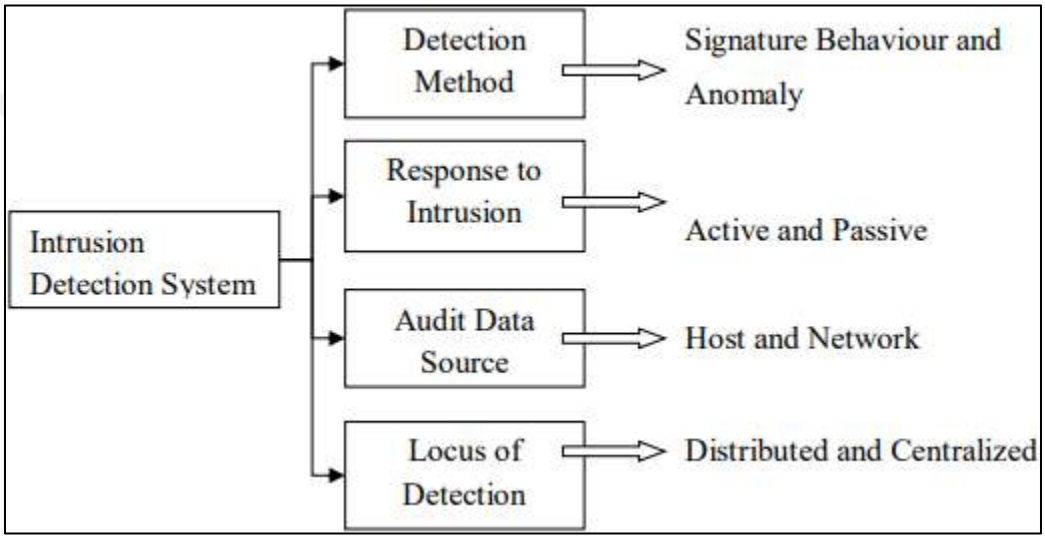

Figure 1.2: Kinds of intrusion detection system

To idly screen network traffic, network-based interference acknowledgment structures will have a central system with an association relationship. They significantly affect system execution and can basically identify network-level exercises once positioned at the association's edge. Network-based ID execution is excessively oversimplified [24-26]. Hostbased IDs in a really significant execution delicate host organization ought to be thoroughly assigned in this way as not to restrict the exhibition of every framework unduly.

1.4 MOTIVATION

The work present in the literature deals with only binary classification of the DDoS attacks. The study by [10] used Bro and Corsaro open-source systems and Classification and Regression Trees(CART) and Naive Bayes Ma- chine learning classifiers to detect DDoS attacks. They had the option to accomplish elite execution without utilizing IP locations or port numbers by using

these AI calculations. On the preparation dataset, they got a precision score of 0.99, a review worth of 1.00, and a F1-Score of 0.97 utilizing Choice Tree, and an exactness score of 0.98, a review worth of 0.93, and a F1-Score of 0.82 utilizing Guileless Bayes. Be that as it may, the calculations created were intensely centered around identifying backscatter traffic and deciding if there was a DDoS assault without respect for the kind of assault.

[11] utilized a Multi-facet Perceptron organization to effectively characterize typical and assault states (twofold) or DDos source, DDos casualty, or normal(3 sorts) of DDoS type assault with a high obvious positive and low misleading positive rate, zeroing in principally on UDP assaults. To settle on a dependable choice, they utilized a blend of measurements that were basically correlative inactive, for example, parcel catching and Netflow-based traffic observing, as well as ROC bends. [12] utilized Multi-facet Perceptron to further develop twofold DDoS location rate with 99.9971% exactness. The review utilized a mixture strategy to choose highlights utilizing the covering Hereditary Calculation (GA) and MLP to characterize the assault..

[13] employs k-Nearest Neighbour (K-NN) in his research to categorize network status into Normal, pre-attack, and attack DDoS class status in order to predict DDoS attacks. It used the packet types(UDP, ICMP, TCP, SYN), source/destination IP addresses and port numbers as features and uses a selection of handlers and agents, communication and compromise to classify the packets into two phases of pre-attack, 3rd phase of Attack and Fourth Normal phase. Other algorithms such as Naive Bayesian, C4.5 and K-Means have been explored to achieve an accuracy of 91.4%, 98.8% and 85.9% respectively in the paper by [14] to binary classify DoS attack majorly targeting layer 3 and layer 4 of OSI 7 layer model.

Enhanced Multi-Class Support Vector (EMCSVM) is used to operate on KD- DCup 99 dataset in [15]. The data set used in the study contains only six types of DDoS attacks hence the classification was only limited to six types. In addition, Linear, Polynomial, Radial Bias kernel functions are used with radial bias giving the best accuracy among the other kernel functions.

To diminish blunders and further develop identification exactness, [17] utilize Hereditary fluffy and Neuro-fluffy techniques as a subsystem of the group. At long last, the NFBoost calculation is proposed to group DDoS assaults as typical (no assault) or assault. The two measurements used to evaluate its presentation were cost per test and location exactness of 99.2%.

[4] highlights the performance of various algorithms like Back Propagation(BP), PCA-BP, Long Short Term Memory(LSTM), PCA-LSTM, PCA- Support Vector Machines(SVM), Recurrent Neural Network (RNN) and Principal Component Analysis-Recurrent Neural Network (PCA-RNN) on basis of accuracy, sensitivity, precision, F1 and prediction time. It concluded that the PCA-RNN method has higher detection efficiency and accuracy using KDD-99

[18] dataset. To understand the network fully, maximum network traffic characteristics were considered, and then PCA was applied for dimensionality reduction to decrease time complexity. However, the dataset used is mostly suitable to detect the wider range of attacks and not particularly DDoS attack [19] because of which the proposed method is only suitable in case of binary classification of DDoS attack.

In addition, [21] explored J48, RF, SVM and KNN to binary classify DDoS attacks and they concluded that J48 is the best classifier in terms of accuracy, sensitivity, specificity, precision, recall and F1-score. The C4.5 calculation was created in [22] to recognize DDoS assaults by framing a choice tree to identify these assaults' marks really. When contrasted with an AI calculation, C4.5 has a precision of 98.8% for accurately grouping a danger as an assault or typical. In any case, past models could foresee whether an assault would happen. They do not detect the types of attacks that are critical in Cybersecurity to effectively deploy countermeasures which is the major distinguishing factor in our study.

1.5 PROBLEM STATEMENT

It is known that the IDS suffers from a high number of false alarms based on anomalies. Efforts are underway to reduce the high rate of false positives. We believe that intrusion detection are a process of data analysis and can be studied as a problem of proper data classification. From this point of view, one can also see that every classification system is as valid as the data presented. The better the data are, the better the results a toore. From the point of view of an IDS based on anomalies, this implies that the rate of false positives can be significantly reduced if we extract characteristics that correctly distinguish normal data from abnormal data. Likewise, we find that most intrusion detection methods based on data mining and machine learning use well-known techniques and tools.

These overall strategies may not be exceptionally powerful in precisely characterizing information as typical or strange. There is a need to fit those strategies to the requirements of interruption location. Aside from the issues referenced over, one of the fundamental worries is the quick location of assaults. With the ongoing degree of intricacy and assortment of assaults, we require a monstrous measure of information to dissect and deliver results. Notwithstanding, the more information there is, the more it takes to break down it, postponing the location of assaults. An IDS will be more valuable in the event that it can produce a caution sufficiently early to moderate the harm that a continuous assault can cause. Thus, there are a need to make IDS as fast as to operate on-line. It is believed that this can be achieved if we can reduce the data, to be analyzed, without degrading its quality.

Up to this point, IDs have had a low degree of flawlessness. There are various weaknesses related with IDS. As these SDIs advance, a few holes are regularly filled by improving and handling existing methods; nonetheless, a couple of these holes are innate in the manner SDIs are planned. Coming up next are the most often experienced imperfections:

i. The construction of Intrusion Detection System (IDS) for computer networks is extremely needed. The work focuses on anomaly based IDS and thus makes a major contribution in computer network society.

ii. Anomaly based IDS is well known to endure from a high rate of false alarms. Continuous efforts are made to bring down the high false negative rate.

iii. We accept that interruption recognition is an information investigation technique that can be considered to be an issue of accurately ordering information. From this vantage point, any grouping plan is just pretty much as great as the information introduced to it as information.

iv. The more spotless the information, the more solid the outcomes are probably going to be. From the point of view of irregularity based interruption identification frameworks, this actually intends that in the event that we can appropriately separate elements that recognize ordinary information from unusual information, the misleading negative rate can be significantly decreased.

v. In this research, we inspect the techniques that support the isolation of normal data from an abnormal data.

vi. We find that most of the intrusion detection approaches mainly based on data mining and machine learning techniques, which make the tools more efficient. This makes these general methods with very high accuracy but classifying data as normal or abnormal is inaccurate.

vii. Machine learning techniques need to be designed according to the intrusion detection requirements. We are also focusing our research work on this point.

In order of achieve objectives this research work uses the machine learning approaches for data reduction, clustering and classification. For data reduction normalization technique are used further features are reduced using co-relation algorithm. Further k-mean clustering is used to reduce the instances of dataset in all the five categories such as DOS, Probe, R2L, U2R and normal. And finally, multilevel classifier based is used to achieve highest accuracy and detection rate

1.6 CONTRIBUTION

Huge server farms are putting away and communicating a rising measure of information. An interruption discovery framework is utilized to guarantee that organization traffic contains no interruptions. Such a framework dissects network information and issues a caution in the event that it identifies an interruption. Since server farms produce such a lot of information, handling everything is troublesome. This is where IP streams enter the image. They are assembled from parcel information yet need data about the payload information. This proposition depicts which assaults can be recognized and the way that IP streams can be utilized to distinguish interruptions. It would likewise be more savvy in the event that an interruption location framework could work consequently and identify assaults with a serious level of sureness. AI can be utilized for this. AI is a subset of man-made brainpower that empowers projects to learn and find designs in information. There are, notwithstanding, various sorts of AI. This proposal gives an outline of various AI calculations, for example, Backing Vector Machines and K-Closest Neighbors, as well as a prologue to AI ideas. A clarification of how these calculations can be utilized in an interruption recognition framework is given. The calculations are tried on different datasets that incorporate both named preparing information and unlabeled certifiable information. Expectations to learn and

adapt and F-scores are utilized in the assessment. In the assessment, it was found that regulated learning produces preferred and more nitty gritty forecasts over solo learning. Among the tried directed learning calculations, K-Closest Neighbors delivers the best outcomes. The discoveries demonstrate that AI is a feasible choice for distinguishing interruptions through IP streams. It ought to likewise be noticed that utilizing additional data, for example, TCP banners is a valuable option that essentially further develops execution.

1.7 THESIS OUTLINE

This thesis includes six chapter inclusive of this chapter comprising the basic introduction of this research work, highlighting preliminaries associated with this research. Rest of the thesis is oriented as follows:

The subsequent part presents a complete outline of the past writing study on different woods fire recognition frameworks, exhibiting cutting edge work in this subject. It additionally talks about the particulars of the distinguished review holes in the current writing, just as the exploration goals we've set up for ourselves.

Part 3 examines the execution of a picture handling strategy for the affirmation of fire identification. This part centres around the picture handling calculation that has been proposed for segregating among fire and non-fire pictures.

Chapter 4 highlights this thesis's research findings as well as key contributions, and examines the future prospects for expanding this study work to make it resilient for real-time exact detection of forest fires.

2. LITERATURE SURVEY

2.1 ATTACK CLASSIFICATION

An interruption identification framework can recognize malignant conduct utilizing various strategies. The exact arrangements of vindictive way of behaving that can be distinguished should be known for the IDS to be pretty much as compelling as could really be expected. Characterization can happen in different ways. They can be characterized in view of comparable way of behaving, for instance, or in light of comparative harm caused. [22]

2.1.1 Classification

Making a differentiation among interior and outside vindictive way of behaving is a valuable characterization. This works with human cognizance. The IDS can work with different arrangements. Be that as it may, the IDS should convey the identifications to a head. Understanding the differentiation among inner and outer pernicious behavior is more straightforward. The exact characterizations don't go against each other. Some vindictive way of behaving can be both interior and outside in nature. The exact grouping should go a lot further than interior and outside. Various qualities recognize each sort of malevolent way of behaving. Realizing these qualities permits you to tweak the IDS to further develop ID. [23].

2.1.2 External Abnormal Behaviour

Outside unusual way of behaving incorporates different sorts of framework assaults. There are various sorts of assaults. Actual assaults, cushion spills over, circulated forswearing of administration assaults, animal power assaults, weakness sweeps, and man in the center assaults are conceivable.

2.1.2.1 DDoS

DDOS assaults, otherwise called Conveyed Disavowal of Administration assaults, endeavor to make an organization asset briefly or forever inaccessible to its clients. Flooding a framework with TCP SYN parcels could bring about an assault. Figure 2.1 portrays a model. Some Expert Hubs are utilized by the aggressor to control various Daemon Hubs. These Daemon Hubs are the PCs that execute the assault.

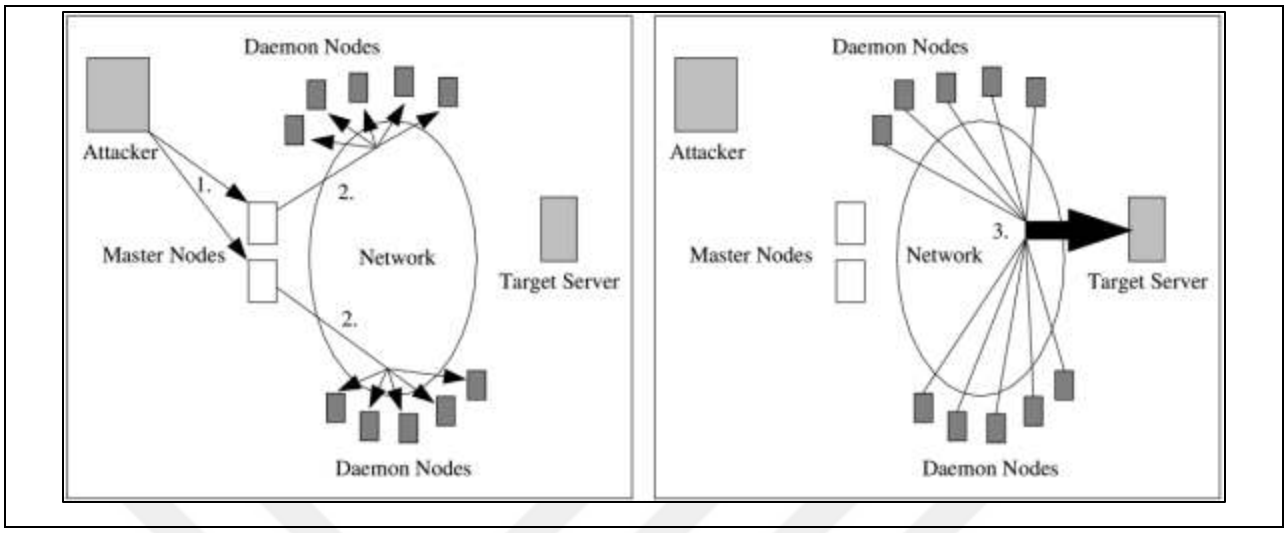

Figure 2.1: DDoS attack. [23]

2.1.2.2 Vulnerability scans

Network filters are information assortment assaults. They cause no harm all alone, yet are generally used to accumulate data about a framework that can be utilized in later assaults. Network filters incorporate organization traffic sniffing and port examining. [13]

2.1.2.3 Buffer overflows

Cradle spills over are one of the most widely recognized kinds of PC or organization assaults. They are utilized to exploit flawed programming. They capability by endeavoring to spill over supports inside the program. At the point when this happens, the spilling over information spills over into adjoining memory. This can ruin information or fundamentally impact the manner in which a program runs. Cradle spills over can be stack-based or load based. Spills over endeavor to spill over the program stack, while load spills over endeavor to spill over powerfully made memory in a program. [22].

2.1.2.4 Brute-force attacks

Beast force assaults are a technique for accessing an objective PC. This can be achieved by, for instance, endeavoring various passwords. These assaults will ultimately succeed, yet it might require a long investment to do them. [22]

2.1.2.5 Physical attacks

Actual assaults are endeavors to hurt PCs genuinely. These assaults can be pretty much as straightforward as cutting a wire. Energy assaults, for example, electro-attractive heartbeats (EMPs) or high and low energy radio recurrence (HERF and LERF) assaults, can likewise happen. [23]

2.1.2.6 Man in the middle

A man in the center (MITM) assault is one in which the assailant endeavors to furtively transfer and perhaps modify correspondence between two PCs or organizations, as displayed in Figure 2.2.

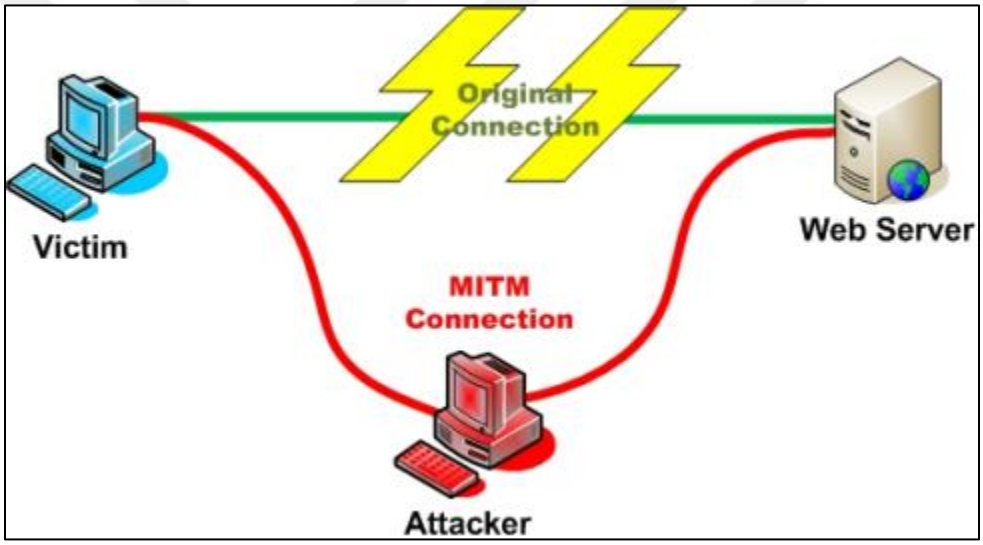

Figure 2.2: Man in the middle attack. [24]

2.1.3 Internal Abnormal Behaviour

Malware alludes to unusual inward way of behaving. Malware arrives in various structures. Malware is characterized into four sorts. Botnets, infections, diversions, and worms all exist. Malware are programs that contaminate a framework and play out a particular undertaking. The malware's assignment figures out which class it has a place in.

2.1.3.1 Trojan horses

Deceptions are programs that have all the earmarks of being innocuous however contain vindictive code. They were named after the Skirmish of Troy. As per legend, the Greeks constructed a

monstrous wooden pony and filled it with officers. The Trojans confused the pony with an acquiescence gift. The pony was brought inside Troy. The Greeks braved around evening time, opened the doors of Troy, and crushed it. This is likewise the way that a deception works. A deception's capacities incorporate keystroke logging, program establishment, and document moves. [23]

2.1.3.2 Viruses

Infections are programs that recreate themselves and taint and spread through different projects. Disease implies that they have invaded another program. At the point when that program is executed, the infection is likewise actuated and executed. [23]

2.1.3.3 Botnets

Botnets are malware that transforms tainted PCs into "slaves" for the expert. The bot-ace controls a tainted PC without the information on the contaminated PC's proprietor. Figure 2.3 shows an illustration of this. The bot-expert can utilize the conveyed organization of "slave" PCs to do extra pernicious undertakings, for example, a DDOS assault. [13]

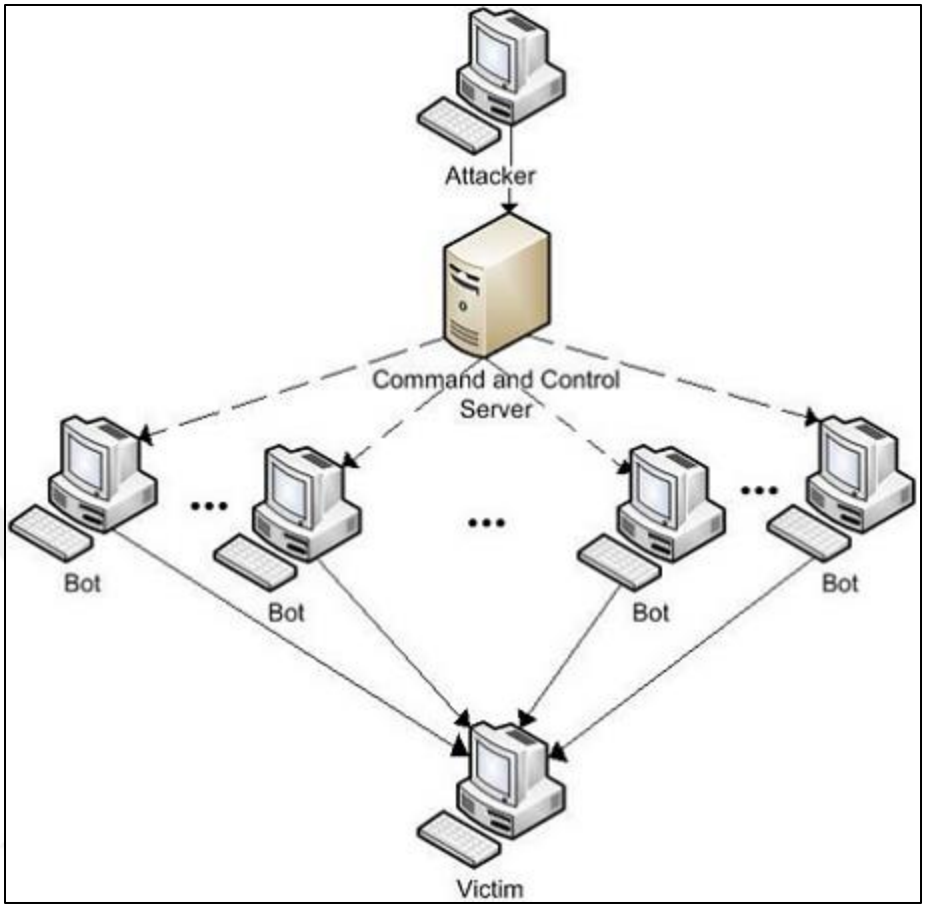

Figure 2.3: Botnet infection. [25]

2.1.4 Detection

An organization interruption identification framework (NIDS) just screens the organization. Accordingly, a NIDS can't recognize each assault. Just goes after that utilize the organization can be identified. Stream based interruption identification frameworks are likewise restricted to utilizing just stream information. This diminishes the quantity of assaults that can be identified.

i. DDOS assaults can commonly be recognized utilizing stream based network interruption location frameworks.

ii. Security filters

iii. Worms

iv. Botnets

Different goes after either don't utilize network correspondence or are not apparent inside network traffic header data. Other recognition frameworks, for example, HIDS or Parcel based NIDS, ought to be utilized to distinguish different assaults, for example, infections, deceptions, Man in the Center, and cradle spills over. Actual assaults are significantly more enthusiastically to identify. Notwithstanding, in specific conditions, different sorts of strange way of behaving can be distinguished. Different sorts of unusual way of behaving can be recognized again data about an organization and the strange way of behaving is known. For instance, When it is known that infections are constantly dispersed from a similar source, this dissemination can be distinguished. This is like different kinds of assaults. Other activities such as phoning home which a keystroke logging Trojan might do, could also be detected. Brute-force attacks can also be detected since they might generate a lot of network traffic. However, unless the Brute-force attack is also a DDoS attack, the traffic mostly looks harmless from a flow-based approach. [22]

2.1.4.1 Distributed denial of service

How much information got can be utilized to identify a disseminated forswearing of administration. DDoS assaults, then again, arrive in various flavors. There are ICMP floods, SYN floods, and different kinds of floods. Traffic examples can be utilized to depict these assaults. A traffic design is characterized by a couple of qualities. These attributes incorporate the quantity of streams and bundles, the size of the parcels, and the all out data transmission utilized during traffic. UDP flooding, for instance, can be recognized by a traffic design that contains countless bundles. During the identification stage, these examples can be looked for. [26]

2.1.4.2 Worms

Worms act contrastingly contingent upon their present status. There are two expresses: the objective revelation state and the exchange state. The worm investigates the organization in the objective revelation state, searching for weaknesses and hosts to taint. The worm really moves itself to the designated have during the exchange state. The Sapphire/Prison worm shows this kind of conduct. [27] In light of the fact that the worm is moved inside the payload information, a stream based NIDS can't identify this state. The objective revelation state is distinguishable. Worms find weak hosts utilizing strategies like organization checks. Worm location methods can in this way be comparable to. [28]

2.1.4.3 Botnets

Botnets are ordinarily comprised of an enormous number of tainted slaves that are constrained by a focal bot-ace. Finding and confining individual tainted slaves is a troublesome issue, yet it is likewise irrelevant given the huge number of outstanding slaves. Identifying and disconnecting the bot-ace is basic to destroying a botnet. Distinguishing botnet conduct, then again, is a definitely more troublesome issue than identifying different sorts of malignant movement. [29] Botnet recognition requires something beyond pernicious way of behaving. Botnets often use IRC channels to impart among slaves and the bot-ace. Streams can be utilized to recognize these on the grounds that they much of the time utilize explicit ports. It is feasible to utilize a technique that doesn't require the utilization of explicit port numbers. This requires streams that incorporate extra data, for example, the quantity of bundles with the PUSH banner set.

2.2 MACHINE LEARNING

A subfield of computer science is machine learning. It is a type of Artificial Intelligence that enables programs to learn and recognize patterns in data. [30] Machine learning investigates algorithms that can learn from and predict data. These are known as machine learning algorithms. Before it can make predictions on data, a machine learning algorithm must learn. Learning requires the algorithm to be shown several examples of data as well as the correct predictions for these examples. The number of examples that must be shown to the algorithm can number in the thousands. Subsequent to gaining from the information, the AI calculation can be utilized to make expectations on new information. For instance, AI can be utilized to screen the pulses of medical clinic patients. The AI calculation is shown a patient's pulse and the ongoing time during the learning stage. Following learning, the AI calculation can anticipate what the patient's heartrate ought to be founded on the ongoing time. By contrasting the anticipated and real pulses, this can be utilized to decide if the patient's pulse is typical. AI calculations are arranged into two kinds. There are two sorts of learning: regulated learning and solo learning. Named information is utilized to prepare regulated learning. Labeled data is made up of input data and the corresponding output data. Unsupervised learning employs the use of unlabeled data. A training set is the data used to train machine learning algorithms. When the data presented to the learning algorithm is labeled, this is referred to as supervised learning. This means that the learning algorithm's user can already make distinctions within the data presented. This is useful for problems involving regression or

classification. Unsupervised learning occurs when the data presented to the learning algorithm is not labeled. The learning algorithm must discover structure in the provided data. A training sample is a single data point from a training set that is used in a predictive modeling task. If machine learning is used to filter spam, for example, the training set is a collection of emails, some of which are spam. A single email would suffice as a training sample. A training example or training instance are alternative names. For predictive modeling, machine learning is used. A specific process or event is modeled in predictive modeling. Using a training set, the model attempts to learn or approximate a specific function, such as distinguishing spam from non-spam email. This is known as the target function. The hypothesis is the function that the model creates to approximate the target function.

At least one highlights are extricated to show the particular interaction or occasion. A component is a particular property. The model of a particular interaction or occasion is characterized by the mix, everything being equal. On account of messages, an element could be the message body or the shipper's email address. Individual characters might show up in the email. The highlights picked are totally reliant upon the issue being settled. This part gives an outline of the numerical groundworks of regulated AI. It starts with linear regression which is a simple category of machine learning algorithms. The principles behind linear regression are the same as for other machine learning algorithms. Afterwards logistic regression is used to explain how classification works within machine learning. It begins by explaining what exactly the hypothesis is and how it can be constructed. [31]

2.2.1 Lineair Regression

Lineair regression is a statistical approach to model the relationship between an "output" value y and one or more "input" values x1...xn. The "input" values are called features. It belongs to the category of supervised learning. An example of this can be seen in Figure 2.4. The black dots represent the data to be modeled. The blue line is the model. This example only has one "input" value. This specific case of regression is called simple linear regression. [32].

2.2.2 Hypothesis

Machine learning relies heavily on a hypothesis which is given the notation H0(x) in equations. This is a function that transforms a given input to the machine learning algorithm into the required output. It is a function that tries to model the target function, it tries to give a function which fits the input data. The input are features, denoted by x. The values for θ are unknown values. These values are chosen during the training of the machine learning algorithm and determine how accurate the algorithm is. When only one feature is considered, a hypothesis is a function of the form [33]:

$$H_0(x) = \theta_0 + \theta_1 * x \tag{2.1}$$

For example, we could use the grades of high school students to predict their chance of success at university. This is shown in Figure 2.1. The x − axis represents the grades (on a scale from 0 to 10) for students and the y − axis is the chance of success. Given is input data (the black points) and from that data a hypothesis (the blue line) is constructed. The hypothesis function can be generalised for N features. It generally has the form:

$$H_0(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n \tag{2.2}$$

x0 always has the value 1. The θ values and the x values can be represented using vectors. Now the hypothesis function can be written in a more convenient way:

$$H_0(x) = \theta^T X = [\theta_1, \theta_2, \dots, \theta_n]^T [x_1, x_2, \dots, x_n] \tag{2.3}$$

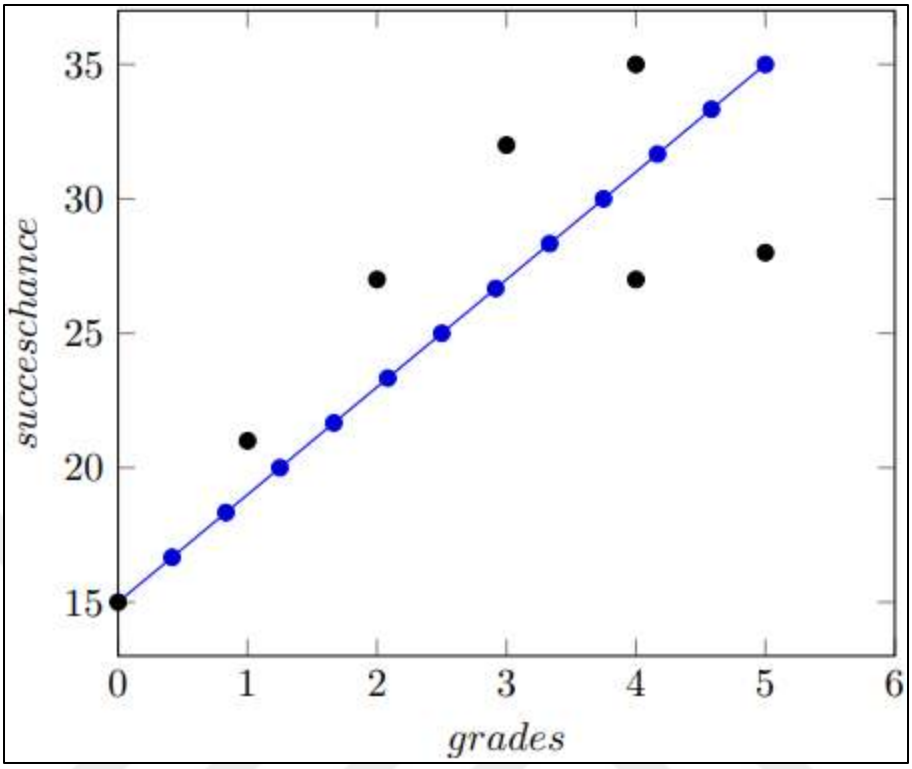

Figure 2.4: Lineair Regression

The hypothesis is the core of a machine learning algorithm. During the training process, the algorithm learns what the hypothesis looks like. More specifically, it tries to find correct values for the vector θ (as seen in equation 2.3. Once the hypothesis has been determined, the difficult work is done. The hypothesis forms the predictive model. The prediction itself is straightforward. A new datapoint is entered into the hypothesis equation and it returns the predicted value.

2.2.3 Cost Function

In order to construct a good hypothesis function, good values of θ have to be chosen. This can be seen as a minimization problem. The difference between any output y and H0(x) has to be minimized. More concretely, the squared difference has to be minimized. This is the MSE, "Minimum Squared Error" function or also called the cost function. The notation x (i) and y (i) is to denote the ith training sample. With dataset size m, the MSE is [33]:

$$J(\theta) = \frac{\sum_{i=1}^{m}\left(H_0(x^{(i)})\right) - y^{(i)})^2}{2m} \qquad (2.4)$$

Seeing this equation already makes it clear that the number of training samples is important. From statistics, it is known that the population of data points, in this case training samples, have to be large enough..

2.2.4 Gradient Descent

To solve the minimization problem, the first step is to start with θ and keep changing the values to minimize J(θ). This is an iterative approach. Gradient descent is such an algorithm that can be used to find a solution to the minimization problem. Gradient descent is an algorithm that uses the gradient or derivative of a function to find a local minimum of that function. In order to easily explain the gradient descent algorithm, a analogy to a ball rolling down a hill can be made. [34] Let's say we have a landscape such as in Figure 2.5. The horizontal location of the ball represents the values of θ, the height of the ball represents the MSE. We want to get the ball to the lowest possible point, in order to minimize the MSE. To do this, the ball rolls down the hill until it comes to rest at the lowest point. Every iteration of gradient descent will bring the ball closer to the lowest point.

Of course, when a ball rolls down a hill, it does not just roll down and immediately stop. The ball takes a path similar as described in Figure 4.2. When the ball finally comes to rest, it can be said that the algorithm converges to that horizontal position.

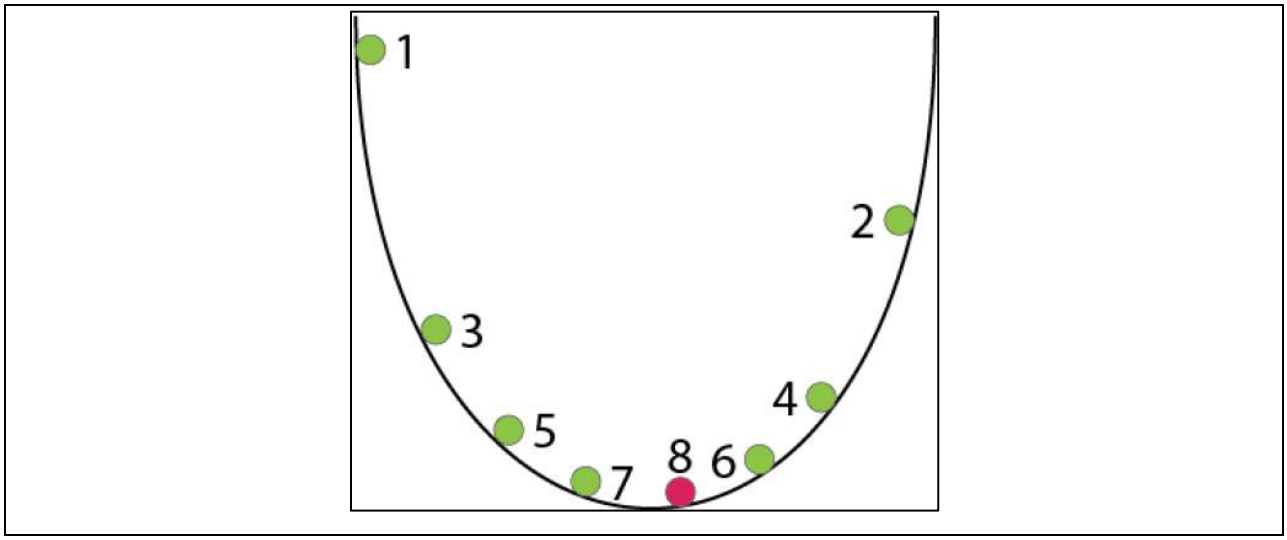

Figure 2.5: Iterations of gradient descent. [35]

To go back to the example of the rolling ball, having α too large, could be seen as having the ball roll down the hill extremely fast. Because of this, the ball could come to rest at another point in the landscape, it might have flown over the hill. The value of α does not need to change during the gradient descent, since the closer the gradient descent gets to the local mimimum, the smaller the derivative becomes, and smaller steps will be taken. If θj is already a local minimum, the derivative is 0 and the gradient descent will not change the value of θj . [36]

2.2.5 Normal Equation

There is another method that could be used in place of gradient descent, a normal equation. This is a method to solve for θ analytically. But this method becomes slow when there are a lot of features. X is a m X n matrix constructed by putting all training samples (vectors of features) together. The notation x (i) and y (i) is used to denote the ith training sample. With dataset size is denoted by m.

$$\theta = (X^T X)^{-1} X^{-1} y \qquad (2.5)$$

But what happens when (XT X) is non-invertible, or singular? This means there are redundant features or more features than training samples. There are mathematical models, such as the pseudo-inverse to still compute a correct result. There are still other methods that could replace gradient descent, such as conjugate gradient, BFGS and L-BFGS. Thinking about matrices and non-invertibility can help to be able to understand how features relate to eachother.

2.2.6 Multi-Class Classification

The above hypothesis and cost function are for binary classification. To compute multi-class classification, the One-vs-all algorithm could be used. This algortihm splits the multiples classes into two groups. One group contains one class, the other group contains all other classes. Using these two new groups, the algorithms for binary classification can be used. In other words, for each class i a different logistic regression classifier Hθ(x) is trained to predict the probability that y = 1. On an input x, the most probable class is chosen. Another method called One-vs-One could be used. This method compares all pairs of classes. It checks, "does the input belong rather to class

A as compared to class B". It does this for all pairs and decides which class most likely contains the input. The One-vs-One method uses nclasses ∗ (nclasses − 1)/2 different logistic regression classifier Hθ(x). Each of these classifiers is trained to fit a pair of classes. [39] To provide an example, Figure 4.9 can be used. On the left, binary classification is shown. The hypothesis and cost function as seen in the above sections can be used to do the classification. On the right, multi-class classification is used. There are three classes: triangles, squares and crosses.

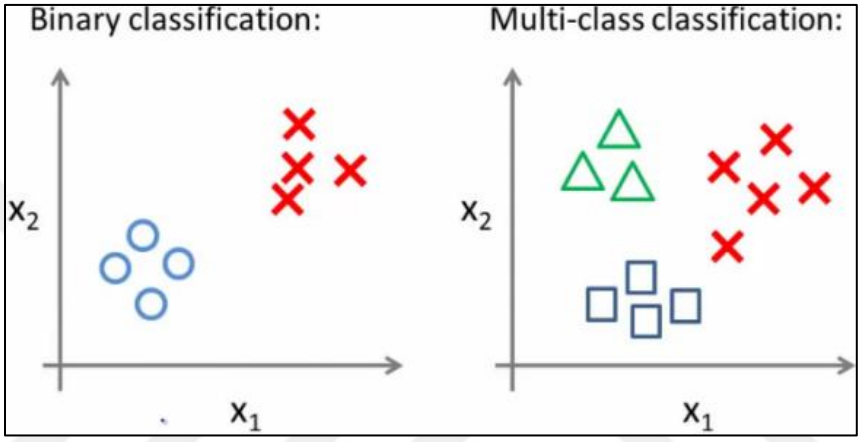

Figure 2.6: On the left, binary classification is shown. On the right, three different classes are used. [40]

One-vs-all classification will have to use three classifiers to predict which class a given input sample belongs to. The first classifier predicts whether the input samples belongs to the triangles or to another class. The second classifier predicts whether the input samples belongs to the squares or to another class and analogous for the third classifier. This method can quite quickly find the result. If the first classifier is used first and it is found that the input sample belongs to the triangles class, the prediction is done. [40] One-vs-One classification will have to use six classifiers. Each classifier is trained to predict whether an input sample belongs to the triangles or the squares, to the triangles or the crosses, and so on. If the classifiers predict that the input samples belongs rather to the triangle class than to the square class and rather to the triangle class than to the crosses class, it can be predicted that the input sample belongs to the triangle class. Because the One-vs-One classification has to use more classifiers than One-vs-All, it is also slower. However, because it compares all pairs it is also more accurate. [40]

2.2.7 Overfitting

When the data is not modeled correctly and the model is too precise for the training data, a problem called overfitting occurs. Models that are overfitted have high variance, there are too many possible hypotheses. In other words, if there are too many features, the hypothesis may fit the training set very well but fails too correctly predict new examples. The hypothesis becomes very complex. An example of this can be seen in Figure 4.10. In a similar way, underfitting may occur.

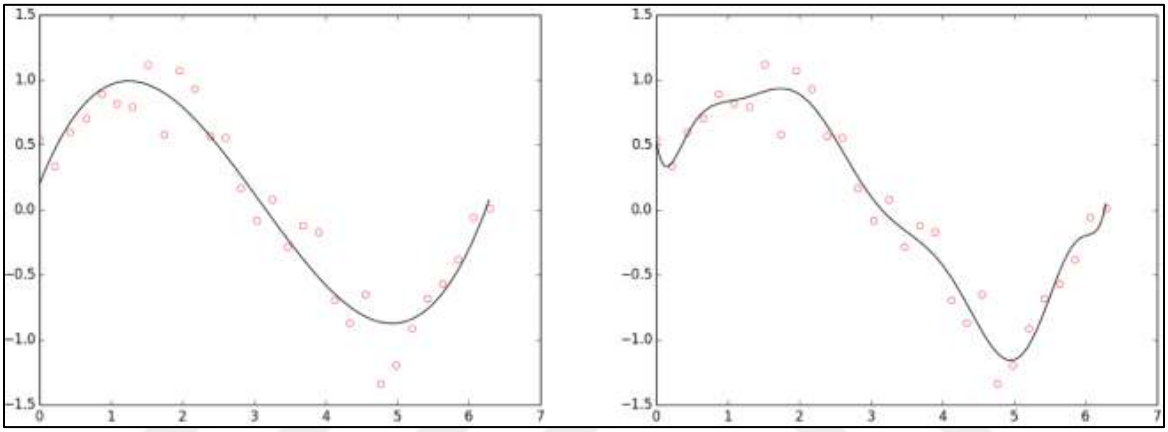

Figure 2.7: The points are generated by a sin function with Gaussian noise. Left: good fit (polynomial of degree 3), Right: overfit (polynomial of degree 10). [41]

There are methods to avoid overfitting. It is possible to reduce the amount of features. This can be done manually or by using a model selection algorithm which is explained in Section 7.2.3. Another option is regularisation. This method keeps all features but manages the values of the parameters of θ. Regularisation works well when there are a lot of features that all contribute to be able to make correct predictions y.

At the core of a machine learning algorithm is the hypothesis. The hypothesis is the function that is used to fit the training data, it is used to predict values for new data. The hypothesis is constructed using the features that represent the data given to the machine learning algorithm and a set of unknown values θ. These values θ are chosen, as such that the hypothesis is the most accurate. In order to do this, a cost function is constructed. The cost function measures the difference between a known output and a predicted output, usually the MSE is used. The values for θ can be found by minimizing the cost function. This is done by using the gradient descent algorithm.

2.3 LITERATURE REVIEW

[15] proposed an intrusion detection model based on BPANN for distinguishing anomalous network traffic from normal traffic, with an accuracy of approximately 93%. [21] proposed a decision tree-based anomaly detection method and compared its performance to that of artificial neural networks, support vector machines, and other decision tree techniques. From result analysis it has been concluded that suggested scheme is better with respect to others. It has been also analyzed that detection rate of U2R and R2L attacks are high whereas most frequent attacks such as DOS, Probe detection rate are low. So, this algorithm is suitable only for identification of low frequent attacks. [22] suggested an intrusion detection system which is based on chi-square feature selection technique and the selected features are classified further with multiclass support vector machine (SVM). The use of multiclass support vector machine reduces the training and testing time essential. In suggested multiclass SVM technique the kernel function is modified for optimized result. This results in increased classification accuracy of the attacks.

[23] designed and analyzed intrusion detection system using machine learning approach such as Logistic Regression, Naive Bayes, Support Vector Machine and Random Forest. The simulation result of this research shows that RF is more effective for predicting and detecting DOS, Probe, U2R and R2L attacks. [24] suggested an anomaly detection model which is hybrid in nature because it combines clustering as well as classification techniques. For clustering k-mean clustering are used and Sequential Minimal Optimization (SMO) are used for classification. The result analysis outperforms better in such hybrid nature. [25] suggested an intrusion detection model which is based on machine learning techniques, such as C5, MLP, and Naïve Bayes. Using these machine learning algorithms a multilevel classification tool are developed. In each level different classification tools are applied respectively and at each stage only one attack are classified. The result analysis shows that the multilevel techniques exhibit higher detection accuracy than a single technique.

Many AI instruments are utilized to lessen the misleading problem pace of abnormality based interruption identification frameworks. [26] utilized outrageous learning machine (ELM) models as well as models that consolidated a few other AI strategies. Each model has its own arrangement of benefits and impediments, with by and large conventional location rates consistently increasing. Out of all other models ELM with SVM outperforms good detection rate for classification of

network data flow into normal or abnormal packets. [27] suggested an intrusion detection system that is based on clustering and classification techniques. For clustering BIRCH algorithm are applied and for clustering support vector machine are applied. The result analysis shows better accuracy rate of about 96% and a false alarm rate of 0.7%. Al-Jarrah et al. [28] proposed a traffic-put together interruption identification framework based with respect to highlight determination procedures. To look at changed calculations and their adequacy, [29] proposed a model for SDI that produces high effectiveness and low misleading up-sides and bogus negatives. Interruption discovery techniques for realized assaults were made sense of by [30]. Future work is expected to keep a high identification rate and a low deception rate. [31] inspected the interruption location framework to refresh it. We want to make a mechanized way to deal with building IDS that gathers review information to catch explicit ways of behaving and gives a system to computing unusual examples and distinguishing inconsistencies.

Our objective, as indicated by [32], is to plan and carry out a scanner that precisely breaks down dubious models and unapproved information. While contrasting acknowledgment models with customary models (in light of marks), the discovery pace of new malignant ways of behaving is multiplied. To distinguish pernicious and unapproved exercises, [33] proposed an organization observation approach or host bundles. This white paper contains information digging procedures for executing interruption location frameworks (IDS) to recognize known and obscure assaults.

[34] explored interruption location and counteraction methods to guarantee an asset's secrecy, respectability, or accessibility. The projects are examined to work on their exactness and security. To recognize assaults and sorts of data set assaults, the calculation comprises of a sensor, an identifier, and an information stockroom. Strategies for distinguishing interruptions were proposed by [35]. Make sense of the construction of IDS, the issue of information grouping, and how to decrease the organization director's responsibility. Labeled information can be utilized to tackle both regulated and solo strategies. [36] researched an interruption recognition framework related study. The design of the characterization techniques and the calculations utilized were made sense of in this article. He showed procedures for the interruption discovery framework and added to conversations and ends for future exploration and factual correlations of different positions in related work. It depicts similar investigations of different classifiers and methods. [37] suggested an intrusion detection system which is based on feature reduction, feature selection and

classification. For feature reduction kernel principal component analysis (KPCA) are applied and for optimal feature selection genetic algorithm (GA) are applied and for classification support vector machine are used. The suggested algorithm was compared with ELM technique and it has been found that suggested technique have better efficiency. The experimental result shows that the average detection rate was 95.26%, whereas the average false alarm rate was 1.03%.

[38] proposed a staggered half breed IDS that consolidates regulated, solo, and exception techniques. This framework was tried utilizing three datasets: constant stream, DDoS, and the KDD Cup 1999 with NSL-KDD datasets. With the revised KDD Cup 1999 dataset, the framework performed well, with a phony problem pace of 3.4%. [39] made a structure with two significant parts: A hereditary calculation (GA) was utilized for highlight choice, and a help vector machine (SVM) was utilized to characterize parcel conduct.

As per [40], interruption discovery frameworks are one of the main organization security highlights. Within the sight of heroes, AI strategies are utilized to dissect network traffic boundaries. This article portrays an outrageous AI strategy for recognizing network interlopers. The exploratory outcomes show that the proposed strategy for distinguishing network traffic assaults is useful. Athanasios [41] fostered a strategy to recognize network interruption in Boston traffic to distinguish these gridlocks, which are as often as possible brought about by an occasion (for instance, a mishap, a path conclusion, and so on.). As per [42], identification of irregularities is a fundamental condition for safeguarding an organization from strikers. Identifies network assaults. Numerous specialists are keen on the examination of social models in IPv4 and IPv6 organizations. It is basic to execute and apply powerful information mining strategies, for example, AI for precise irregularity discovery. In this article, we took a gander at an oddity location model that uses AI calculations for information extraction in the organization to recognize peculiarities whenever. On IPv4 and IPv6 organizations, the proposed model is tried against Disavowal of Administration (DOS). To enhance revelation, pick the most widely recognized and clear IPv6 and IPv4 organizations. The outcomes additionally demonstrate the way that the proposed framework can identify most of IPv4 and IPv6 assaults.

3. PROPOSED SYSTEM

DDoS assaults [1] are flooding dangers that keep genuine clients from getting to their expected assistance. As indicated by late statistical surveying, it is perhaps of the most predominant danger that the Network safety industry is confronting. Therefore, it is basic to comprehend and identify different sorts of DDoS dangers. As expressed in [2,] there are two significant kinds of DDoS assaults: Refection Based and Double-dealing Based, and each DDoS assault can be categorized as one of the two classes. Reflection-based DDoS assaults are those where the programmer covers its personality by using outsider instruments and parts. The assault starts by sending information parcels to a reflector server with the person in question/source target's and IP address. These assaults are completed through the Application layer convention, which utilizes either the Vehicle Control Convention (TCP), the Client Datagram Convention, or both simultaneously. MSSQL and SSDP are instances of TCP-based assaults, though NTP and TFTP are instances of UDP-based assaults, and TCP/UDP-based consolidated assaults incorporate DNS, LDAP, NETBIOS, and SNMP.

Abuse based assaults are like Reflection-based assaults in that the two of them utilize outsider programming and parts to stay mysterious while sending off the assault. It has TCP-based and UDP-based assaults. TCP-based assaults incorporate SYN Flood, while UDP-based assaults incorporate UDP-Flood(UDP) and UDP-slack. The hacker initiates the UDP threat by sending a huge amount of UDP packets at a very high rate on the random ports of the target machine which leads to the exhaustion of bandwidth and thus the system crashes. UDP-lag works by disrupting client-server connection which is carried out by a lag(hardware) switch or software that hogs the bandwidth of the network thereby slowing the user. SYN attack works by exploiting the TCPthree-way-handshake which includes sending out SYN packets continuously until the system crashes.

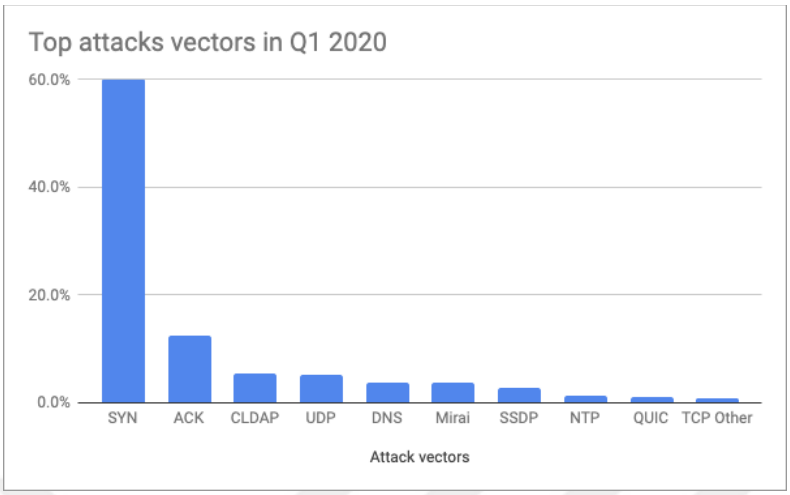

Figure 3.1: Top attack vectors in Quarter 1 2020 [3]

The figure 1, taken from the recent 2020 article by Cloudflare [3] depicts the seriousness to classify the type of DDoS attack as SYN type took up to 60% of attack vectors in first Quarter of 2020, so a need to correctly detect them as early as possible arises. However, this source didn't capture the DDoS attack on Amazon in February 2020 in which the company experienced the largest DDoS attack in history. The Company experienced an attack of 2.3 Terabit per second (Tbps) DDoS attack which lasted more than the average duration for these types of attacks.

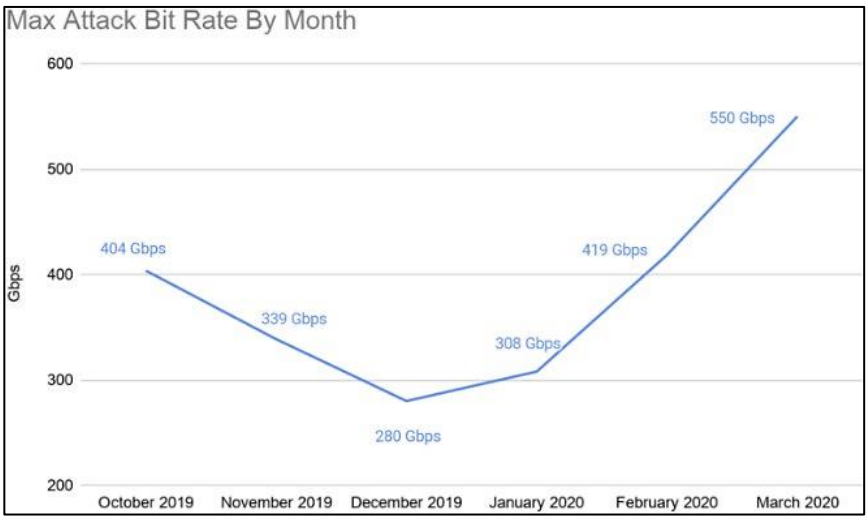

Figure 3.2: Maximum Attack Bit Rate by Month [3]

The criticality of the circumstance makes this one of the serious issues of web security, and numerous factual identification strategies, for example, wavelet-based, port entropy-based, objective entropy-based, etc, can be tracked down in the writing, as featured in [[5], [6], [7], [8], and [9]]. These techniques, in any case, are tedious and ineffectual in light of the fact that the web

is a profoundly powerful field that is continually evolving. Accordingly, to resolve these issues, numerous scientists went to AI and Man-made consciousness systems to distinguish DDoS assaults.

The study applies the baseline models and advanced models described in detail in section 4. In addition, the study presents an Ensemble Classifier MV-4 which combines the performance of the top 4 AI/ML models to give a good performing classifier for this dataset. The study is divided into the following section. Section 3 gives an overview of relevant research and study conducted for DDoS Cyberthreat detection, section 4 deals with the description of the methods and parameters used in this study, section 5 gives information about the Dataset and software used along with Metrics used for evaluating AI and ML models. Section 6 discusses testing techniques and section 7 consists of results and analysis for the used model, in the final section we give conclusions, and possible future works.

3.1 PROPOSED SYSTEM

Since IDS gains designs from preparing information, it can distinguish known assaults; new goes after can't be recognized. This review depends on planning a streamlined component based classifier and breaking down three different datasets. The proposed half and half model for interruption discovery is portrayed in this examination paper. The proposed model's exhibition is assessed utilizing the CICDDoS2019 dataset as a benchmark.

Main goal of the study to reduce or remove unwanted feature using statistically analysis methods and selects the significant features, identify the attacks, choose appropriate classifier to reduce false negative rate, predication time and increase the accuracy. The f-score value is between ranges of 0 to 1. If nearest above 0.9 to below 1 the model is good.

The framework is assessed on CICDDoS2019 datasets and accomplished promising location precision 99%. From dataset there are five different classifiers including, Decision Tree, Logistic Regression, k-Nearest Neighbours and Gradiant Boosting classifier. Out of which one is normal traffic and the other classifiers are four types of attacks listed below: User to Root Attacks (U2R), Denial of Service (DoS), Remote to Local Attacks (R2L) and Probe attacks. It involves various machine learning classifiers.

3.2 METHODOLOGY

The rise of online applications, such as online shopping, hospital record storage, educational course, money transfer, infotainments, etc., has increased the use of interest, which has simplified the work. The rise in internet usage also tends to cause more malicious attacks Protection mechanisms such as firewalls and antivirus detect most of the malicious attacks. There are few unidentified malicious attacks which can cause severe loss of valuable or personal information. The fast-growing computer and wireless networks are vulnerable to malicious violence. This problem is solved by intrusion detection system. Intrusion Detection System (IDS) plays a vital role in shielding data from malicious attacks on computer networks. The main challenge for the development of an IDS model is to optimize accuracy, predict time and increase the false negative rate. Existing IDS model methods have disadvantages such as higher prediction time, higher false negative rate, lower recall and less accuracy. The IDS approach is best suited for the development of the IDS platform, based on data mining and machine learning methods. 10% CICDDoS2019 dataset is different percentages considering all possible attacks was utilized, preprocessing was performed using one hot encoding and we proposed statistical methods are used obtain Z-scores, mean, median and mode for determining the significant features, training was performed for 80% of the dataset. The remaining 20% dataset was tested and validation using decision tree, K-NN and Gradient boosting algorithm. The efficiency of these algorithms was analyzed and discussed it also found that Gradient boosting algorithm is best suitable algorithm to be utilized for development of IDS. Different parameters were interacted based on the testing and are false negative rate, accuracy, f-score, precision time. These metrics are vulnerable to the efficiency of the IDS model that was developed.

The proposed methodology divided three parts of simulation initially 1%, similarly 10% and 100% from 80% training data using Machine Learning algorithm such as Gradient Boosting algorithm, KNN and Decision Tree algorithm from training and validate the metrics like accuracy, F-score, Prediction time and False Negative rate, t-Test and P-Values shown in figure 3.1.

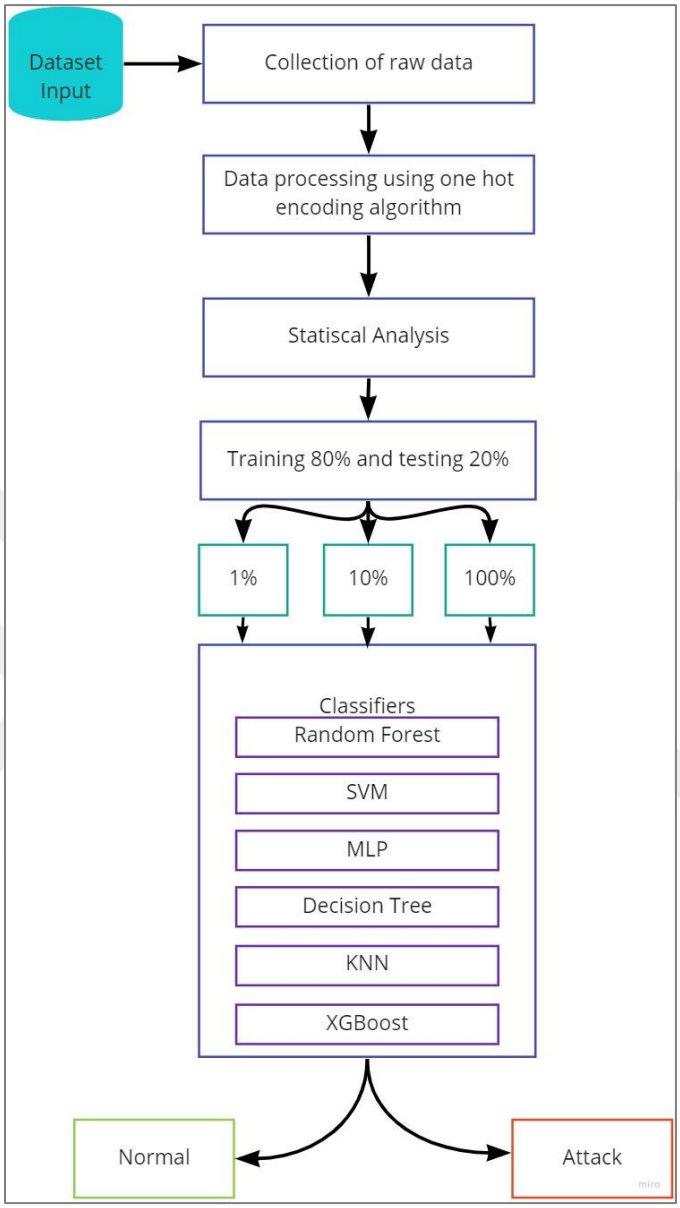

Figure 3.3: Methodology diagram

3.2.1 Preprocessing Phase

The goal of this stage is to preprocess the database file by converting symbolic attributes such as protocol, service, and flag to numerical values. Additional data is normalized. This report is focuses on attribute normalization which are further essential for intrusion detection. Besides the original attributes, in this report, data attributes are normalized for further processing.

3.2.2 Statistical Analysis

The objective of measurable standardization is to change over information from any Normal dissemination into a standard Normal appropriation with a mean of nothing and a difference of one. The measurable standardization process is characterized as follows:

$$Data_i = \frac{x_i - \mu}{\sigma} \quad (3.1)$$

Where, xi = original data of the feature or attribute

μ= mean of data value

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (3.2)$$

σ=standard deviation

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)} \quad (3.3)$$

Utilizing measurable standardization, notwithstanding, the informational index ought to have a Normal dispersion, and that implies that the quantity of tests n ought to be huge as indicated by as far as possible hypothesis. The factual standardization doesn't scale the property's estimation into the reach [0,1].

3.3 SELECTED ALGORITHMS

The calculations for various sorts of DDoS Cyberthreat location are picked in light of the simulated intelligence and ML calculations' computational intricacy [23]. At the point when the exhibition of various calculations is tantamount, this model is significant in choosing a low Computational Intricacy calculation.

3.4 MACHINE LEARNING TECHNIQUES

The objective of AI calculations is for them to learn all alone, without the requirement for human intercession. Since learning is at the core of insight, AI is the essential field of man-made brainpower.

3.4.1 k-Nearest Neighbor (kNN)

For grouping and relapse issues, kNN is utilized. This is quite possibly the most clear characterization calculation. The boundary k, which is the quantity of closest, not entirely set in stone. When a new data point is to be classified, the training data determines its nearest neighbors. The distance is measured using one of the Euclidean distances, the Minkowski distance, or the Mahalanobis distance. The greater the k, the more accurate the classification.

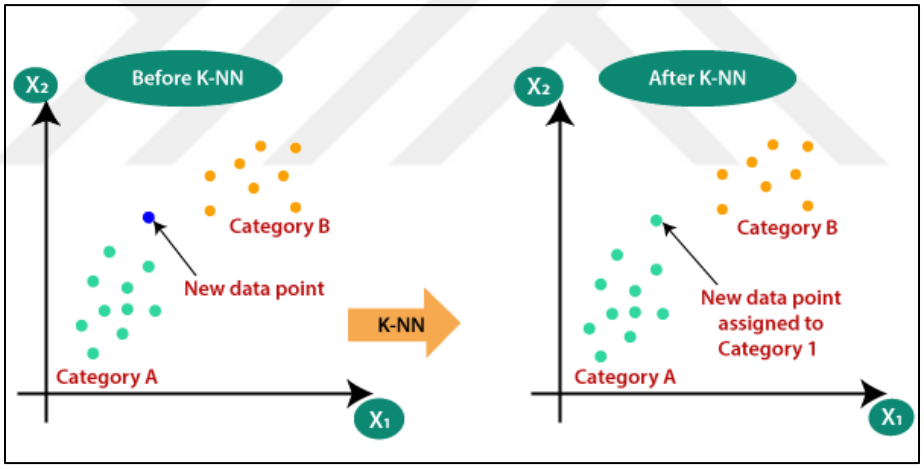

Figure 3.4: KNN Algorithm

3.4.2 Decision Tree Algorithms

The choice model is upheld by the particular upsides of the characteristics contained in the information. The choice span is stretched out until a forecast choice is gone after a particular record. It accompanies a predefined objective variable. For characterization and relapse issues, choice trees are prepared inside the information. Choice trees are famous in AI since they are normally quick and exact. It is material to both all out and consistent information and result factors. The populace or test is partitioned into at least two homogeneous subpopulations or extra upheld the main piece inside the info factors in this strategy. The strategic division makes the decision

based on the tree. This has a significant impact on the tree's accuracy. Figure 3.5 shows how this decision criterion varies for classification and regression trees. Different algorithms are used by decision trees to determine whether to divide a node into two or more sub nodes. The trees divide the nodes based on all of the available variables, then choose the division that results in the most homogeneous sub-nodes. C4.5 and C5.0, as well as CART, are the most widely used decision tree algorithms (Classification and Regression Tree).

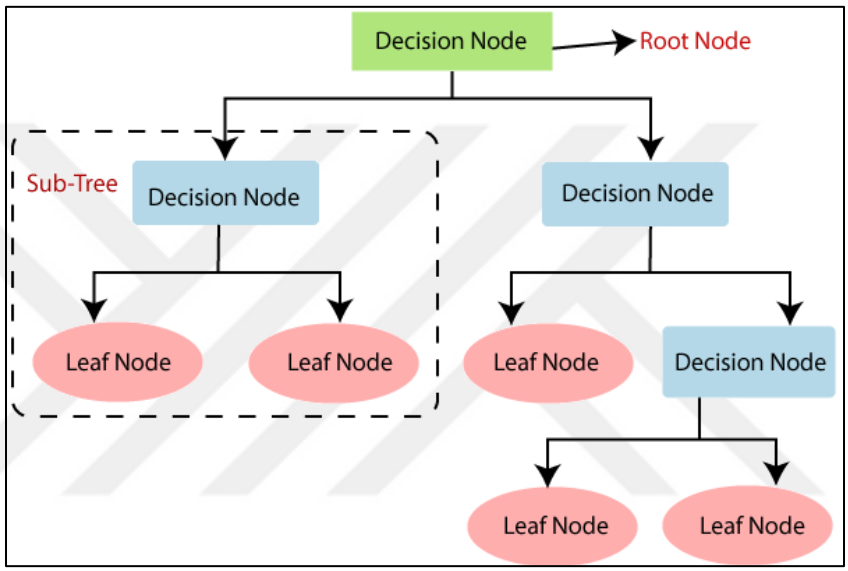

Figure 3.5: Decision Tree Algorithm

3.4.3 Gradient Boosting Classifier

An inclination helping classifier has boundaries that can be tuned to stay away from overfitting to test information and, thus, sums up well. It predicts rapidly and performs best when its boundaries are calibrated, and in light of the fact that an added substance supporting troupe procedure utilizes a relapse tree at each stage, inclination helping classifier ought to beat choice trees on informational indexes where choice trees perform best.

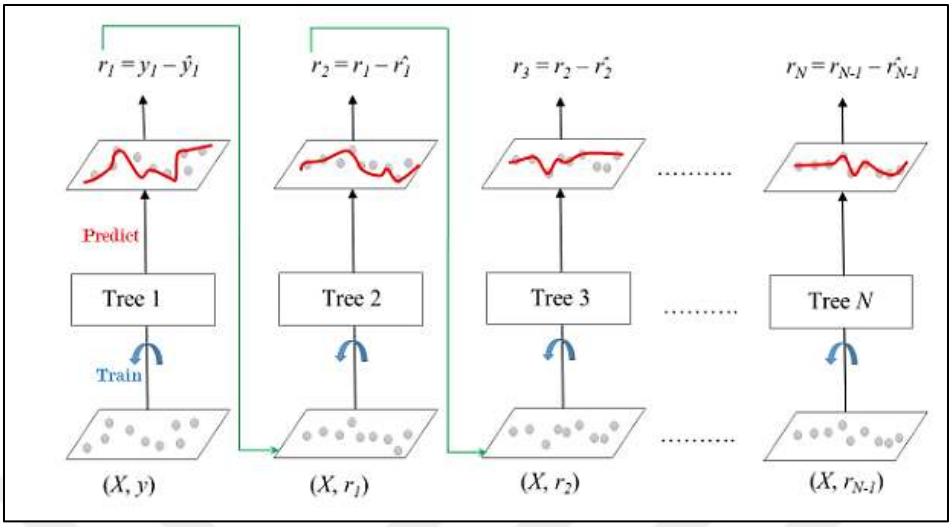

Figure 3.6: gradient boosting classifier

3.4.4 Support Vector Machines (SVMs)

SVM divides training data into classes by identifying a hyperplane (line) that divides training data into classes. The probability of generalizing invisible data increases when the hyperplane that maximizes the distance between classes is identified. SVM provides the best classification performance, i.e. the training set's accuracy. It does not cause the data to overflow. SVM is primarily used for time series forecasting. SVM does not make any significant assumptions about the data. Show greater efficiency in future data classification. SVM is divided into two types: linear and non-linear. A line, i.e. a hyperplane, is used to represent training data in a linear approach.

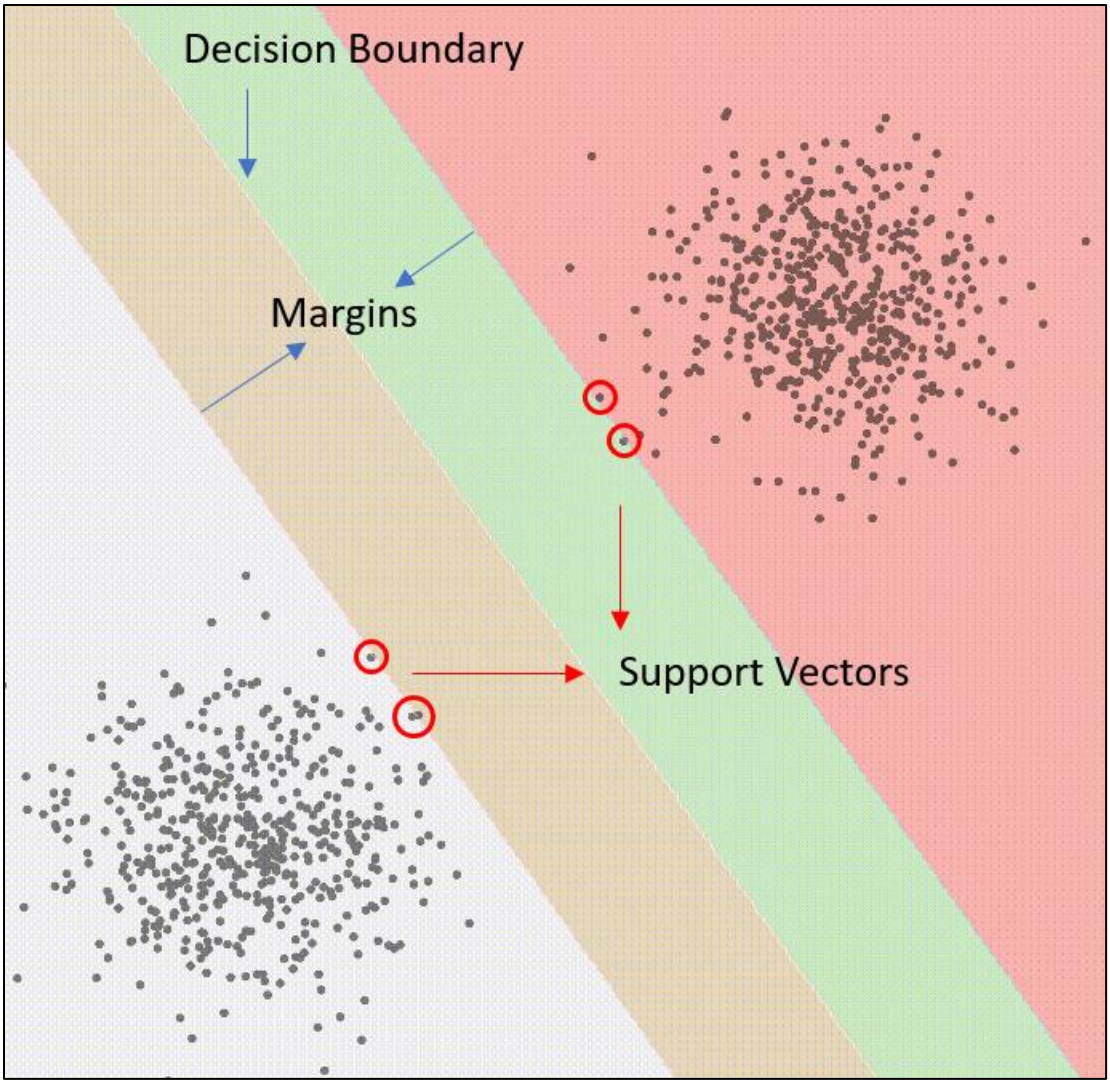

Figure 3.7: Support Vector Machines

3.4.5 Multi-Layer Perceptron (MLP)

Multilayer Perceptron [27] (MLP) is a supervised learning algorithm that learns a function f(): Rf Ro by training on a dataset using the Backpropagation learning method [15]. Thusly in this review, enhancer utilized is ad delta alongside clear cut cross entropy as the capacity of misfortune To accomplish greatest exactness in multiclass arrangement, the Softmax actuation work on its result layer and the relu initiation work in the secret layers are utilized.

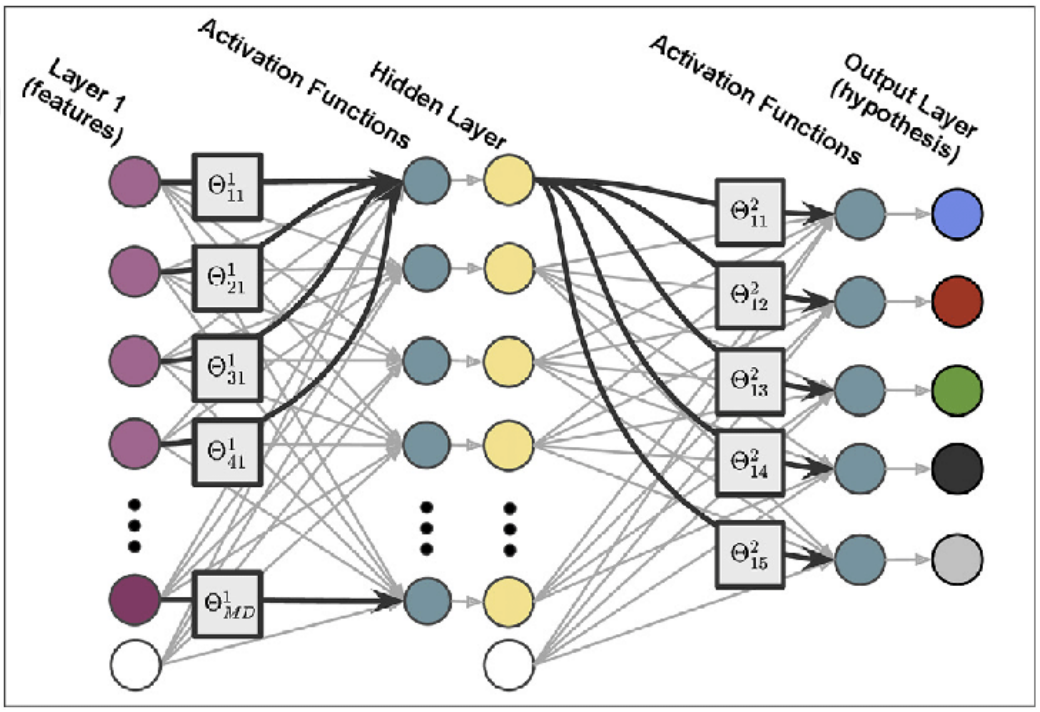

Figure 3.8: Multilayer Perceptron (MLP)

3.5 EVALUATION METRICS

The following metrics are used to evaluate Artificial Intelligence and Machine Learning (AI and ML) models:

Accuracy Score: The accuracy score indicates how close a value is to being correct. The Accuracy score formula from the Sci-Kit learn [24] manual is as follows.:

$$Accuracy(y, y') = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y'_i = y_i) \quad (3.4)$$

In equation 3.4, y addresses the genuine name, while y 0 addresses the anticipated mark, which is given by a calculation after expectation. The Precision Score for multilabel grouping provides us with the exactness of the subset. At the point when the calculation's whole arrangement of anticipated values matches the genuine mark, the precision score is 1.0; in any case, it is 0.0.

F1-Score: F1-Score is moreover insinuated as F-Score or F-Measure. It is an extent of the precision of the test set. To choose the score, F-Measure processes the consonant mean of exactness (P) and audit (R). The F1-Score condition is given by:

$$F1\ Score = 2 * \frac{P*R}{P+R} \quad (3.5)$$

Receiver Operating Characteristic Curve: The Beneficiary Working Trademark bend (RoC) is a helpful device for assessing characterization models in view of their exhibition by taking the Bogus Positive Rate (FPR) and Genuine Positive Rate (TPR) into account (TPR). The TPR and FPR values are determined by changing the classifier's choice limit.

4. RESULTS AND DISCUSSION

In this section we present the excremental founding and the performance analysis of the proposed machine learning algorithms. Every one of the projects for every one of the AI calculations are written in Python 3.7, with the assistance of well known libraries like the Numpy and Pandas Python modules. Moreover, Keras [30] is utilized as the application layer for carrying out simulated intelligence models, and the Tensorflow library [31] is utilized as backend support on Python 3.7.

4.1 DATASET

The dataset used in this study is CICDDoS2019 [2], which is an improvement over its predecessor and addresses the majority of the shortcomings of the previous dataset. The primary advantage of utilizing this dataset is that it has broken down attacks that are primarily done at the Application layer utilizing TCP/UDP conventions. Notwithstanding the new dangers, the dataset focuses on profiling the theoretical way of behaving of human cooperation's using the B-Profile System, which produces reasonable harmless foundation traffic. It generated the abstract behavior of 25 users using the HTTP, HTTPS, FTP, SSH, and email protocols, as described in [2]. The dataset [2] categorizes various types of attacks primarily into two groups. Assaults that are Reflective and Exploitative. It is grouped into different classes in view of the conventions utilized. PortMap, NetBios, LDAP, MSSQL, UDP, UDP-slack, SYN, NTP, DNS, and SNMP are instances of present-day intelligent assaults. Utilizing this dataset for multi-class arrangement for AI and ML calculations is a troublesome assignment because of the enormous number of assault classifications, and no exploration paper has yet utilized this dataset for multi-class grouping for different kinds of DDoS assault identification. Moreover, the dataset is almost 26 GB in size, and that implies that it can't be straightforwardly utilized for this review on standard machines since handling this measure of information requires more calculation assets.

4.2 EXPERIMENT RESULTS

This section presents the results obtained on the test dataset for each of the AI and ML algorithms. The accuracy score for each of the algorithms obtained on the multilabel test dataset is shown in figures 4.1-4.5. Machine Learning algorithms include, the SVM classifier has an accuracy score of 92.75% on a multilabel dataset which is the lowest accuracy score among all the algorithms given in section 3.1. Also, the precision score for Decision Tree is 98.99% which is higher than that of SVM. In the case of AI algorithms, the performance of LSTM is slightly better as it has better accuracy score than that of Multilayer Perceptron (MLP) on the test dataset. The accuracy score for LSTM is 98.17% whereas that of MLP is 97.33% which is 0.84% higher than MLP. The XGBoost calculation has a lower precision score of 98.59% than Adapive Supporting and Irregular Timberland. Moreover, the grouping precision of the Larger part Casting a ballot (MJV-4) calculation is 99.01%. Aside from Arbitrary Backwoods, the precision score acquired on the Test dataset is higher than that of some other calculation introduced in area 4. The precision score for every calculation relates to the F1-Score and the RoC Bend, as displayed in Figure 4.1.

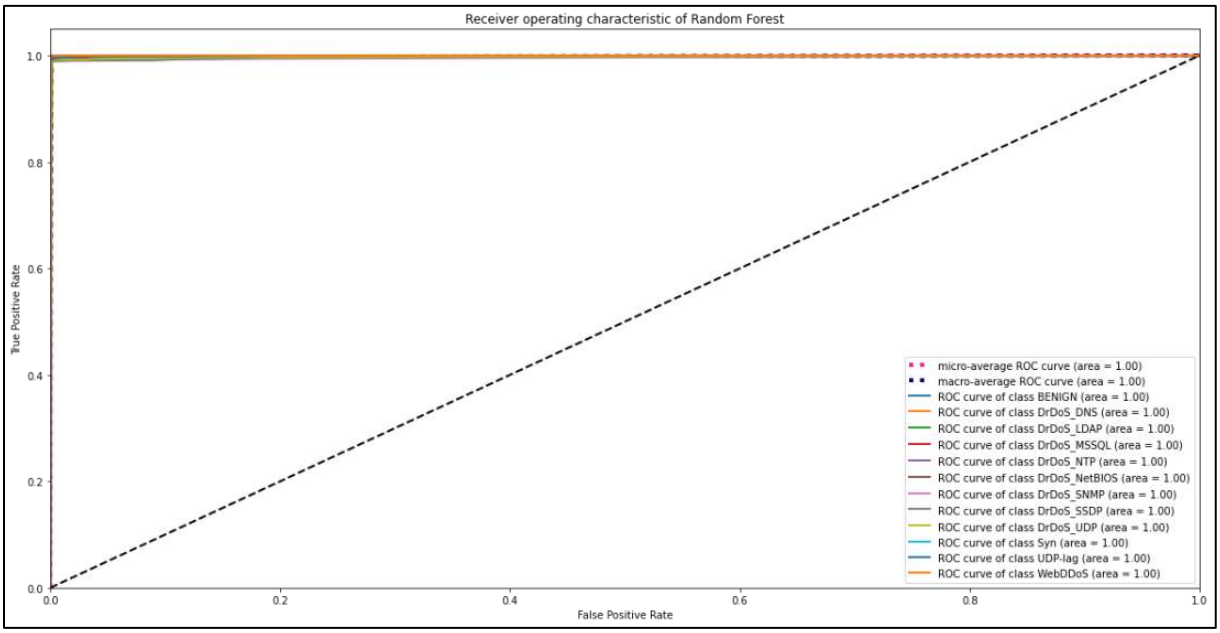

Figure 4.1 : RoC Curve for Random Forest

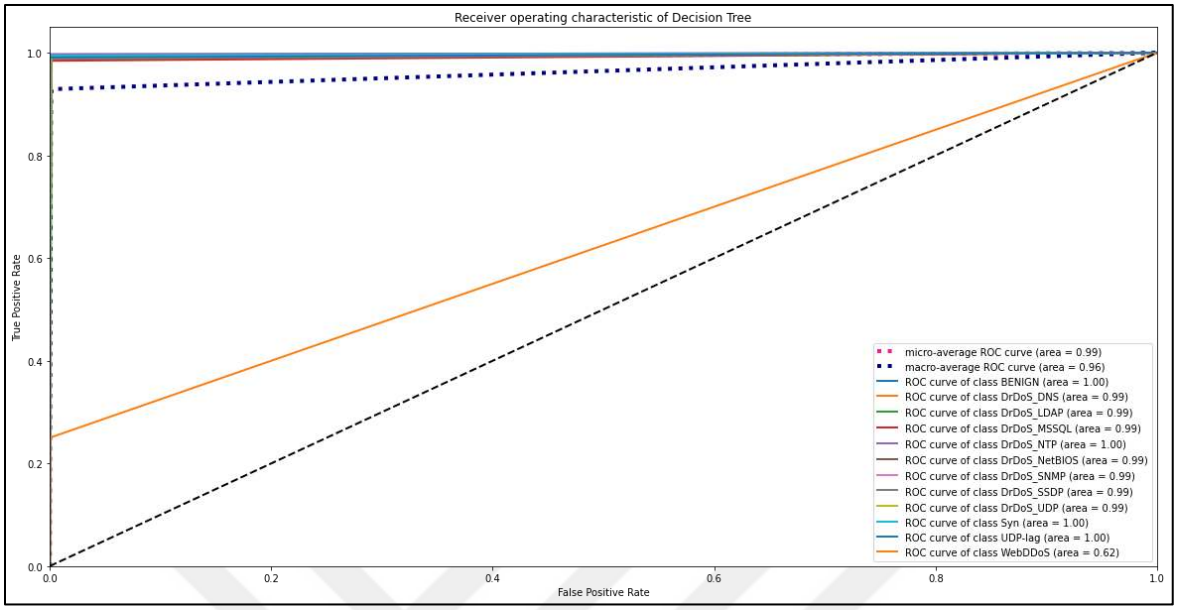

Figure 4.2 : RoC Curve for XGBoost Classifier

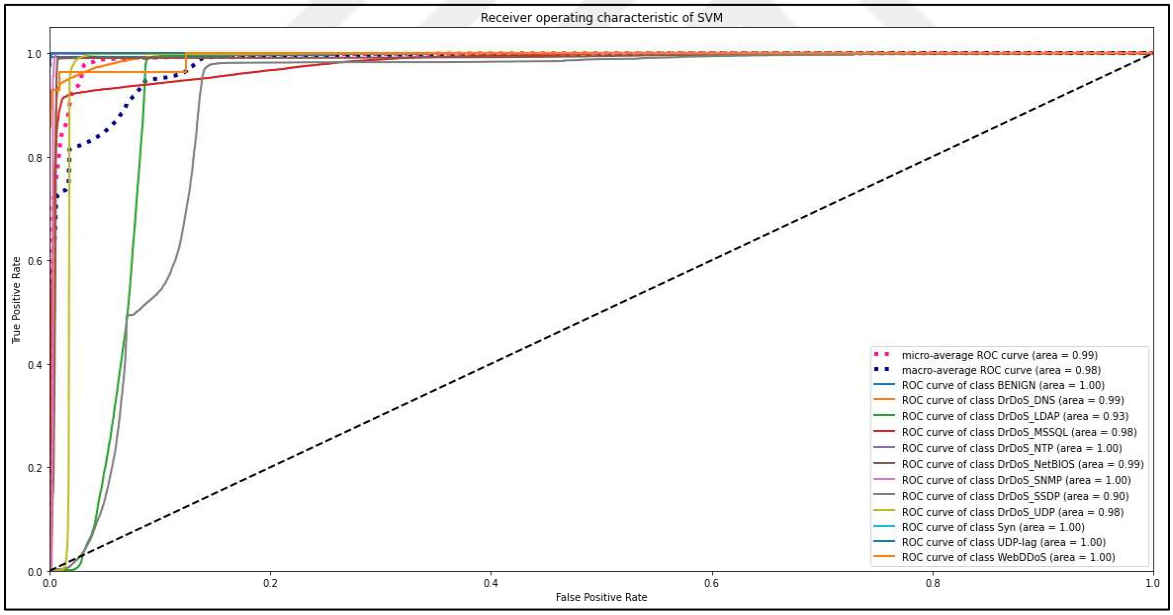

Figure 4.3 : RoC Curve for Support Vector Machine (SVM)

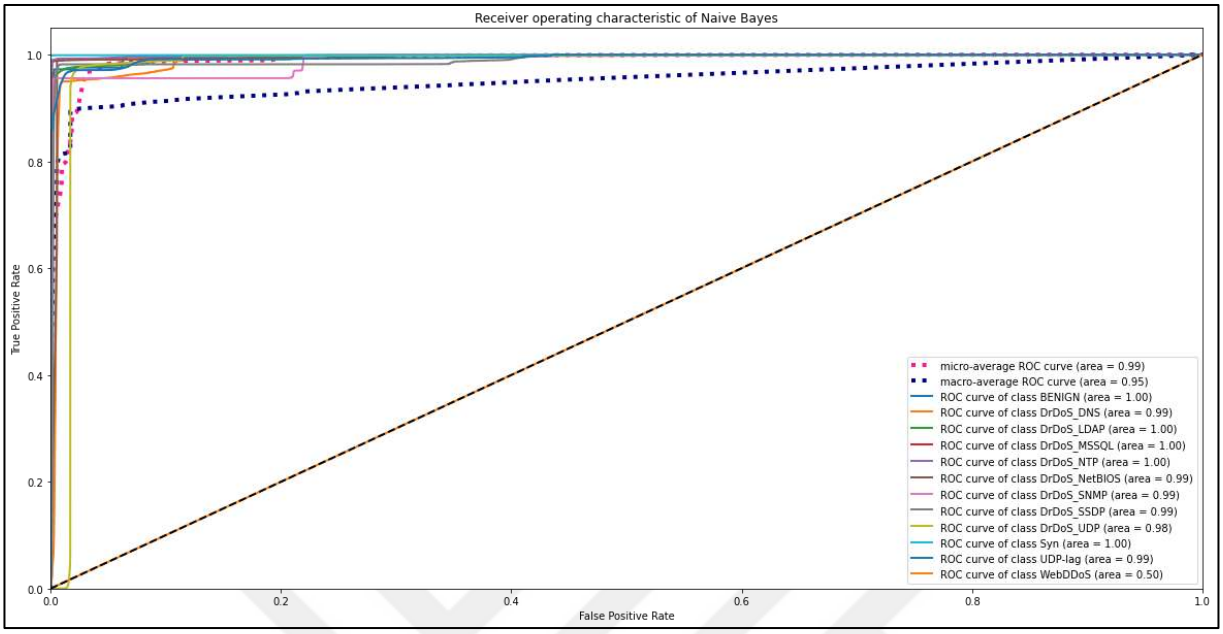

Figure 4.4 : RoC Curve for Decision Tree

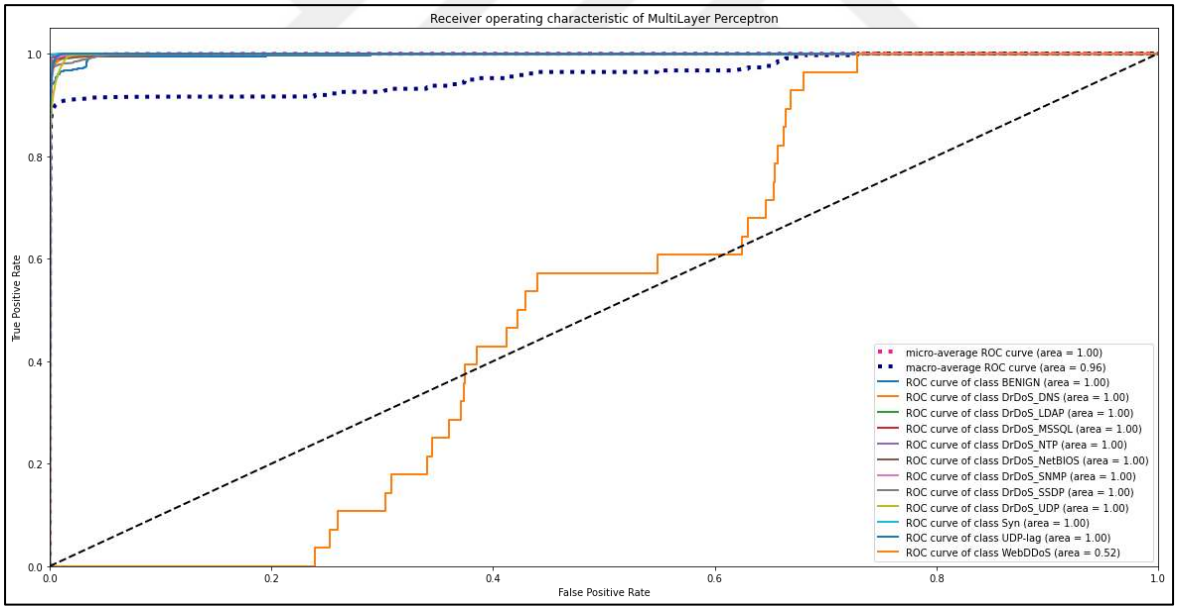

Figure 4.5 : RoC Curve for MultiLayer Perceptron (MLP)

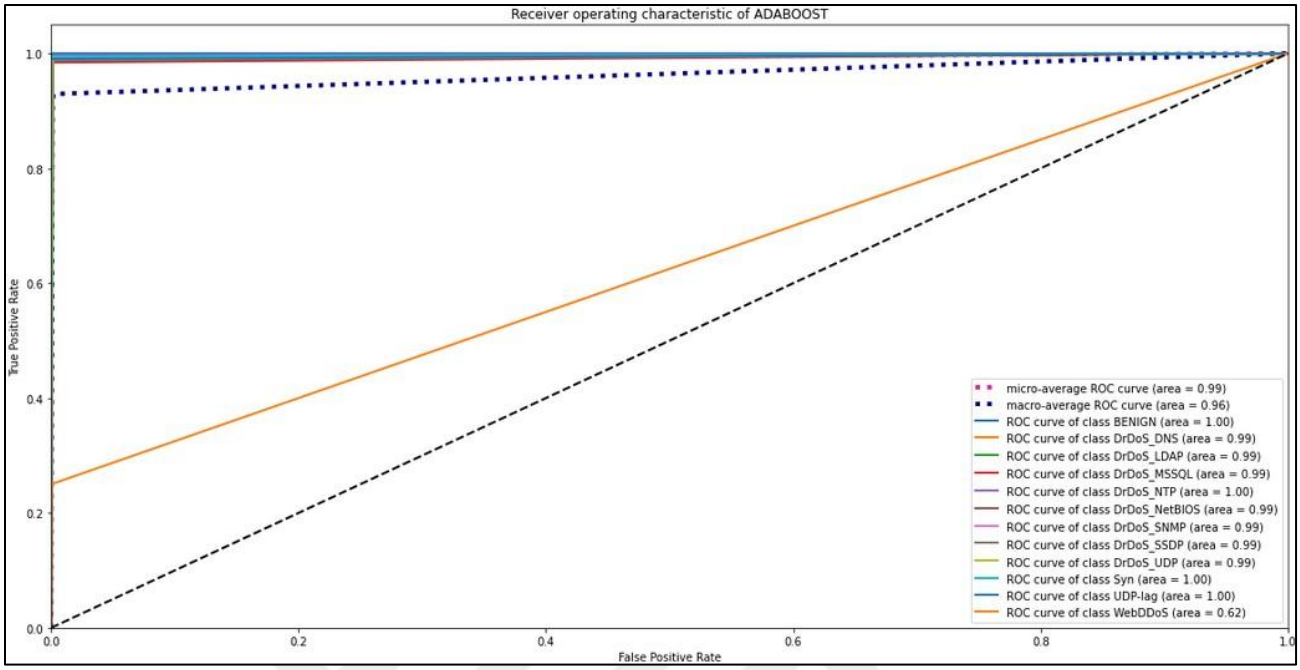

Figure 4.6: RoC Curve for Adaptive Boosting Classifier (Adaboost)

The DDoS Cyberthreat multiclass arrangement was performed utilizing different simulated intelligence and ML calculations, and every danger was exclusively distinguished and approved utilizing different measurements. Moreover, a far reaching investigation of different simulated intelligence and ML calculations for DDoS multiclass Cyberthreat recognition was led, and Irregular Backwoods Classifier has the most elevated exactness score of 99.24% among all calculations, trailed by MV-4 and AdaBoost Classifier, the two of which have a precision score of 99.01%. However, the F1-Score and results from the RoC curve show that AdaBoost lags behind Random Forest and MV-4 in terms of not detecting a few threats correctly. The F1-Score of Random Forest and MV-4 is very similar to Random Forest having a slight advantage as it detects 3 threats perfectly. The RoC Bend for every calculation was introduced for a careful examination of every calculation. As indicated by the RoC Bend, the exhibition of the MV-4 and Irregular Backwoods classifiers acts as an optimal classifier, as the region under the bend for the two classifiers for a wide range of Cyberthreats is 1.00.

4.3 PERFORMANCE EVALUATION

In this section, we discuss and analyze the results obtained from each of the algorithms on the CICDDoS2019 [2] dataset.

According to Table 4.1, the Help Vector Machine (SVM) has an exactness score of 92.75%. It is approximately 6.24% less than Choice Tree and 5.84% less than XGBoost. Furthermore, SVM has the lowest exactness score of any of the calculations in Table 4.1. Table 4.2 also displays the F1-Score for each DDoS Cyberthreat. The F1-score for SYN assault is 1.00, indicating that it can accurately distinguish SYN Cyberthreats on the framework. WebDDoS assaults have a F1-score of 0.52, demonstrating that SVM doesn't recognize them appropriately. The F1-Score for Harmless Traffic is 0.90, demonstrating that the calculation can recognize typical and strange traffic. Moreover, the SSDP assault has a F1-Score of 0.80, which is lower than the F1-Score of other assault types, which is more noteworthy than 0.89. According to the examination of the RoC Bend for SVM shown in Figure 4.3, a few follows like DNS, SNMP, and so on have a higher region under the bend, while others like LDAP, SSDP, and so on have a lower region under the bend, which is consistent with the F1-Score obtained in Table 4.2. SVM has a higher large scale normal of 0.98 than Choice Tree and XGBoost.

Table 4.1: Accuracy Score of Artificial Intelligence and Machine Learning Models

Algorithm	Accuracy Score	Design Parameters
SVM	92.75%	C = 1.0, penalty='l2', loss='squared hinge'
Decision Tree	98.99%	criterion='gini', samples split=3
Random Forest	99.24%	criterion='gini', samples split=2
XGBoost	98.59%	criterion='friedman mse', loss='deviance'
AdaBoost	99.01%	criterion='gini', splitter='best'
MJV-4	99.01%	Random Forest, AdaBoost Decision Tree, XGBoost
MLP	97.33%	activation='relu', optimizer='adadelta'
LSTM	98.16%	activation='softmax', optimizer='adam'

Figure 4.4 portrays the RoC Bend for every one of the assault types for the Choice Tree calculation. It is obvious from the figure that the region under the bend for the majority of the assault types is 0.99, as found in the upper left corner, and the full scale normal is 0.96, which is marginally lower

than that of SVM, attributable to the WebDDoS assault having an extremely low region under the bend of 0.62. Moreover, for by far most of assaults, the bends are situated in the upper left corner, and measurements examination shows that it outflanks SVM on this dataset.

The XGBoost calculation has an exactness score of 98.59%, which is higher than the SVM calculation however somewhat lower than the Choice Tree calculation. Table 4.2 shows the F1-Score for this calculation, and obviously it doesn't distinguish WebDDoS and Harmless dangers appropriately. The F1-Score of Harmless Traffic is 0.46, the least, everything being equal, showing that XGBoost doesn't recognize typical traffic appropriately. In any case, the SYN danger is impeccably identified, though SVM and Choice Tree were not. Other than that, any remaining dangers have a F1-Score of 0.98 or 0.99. Figure 5 portrays the RoC Bend for the XGBoost calculation. Harmless traffic plainly has a lower region under the bend, with a worth of 0.45. Moreover, the region under the bend for WebDDoS Cyberthreat is 0.00, showing that it doesn't distinguish this assault type by any means. The region under the bend for the large scale normal is 0.86, which is lower than the region under the bend for SVM and Choice Tree.

Table 4.2: F1-Score of Artificial Intelligence and Machine Learning Models

Threats	SVM	Decision Tree	Random Forest	XGBoost	AdaBoost	MLP
BENIGN	0.90	0.90	0.95	0.46	0.90	0.92
DNS	0.90	0.99	0.99	0.98	0.99	0.97
LDAP	0.95	0.99	0.99	0.98	0.99	0.98
MSSQL	0.89	0.99	0.99	0.98	0.99	0.99
NTP	0.98	0.99	1.00	0.99	0.99	0.99
NetBIOS	0.93	0.99	0.99	0.99	0.99	0.97
SNMP	0.97	0.99	0.99	0.99	0.99	0.97
SSDP	0.80	0.99	0.99	0.98	0.99	0.97
UDP	0.89	0.99	0.99	0.99	0.99	0.95
Syn	1.00	1.00	1.00	1.00	1.00	1.00
UDP-lag	0.95	0.99	1.00	0.99	0.99	0.97
WebDDoS	0.52	0.38	0.53	0.00	0.32	0.00

Arbitrary Backwoods has the most elevated exactness score of 99.24% of all the artificial intelligence and ML calculations introduced in this review. This accuracy score is better than the SVM algorithm by approximately 6.50% and by 0.65% from XGBoost. In any case, the exhibition of Arbitrary Timberland and Choice Tree is firmly coordinated and just contrast by 0.25%. The F1-Score for Irregular Woods is given in Table 4.2 and it is clear that it has the best F1-Score

among the wide range of various calculations present in Table 4.2. It has three dangers SYN, UDP-Slack and NTP which are impeccably identified which is apparent from the F1-Score of 1.00. Moreover, it has a F1-Score of 0.95 for Harmless traffic which is superior to any remaining calculations which brings up that it can really separate among ordinary and strange traffic. It has a F1-Score 0.53 for WebDDoS assault which is superior to the above-examined calculations. Figure 4.1 portrays the RoC Twist for Random Forest, which seems, by all accounts, to be an ideal classifier for this dataset in light of the fact that the locale under the curve for recognizing all attack types is 1.00. The smaller than usual and enormous scope typical both have a score of 1. 00

The Versatile Supporting calculation is a Group learning approach that utilizes Choice Tree as the base assessor in this review. The F1-Score for harmless traffic is 0.90, which is 0.44 higher than XGBoost and is very much identified in this calculation. Moreover, most of dangers have a F1-Score of 0.99, while SYN Cyberthreat has an ideal F1-Score of 1.00, showing that they are very much distinguished by this calculation.

The Greater part Casting a ballot Classifier (MV-4) calculation joins the presentation of the Irregular Timberland, Adaboost, Choice Tree, and XGBoost calculations, yielding a precision score of 99.01% for MV-4, which is equivalent to AdaBoost and somewhat lower than Arbitrary Backwoods. Table 3 shows that most of the dangers have a F1-Score of 0.99, though SYN has a score of 1.00. Moreover, the WebDDoS danger has a lower score of 0.38, and Harmless has a lower score of 0.90 than Irregular Timberland [2]. When contrasted with AI calculations, the exhibition of simulated intelligence calculations is marginally lower. Figure 4.5 portrays the RoC Bend for MLP. The region under the bend for WebDDoS in this figure is 0.52, which lessens the region under the bend for full scale normal to 0.96. With the exception of the WebDDoS Cyberthreat, any remaining Cyberthreats have a score of 1.00, demonstrating that this calculation accurately identifies these dangers.

4.4 DISCUSSION

In this section, we discuss and analyze the results of each algorithm on the CICDDoS2019 [2] dataset mentioned in section 4. 1. The accuracy of the Support Vector Machine (SVM) is 92.75 percent. It is approximately 6.24 percent lower than Decision Tree and 5.84 percent lower than XGBoost. Besides, the accuracy score of SVM is the lowest among all the algorithms. In addition,

the F1-Score for each of the DDoS CyberThreats. The F1-Score for SYN attack is 1.00 which means it can perfectly detect the SYN Cyberthreats on the system. WebDDoS assaults have a F1-Score of 0.52 which shows that SVM doesn't distinguish this assault appropriately. For Benign Traffic, the F1-Score is 0.90 which shows that the calculation can separate well among Normal and unusual traffic. Also, the SSDP attack has an F1-Score of 0.80 which is lower than other attack types which have a good F1-Score of more than 0.89.

The DDoS Cyberthreat multiclass classification was performed using various AI and ML algorithms, and each threat was individually identified and validated using various metrics. An Ensemble Classifier MV-4 with an accuracy score of 99.01 percent was presented for multiclass DDoS Cyberthreat detection. Furthermore, a comprehensive study of various AI and ML algorithms for DDoS multiclass Cyberthreat detection was conducted, also, Random Forest Classifier has the most elevated precision score of 99.24 percent among all calculations, trailed by AdaBoost Classifier, the two of which have an exactness score of 99.01 percent. In any case, the F1-Score and results from the RoC bend show that AdaBoost falls behind Random Forest and MV-4 as far as not distinguishing a couple of dangers accurately. Random Forest and MV-4 have very similar F1-scores, with Random Forest having a slight advantage because it detects three threats perfectly. The RoC Curve for each algorithm was presented in order to provide a comprehensive analysis of each algorithm.

5. CONCLUSIONS AND FUTURE WORKS

5.1 CONCLUSIONS

The multiclass characterization for DDoS Cyberthreat was performed utilizing different simulated intelligence and ML calculations, and every danger was exclusively distinguished and approved utilizing different measurements. A Group Classifier MV-4 with a 99.01% precision score was introduced for multiclass DDoS Cyberthreat recognition. Besides, a far reaching investigation of various simulated intelligence and ML calculations for DDoS multiclass Cyberthreat location was directed, and Irregular Backwoods Classifier has the most noteworthy precision score of 99.24%, trailed by AdaBoost Classifier, the two of which have an exactness score of 99.01%. The RoC Bend for every calculation was introduced for an exhaustive investigation of every calculation. As indicated by the RoC Bend, the exhibition of the Irregular Woods classifiers acts as an optimal classifier, as the region under the bend for the two classifiers for a wide range of Cyberthreats is 1.00.

5.2 FUTURE WORKS

In future work the system will be designed for classification of real time attacks with live packets. For that new rules and features will be extracted apart from KDD features. Apart from the experimented combination of data mining techniques, further combinations such as artificial intelligence, soft computing and other clustering algorithms can be used to improve the detection accuracy and to reduce the rate of false negative alarm and false positive alarm.

REFERENCES

[1] Fleuret, F., "Fast binandary feature selection with conditional mutual information", Journal of Machine Learning Research, vol. 5, 2004, pp. 1531– 1555.

[2] Chebrolu, S., Abraham, A., Thomas, P. j., "Feature deduction and ensemble design of intrusion detection systems", Computer and Security, vol. 24, issue 4, 2005, pp. 295–307.

[3] Mukkamela, S., Sung, A. H., "Significant feature selection using computational intelligent techniques for intrusion detection", Advanced Information and Knowledge Processing, vol. 24, 2006, pp. 285–306.

[4] Horng, S. J., Su, M.-Y., Chen, Y. H., Kao, T. K., Chen, R. J., Lai, J. L., "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, vol. 38, issue 1, 2011, pp. 306-313.

[5] Amiri, F., Yousefi, M. M. R., Lucas, C., Shakery, A., and Yazdani, N., "Mutual information based feature selection for intrusion detection", Network and Computer Application, vol. 34, issue 04, 2011, pp. 1184–1199.

[6] Dwivedi, A., Rana, YK, Patel, BP, "A literature review on agent based intrusion detection system," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, Issue 10, 2014, pp. 140- 149.

[7] https://en.wikipedia.org/wiki/Host-based_intrusion_detection_system. Date: 21/02/2019.

[8] Uguz, H., "Two stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", Journal of Knowledge Based Systems, vol. 24, issue 07, 2011, pp.1024–1032.

[9] Mukherjee, S., Sharma, N., "Intrusion detection using Naïve Bayes classifier with feature reduction", Procedia Technology, vol. 4, 2012, pp. 119–128.

[10] Li, Y., Xia, J., Zhang, S., Yan, J., Chuan, X., Dai, K., "An efficient intrusion detection system based on support vector machine and gradually features removal method", Expert System with Applications, vol. 39, issue 01, 2012, pp. 424–430.

[11] Karimi, Z., Mansour, M., Harounabadi, A., "Feature ranking in intrusion detection dataset using combination of filter methods", International Journal of Computer Application, vol. 78, issue 04, 2013, pp. 21–27.

[12] Al-Jarrah, O. Y., Siddiqui, A., Elsalamouny, M., Yoo, P. D., Muhaidat, S., Kim, K., "Machine learning based feature selection techniques for large scale intrusion detection", 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW), Madrid, 2014, pp. 177-181.

[13] Manzoor, I., Kumar, N., “A feature reduce d intrusion detection system using ANN classifier”, Expert Systems with Applications, vol. 88, issue C, 2017, pp. 249–257.

[14] Garcia-Teodoro, P., “Anomaly-based network intrusion detection: techniques”, systems and challenges. Computer Security, vol. 28, issue 1-2, 2009, pp. 18–28.

[15] Sufyan, T., Faraj, Al. J., Hadeel, A. S., “A neural network-based anomaly intrusion detection system”, 2011 Developments in E-systems Engineering, Dubai, 2011, pp. 221-226.

[16] Devaraju, S., Ramakrishnan, S., “Performance Comparison for Intrusion Detection System using Neural Network with KDD dataset,” ICTACT Journal on Soft Computing, vol. 04, Issue 03, 2014, pp. 743-752.

[17] Agarwal, M., Member, S., Purwar, S., & Biswas, S. (2017). Intrusion Detection System for PS-Poll DoS. 4(4), 792–808.

[18] Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD Dataset Attributes - Class wise for Intrusion Detection. Procedia Computer Science, 57, 842–851. https://doi.org/10.1016/j.procs.2015.07.490

[19] Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. IEEE Access, 6, 33789–33795.

[20] Ahmed, P. (2017). a Study of Feature Selection Methods in Intrusion Detection System : a Survey. International Journal of Computer Science Engineering and Information Technology Research, 2(June), 1–25.

[21] Al-Mamory, S. O., & Zhang, H. (2010). New data mining technique to enhance IDS alarms quality. Journal in Computer Virology, 6(1), 43–55.

[22] Almseidin, M., Alzubi, M., Kovacs, S., & Alkasassbeh, M. (2017). Evaluation of machine learning algorithms for intrusion detection system. SISY 2017 - IEEE 15th International Symposium on Intelligent Systems and Informatics, Proceedings, Iv, 277–282.

[23] Amro, S. Al, Elizondo, D. A., Solanas, A., & Martínez-Ballesté, A. (2012). Evolutionary computation in computer security and forensics: An overview. Studies in Computational Intelligence, 394, 25–34.

[24] Azad, C., & Kumar Jha, V. (2015). Genetic Algorithm to Solve the Problem of Small Disjunct In the Decision Tree Based Intrusion Detection System. International Journal of Computer Network and Information Security, 7(8), 56–71.

[25] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network anomaly detection: Methods, systems and tools. IEEE Communications Surveys and Tutorials, 16(1), 303–336.

[26] Chandolikar, N. S., & Nandavadekar, D. (2012). Selection of Relevant Feature for Intrusion Attack Classification by Analyzing KDD Cup 99. MIT International Journal of Computer Science & Information Technology, 2(2), 85–90.

[27] D. Gadze, J., Cobbah, M., & S. Klogo, G. (2016). Network Intrusion Detection And Countermeasure Selection In Virtual Network (NIDCS). International Journal of Security, Privacy and Trust Management, 5(1), 25–39

[28] El-Sappagh, S., Mohammed, A. S., & AlSheshtawy, T. A. (2019). Classification Procedures for Intrusion Detection Based on Kdd Cup 99 Data Set. International Journal of Network Security & Its Applications, 11(03), 21–29.

[29] Gabra, H. N., Bahaa-Eldin, A. M., & Korashy, H. (2014). Classification of IDS Alerts with Data Mining Techniques. International Journal of Electronic Commerce Studies, 5(1), 1–6.

[30] Haddadi, F., Khanchi, S., Shetabi, M., & Derhami, V. (2010). Intrusion detection and attack classification using feed-forward neural network. 2nd International Conference on Computer and Network Technology, ICCNT 2010, 3(3), 262–266.

[31] Kalmegh, S. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. IJISET -International Journal of Innovative Science, Engineering & Technology, 2(2), 438–446. www.ijiset.com

[32] Keegan, N., Ji, S. Y., Chaudhary, A., Concolato, C., Yu, B., & Jeong, D. H. (2016). A survey of cloud-based network intrusion detection analysis. Human-Centric Computing and Information Sciences, 6(1). https://doi.org/10.1186/s13673-016- 0076-z

[33] Mabu, S., Chen, C., Lu, N., Shimada, K., & Hirasawa, K. (2011). An intrusiondetection model based on fuzzy class-association-rule mining using genetic network programming. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 41(1), 130–139.

[34] Malik, A. J., & Khan, F. A. (2017). A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection. Cluster Computing, 1–14. https://doi.org/10.1007/s10586-017-0971-8

[35] Marir, N., Wang, H., Feng, G., Li, B., & Jia, M. (2018). Distributed abnormal behavior detection approach based on deep belief network and ensemble SVM using spark. IEEE Access, 6, 59657–59671.

[36] McHugh, J. (2000). Testing Intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Transactions on Information and System Security, 3(4), 262– 294.

[37] Meena, G., & Choudhary, R. R. (2017). A review paper on IDS classification using KDD 99 and NSL KDD dataset in WEKA. 2017 International Conference on Computer, Communications and Electronics, COMPTELIX 2017, 553–558.

[38] Ponmaniraj, S., Rashmi, R., & Anand, M. V. (2019). IDS based network security architecture with TCP/IP parameters using machine learning. 2018 International Conference on Computing, Power and Communication Technologies, GUCON 2018, 111–114.

[39] Relan, N. G., & Patil, D. R. (2015). Implementation of network intrusion detection system using variant of decision tree algorithm. 2015 International Conference on Nascent Technologies in the Engineering Field, ICNTE 2015 - Proceedings, 3–7.

[40] Reshamwala, A., & Mahajan, S. (2012). Prediction of DoS attack sequences. Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012, 1–5.

[41] Yao, H., Fu, D., Zhang, P., Li, M., & Liu, Y. (2019). MSML: A novel multilevel semi-supervised machine learning framework for intrusion detection system. IEEE Internet of Things Journal, 6(2), 1949–1959