THESIS PAPER

Cyber Threat Detection Using Machine Learning Algorithms On Heterogeneous VHS-22 Dataset

Prepared by

Name: Faiaz Rahman

Student ID: 2019-1-50-049

Name: Rafee Zunaied Tanna

Student ID: 2019-1-50-059

Name: Umme Habiba

Student ID: 2019-1-50-048

Supervised by

Zahidur Rahman

Lecturer

This Thesis Paper is Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Information and Communications Engineering

DEPARTMENT OF ELECTRONICS & COMMUNICATIONS ENGINEERING

EAST WEST UNIVERSITY

APPROVAL

The thesis paper titled "Cyber Threat Detection Using Machine Learning Algorithms On Heterogeneous VHS-22 Dataset" submitted by Faiaz Rahman (Student ID: 2019-1-50-049), Rafee Zunaied Tanna (Student ID: 2019-1-50-059) and Umme Habiba (Student ID: 2019-1-50-048) to the Department of Electronics and Communications Engineering, East West University, Dhaka, Bangladesh has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Information and Communications Engineering and approved as to its style and contents.

Approved By

(Supervisor)

Zahidur Rahman
Lecturer
ECE Department
East West University
Dhaka, Bangladesh

Approved By

(Chairperson)

Mohammad Arifuzzaman
Chairperson, Associate Professor
ECE Department
East West University
Dhaka, Bangladesh

DECLARATION

We declare that our work has not been previously submitted and approved for the award of a degree by this or any other University. As per of my knowledge and belief, this paper contains no material previously published or written by another person except where due reference is made in the paper itself. We hereby, declare that the work presented in this thesis paper is the outcome of the investigation performed by us under the supervision of Zahidur Rahman, Lecturer, Department of Electronics & Communications Engineering, East West University, Dhaka, Bangladesh.

Countersigned

**---------- **

(Supervisor)

Zahidur Rahman

Signature

**---------- **

Faiaz Rahman

Student ID: 2019-1-50-049

Signature

**---------- **

Rafee Zunaied Tanna

Student ID: 2019-1-50-059

Signature

**---------- **

Umme Habiba

Student ID: 2019-1-50-048

Signature

**---------- **

DEDICATION

This paper is dedicated

To

Our beloved parents and honorable teachers

ACKNOWLEDGEMENT

We would like to extend our gratitude to our supervisor, Zahidur Rahman for his inspiring words and sincere guidance to complete this work in an efficient way. His constant support gave us the assurance we needed to work on this paper. Our completion of this paper could not have been accomplished without the support of our honorable supervisor. We are truly honored and thankful to have him as a supervisor in the journey of our thesis. We are also grateful to our honorable faculty members and office stuffs who rendered their help during the period of our thesis work.

ABSTRACT

Large and varied amounts of data are needed for the research of emerging machine learning (ML) techniques for detecting network threats, such as malware-related threats. The research community has been using a number of network traffic datasets that have been proposed in recent years. The majority of these datasets contain, however, only a few classes of bot and malware, lacking significant diversity and generalization to identify threats. In this work, we considered a heterogeneous dataset of 27.7 million data named VHS-22. This dataset contains flow parameters extracted using a software network probe from four datasets and a network traffic malware monitoring website. Our methodology evaluates different machine learning techniques and the ensemble classifiers. More than 99% of the threats associated with malware are successfully identified by the Bagging Decision Tree, Random Forest, Extremely Randomize Tree, Decision Tree, Histogram Based Gradient Boosting etc. Additionally, we constructed a prototype dataset named MiniVHS-22 from the original VHS-22 dataset to reduce the computational burden for the future researchers on model training and evaluation. We calculated the ratio of normal and attack data in the original dataset and maintained the same ratio in the MiniVHS-22 dataset of 1M data and used different dimensionality reduction techniques such as the Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) with varying numbers of principal component values on it and explained our analysis in result section. Sophisticated network traffic threat detection systems can be developed using the results of our investigation.

Table of Contents

List of Figures

CHAPTER ONE

Introduction

1.1 Introduction

The prevalence of cyberattacks has increased as hackers use system flaws to steal intellectual property, achieve the financial benefit, or even completely destroy network infrastructures [1]. Malware deployed in such attacks is intended to harm the machine it runs on or the network it communicates over [2]. The signature-based approach, which is used by traditional security tools such as firewalls, intrusion detection systems (IDSs), and virus scanners, compares a payload hash to a database of known dangerous signatures [1].

Since hackers using the most recent attack tactics can readily get beyond such security safeguards, those are by no means sufficient defense. The condition of cyber threats has been rapidly evolving in recent years. The number and sophistication of cyber-attacks have increased, making it more difficult for organizations to protect themselves. In particular, ransomware attacks have become more prevalent, with attackers using this tactic to encrypt a victim's files and demand a ransom payment in exchange for the decryption key. The COVID-19 pandemic has also led to an increase in cyber threats, as many people have been working remotely and using personal devices to access company networks. Overall, the condition of cyber threats is continually changing, and organizations need to stay vigilant in order to protect themselves. On the other hand, anomaly detection systems search for anomalous acts, including those that have never occurred before, because they view any unusual event as a potential attack. As a result, when common activities are labeled aberrant, false positives may result [3]. Dynamic analysis using machine learning algorithms [4]-[5], a promising alternative to traditional malware detection techniques, can be implemented by executing a program and closely observing its activities [6].

A variety of network traffic datasets, including UNSWNB15 [7], CTU-13 [8]-[9], NSL-KDD [10], KDDMTA'19 [11], have been heavily used by the scientific community in recent years. However, the majority of them are rather homogeneous, and, therefore, lack generalization in identifying threats. As a result, they are not very effective in detecting diverse cyber threats. The dataset, VHS-22 [12], used for this study is a heterogeneous dataset made up of ISOT, CICIDS 2017, CTU-13, Booters, and traffic samples from the Malware Traffic Analysis (MTA) website.

So far, many studies and investigations have been made for detecting the cyber threats and also the challenges of its. There are various ways to detect the cyber threats where Machine learning is one of them. The main task of cyber threat detection is to classify the normal and malware attack.

In this study, we have tried to detect the normal and malware attack by applying some Machine Learning Algorithms. Here, we tried to show some comparison between the algorithms for computing the accuracy for our dataset. Machine Leaning techniques provide a big contribution to the detection of the cyber threats which will help the researcher who are interested in the field of cyber threat detection, can choose the best Machine Learning method. Also, this paper will help the reader to easily understand the field of cyber threat detection and the significance of working with this big amount of data.

1.2 Background

The digitalization of the world has led to an unprecedented increase in the volume and complexity of cyber threats. As more and more of our lives and businesses are conducted online, the potential attack surface has grown, making it increasingly difficult to detect and respond to cyber threats.

In the early days of computer security, threat detection was largely based on signature-based detection, which involves looking for specific patterns or "signatures" of known malware in order to identify and block it. However, this approach was limited in its effectiveness, as it could only detect known threats and was easily bypassed by attackers who modified the malware code to evade detection.

With the emergence of new technologies such as artificial intelligence and machine learning, there has been a significant advancement in the field of cyber threat detection. Machine learning algorithms are used to detect malicious activities and threats automatically and with higher accuracy, using techniques such as anomaly detection, supervised learning, and unsupervised learning.

Additionally, with the advent of cloud computing and increased use of mobile and IoT devices, the threat landscape has become increasingly complex, requiring a new approach in threat detection. This has led to the development of threat intelligence, which is an approach that uses various sources of data and knowledge to detect and respond to advanced threats.

Overall, the field of cyber threat detection continues to evolve as new technologies and techniques are developed and as the threat landscape changes. The goal of cyber threat detection is to provide organizations and individuals with the ability to detect and respond to cyber threats in a timely and effective manner, in order to protect sensitive data and maintain business continuity.

1.3 Problem Statement

The problem of cyber threat detection is to identify and respond to potential security breaches and other malicious activities that can cause harm to a computer system or network. This problem is becoming increasingly complex as the volume and sophistication of cyber threats continue to grow, making it difficult to detect and respond to them in a timely and effective manner.

With the widespread use of digital technologies, organizations are generating and collecting vast amounts of data from various sources. Sifting through this data to identify potential threats can be a daunting task, especially as the number and complexity of cyber threats continue to grow.

Another challenge is the evolving nature of cyber threats. Attackers are constantly developing new techniques and tactics to evade detection, making it difficult for organizations to keep up. Also, the lack of preparedness and capability of organizations to detect and respond to cyber threats in a timely and effective manner. Many organizations are defenseless against attacks because they lack the tools, knowledge, and technology necessary to identify and counter the most recent threats.

To come up with practical solutions, extensive study has been done in the area of cyber threat detection. We used machine learning methods in our study to model our data using dataset. Our final objective for this study in the area of cyber threat detection can be summarized as follows:

I. Choosing a dataset that is relevant to this research area.

II. Creating a smaller dataset called Mini VHS-22 using the VHS-22 dataset.

III. Pre-processing the targeted data before applying different techniques.

IV. Applying machine learning techniques to dataset.

V. Comparing various machine learning classification outcomes.

VI. Analyzing the outcomes and concluding from our research's findings.

1.4 Motivation

The condition of cyber threats has been rapidly evolving in recent years. The number and sophistication of cyber-attacks have increased, making it more difficult for organizations to protect themselves. The COVID-19 pandemic has also led to an increase in cyber threats, as many people have been working remotely and using personal devices to access company networks. Attackers use malicious software, such as viruses, worms, or trojan horses, to gain unauthorized access to a computer or network and steal sensitive information or disrupt operations. Sometimes a person can get affected by a cybercriminal who can steal personal information such as names, addresses, Social Security numbers, and credit card numbers, which can be used for identity theft or financial fraud. Also, malware can cause damage to a person's computer or mobile device, making it inoperable or causing it to run slowly.

In this research we are going to focus on the various types of cyber threats also the detecting method of cyber threat detection using machine learning in order to demolish the risk of cyber threat.

1.5 Thesis Organizations

Our whole thesis work consists of total nine chapters. In our first chapter we have included introduction, background, problem statement, motivation, and thesis organization of our paper. In the second chapter we have presented the literature review where we have briefly discussed similar works have been before regarding our work. In the third chapter we have given the basic idea about cyber threat detection. In our fourth chapter we have discussed the implemented machine learning algorithms. In the fifth chapter we have discussed the differences between balanced and imbalanced dataset. Then in our sixth chapter we have presented our research methodology. The seventh chapter represents our research findings and analysis based on machine learning algorithms. Then we have given the advantages and challenges of machine learning in chapter eighth. Finally, in chapter nine we have given the concluding words regarding our whole thesis work.

CHAPTER TWO

Literature Review

An important source of cyber-attacks is malware. The malware typically looks for vulnerable devices across the Internet, rather than targeting specific individuals, companies, or industries. It attempts to infect as many connected devices as possible, using their resources for automated tasks that may cause significant economic and social harm while being hidden to the user and device. Various research and

Alazab [13] examined the evolution of malware by the nature of its activity and variants. The paper investigated malware implication on the computer industry and provided a framework using feature S. Garcia [14] proposes three methods for botnet detection. The third proposal is converted to a testing framework named Stratosphere and is available for both Windows and Linux systems. The dataset used in this proposal is one of the most credible datasets available for botnet detection research.

In 2017, Muhammad Ejaz Ahmed and Hyoungshick Kim [15] investigated proposed an analysis of DNS DDos by the total number of packets transmitted (ToP), ratio of the source and destination bytes. These works need a threshold to measure occurred of attack. Threshold is more reliant on specific network environment and frequently generate a large number of false positives.

Strayer et al. [16] were one of the first to demonstrate the use of supervised machine learning for identifying botnet traffic. The authors devised the detection approach that targets IRC botnets, by performing multi-phase traffic analysis, where the classification of TCP flows using supervised MLAs plays a key role. For the classification of traffic flows the authors considered three MLAs: Naive Bayesian, Bayesian network and C4.5 decision tree, providing relatively low false positive and false negative rates (under 3%). The main disadvantage of the approach is the fact that it is only modeling TCP traffic as the main carrier of IRC communication.

In the same context, authors in [17] investigated the effective-ness of different machine learning algorithms in securing IoT devices against DoS attacks. This study tried to suggest appropriate methods for developing IDSs using ensemble learning for IoT applications. The assessed classifiers are Random Forest (RF), AdaBoost (AB), Extreme gradient boosting (XGB), Gradient boosted machine (GBM), and extremely randomized trees (ETC)

Malowidzki et al. [18] discuss missing data sets as a significant problem for intrusion detection, set up requirements for good data sets, and list available data sets.

As computing power increases and cost drops, Machine Learning is seen as an alternative method or an additional mechanism to defend against malwares, botnets, and other attacks. Antoine(2019) [9] classified malicious traffic in a network by examining its capabilities using Machine Learning.

CHAPTER THREE

Introduction to Cyber Threat Detection

3.1 Cyber Threat Detection

Cyber threat detection is important because it helps to identify and prevent potential security breaches, such as unauthorized access to sensitive data or malware infections. This can help to protect an organization's assets and reputation, as well as comply with regulations and laws related to data protection. Additionally, early detection of cyber threats can help to minimize the damage caused by an attack and facilitate a quicker response and recovery [19].

3.1.1 Understanding a Cyber Threat

A cyber or cybersecurity threat is a malicious act that seeks to damage data, steal data, or disrupt digital life in general. Cyber threats include computer viruses, data breaches, Denial of Service (DoS) attacks, and other attack vectors. Cyber threats also refer to the possibility of a successful cyber-attack that aims to gain unauthorized access, damage, disrupt, or steal an information technology asset, computer network, intellectual property, or any other form of sensitive data. Cyber threats can come from within an organization by trusted users or from remote locations by unknown parties [19].

3.1.2 Types of Cyber Threat

Cyber Threat can be classified by the attacks in cyber world on the base of its process and resource impact. Their varieties, in such a scenario, will be numerous.[20] Here is a quick overview of key cyber-attack types:

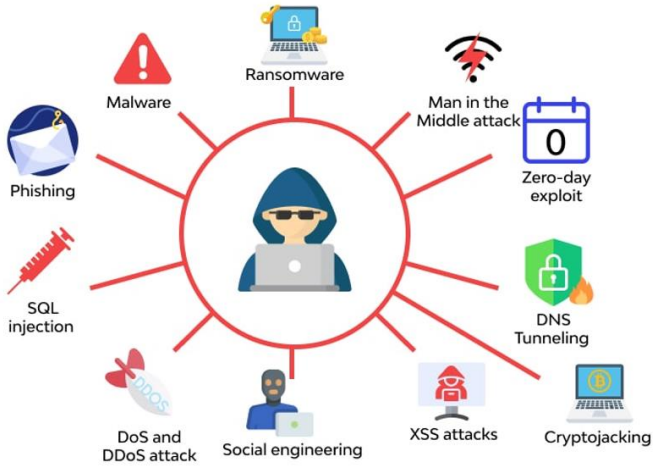

Figure 3. 1: Types of Cyber Attacks

1. **Man-in-the-middle attack:**

By using the MitM technique, the threat actor is introduced as a trustworthy resource between two entities, such as a computer system and a server or a server and a web application. The attack becomes a part of information exchange and other procedures with the forced introduction between two parties and steals vital information.

2. **DoS and DDoS attack:**

By delivering an excessive number of access requests, denial of service (DoS) and distributed denial of service (DDoS) prevent verified resources from accessing a certain system or website. For instance, an attacker can bombard a company's CRM software with access requests in order to keep it busy and prevent real professionals from using it when they are in need. It mostly acts as a planning for future, more destructive attacks.

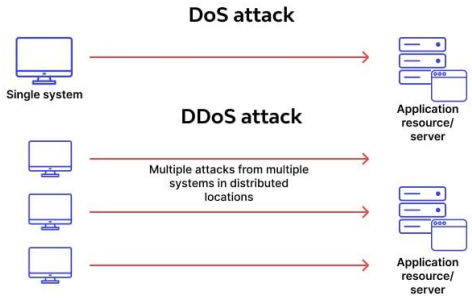

Figure 3. 2: DoS and DDoS Attack

3. **SQL injection:**

These attacks are made through the SQL-based ill-intended codes introduced to the vulnerable system/applications. Upon successful introduction, a SQL injection can collect the query results, give new commands to the systems, and perform prohibited actions on success.

4. **Zero-day exploit:**

This phrase refers to cyberattacks that go unreported for weeks or even months at a time. Zero-day exploits often work by exploiting any flaws in hardware or software. Generally speaking, 0-day attacks start out mildly and persist a longer period of time.

5. **DNS Tunneling:**

Cyber-attacks happening via exploiting the DNS tunneling, a well-known transactional protocol, are not very uncommon. Attackers can use them for their gains and can steal crucial information. As the involved protocol deals with data exchange processes of the application, organizations need to be very careful against it.

6. **Phishing:**

Phishing is a type of cyberattack that uses compromised emails to steal sensitive data and causes significant annoyance. To lure the target of a cyberattack, threat actors will send enticing emails with messages like "you have won a prize," "you got an offer," "a loan is approved," and many others. These emails will ask the recipient to click on a specific link and share information like credit card numbers, bank account information, CVV data, and many other things. The emails are so expertly written that it appears they are from reliable sources. Nearly 50% of all cyberattacks that occur worldwide involve phishing.

7. **Malware**

The goal of a malware cyberattack is to steal information from the targeted system or cause it to entirely malfunction. These assaults make use of a variety of software, including Trojan, Remote Access Trojan, spyware, worms, and ransomware.

8. XSS attacks

XSS, or cross-scripting attack, is essentially a security flaw that affects the entire online application. If an XSS attack is successful, the attacker will be able to add client-side scripts to the targeted web application page. The exploit is frequently used to get around access control restrictions placed on an online application.

9. Social engineering

It is a sort of cyber-attack that depends on manipulating the target's mind. Unlike other cyber-attacks, it requires special knowledge to manipulate people, exploit their emotional tendencies, and track personal or sensitive data. The most frequent application of this method, which has a very high success rate, is intrusions.

10. Ransomware

Ransomware attacks are a subset of malware attacks that threaten the victim with leaking or publishing sensitive information in the public domain if the demanded ransom is not paid. The hacker infects the victim's system with ransomware at the start of the attack, which decrypts the victim's data and sends it to the hacker. Phishing, adware, and USB sticks are a few of the methods used most frequently to spread ransomware.

Figure 3. 3 Ransomware Attack

11. Cryptojacking

Crypto jacking, one of the most recent and troublesome cyberattacks, targets only cryptocurrency owners. Hackers gain access to your resources and begin mining cryptocurrency. The resources and network of the victim will now be used to pay for this resource-intensive task, with the gain going to the intrusion.

3.2 Sources of Cyber Threat

Cyber threats come from a variety of places, people, and contexts. Understanding threat actors and their tactics, techniques, and procedures is essential to respond effectively to any cyberattack Malicious actors include:

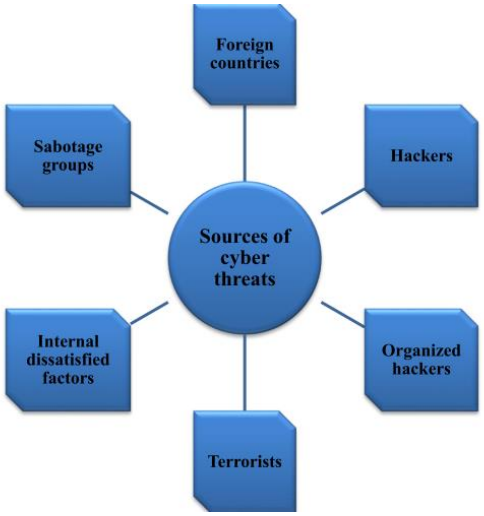

Figure 3. 4 Sources of Cyber Threats

- **Nation-states.** Cyber-attacks by a nation can disrupt communications, military activities, and everyday life.
- **Organized crime.** Criminal groups aim to infiltrate systems or networks for financial gain. These groups use phishing, spam, spyware, and malware to conduct identity theft, online fraud, and system extortion.
- **Hackers.** In order to circumvent security measures and take advantage of holes in a computer system or network, hackers research numerous cyber methods. Typically, their motivations include selfish interests, retaliation, stalking, monetary gain, or political activism. For the thrill of a challenge or for bragging rights within the hacker community, hackers may create new risks.

- **Terrorist groups.** Using cyberattacks, terrorists threaten national security, compromise military hardware, disrupt the economy, and inflict massive deaths by destroying, infiltrating, or exploiting vital infrastructure.
- **Insiders with malicious intent.** Insiders can be employees, independent contractors, outside vendors, or other business partners who legitimately have access to firm resources but misuse it to steal or destroy data for their own financial or personal gain or the benefit of others [20].

3.3 Impact of Cyber Threats

- **The Impact of Cyberattacks on The Government**

According to IBM's 2021 Cost of a Data Breach Report, cyberattacks against government organizations increased considerably in 2021. Simply identifying the breach requires an average of 287 days, and managing it requires an additional 93 days.

The sheer amount of data that may be lost in a cyberattack on a government agency is the biggest worry, ranging from issues of national security and military data that can be utilized by terrorist organizations to information about specific people that can be sold on the dark web [21].

- **The Negative Effects of Cyberattacks on Society**

Every part of our society is impacted by cyberattacks, whether they target a large corporation or a small business.

Considering the Colonial Pipeline ransomware attack once again — This company's pipeline accounts for roughly 45% of the petroleum used along the East Coast and transports gasoline, diesel, and jet fuel from Texas to New York. Millions of people endured fuel shortages when the pipeline was closed, and even airlines had to cancel and reroute flights to accommodate for the scarcity.

- **Rising Costs**

When businesses, companies, and even non-profit service providers such as hospitals, have to pay the costs of a cyberattack, those costs are passed to the consumer.

- **Shortages Of Products or Services**

There will be shortages for the customer, whether it be in the gas pipeline or healthcare, when a company is forced to stop providing its service as a result of a cyber-attack or data breach.

3.4 Consequences of Cyber Threats

Due to the greater number of security flaws and loopholes that enterprises and companies tend to have, cybercriminals are more inclined to target them. All types of enterprises can be targeted, and they may experience the short- and long-term repercussions listed below. The most heavily targeted industries are those in the energy, finance, and technological sectors [22].

- **Loss of Productivity**

When malware is introduced into a company, it frequently causes activities to be suspended for hours or even days. Take the May 2021 Colonial Pipeline ransomware attack as an example. They had to shut down the entire system controlling the pipeline even though only the piece of their computer system that was related to the billing infrastructure was compromised. Although this is one of the more extreme cases, almost every company that experiences a cyberattack must suspend some or all of its operations until the attack is resolved, whether this is done by paying a ransom, removing the malware from the device, network, or system, or by restoring a backup copy of the system.

- **Loss of Revenue**

The expenses of a cyberattack can destroy a company. According to a report from Kaspersky Labs, the average cost of a data breach for a small to medium-sized business is $117,000 regardless of whether operations must be shut down for several days, money must be paid as ransom, data must be lost, devices must be replaced, or money must be paid to a security expert to remove all malware from the system or network. Later costs, such as paying for customer notice, credit monitoring, or even regulatory penalties, are not covered by this study.

- **Loss of Reputation**

The loss of reputation is the most detrimental result of a cyberattack. Think about the recent data breaches at Equifax, Target, and J.P. Morgan Chase, where each of these businesses lost client personal data including credit card numbers, bank account information, and social security numbers. Most organizations don't recover from security breaches even when they have the capacity to do so because they lose their customers' trust and therefore their business.

3.5 Examples of Damages to Companies Affected by Cyber Attacks

The number of cyber-attacks in recent years is staggering and it's easy to produce a laundry list of companies that are household names that have been affected.

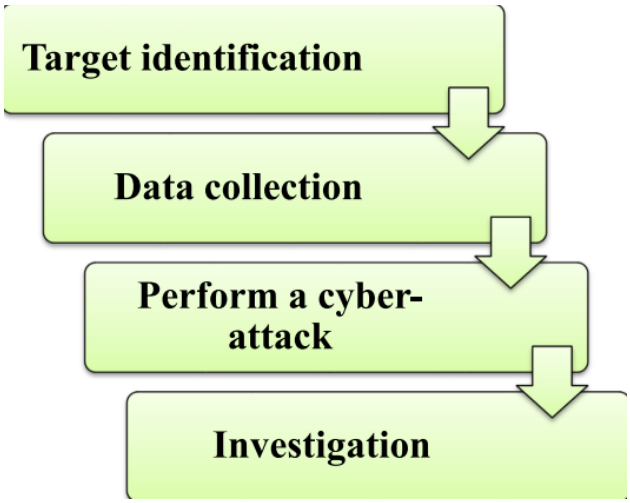

Figure 3. 5 An Anatomy of Cyber Attack

Here are just a few examples.

- **Equifax**

The Equifax cybercrime identity theft event affected approximately 145.5 million U.S. consumers along with 400,000-44 million British residents and 19,000 Canadian residents. Equifax shares dropped 13% in early trading the day after numerous lawsuits were filed against Equifax as a result of the breach. Not to mention the reputational damage that Equifax suffered. On July 22, 2019, Equifax agreed to a settlement with the FTC which included a $300 million fund for victim compensation, $175m for states and territories in the agreement, and $100 million in fines.

- **eBay**

Between February and March 2014, eBay was the victim of a breach of encrypted passwords, which resulted in asking all of its 145 million users to reset their passwords. Attackers used a small set of employee credentials to access this trove of user data. The stolen information included encrypted passwords and other personal information, including names, e-mail addresses, physical addresses, phone numbers, and dates of birth. The breach was disclosed in May 2014, after a month-long investigation by eBay.

- **Adult Friend Finder**

In October 2016, hackers collected 20 years of data on six databases that included names, email addresses, and passwords for The FriendFinder Network. The FriendFinder Network includes websites like Adult Friend Finder, Penthouse.com, Cams.com, iCams.com, and Stripshow.com. Most of the passwords were protected only by the weak SHA-1 hashing algorithm, which meant that 99% of them had been cracked by the time LeakedSource.com published its analysis of the entire data set on November 14.

- **Yahoo**

Yahoo disclosed that a breach in August 2013 by a group of hackers had compromised 1 billion accounts. In this instance, security questions and answers were also compromised, increasing the risk of identity theft. The breach was first reported by Yahoo on December 14, 2016 and forced all affected users to change passwords and to reenter any unencrypted security questions and answers to make them encrypted in the future. However, by October of 2017, Yahoo changed the estimate to 3 billion user accounts. An investigation revealed that users' passwords in clear text, payment card data, and bank information were not stolen. Nonetheless, this remains one of the largest data breaches of this type in history.

3.6 Malware Threat

Malware (short for “malicious software”) is a file or code, typically delivered over a network, that infects, explores, steals, or conducts virtually any behavior an attacker wants [23]. And because malware comes in so many variants, there are numerous methods to infect computer systems. Though varied in type and capabilities, malware usually has one of the following objectives:

- Enable remote control of an infected machine for an attacker.
- Send spam to unknowing recipients via the infected device.
- Examine the user's local network for threats.
- Steal private information.

3.6.1 Types of Malwares

There are various types of malwares. They are in the following:

- **Virus**

Malware includes viruses as a subcategory. A virus is a piece of malicious software that is attached to a file or document and uses macros to run its code and spread from one host to another. The virus will remain dormant after it has been downloaded until the file is used and opened. Viruses are made to interfere with a system's functionality. Consequently, infections might result in serious operational problems and data loss.

- **Worms**

Worms are a type of malicious software that spreads quickly to all devices connected to a network. Worms can spread without a host application, unlike viruses. A worm enters a system through a network connection or a downloaded file, where it then multiplies and spreads at an exponential rate. Worms, like viruses, can seriously impair a device's functionality and destroy data.

- **Trojan virus**

Trojan viruses are concealed in useful software. However, after the user downloads it, the Trojan virus can access private information and change, block, or destroy the information. The device's performance may be severely harmed by this. Trojan viruses are not made to propagate themselves like other viruses and worms are.

- **Spyware**

Malicious software called spyware operates covertly on a computer and sends information to a remote user. Spyware targets sensitive information and can provide predators remote access rather than just interfering with a device's functionality. Spyware is frequently used to steal personal or financial data. Keyloggers are a particular kind of spyware that track your keystrokes and leak passwords and other private data.

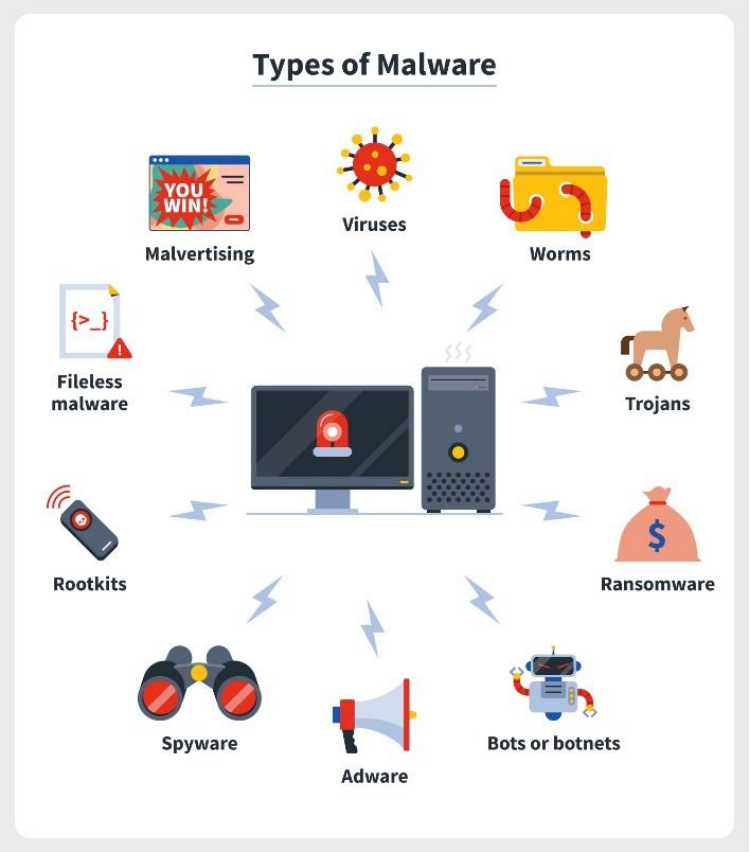

Figure 3. 6 Types of Malware

- **Adware**

Adware is harmful software that tracks how you use your computer so it can show you relevant advertising. While adware is not always harmful, it can sometimes interfere with your system. Adware can reroute your browser to risky websites and even include spyware and Trojan horses. A substantial amount of adware can also dramatically slow down your PC. It is crucial to have security that continuously and intelligently monitors these programs because not all adware is dangerous.

- **Ransomware**

Ransomware is malicious software that accesses private data on a computer, encrypts it so the user cannot access it, and then demands payment in exchange for the data's release. A ransomware attack frequently includes a phishing fraud. The consumer downloads the malware by accessing a fake link. The attacker then goes about encrypting particular data that can only be decrypted using a mathematical key they are aware of. The data is opened once the attacker has been paid.

- **Fileless malware**

A type of memory-resident malware is fileless malware. As the name implies, this malware runs from the computer's memory rather than from data on the hard drive. It is more difficult to find than conventional malware because there aren't any files to scan. Due to the malware's tendency to vanish after a reboot of the target computer, forensics are also made more challenging. The Cisco Talos threat intelligence team published DNSMessenger as an example of fileless malware in the latter part of 2017.

3.7 Importance of cyber security against malware threats

The necessity and requirement to keep information, data, and devices secure essentially sums up the significance of cyber security. People keep enormous amounts of data on computers, servers, and other linked gadgets in today's society. A large portion of material is delicate, including Personally Identifiable Information (PII), which includes passwords or financial information. There is also intellectual property (IP). A cybercriminal might wreak havoc if they were to get their hands on this information. Sensitive information can be shared, passwords can be used to steal money, or even data can be changed to the attacker's advantage. In order to be compliant, organizations must have security solutions [24].

Cybersecurity assists in ensuring that the community can continue to rely on public services or governmental institutions. A city-wide blackout might result from a cyberattack that targeted the energy sector, such as a power plant. It could steal from millions of individuals if it targeted a bank.

3.8 Challenges of Cyber Security

The biggest threat currently facing the digital world is ransomware. In 2021, there were unheard-of ransomware attacks, and 2022 is predicted to see more of the same. The greatest risk to data security exists with the Internet of Things, or IoT. Any digital, mechanical, computer-smart gadget, like a laptop or a phone, is referred to as the Internet of Things (IoT). Hackers use nearby devices to access your own device, which contains sensitive information, like wearable smartwatches, baby monitors, smart fridges, and smart lighting. This is another major challenge. Lack of encryption, authentication, and inadequate cloud settings are some significant factors that contribute to compromised data security [25].

CHAPTER FOUR

Introduction to Machine Learning Algorithms

A machine learning algorithm is a set of instructions that a computer uses to learn from data and make predictions or decisions without being explicitly programmed to do so. The two main processes of machine learning algorithms are classification and regression [29].

Machine Learning Algorithm can be broadly classified into three types:

1. Supervised Learning Algorithms
2. Unsupervised Learning Algorithms
3. Reinforcement Learning algorithm

As we saw in the last chapter, cyber-attacks are a significant issue for cyber security. Therefore, it is important to recognize and stop those cyberattacks. To find cyberattacks, we can utilize machine learning techniques. For the purpose of detecting cyberattacks, we apply machine learning algorithms in our work. In our paper, we worked with supervised learning for our detection techniques [26].

Supervised learning is a sort of machine learning in which the output is predicted by the machines using well-labeled training data that has been used to train the machines. The term "labelled data" refers to input data that has already been assigned the appropriate output. The method of supervised learning involves giving the machine learning model the right input data as well as the output data. Finding a mapping function to link the input variable (x) with the output variable is the goal of a supervised learning algorithm (y) [26]

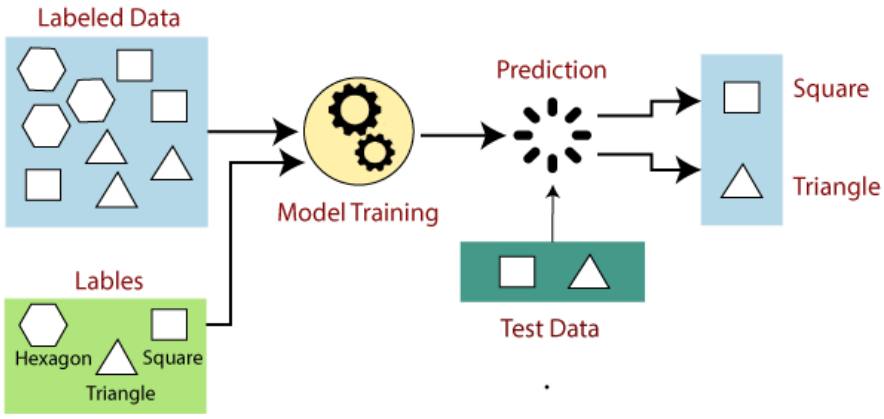

Figure 4. 1 Supervised Learning

Supervised learning can be further divided into two types of problems:

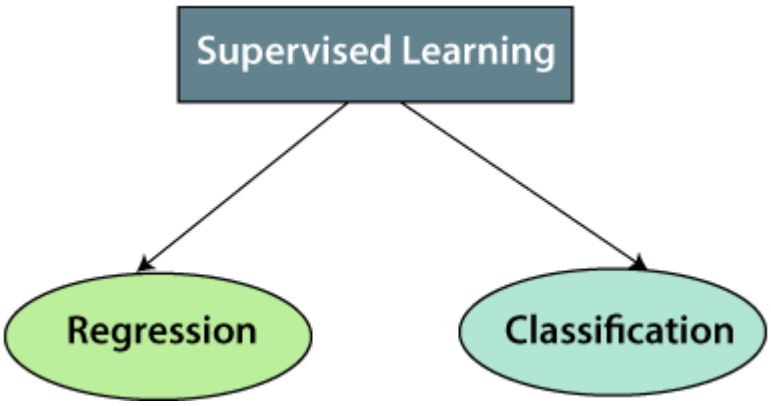

Figure 4. 2 Types of Supervised Learning

1. **Classification** − Classification-based tasks predict the categorical output responses or labels for the given training data. This output, which will belong to a specific discrete category, is based on what our ML model learns in the training phase.

2. **Regression** − Regression-based tasks predict the continues numerical output responses or labels for the given training data. This output will also be based on what our ML model learns in the training phase.

We used Classification technique in which the algorithm learns from the input given to it and then uses this learning to classify new observation. We will use the following algorithms:

1. Random Forest
2. Decision Tree
3. Naive Bayes
4. Logistic Regression
5. Bagging Decision Tree
6. Bagging Naive Bayes
7. Bagging Logistic
8. ADABoost
9. Extremely Randomized Trees
10. Gradient Tree Boosting
11. Histogram-Based Gradient Boosting

4.1 Random Forest Classifier

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression.

Random Forest is a classifier that uses many decision trees on different subsets of the input dataset and averages the results to increase the dataset's predicted accuracy. Instead of depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. More trees in the forest result in increased accuracy and mitigate the overfitting issue [27].

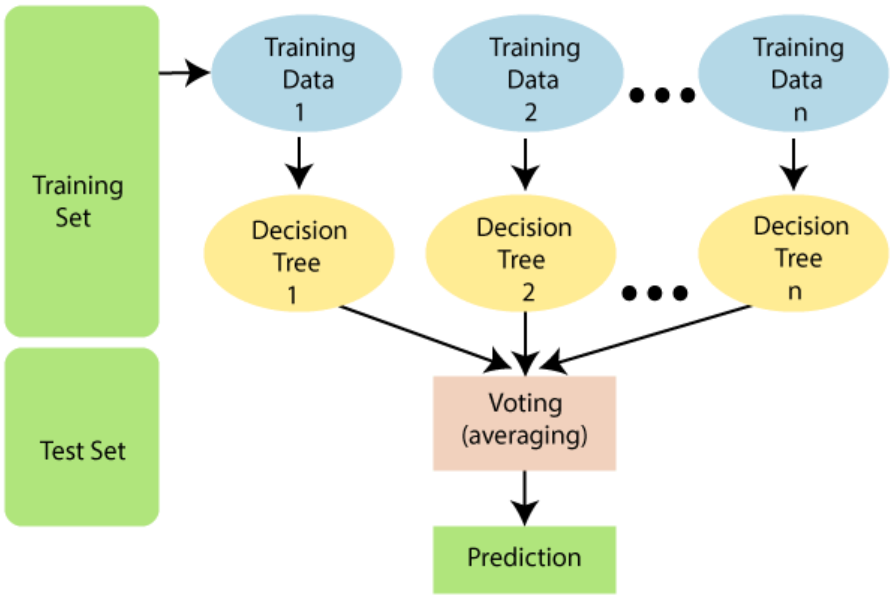

Figure 4. 3 Random Forest Classifier

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

4.2 Decision Tree

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. Given certain parameters, it provides a graphic representation of all potential answers to an issue or decision [28].

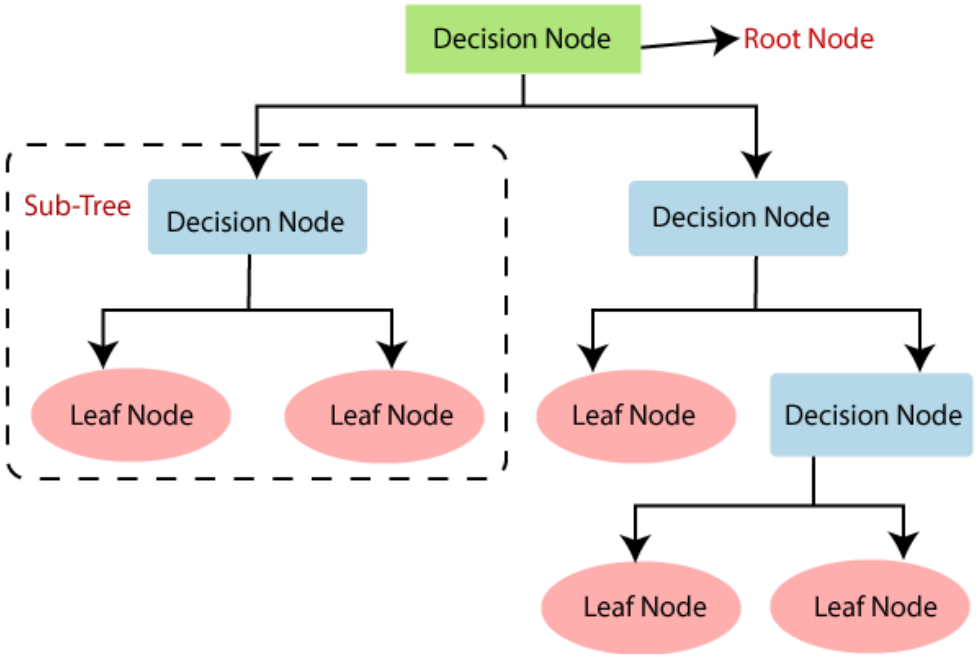

Figure 4. 4 Decision Tree Classifier

Step-1: Begin the tree with the root node, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

Step-3: Divide the root node into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

4.3 Naïve Bayes

The Naïve Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. The Naïve Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. Because it is a probabilistic classifier, it makes predictions based on the likelihood that an object exists [30].

The Working of Naïve Bayes' Classifier is given below:

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

Bayes' Theorem:

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{\mathrm{P(B|A)} * \mathrm{P(A)}}{\mathrm{P(B)}}$$

where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Gaussian Naïve Bayer and Bernoulli Naïve Bayes are the types of Naïve Bayes Classifier.

4.3.1 Gaussian Naïve Bayes

Continuous values connected to each feature in Gaussian Naive Bayes are presumptively distributed in a Gaussian manner. Normal distribution is another name for a Gaussian distribution. Plotting it results in the bell-shaped curve below, which is symmetric about the mean of the feature values:

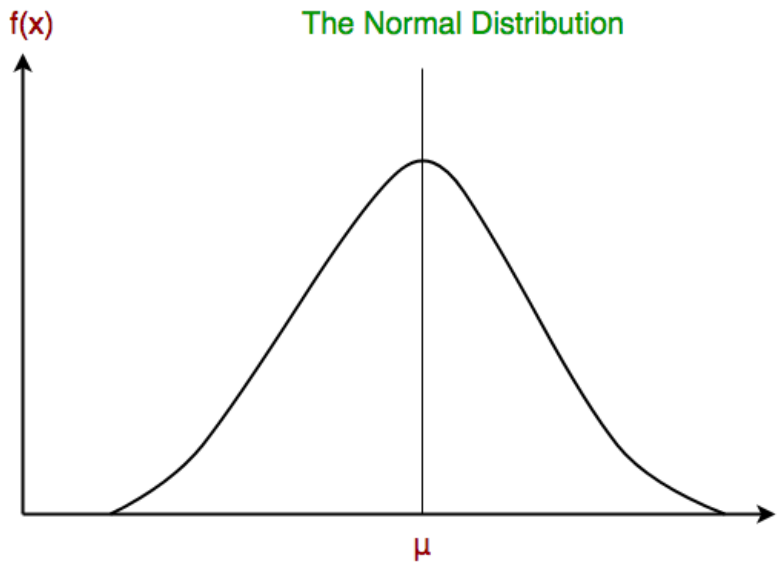

Figure 4. 5 Gaussian Distribution (Normal Distribution)

4.3.2 Bernoulli Naïve Bayes

The predictors in Bernoulli Naive Bayes are Boolean variables. Given the class name, this algorithm operates under the "naive" presumption that all features are independent of one another. Given a collection of inputs, each of which is represented as a binary feature (a feature with just two potential values), The algorithm computes the chance that each input belongs to each class using the Bayes theorem, and then classifies the input into that class with the highest probability.

4.4 Logistic Regression

One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. It can be either True or False, Yes or No, 0 or 1, etc., but rather than providing the exact values of 0 and 1, it provides probabilistic values that fall between 0 and 1 [31].

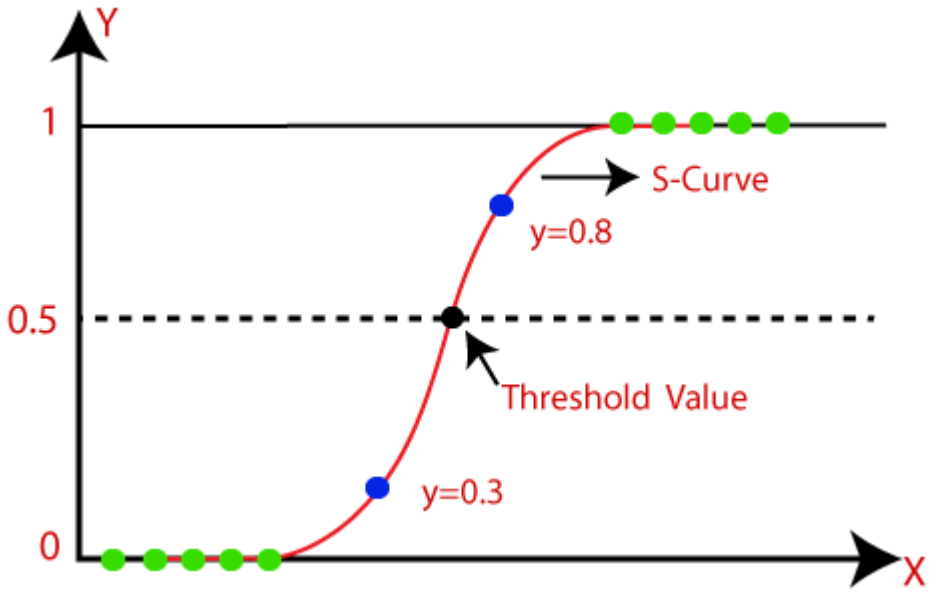

Figure 4. 6 Logistic Regression

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

4.5 Ensemble Classifier

An ensemble classifier is a machine learning model that enhances performance by combining the predictions of various base models [32]. Ensemble classifiers work by combining the advantages of various base models to compensate for the limitations of each one alone. Two primary categories of ensemble classifiers exist. They are given below:

4.5.1 Bagging

Bootstrap Aggregating is referred to as bagging. Using this technique, various base model iterations are created, each trained using a different random subset of the training data. Using a majority vote or averaging procedure, the combined forecasts of all the basis models get the final prediction [32].

4.5.2 Boosting

Boosting is an ensemble technique that adapts the training instance weights. This strategy involves training the basic model sequentially, with each stage focusing on the instances that the prior models misclassified. A weighted voting approach is used to combine the predictions of all the base models, with the weights being based on the misclassification rate of each base model [32].

- **Bagging Decision Tree**

By combining numerous Decision Tree models that have been trained on various subsets of the data, Bagging Decision Tree classification is an ensemble technique that enhances prediction accuracy and stability by minimizing overfitting.

- **Bagging Naïve Bayes**

An ensemble technique called bagging Naive Bayes classification combines numerous Naive Bayes models that have been trained on various subsets of the data to increase the predictions' overall accuracy and stability.

- **Bagging Logistic**

In order to increase the overall stability and accuracy of the predictions, the ensemble method known as bagging logistic classification mixes numerous logistic regression models that have been trained on various subsets of the data.

- **ADABoost**

AdaBoost (Adaptive Boosting) classification is an ensemble method that combines multiple weak classifiers to create a strong classifier with improved accuracy and stability by adjusting the weights of incorrectly categorized observations.

- **Gradient Tree Boosting**

To increase the overall accuracy and stability of the predictions by eliminating overfitting, the Gradient Tree Boosting classification ensemble approach combines numerous decision trees using the gradient descent optimization algorithm.

- **Histogram-Based Gradient Boosting**

Histogram-based Gradient Boosting classification is a tree-based ensemble method that enhances prediction accuracy and minimizes overfitting by using histograms to estimate the decision tree leaf value rather than a single scalar value.

4.5.3 Extremely Randomized Tree

Extremely Randomized Trees Classification is an ensemble approach that combines a number of decision trees created using random feature and threshold selections in order to reduce overfitting and increase prediction accuracy and stability.

4.6 Evaluation Metrics

To understand classifier model's performance, we need to be familiar with some evaluation parameters. A confusion matrix is a table that is used to describe the performance of a classifier algorithm by evaluating the accuracy of it. The elements of confusion matrix are:

True Positive (TP): Which results when classifier model correctly predicts the positive class. **True Negative (TN):** Which results when classifier model correctly predicts the negative class. **False Positive (FP):** Which results when classifier model incorrectly predicts the positive class. **False Negative (FN):** Which results when classifier model incorrectly predicts the negative class.

Based on the data of confusion matrix, precision, recall, F-measure, and accuracy are the evaluation measures used for evaluating performance of classifier [33]

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{.....(1)}$$

Recall: The recall is calculated as the ratio between the numbers of positive samples correctly classified as positive to the total number of positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples are detected.[19]

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \text{..... (2)}$$

Accuracy: For binary classification, accuracy can be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{..... (3)}$$

F1 Score: F1 Score is the harmonic mean of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$\text{F1 Score} = \frac{2*(\text{Recall}*\text{Precision})}{(\text{Recall+Precision})} \quad \text{..... (4)}$$

CHAPTER FIVE

Blanced Vs Imbalanced Datatset

A balanced dataset is one in which the number of samples belonging to each class is roughly equal. For example, in a binary classification task with two classes, a balanced dataset would have roughly the same number of samples labeled as class 0 and class 1. Then we can say our dataset in Imbalance Dataset [34].

Figure 5. 1 Balanced Dataset

Consider Orange color as a positive value and blue color as a Negative value. We can say that the number of positive values and negative values in approximately same.

An imbalanced dataset is one in which the number of samples belonging to one class is significantly different from the number of samples belonging to the other class. If there is the very high different between the positive values and negative values [34]. For example, in a binary classification task, an imbalanced dataset would have many more samples labeled as class 0 than class 1. Then we can say our dataset in Imbalance Dataset.

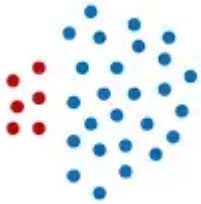

Figure 5. 2 Imbalanced Dataset

Consider red color as a positive value and blue color as a Negative value. We can say that the number of positive values and the negative value ratio are not the same.

CHAPTER SIX

Research Methodology

6.1 Data Analysis

6.1.1 VHS-22 Dataset

In this paper, we have worked with VHS-22 dataset created by a group of researchers from Warsaw University of Technology in 2022 [12]. The datasets used to create the VHS-22 dataset are shown in the below diagram. There are 27.7 million flows (20.3 million legitimate and 7.4 million attacks) in it. The flows are represented by 45 features, including network-level features and statistical data in addition to the traditional NetFlow features. This makes the VHS-22 dataset more effective in detecting network threats.

VHS-22 is a heterogeneous, flow-level dataset which combines ISOT, CICIDS-17, Booters and CTU-13 datasets, as well as traffic from Malware Traffic Analysis (MTA) site, to increase a variety of malicious and legitimate traffic flows.

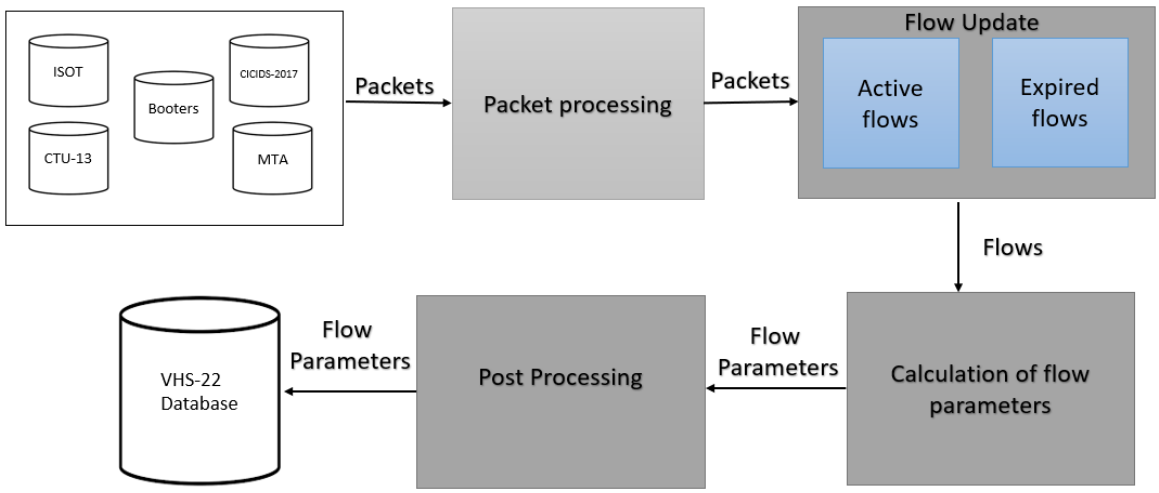

Figure 6. 1 Creation of VHS-22 Dataset

A dataset that is made up of different or various types of data is said to be heterogeneous. This means that the data in the dataset may come from different sources, be in different formats, or have different characteristics.

As we have seen, VHS-22 is a dataset which combines four individual datasets as well as traffic from Malware Traffic Analysis (MTA) together. These attacks of datasets are contained with different - attacks with legitimate traffic.

Table 6. 1 DIFFERENT ATTACKS IN VHS-22

Attack type	Source dataset	Starting date
botnet	ISOT	2022-01-01
various	MTA	2022-01-02
webattacks	CICIDS-17	2022-01-03
bruteforce	CICIDS-17	2022-01-04
botnet	CICIDS-17	2022-01-05
DDoS	CICIDS-17	2022-01-06
DDoS	Booters	2022-01-07
botnet	CTU-13	2022-01-12

We can see that we have collected several attack types from various sources. The data has various characteristics and is presented in various formats. This indicates that the data is said to be heterogeneous.[12]

Types of Attack:

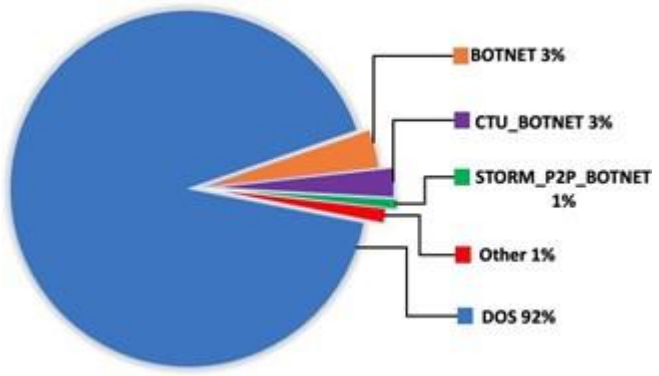

Figure 6. 2 Types of Attacks

Here in Fig. 1, we can observe that the VHS-22 dataset contains approximately 92% of DOS attack, 3% Botnet attack, 3% of CTU Botnet attack, 1% STORM P2P Botnet attack and the remaining 1% attacks contain Web attack, Brute Force, SSH Brute Force and other 108 types of attacks. Our target is to detect these versatile attacks using machine learning algorithms.

Label:

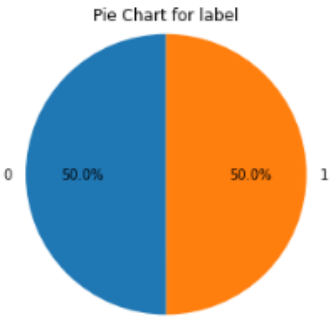

Figure 6. 3 Attack Vs Normal Data Ratio from Label

The graph in this case shows that the ratio is 50:50. We kept the ratio of attack to normal data at 50:50 to maintain the target label balance in our dataset. Additionally, 1 is labeled as attack data and 0 is labeled as normal data.

Timestamps:

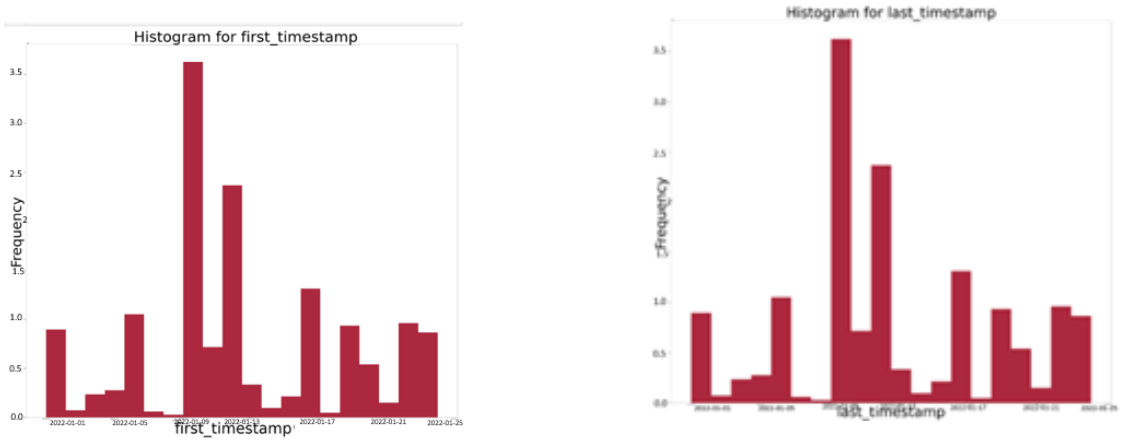

Figure 6. 4 First Timestamp And Last Timestamp.

In figure 6.4, we have two related features called First Time Stamp and Last Time Stamp that contain the beginning and ending times of the network flow. It's because every network transaction took place on the same day and at the same hour. Because of this, the data in these two graphs is similar.

6.1.2 Network traffic parameters

Parameter	Description
ip_src_str	source IP address
ip_dst_str	destination IP address
ip_protocol	fourth layer protocol
sport	source port
dport	destination port
in_packets	sum of packets in flow
b_packet_total	sum of bytes in flow
first_timestamp	first timestamp in flow
last_timestamp	last timestamp in flow
duration	duration of flow
flags_sum	sum of TCP flags (URG-32, ACK-16, PSH-8,RST-4, SYN-2, FIN-1)
urg_nr_count	URG flag count in flow
ack_nr_count	ACK flag count in flow
rst_nr_count	RST flag count in flow
fin_nr_count	SYN flag count in flow
psh_nr_count	PSH flag count in flow
syn_nr_count	SYN flag count in flow
b_packet_max	size of the largest packet
b_packet_min	size of the smallest packet
b_packet_mean	mean packet size

b_packet_median	median packet size
b_packet_first_q	1^{st} quantile of packet size
b_packet_third_q	3^{rd} quantile of packet size
b_packet_std	std. deviation of packet size
b_packet_total	total size of flow in bytes
iat_min	lowest IAT
iat_max	highest IAT
iat_first_q	1^{st} quantile of IAT
iat_third_q	3^{rd} quantile of IAT
iat_std	standard deviation of IAT
iat_mean	mean of IAT
iat_median	median of IAT
iat_var	variance of IAT

connections_from_this_host	No. of connections from given host
connections_to_this_host	No. of connections to given host
connections_rst_to_this_host	No. of connections to host ended with RST flag
connections_rst_from_this_host	No. of connections from host ended with RST flag
connections_to_this_port	No. of connections with the same destination port number
connections_from_this_port	No. of connections with the same source port number
connections_ratio_from_this_host	% of connections to the host with the same destination address
connections_ratio_to_this_host	% of connections from the host with the same source address
connections_ratio_rst_to_this_host	% of connections to host ended with RST flag
connections_ratio_rst_from_this_host	% of connections from host ended with RST flag
connections_ratio_to_this_port	% of connections to host with the same destination port
connections_ratio_from_this_port	% of connections from host with the same source port

6.1.3 Existing Network Traffic Datasets

An ideal dataset for network threat detection research should be relatively up to date, correctly labeled, and possibly contain real network traffic with various kinds of attacks. Each of the five datasets includes genuine traffic from everyday activities as well as network threats. Therefore, we can say that it may be advantageous to combine different dataset types that contain both simulated and actual traffic.[12]

Table 6. 2 LIST OF DATASETS USED TO CREATE VHS-22

Dataset Name	Description
ISOT [39]	Blend of four other datasets, and contains, among other things, nor- mal traffic generated by a variety of everyday activities.
CICIDS-2017[40]	Web attacks, Web scanning, DoS and DDoS attacks, botnet traffic, brute force attacks, infiltration, and heart bleed vulnerability usage.
CTU-13 [9]	It includes non-peer-to-peer traffic, which is uncommon among other datasets.
Booters [41]	These are representative examples of DDoS attacks because each dataset in Booters has over 100 attack logs.
MTA-Traffic Samples [42]	The MTA website's traffic is recent traffic produced by real malware samples.

6.2 MiniVHS-22

VHS-22 dataset contains 27.7 million flows, among which 20.3 million are normal flow and 7.4 million are attack flow [12]. 27.7 million is a huge amount of data to work with, and evidently needs a great amount of computational power. To overcome this situation and further data processing, we constructed the MiniVHS-22 dataset from the original VHS-22 dataset in order to reduce the computational burden for our model training and evaluation. We aimed to create a smaller version of VHS-22 dataset which would have 1 million(M) data but also should have all the same characteristics of VHS-22 dataset. To that end, we employed stratified sampling: we calculated the ratio of normal and attack data in the original dataset and maintained the same ratio in the MiniVHS-22 dataset of 1M data. Note that, attack data contains 115 different types of attacks. We calculated the ratios of each attack type of VHS-22 dataset and maintained the same ratio in the new MiniVHS-22 dataset. We have randomly selected 0.73M normal flow data from 20.3M normal flow data of VHS-22 in the MiniVHS-22 dataset.

Table 6. 3 VHS-22 and MiniVHS-22

Dataset Name	Total Data	Normal Data	Attack Data
VHS-22	27.7 M	20.3 M	7.40 M
MiniVHS-22	1M	0.73 M	0.27 M

Due to our stratified sampling approach, the results that we obtain with the ML algorithms on the MiniVHS-22 dataset should serve as an excellent estimate of the same on the original VHS-22 dataset

The advantages of MiniVHS-22:

1. It's a smaller version which contains less data compared to VHS-22
2. It contains the same characteristics as VHS-22 dataset.
3. We will need less computational power to simulate MiniVHS-22 dataset.
4. The result of MiniVHS-22 is a estimation of VHS-22 dataset.
5. Any researcher can utilize the MiniVHS-22 dataset and the methods to anticipate the results of the VHS-22.

6.3 Data Preprocessing

6.3.1 Balanced Dataset

The VHS-22 dataset contains 27.7 million flows (20.3 million legitimate and 7.4 million attacks) in it which is why the dataset is said to be imbalanced. If a dataset is imbalanced, it means that there is a disproportionate ratio of samples in different classes. Thus, the model may be more likely to predict the majority class, resulting in poor performance on the minority class. This can lead to a high number of false negatives, where the model fails to identify the minority class examples.

Then, to balance the dataset and create a balanced ratio of normal
and attacks, we applied the under-sampling technique. This technique involves removing data from the majority class to make it more similar in size to the minority class. Now, our dataset contains 14.8 million flows (7.4 million normal traffic and 7.4 million attacks).

6.3.2 Feature Extraction

Feature extraction is a process of extracting relevant information from raw data and converting it into a set of features or characteristics that can be used as input for machine learning algorithms. The goal of feature extraction is to reduce the dimensionality of the data while preserving the most important information. This can help improve the performance and efficiency of machine learning models, as well as make the data more understandable to humans [35].

In our paper, we manually extracted the features. The weekday of the attack and the hour of the attack were extracted as features. In order to extract the feature, we used the object formatted timestamp feature. We converted it to date-time format in order to use the timestamp capability. After that, we manually removed the year, month, minute, and seconds while keeping the weekday and hour of the day.

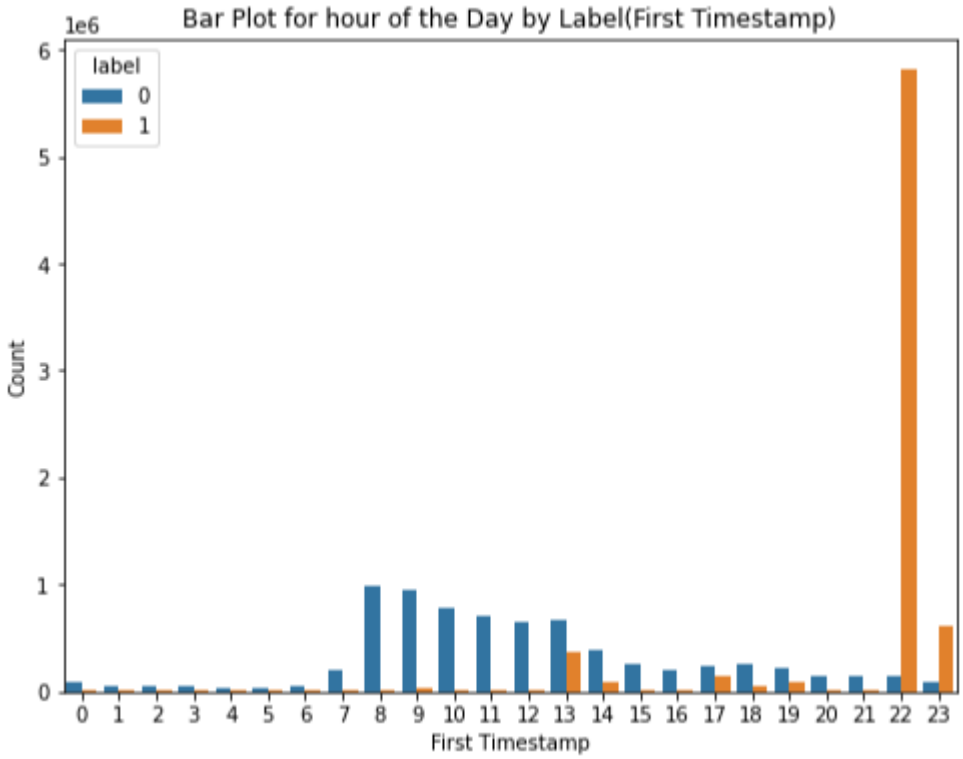

Figure 6. 5Flow of attack and normal data of a day

In the graph above, 0 denotes normal data, and 1 denotes attack data. The network flow is shown in this graph for each hour of the day. The attack network flow increases on the x-axis at 22nd and 23rd hours (11:00 and 11:59 p.m.), and normal data continues to increase throughout the remainder of the day. On the other hand, in 13th hour (1:00pm) we see a peak value of attacks.

In our analysis, the first timestamp and the last timestamp show the same kind of data for both timestamps. Thus, the graph of the attack and the normal data of the day will be similar. For this reason, we only displayed the first timestamps graph, which also represents the last timestamp graph.

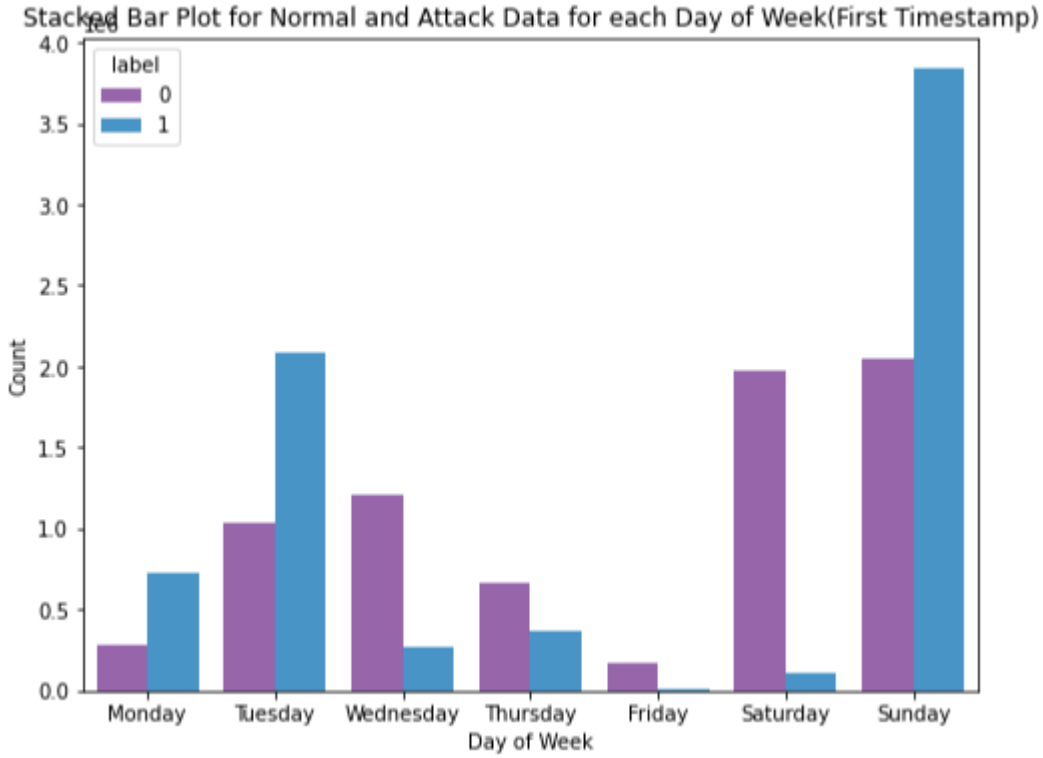

Figure 6. 6 Flow of attack and normal data of a week

The numbers 0 and 1 in the graph above represent normal and attack data, respectively. This graph demonstrates the network flow for each day of the week. The graph shows that the network flow of attacks is greatest on Sunday, Monday, and Tuesday while the network flow of normal data is greatest the rest of the week.

In our analysis, the first timestamp and the last timestamp show the same kind of data for both timestamps. Thus, the graph of the attack and the normal data of the weekday will be similar. For this reason, we only displayed the first timestamps graph, which also represents the last timestamp graph.

6.3.3 IP Encoding

IP encoding (also known as IP address encoding) is a technique used to represent IP addresses (Internet Protocol addresses) as numerical values. IP addresses are used to identify and locate devices connected to a network, such as computers and servers.

We have both normal data and attack IP addresses in our dataset. Before feeding our model, we tried to apply label encoding on IP addresses. However, when the model is tested, it does not produce satisfactory encoding results. It's because IP addresses are dynamic and don't have a set or similar quantity. As a result, the model is receiving data that it has never seen before while testing data. Due to

this, we decided to remove the IP address feature because an IP address can be any number and has no restrictions.

6.3.4 Feature Selection

Feature Selection is a technique which is used to improve the accuracy and interpretability of machine learning models by removing irrelevant, redundant, or highly correlated features. It can decrease the dimensionality of the data and decrease the computational cost and memory required to train the model [36].

6.3.4.1 Feature Importance

Scikit-learn has a built-in attribute (feature_importances_) that returns an array of values, with each value representing the measure of importance of a feature. The higher the importance value, the more significant that feature is. Figure 6.7 illustrates the importance for all of the extracted features.

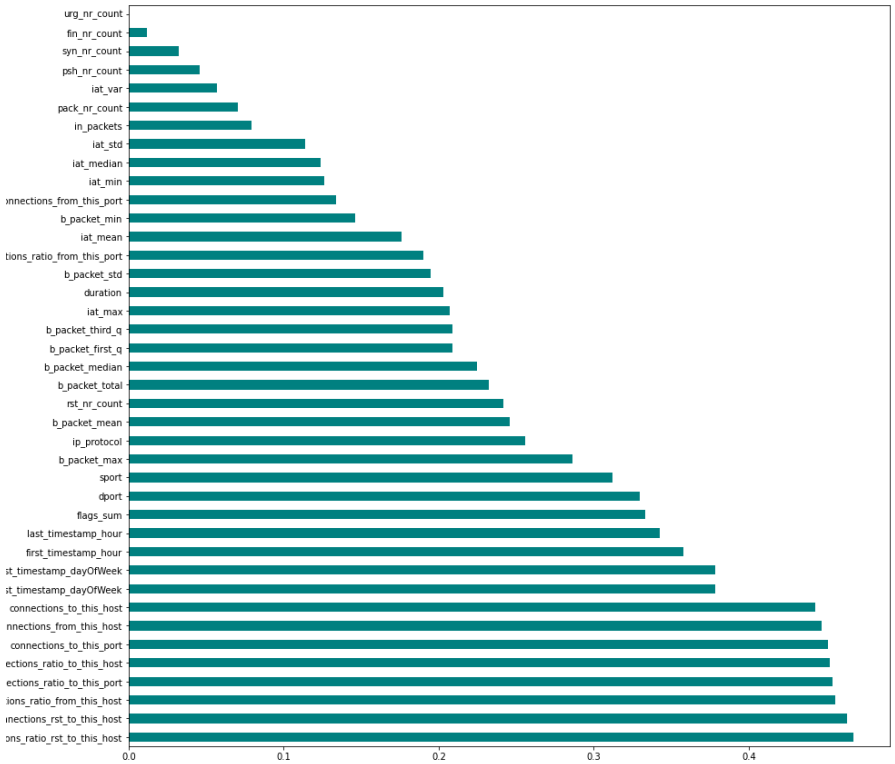

Figure 6. 7 Feature Importance

From Figure 6.7, we can conclude that the most important features based on the graph are connections_ratio_rst_to_this_host, connections_rst_to_this_host, connections_ratio_from_this_host, connection_ratio_to_this_port, connection_ratio_to_this_host.

6.3.4.2 Feature Correlation

In order to avoid using redundant features, which may contribute toward overfitting our model to the training data, we intend to investigate the feature correlation matrix and measure how each feature relates to other features. Feature correlation matrix is useful to identify how each feature relates to one another.

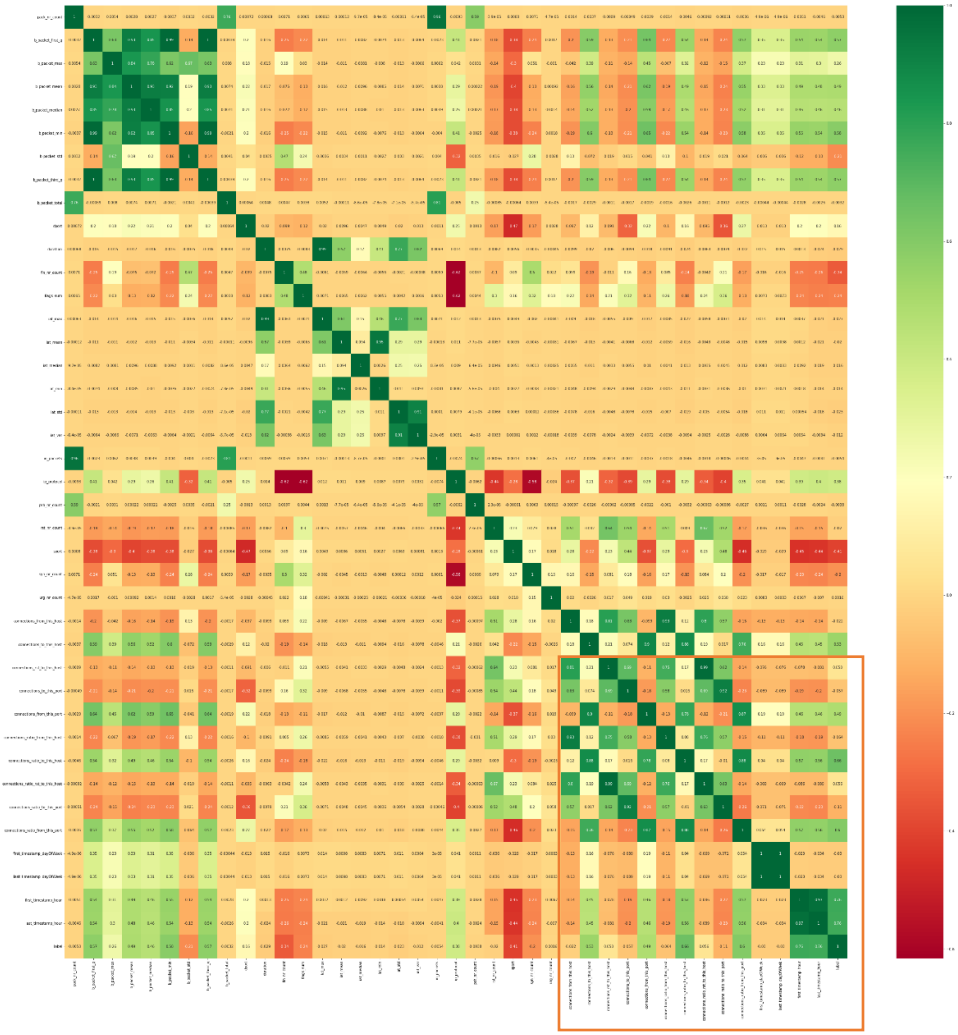

Figure 6. 8 Correlation between a subset of the features

VHS-22 dataset contains 45 features. Visualizing all extracted features in a single correlation matrix for better visualization. To that end, we intend to choose the features with the strongest association with the target label for presenting in the correlation matrix. In Figure 6.8, we observe that features ‘ip protocol (fourth layer protocol)’, ‘dport (destination port)’, ‘connections to this port (No. of connections with the same destination port number)’, and ‘connections ratio rst to this host (percent of connections to host ended with RST flag)’ are highly correlated with ‘label’. Feature selection offers a number of advantages, including faster classification models by eliminating irrelevant or noisy features [17]. In this work, we dropped 8 features from the dataset based on whether there is variation in the data of that feature, whether the feature is made of non-numerical and/or non-categorical data.

6.3.4.3 Variance threshold

Variance threshold is a method used in feature selection to remove features with low variance from a dataset. It is based on the idea that features with low variance do not contain much information and can be removed without significantly affecting the performance of a model. This method can be used as a preprocessing step before training a machine learning model to improve its performance by removing unnecessary features and reducing the complexity of the model.

We discovered that the variance threshold only detects low variance for 4 features when applied to our dataset. These features are

- iat_first_q
- iat_third_q,
- Connections_rst_from_this_host
- Connections_ratio_rst_from_this_host

To increase the accuracy, we deleted these following features of low variance.

6.3.4.4 Handling Non-Numeric Data

Dropping non-numeric data is a technique used in data preprocessing to remove non-numeric columns (also known as categorical or string data) from a dataset. This is often done because many machine learning algorithms are designed to work with numerical data, and they may not be able to handle non-numeric data.

Dropping non-numeric data can also help to reduce the dimensionality of the dataset, which can make it easier to work with and improve the performance of machine learning models. We manually dropped some non-numeric data which has no use in our model and the importance of these data is very low. These data are

- Attack_label
- Attack_file
- Last_timestamp
- Ip_src_str
- Ip_dst_str
- First_timestamp
- Last_timestamp_dayNameOfWeek
- First_timestamp_dayNameOfWeek

6.4 Feature Scaling

Feature scaling is the process of normalizing or standardizing the features of a dataset. This is done to ensure that all features are on a similar scale and have similar properties, which can improve the performance of machine learning algorithms. It is very difficult for machine learning algorithms to compute with different ranges of values as it is time-consuming, the predictions are less accurate, and the chance of overfitting. One way to handle this situation is to scale each element of these applicable features. In our case, we have applied standardization feature scaling as the following:

$$\mathrm{X}_{\mathrm{stand}} = \frac{\mathrm{x} - \mathrm{mean}(\mu_{\mathrm{x}})}{\text{standard deviation}(\sigma_{\mathrm{x}})}$$

We applied the standardization feature scaling technique to all the features except the target label.

6.5 Dimensionality Reduction:

Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much of the relevant information as possible. High-dimensional data can be difficult to visualize and interpret. Dimensionality reduction can help to make the data more interpretable by reducing the number of features and making patterns and relationships in the data clearer.

In our previous work on MiniVHS-22, we applied the two techniques called Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).In addition, without PCA and LDA, the machine learning algorithms produce the best accuracy and F1-score of all evaluation measures. As a result, we saw that any PCA or LDA values for the machine learning algorithms did not reach the value of the F1 score. The result part of MiniVHS-22 displays the outcomes of PCA and LDA [37].

6.6 Splitting Training and Testing data

For the evaluations of our datasets, we divide it into two sets of data: training set and testing set. For training the machine learning model, we use 70% of the comprises data and for testing, we use 30% of the comprises data. So, The VHS-22 dataset consists of 10.3 million data for training and 4.4 million data for testing.

6.7 Model Training and Evaluation Technology

We divided the VHS-22 data into training and testing subsets in the proportion of 70:30. For building a model we have used Random Forest, Gaussian Naïve Bayes, Bernoulli's Naïve Bayes, Decision Tree (DT), and Logistic Regression classifier. Besides that, we have also used ensemble classifier. In order to select relevant features, we used feature extraction, variance threshold and handled non-numeric data manually. We compared the evaluation metrics scores such as precision, recall, f1 score and accuracy. For implementation of these classifiers, we used the scikit-learn library and Python 3. The classification results are described in the next section.

CHAPTER SEVEN

Result and Analysis

7.1 Result of MiniVHS-22

In our paper, Table 7.1 represents the results of initial attack detection experiments run on our MiniVHS-22 dataset. From the table, we observe that Random Forest classifier obtained the maximum F1-Score at a level of 99.6% without applying the dimensionality reduction method to it. Decision Tree, KNN, and MLP also obtained higher F1-score at levels of 99.5%, 99.2%, and 99.1% respectively without applying the dimensionality reduction method.

Table 7. 1 PERFORMANCE OF THE CLASSIFICATION MODELS ON MiniVHS-22 (PCA=PCA-20)

Classifier Name	Dimensionality Reduction Method	TP Rate	FP Rate	Precision	Recall	Accuracy	F1 Score
Random Forest	None	0.9949	0.0019	0.9978	0.9949	0.9980	0.9964
	PCA	0.9887	0.0042	0.9963	0.9887	0.9960	0.9925
	LDA	0.9003	0.0367	0.9175	0.9003	0.9512	0.9088
Decision Tree	None	0.9960	0.0015	0.9951	0.9960	0.9976	0.9955
	PCA	0.9899	0.0037	0.9889	0.9899	0.9943	0.9894
	LDA	0.9022	0.0362	0.9034	0.9022	0.9475	0.9028
SGD	None	0.9133	0.0314	0.9729	0.9133	0.9697	0.9421
	PCA	0.9007	0.0357	0.9778	0.9007	0.9676	0.9377
	LDA	0.9056	0.0344	0.9452	0.9056	0.9603	0.9249
MLP	None	0.9909	0.0034	0.9917	0.9909	0.9953	0.9913
	PCA	0.9864	0.0050	0.9878	0.9864	0.9930	0.9871
	LDA	0.8930	0.0383	0.9838	0.8930	0.9671	0.9362
KNN	None	0.9929	0.0026	0.9923	0.9929	0.9960	0.9926
	PCA	0.9913	0.0032	0.9910	0.9913	0.9952	0.9912
	LDA	0.9066	0.0338	0.9662	0.9066	0.9662	0.9355
Naive Bayes	None	0.9894	0.0061	0.5024	0.9894	0.7323	0.6664
	PCA	0.9689	0.0340	0.3476	0.9689	0.5002	0.5117
	LDA	0.9124	0.0322	0.9338	0.9124	0.9588	0.9230
Bernoulli Naive Bayes	None	0.9065	0.0344	0.9229	0.9065	0.9543	0.9146
	PCA	0.7822	0.0778	0.8702	0.7822	0.9096	0.8238
	LDA	0.9237	0.0295	0.8265	0.9237	0.9270	0.8724
Logistic Regression	None	0.9170	0.0303	0.9573	0.9170	0.9665	0.9367
	PCA	0.9082	0.0335	0.9476	0.9082	0.9616	0.9275
	LDA	0.9069	0.0340	0.9435	0.9069	0.9602	0.9248

In our previous work, in table 7.1, we recorded the results of the evaluation metrics without reducing the dimension of the dataset and applying dimensionality reduction methods PCA and LDA. From the table, we observe that Random Forest classifier obtained the maximum F1-Score at a level of 99.6% without applying the dimensionality reduction method to it. But when we reduced the dimension of the dataset using Principal Component Analysis (PCA) with 20 components and Linear Discriminant Analysis (LDA) and then, applied the classifiers to the reduced dataset, we can notice that the F1-Score is obtained from Random Forest classifiers is at the level of 99.2%. When dimensionality reduction techniques are used, the result of the F1 Score in the Random Forest Classifier is lower than the result of the F1 Score in the Random Forest Classifier without using dimensionality reduction techniques. The able 7.1 reveals that LDA consistently produces the worst outcomes.

7.1.1 Comparison of F1 Score and Accuracy of Random Forest Classifier with different PCA values

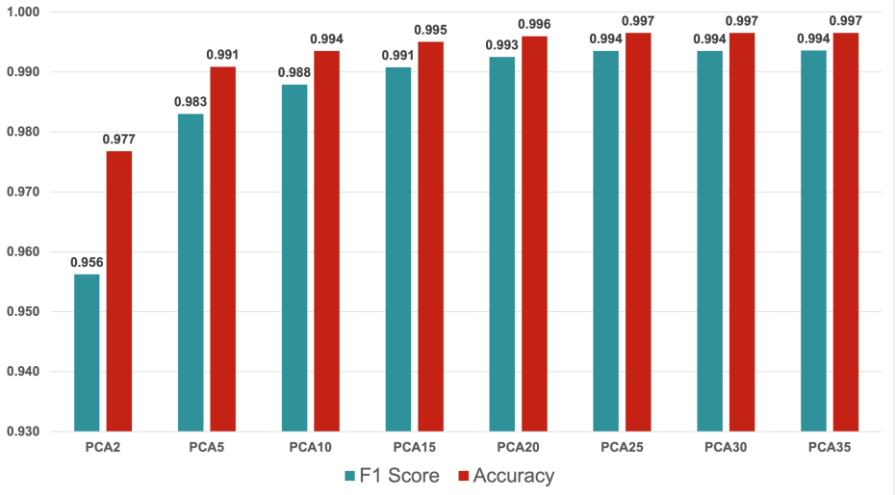

Figure 7. 1 Results of Evaluation Metrics for Random Forest classifier with varying numbers of principal components for PCA.

In the case of PCA, we have recorded values of the Precision, Recall, Accuracy, and F1-Score for different numbers of principal components in Table 7.1. Then, we compared the values of Accuracy and F1-Score in Fig. 7.1. According to the graph, the PCA with 35 principal components produces a better outcome than the others. From PCA 2 to PCA 35, we experimented with a range of PCA component values. By examining the F1 Score, we can observe from the graph that any PCA values for the Random Forest Classifier do not exceed the value of the F1 Score in the Random Forest Classifier without using the dimensionality reduction method. However, the dataset without PCA and LDA yields the best accuracy and F1-score of all evaluation measures.

7.2 Result of VHS-22

In our paper, table 7.2 represents the performance of various machine learning classifiers on VHS-22 dataset.

Table 7. 2
PERFORMANCE OF THE CLASSIFICATION MODELS ON VHS-22

Classifier Name	TP Rate	FP Rate	Precision	Recall	F1 Score	Accuracy
Decision Tree	0.997893529	0.002103063	0.997756321	0.997893529	0.99782492	0.997826681
Regular Random Forest	0.997380437	0.002610808	0.998973983	0.997380437	0.998176574	**0.998179629**
Gaussian Naïve Bayes	0.992956937	0.023235059	0.584561286	0.992956937	0.735895377	**0.643951895**
Bernoulli Naive Bayes	0.825862619	0.153785263	0.949903646	0.825862619	0.883550868	0.891249605
Logistic Regression	0.937854582	0.061567778	0.945059513	0.937854582	0.941443263	0.941717717
Bagging Decision Tree	0.997643295	0.002349977	0.998746545	0.997643295	0.998194615	**0.998197197**
Bagging Naive Bayes	0.992976775	0.023157935	0.58462557	0.992976775	0.735951762	0.644048071
Bagging Logistic	0.937821669	0.061588611	0.945210148	0.937821669	0.941501413	0.941781234
ADABoost	0.96313224	0.036133244	0.981355614	0.96313224	0.972158534	0.972441266
Extremely Randomized Trees	0.997399373	0.002592547	0.998755708	0.997399373	0.99807708	0.998080074
Gradient Tree Boosting	0.940861895	0.058652684	0.947034474	0.940861895	0.943938094	0.94416988
Histogram-Based Gradient Boosting	0.992479013	0.007462436	0.998557901	0.992479013	0.995509177	0.995526779

From the table 7.2 and figure 7.2, we can see that Gaussian Naïve Bayes performs comparatively worse than other classifiers with accuracy of 64.3951% and f1 score of 88.3550%, recall score of 82.5862% and lastly the precision of 94.9903%. Also, we can state that, the performance of Bagging Naïve Bayes with accuracy of 64.4048% and f1 score of 73.5951%, recall score of 99.2976 and lastly the precision of 58.4625% is not satisfactory.

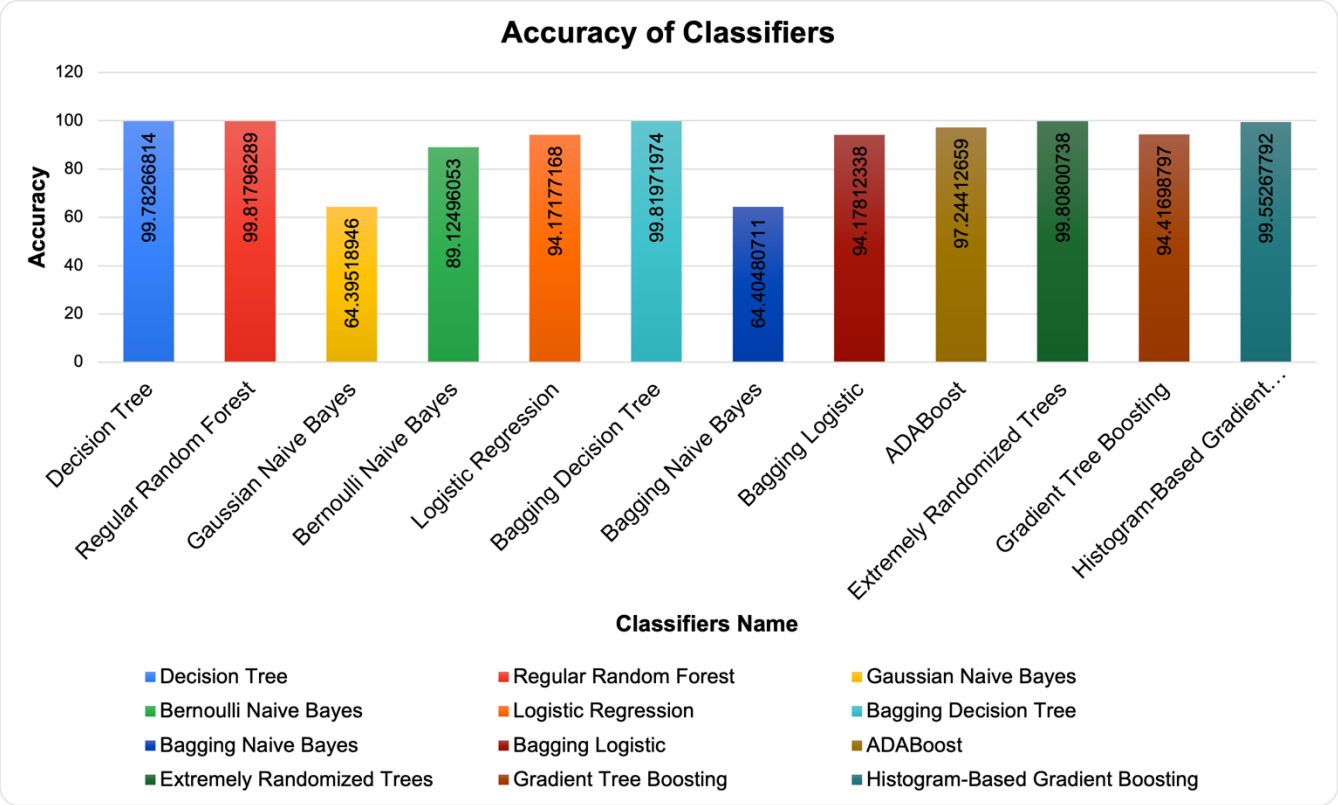

Figure 7. 2 Chart of the Accuracy of the different classifiers in %

On the other hand, we can see that Bagging Decision Tree worked better than all other classifiers we experimented with the accuracy of 99.8197% and f1 score of 99.8194%, recall score of 99.7643% and lastly the precision of 99.8746%. It is also noticeable that Regular Random Forest also performed well with accuracy of 99.8179% and f1 score of 99.8176%, recall score of 99.7380% and lastly the precision of 99.8973%.

7.2.1 Applying Confusion Matrix for Machine Learning Algorithms

Here, we have applied the confusion matrix on our machine learning classifier models to observe the performance of the models which have achieved best accuracy and worst accuracy compared to other classifiers.

Figure 7. 3 Confusion Matrix of Bagging Decision Tree

The confusion matrix of Bagging Decision Tree shows that the classifier has correctly predicted the True positive class and True negative class. Also, we can say that the classifier incorrectly predicted the False positive and False negative class with minor errors for the dataset. Based on these figures, we can see that we have created acceptable machine learning models.

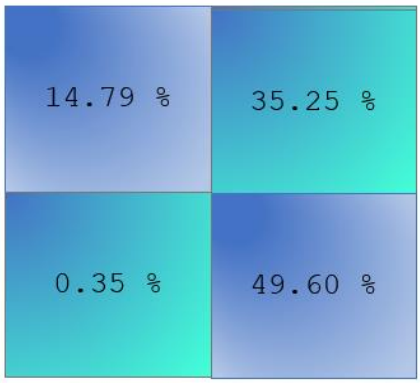

Figure 7. 4 Confusion Matrix of Gaussian Naïve Bayes

The confusion matrix of Gaussian Naïve Bayes shows that the classifier has incorrectly predicted the True positive class with major errors and almost correctly predicted True negative class. Also, we can say that the classifier predicted the False positive and False negative class with major errors for the dataset. Based on these figures, we can say that Gaussian Naïve Bayes performs worst in this case.

CHAPTER EIGHT

Machine Learning

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data as input and algorithms to imitate the way humans learn and gradually becomes more accurate at predicting the outcomes [38]. Machine learning has applications in deferent area like Natural Language processing [43][44][45][46], medical [47], cyber security[48][49][50][51].

The advantages and challenges of machine learning and deep learning are given below:

8.1 Advantages of Machine Learning

- One of most important advantages of machine learning is it able to discover specific trends and patterns by analyzing large amount of data. Often, these patterns would not be apparent to human eyes. In most cases, a well build machine learning model can accurately identify those patterns and trends. For example, e-commerce companies like Amazon, Alibaba, eBay uses machine learning techniques to understand the browsing behaviors and purchasing histories of its user and predicts which products and deals that the users may find useful and push those products and deals.
- Machine learning models are capable of continuous accuracy improvement with the increasement of refined data as it trains more data. Internet companies like Amazon, Walmart collects huge volume of refined data every day and by training those data it improves the accuracy of recommended products and deals to its customers.
- An important feature of machine learning is the ability of automation on various decision-making tasks as it automatically performs many time-consuming and repetitive tasks to help to improve the model without human intervention. A common example is anti-virus software, which learn to filters new threats as they recognize with the help of machine learning techniques.
- Now these days, machine learning is being used by every industry possible. From state defense to small retail shops, it is being used as it helps to generate profit by reducing cost, automation, less human resources, and the ability to analyze patterns from previous data. It is being used in real-life applications such as image and speech recognition, fake review detection, self-driving cars,etc.

8.2 Challenges of Machine Learning

- Data acquisition is one of the most important parts of machine learning, which can also be problematic. Data acquisition means collecting the data from a relevant source through surveys, real-life physical conditions, etc. before it can be used. In this process, there is a chance, it may contain imbalanced data, inaccurate data, or data full of errors. This can lead to poor accuracy in model building. Also, often time, to collect data, an organization have to pay for it. All this makes data acquisition a massive disadvantage.
- It is always important to remember that we need to provide well-cleaned data and apply feature engineering before we train it in the machine learning model. A dataset full of errors or imbalanced can causes incorrect results.
- There are different kinds of machine learning algorithms, and we need to identify which algorithm works best for the dataset. This is a manual process and also a disadvantage.
- When we process large volume of data in machine leaning model, the time complexity of that model increases so that we can have consideration amount of accuracy. It also needs massive amount of computing resources to process that dataset [38].

CHAPTER NINE

Conclusion

9.1 Research Challenges

There are a lot of challenges occur for detecting the cyber threats. From our point of view, dataset plays a major role for getting the better accuracy. In our previous work, we constructed the MiniVHS-22 dataset from the original VHS-22 dataset in order to reduce the computational burden for our model training and evaluation. We have applied the algorithms on the dataset which contains total 1 million attacks. After applying the machine learning algorithms on the dataset, Random Forest classifier obtained the maximum F1-Score without applying any dimensionality reduction techniques. But after applying dimensionality reduction techniques, we got less result of F1 Score in Random Forest Classifier. The best outcome of all comes from our work with VHS-22 as our extension project. Without utilizing the dimensionality reduction techniques, we achieved the highest accuracy. However, the dataset without PCA and LDA yields the best accuracy and F1-score of all evaluation measures. Besides that, there are some other challenges occur while detecting the cyber threats such as: there were 27.7 million flows where 20.3 million legitimate and 7.4 million of attacks. It is difficult to distinguish the prediction of attack label in the imbalanced dataset. Machine learning models tend to perform better on the majority class, leading to suboptimal performance on the minority class. Because the minority class has fewer observations, it can be difficult to find meaningful patterns in the data, leading to poor performance. Imbalanced datasets can be a major challenge in machine learning, and it is important to be aware of this issue when working with such datasets and take appropriate actions to address it. Also, before starting the training and testing, the classification and the preprocessing parts put a major impact on detecting the cyber threats.

9.2 Future Work

For improving the performance of the techniques which we have used, we will continue our research work in future, and we planned to propose some algorithms for detecting the cyber threats. In this study, to determine whether any feature combination works well with any machine learning model, we can use wrapper methods that use forward or backward elimination. Along with the wrapper method, we can also use the Regularization Method (l2), which is another feature selection technique. Additionally clustered algorithm can be a good option to analyze the result. Before taking any measures to prevent a cyber threat, it is crucial to identify the malware that is causing the system to malfunction. We can utilize multiclass classification for this reason, which enables us to determine what kind of malware or cyber threat it is. Alternatively, hyperparameter settings can be tuned to evaluate different neural network architectures and binary classification algorithms. Training with cross-validation is an additional suggestion. In order to detect the behavior of botnets without using the data labels, unsupervised learning can also be explored.

9.3 Conclusion

In this era of the digitalization, rapid development of the internet makes us involve with different ioT devices. The digitalization of the world has brought many benefits, such as increased efficiency, convenience, and connectivity. However, it has also led to new challenges, including cyber threats. The cyber threats can have serious consequences, such as financial loss, reputational damage, and disruption of critical services. It is important for individuals, organizations, and governments to be aware of the risks and take steps to protect against cyber threats. This may include implementing security measures such as firewalls, antivirus software, and intrusion detection systems, as well as educating employees and users on how to identify and avoid cyber threats. That is why, cybersecurity is a critical component in the fight against cyber threats. Organizations should implement a multi-layered approach that includes a combination of technical and non-technical measures to protect against cyber threats. In this thesis paper, we used the recently released VHS-22 heterogeneous flow-level network traffic dataset to train machine learning models for cyber threat detection. The combination of ISOT, CICIDS 2017, CTU-13, and traffic samples from MTA and Booters produced a dataset that was genuinely heterogeneous and diverse, which made it challenging for machine learning models detecting cyberattacks. This thesis paper represents machine learning algorithms such as Gaussian Naive Bayes, Bernoulli Naive Bayes, Random Forrest, k-Nearest Neighbors, Logistic Regression, Decision Tree and Ensemble Classifiers and we have showed the comparison between them to identify the attacks. The machine learning algorithms such as Random Forest and Bagging Decision Tree provides better accuracy of all machine learning algorithms for the VHS-22 dataset. The Bagging Decision Tree and Regular Random Forest classifiers have the highest accuracy of 99.81%. We have applied the Confusion Matrix for the best and worst algorithms to observe the differences for building the model. In this case, we have also achieved high f-1 score. We conclude that practical cyberattack defense systems can benefit from our study.

REFERENCES

[1] R. Singh, H. Kumar, R. K. Singla, and R. R. Ketti, "Internet attacks and intrusion detection system," *Online Information Review*, vol. 41, no. 2, p. 171–184, 2017. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/OIR- 12-2015-0394/full/html

[2] M. D. Preda, M. Christodorescu, S. Jha, and S. Debray, "A semantics-based approach to malware detection," *SIGPLAN Not.*, vol. 42, no. 1, p. 377–388, jan 2007. [Online]. Available: https://doi.org/10.1145/1190215.1190270

[3] P. Kaur, M. Kumar, and A. Bhandari, "A review of detection approaches for distributed denial of service attacks," *Systems Science & Control Engineering*, vol. 5, no. 1, pp. 301–320, 2017. [Online]. Available: https://doi.org/10.1080/21642583.2017.1331768

[4] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, "Enhanced network anomaly detection based on deep neural networks," *IEEE Access*, vol. 6, pp. 48 231–48 246, 2018.

[5] W. Wang, Y. Sheng, J. Wang, X. Zeng, X. Ye, Y. Huang, and M. Zhu, "Hast-ids: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection," *IEEE Access*, vol. 6, pp. 1792–1806, 2018.

[6] M. Apel, C. Bockermann, and M. Meier, "Measuring similarity of malware behavior," in *2009 IEEE 34th Conference on Local Computer Networks*, 2009, pp. 891–898.

[7] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.

[8] S. Garc´ıa, M. Grill, J. Stiborek, and A. Zunino, "An empirical compari- son of botnet detection methods," *Comput. Secur.*, vol. 45, pp. 100–123, 2014.

[9] A. Delplace, S. Hermoso, and K. Anandita, "Cyber-attack detection thanks to machine learning algorithms," 2020. [Online]. Available: https://arxiv.org/abs/2001.06309 (review+intro)

[10] D. H. Deshmukh, T. Ghorpade, and P. Padiya, "Improving classifi- cation using preprocessing and machine learning algorithms on nsl- kdd dataset," in *2015 International Conference on Communication, Information Computing Technology*

(ICCICT), 2015, pp. 1–6.

[11]I. Letteri, G. Della Penna, L. Di Vita, and M. T. Grifa, "Mta-kdd'19: A dataset for malware traffic detection." in *ITASEC*, 2020, pp. 153–165.

[12]P. Szumelda, N. Orzechowski, M. Rawski, and A. Janicki, "Vhs-22 – a very heterogeneous set of network traffic data for threat detection," in *Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference*, ser. EICC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 72–78. [Online]. Available: https://doi.org/10.1145/3528580.3532843

[13]Alazab, M. Profiling and classifying the behavior of malicious codes. *J. Syst. Softw.* **2015**, *100*, 91–102. [Google Scholar] [CrossRef]

[14]G. Sebastian. 2014. Identifying, Modeling and Detecting Botnet Behaviors in the Network. Ph.D. Dissertation. Universidad Nacional del Centro de la Provincia de Buenos Aires.

[15]ME Ahmed, H Kim, M Park. "Mitigating DNS query-based DDoS attacks with machine learning on software-defined networking" in proceedings of Milcom IEEE Military Communications Conference , 2017 :11-16

[16]W. T. Strayer, D. Lapsely, R. Walsh, and C. Livadas, "Botnet detection based on network behaviour," in Botnet Detection, ser. Advances in Information Security, W. Lee, C. Wang, and D. Dagon, Eds. Springer, 2008, vol. 36, pp. 1–24.

[17]Verma, Abhishek and Ranga, Virender, "Machine learning based intrusion detection systems for iot applications," Wireless Personal Communications, vol. 111, no. 4, pp. 2287–2310, 2020.

[18]M. Ma lowidzki, P. Berezinski, M. Mazur, Network Intrusion Detection: Half a Kingdom for a Good Dataset, in: NATO STO SAS-139 Workshop, Portugal, 2015.

[19]Shaukat, K., Luo, S., Chen, S., & Liu, D. (2020, October). Cyber threat detection using machine learning techniques: A performance evaluation perspective. In *2020 International Conference on Cyber Warfare and Security (ICCWS)* (pp. 1-6). IEEE.

[20] Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. *Energy Reports*, *7*, 8176-8186.

[21]Ramsdale, A., Shiaeles, S., & Kolokotronis, N. (2020). A comparative analysis of cyber-threat intelligence sources, formats and languages. *Electronics*, *9*(5), 824.

[22] David Morris, Garikayi Madzudzo, Alexeis Garcia-Perez, Cybersecurity threats in the auto industry: Tensions in the knowledge environment, Technological Forecasting and Social Change, Volume 157, 2020, 120102, ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2020.120102.

[23]Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., & Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. *Applied Sciences*, *12*(17), 8482.

[24]Martens, M., De Wolf, R., & De Marez, L. (2019). Investigating and comparing the predictors of the intention towards taking security measures against malware, scams and cybercrime in general. *Computers in Human Behavior*, *92*, 139-150.

[25]Hamid, B., Jhanjhi, N. Z., Humayun, M., Khan, A., & Alsayat, A. (2019, December). Cyber security issues and challenges for smart cities: A survey. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (pp. 1-7). IEEE.

[26]Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381-386.

[27]Liu, Y., Wang, Y., & Zhang, J. (2012, September). New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer, Berlin, Heidelberg.

[28]Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, *3*(2), 334-337.

[29]Ayodele, T. O. (2010). Types of machine learning algorithms. New advances in machine learning, 3, 19-48.

[30]Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361.

[31]Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281-299.

[32]Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.

[33]M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression- related posts in

reddit social media forum," *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019. Evaluation metric

[34]Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2010). Comparison of balancing techniques for unbalanced datasets. *Mach. Learn. Gr. Univ. Libr. Bruxelles Belgium*, *16*(1), 732-735.

[35]Levine, M. D. (1969). Feature extraction: A survey. *Proceedings of the IEEE*, *57*(8), 1391-1407.

[36]Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3-26.

[37]Abdulhammed, R., Musafer, H., Alessa, A., Faezipour, M., & Abuzneid, A. (2019). Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electronics*, *8*(3), 322.

[38] Kour, H., & Gondhi, N. (2020). Machine learning techniques: a survey. In *International Conference on Innovative Data Communication Technologies and Application* (pp. 266-275). Springer, Cham.

[39]S.Saad,I.Traore, A. Ghorbani, B. Sayed, D. Zhao, W. Lu, J. Felix, and P. Hakimian, "Detecting p2p botnets through network behavior analysis and machine learning," in 2011 Ninth Annual International Conference on Privacy, Security and Trust, 2011, pp. 174–180.

[40]M. A. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, and H. Janicke, "Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks," Future Internet,vol. 12, no. 3, 2020. [Online]. Available: https://www.mdpi.com/1999-5903/12/3/44

[41]J. J. Santanna, R. van Rijswijk-Deij, R. Hofstede, A. Sperotto, M. Wierbosch, L. Z. Granville, and A. Pras, "Booters — an analysis of ddosas-a-service attacks," in 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), 2015, pp. 243–251.

[42] Traffic-analysis.net." [Online]. Available: https://www.malware-trafficanalysis.net/

[43] Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019, July). Sentiment analysis with NLP on Twitter data. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.

[44] Bhowmik, N. R., Arifuzzaman, M., Mondal, M. R. H., & Islam, M. S. (2021). Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary. *Natural Language Processing Research*, *1*(3-4), 34-45.

[45] Bhowmik, N. R., Arifuzzaman, M., & Mondal, M. R. H. (2022). Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, *13*, 100123.

[46] Gope, J. C., Tabassum, T., Mabrur, M. M., Yu, K., & Arifuzzaman, M. (2022, February). Sentiment Analysis of Amazon Product Reviews Using Machine Learning and Deep Learning Models. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (pp. 1-6). IEEE.

[47] Basunia, M. R., Pervin, I. A., Al Mahmud, M., Saha, S., & Arifuzzaman, M. (2020, June). On predicting and analyzing breast cancer using data mining approach. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 1257-1260). IEEE.

[48] Ripa, S. P., Islam, F., & Arifuzzaman, M. (2021, July). The emergence threat of phishing attack and the detection techniques using machine learning models. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)* (pp. 1-6). IEEE.

[49] Toma, T., Hassan, S., & Arifuzzaman, M. (2021, July). An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection. In 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI) (pp. 1-5). IEEE.

[50] Arifuzzaman, M., Yu, K., & Sato, T. (2014, June). Content distribution in Information Centric Network: Economic incentive analysis in game theoretic approach. In Proceedings of the 2014 ITU kaleidoscope academic conference: Living in a converged world-Impossible without standards? (pp. 215-220). IEEE.

[51] Siddiq, M. A. A., Arifuzzaman, M., & Islam, M. S. (2022, March). Phishing Website Detection using Deep Learning. In Proceedings of the 2nd International Conference on Computing Advancements (pp. 83-88).