A MACHINE LEARNING APPROACH TO DETECT INTRUSION

BY

Shahriar Parvez Shaon
ID: 201-15-13803

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Touhid Bhuiyan
Professor
Department of CSE
Daffodil International University

Co-Supervised By

Ms. Tania Khatun
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

25 JANUARY 2024

APPROVAL

This Project titled **"A MACHINE LEARNING APPROACH TO DETECT INTRUSION"**, submitted by **Shahriar Parvez Shaon, ID No: 201-15-13803** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *date.*

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque (SMAH)
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sazzadur Ahamed(SZ)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Amatul Bushra Akhi (ABA)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Ahmed Wasif Reza (DWR)
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that; this project has been done by me under the supervision of **Dr. Touhid Bhuiyan,**Professor, Department of Computer Science and Engineering (CSE),Daffodil International University (DIU). I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:

Fr

Dr. Touhid Bhuiyan
Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Tania Khat

Ms. Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Shaon

Shahriar Parvez Shaon
ID: 201-15-13803
Department of CSE
DaffodilInternationalUniversity

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish my profound my indebtedness to **Dr. Touhid Bhuiyan, Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Machine learning and Deep learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Professor and Head,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

The rapid growth of interconnected digital device and the advancement of technology, the data security become an issue now a days. This leads various type of cyberattacks. In order to detect and effectively analyze malicious activities in a system or network, the implementation of an intrusion detection system is needed. It’s a system in a form of hardware or software that searches the network system for unusual behavior. Intrusion detection becomes paramount in ensuring network security as computer becoming more interconnected. Therefore, intrusion detection systems actively monitor the traffic of computers on the network to detect and alert to threats or malicious activities. In this study I develop intelligent detection system that can detect intrusion using various machine learning technique. I also use different metrics to evaluate the effectiveness of our solutions and make comparisons to determine the best intrusion detection network. Results from various studies were carefully analyzed and compared; this provided insight and direction for future work in this area. Data breaches often lead to unauthorized access to, alteration or deletion of sensitive data, leading to privacy and confidentiality issues. This can impact service because denial of service can cause outages and prevent important operations. The consequences of such incidents can be devastating, including security breaches and legal and financial damages to the organization's reputation.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

TABLES **PAGE NO**

CHAPTER 1
INTRODUCTION

1.1 Introduction

Intrusion detection systems (IDS) play an important role in network security and are designed to identify and prevent malicious behavior from occurring on a computer or system. The main purpose of IDS is to improve overall security through timely detection and reporting of any malicious intent or security breach. There are many ways to find access, of course it's important. However, the rate at which problems are detected and the occurrence of false positives also increases the overall value. A good understanding of fact checking is important because it is a simple guide to accessing research. To improve the process, the search process must be improved to minimize downside and maximize value. [1] There have been many developments in the field of communication in the last few years. The rise of various communication methods and the development of digital communications and devices have caused serious concerns about network security. It is important to protect information and communication technologies from threats. Attackers are constantly detecting and developing new attack methods, so it is important to identify an intelligence system (IDS) that will detect these attacks. The main purpose of IDS is to provide appropriate information when detected. Cybersecurity plays an important role in this process and consists of three main components: data collection, custom selection/transformation, and decision engine [2] Intrusion detection systems (IDS) are intelligent software or hardware devices embedded in a network. They passively monitor traffic over the network and on the device itself. IDS, an intrusion prevention system, has two components. One interface is used to monitor network traffic, while the other interface is used for management and reporting purposes. IDS have many advantages, especially when monitoring large networks, because they can be tailored to the specific needs of the network. In addition, IDS is generally invisible to potential attackers, providing an additional layer of protection. However, IDS may encounter problems in detecting and analyzing the attack during heavy traffic [3]

1.2 Motivation

The use of intrusion prevention devices (IDS) is driven by the urgent need to strengthen network security and protect digital assets against various threats. IDS is a method that focuses on the early detection of security incidents and unauthorized activities in a computer network or system. Unlike traditional security systems that focus on protection, IDS monitors and analyzes real-time network or system activity to identify suspicious and potentially threatening ones. This strategic approach is particularly important in the face of a changing digital environment where complex and ever-changing cyber threats pose serious challenges. IDS also plays an important role in mitigating insider threats by monitoring user activities and detecting malicious behavior. Compliance regulations continue to encourage the use of IDS, enabling organizations to meet regulatory standards and avoid penalties. IDS helps protect data, improve incident response, and reduce overall risk by providing timely alerts and simplifying detection and mitigation security conditions. Its presence increases security awareness and promotes a culture of cybersecurity in the organization. With their ability to adapt to new threats, IDS remains an important part of ongoing efforts to protect the confidentiality, integrity and availability of critical digital assets.

1.3 Rationale of Study

Intrusion detection systems (IDS) play an important role in network security by detecting and responding to unauthorized or malicious activity in the digital environment. Unlike traditional security measures, IDS acts as a watchful observer by constantly monitoring and analyzing connections or activities in real time. It uses a variety of detection methods, including signature detection, anomaly detection, and behavior analysis, to identify patterns that indicate threats or deviations from good character. The main purpose is to detect security incidents early, allowing rapid interventions to reduce risks and prevent developing cyber threats. Deploying an IDS is crucial for organizations looking to strengthen their cybersecurity and ensure the integrity, confidentiality and ownership of their digital assets. Motivations for IDS use are diverse and stem from personal and social concerns. For businesses, data breaches can lead to financial loss, reputational damage, and legal issues. IDS ensures national security and public health by protecting important systems in the country, such as the electrical grid and economy. Cyber-attacks are undermining trust in the digital world, so it is necessary to create a safer online environment where people and businesses can work safely knowing that their data and systems will not be damaged. The use of IDS has a significant impact on cybersecurity, improving incident response capabilities, regulatory compliance, risk reduction and incident awareness in the digital environment. Finally, IDS serves as a dynamic and adaptable tool to protect sensitive data and respond to emerging threats in the changing digital environment.

1.4 Research Question

With hard work and determination, this study is completed successfully. Completing this task was not easy. I faced many challenges during the project but I overcame them to ensure the project was fair, real and precise. Many questions are asked to understand basic concepts and provide good answers:

1. From where did I collect my data for this project?
2. Is there any need of data pre-processing?
3. What was the feature selection method?
4. Which algorithm perform the best?
5. What’s the best model which carried out best accuracy?
6. Does feature selection method impact on prediction result?

1.5 Expected Outcomes

- This project outcome is to analyze and understand the dataset.
- To make an efficient way for Intrusion data preprocessing.
- Different Features selection technique and its importance.
- Make comparison between multiple models
- Our main outcome is to detect intrusion accurately

1.6 Report Layout

In this report, there are 6 chapters.

- In Chapter 1, I review and subdivide the principle of our research. This includes the introduction, motivation, vision, research content and expected results of our project.
- In Section 2, I review past research on Intrusion Detection, the nature of the problem, and the challenges faced in this research.
- Chapter 3 will provide an in-depth dive into our work, methods, and techniques to build an IDS.
- In Section 4, I analyze the experimental results and provide a general discussion of our design.
- In Chapter 5, I discuss the social, environmental, ethical and future impacts of our work.
- In Chapter 6, I will review the results, implications, and further research of this study.

CHAPTER 2
BACKGROUND

2.1 Introduction

Intrusion detection system uses machine learning. I collected data from online for this project. The main objective of our work is to detect malicious activities over a network that high prediction accuracy method models may be implemented for better performance. Here I bring new formulations and results of the ideal model to be implemented. The review of earlier research to support the proposed work, as well as future goal. Intrusion detection is a technique that looks for several attributes in the data that the system fetch from a network and looks for malicious activities. To mitigate suspicious behavior, create a framework for classifying intrusion. I have seen that a lot of effort has been put into finding and using different methods and techniques to solve these attacks and secure the network without affecting legitimate network users. I can train the model utilizing ML with the aid of current technologies.

2.2 Related Works

A lot of research has been done on the use of machine learning in attack detection. These studies discuss various methods and demonstrate their collaboration. In our work, I also reveal the advantages that make us unique. Agrawal et al. [1] investigated the vulnerability in intrusion detection using data mining. They divided this model into three types: integration, classification-based, and hybrid methods. Clustering-based methods include K-means, K-Meoids, EM clustering, and anomaly detection techniques. Classification-based methods include Naive Bayes, genetic algorithms, neural networks, and support vector machines. Hybrid methods describe the combination of different learning systems. They also provide a concise comparison of data using clustering techniques. Haq et al. [2] investigated the application of machine learning in access detection. They roughly divide these ideas into three categories: supervised learning, unsupervised learning, and extended learning. Supervised learning involves training a class on a dataset, while unsupervised learning is used when there is no dataset. To support learning, experts can record anonymous events. They provide a brief description of various separation and clustering algorithms for access control and document machine learning, but do not include analysis or critical analysis. Ahmed et al. [3] performed a comprehensive evaluation of the method for detecting anomalies. They used the KDD'99 dataset to classify attacks into four groups: DoS, probe, U2R, and R2L. Each category corresponds to a different type of inequality. The authors discuss various machine learning methods, including classification-based, cluster-based, statistical, and process theory. Using these techniques, they can distinguish between normal and abnormal situations in intrusion detection. The survey also briefly discusses the challenges associated with multi-data network access analysis. As a future study, the authors suggest the use of IDS collaboration. However, their survey lacks a detailed and descriptive analysis of the current education system through IDS. Additionally, the authors do not provide future directions for machine learning algorithms. Buzak et al. [4] conducted an extensive discussion on intrusion detection using machine learning and data mining techniques. Their research focused on misuse and misdiagnosis of this technology. The authors highlight the differences between machine learning (ML) and data mining (DM), noting that ML is a more mature concept compared to DM. However, since both ML and DM adopt similar methods for data classification and knowledge discovery, the authors collectively refer to these as ML/DM learning opportunities. The survey provides a detailed description of various ML/DM methods and their relationship to imputation, uncertainty, and hybrid search. The authors also discuss the time complexity of the algorithms used in this model. They note that KDD'99 and DARPA data are often used for research because it allows meaningful comparisons between different studies. However, some researchers have also used NetFlow and tcpdump datasets. Kumar S. [5] analyzes that monitor machine learning classifiers by analyzing data containing labels of the characteristics of network traffic from real and unreal applications. The main purpose is to ensure compatibility with the Network Intrusion Detection System (NIDS). The researchers used the ISCX Android botnet dataset, which contains 1,929 examples of botnet families spanning four years. Initially, the information in the ISCX Android botnet configuration file was processed by legitimate and malicious applications. This process involves filtering and selecting features to create a domain name. The recorded data is divided into a testing and a training set. The training process is used to create a machine learning algorithm classifier that is used to evaluate test data. But this system has indisputable advantages. To classify data, researchers used random forest classification because it was actually more efficient compared to other learning machines. However, their false positive rates are slightly higher. They also use weka's data mining tool, which puts more load on the system. Divyatmika & Manasa Sreekesh [6] suggestions regarding two layers for network access analysis. The researchers used the weka database search tool and the NSL-KDD database. They first used hierarchical aggregative clustering to create an autonomous model of the training set that would process the data. KNN classification was used to classify the data and finally multi-layer perceptron and reinforcement algorithms were used for error detection and detection. By using unsupervised learning, especially hierarchical clustering, the system itself can be created by repeatedly designing the data warehouse. However, using Weka as a data search tool increases the overhead. Additionally, the false positive rate is high. Bhavani et al. The accuracy of the random classifier reaches 95.323%, which is the best result. However, the proposed work [7] does not solve the problem of low values and false alarms. Ponthapalli et al. also used a machine learning algorithm such as decision tree, logistic regression, random forest, and support vector machine [8] to identify network interactions using the KDD-NSL dataset. Their work showed that the random forest distribution was the most efficient and had the shortest processing time. However, the application functionality is limited to optimization using only a single file. Marzia Z. and Chung-Horng L. [9] used an integrated IDS-based method that uses polling methods to record the results of learning algorithms of multiple supervised and unsupervised systems. This approach improves the accuracy and performance of existing intrusion prevention devices. They used the Kyoto 2006+ dataset, which is considered more reliable than the widely used KDDCup '99 dataset due to its relative age. Although their work has achieved a certain level of accuracy, sometimes results are returned lower, indicating negative probability (FPR).

Verma et al. [10] conducted a study showing that the detection of anomalies can be improved, especially in terms of reducing false alarms. They used Extreme Gradient Boosting (XGBoost) and Adaptive Boosting (AdaBoost) learning algorithms on the NSL-KDD dataset. Although their accuracy reached 84,253, improvements can still be made by using a combination of hybrid or learning machines. Previous studies have faced limitations due to the inability to use a specific selection of reference material, again leading to the inclusion of irrelevant, redundant and redundant features. Kazi Abu Taher et al. [11], various machine learning models with different algorithms were evaluated using the NSL-KDD dataset and feature selection was applied using the wrapper method. This approach improves accuracy compared to previous studies using the same data. However, the model still has difficulty detecting zero-day attacks due to false positives, and this work only focuses on signature-based attacks and does not detect a new attack.

2.3 Comparative Analysis and Summary

The main focus of this work is the opportunities to analyze more of new intrusion dataset. Different methods are applied and added various algorithms to our dataset. The source of the data in this project is EDGE IIoT data, which is available online. As mentioned before, our collection includes newly acquired information and previously used information. I will be able to evaluate the effectiveness of all the methods I use and the impact of other data I provide from the same source. There are similar kind of classes and labels. Machine learning techniques and feature extraction are used in the extractor process, and the data preparation process is mainly implemented in Python. My favorite feature extraction engine is Python and I use the SVM classifier approach to classify web applications.

2.4 Scope of the Problem

Essentially, our research entails assessing the available data and constructing a model through the application of machine learning methods. Our method may enable the diagnosis of malicious activities. This endeavor will have a profound impact on the people inside this community. Intrusion detection system (IDS) problems are common and play an important role in network security. IDS operates in an ever-changing cyber threat landscape, including malware, phishing and persistent threats. These threats pose risks to the confidentiality, integrity and availability of digital assets. IDS solutions use various detection methods such as signature-based detection and conflict detection. They include network and host-based monitoring to detect unauthorized activity. Real-time monitoring is important for timely detection of threats and rapid response to minimize potential disruptions. But it is difficult to balance good and evil. The integration of machine learning and artificial intelligence into an IDS improves the ability to identify complex and evolving threats. Using machine learning models, the project will analyze it to identify malicious activities over network. In order to uncover terms that could be suggestive of IDS, I scanned through some previous work for this project.

2.5 Challenges

Considering the difficulties of analyzing such a large number of features, the main problem in this research seems to be the synthesis and evaluation of all this data. A number of tools and methods is used to clean and standardize the dataset. The processing and relationship of information is difficult due to the large amount of information and the rapid creation of information from various sources such as network and system log. To ensure the accuracy of the analysis, it is important to pre-process the data and make sure it is good. Developing detection algorithms that can detect complex and evolving strategies used in cyber-attacks is a difficult task. Striking the right balance between negatives and minimizing negatives is a constant effort, as is performing immediate audits in a fast-paced environment. Additionally, factors such as lack of understanding of the content, privacy issues, and the need to adapt to emerging threats also increase the number of challenges. I had to work hard on all the important tasks, making it difficult to think of the best answer quickly. Intrusion is a common problem in this digital age and can have serious consequences for the target. There is interest in developing intrusion detection techniques to obtain classifications of cybersecurity cases; because this can help identify and prevent these activities. However, there are many difficulties in this project. Classification may be more difficult when classes have different data type.

- Data collection
- Data handling
- Preprocessing
- Feature extraction
- Maintaining high accuracy
- Making decision

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

Here, I will describe the methodology and procedure of the project. Tools I used for the project, data collection, preprocessing, analysis, model selection and implementation all step will be discussed.

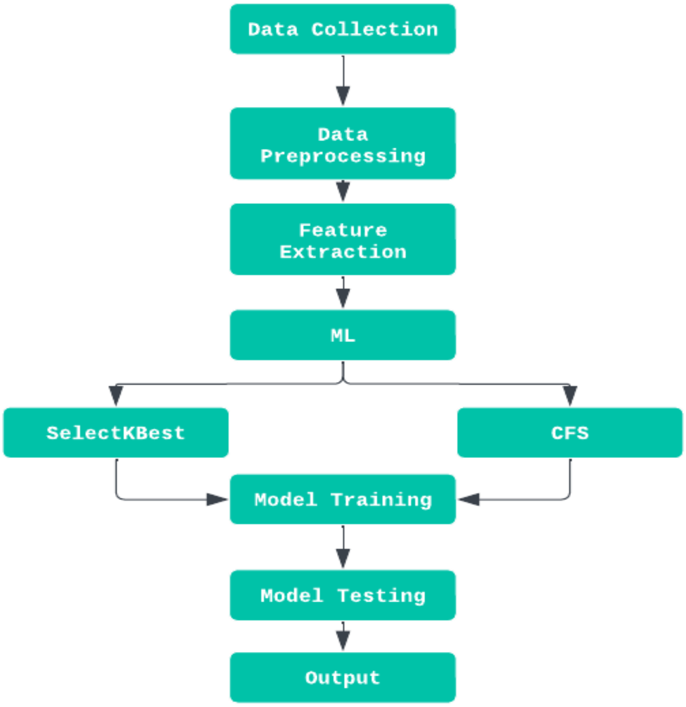


Figure 3.1: Methodology

This chapter aims to provide an overview of all the methods and techniques used to identify malicious activities on the network. The entire research process is presented in this section. Each review can be solved in various ways. The next step is to choose the ML method. As I said before, since I use five different learning methods of the machine, data collection needs to be created to create the model and the adaptation algorithm. Then use this information to specify the model. Feature selection is done on this basis. The data is divided into training and testing. This is also called test data and training data. After fitting data into multiple machine learning and training models using training data, I only get as much data as I need to evaluate the model. The model's accuracy is then measured. I explained using our simple recipe; To better understand some processes, I will examine them with equations and diagrams. Details of all research studies and research methods are given below. Data collection, analysis and design are some of the basic concepts. With proper balance, drawing, design and expression, this concept becomes clearer. This study used the most accurate classification machine learning model to make predictions and collect real-world data.

3.2 Research Subject and Instrumentation

A research topic is an area of study that is analyzed and analyzed to clarify the content. I discuss the tools and techniques I use in the instrumentation industry. It supports Python programming language, Windows operating system and NumPy, SkLearn, OpenCV etc. Google Colab platform was used in all steps of the training and testing process. Python programmers can develop code for machine learning and data science algorithms at Google Colab. These algorithms use statistical techniques associated with machine learning to define categories such as intrusion and normal. I also need to review the relevant tests. Then I need to make a few choices:

1. Is the dataset is updated or trusted?
2. How’s the data collected?
3. How the data should be organized and processed?
4. How the data should be labeled?
5. Which model should I choose for the project?

3.2.1 Data Collection Procedure

I present a novel and extensive cyber security dataset for IoT and IIoT applications, named Edge-IIoTset, in this project. The final data used for this project was created by combining with source data. The resulting file contains a total of 157800 data points and contains 63 columns. The data is divided into two parts: training and testing. After removing duplicate data and dividing the dataset, there are 30352 (20%) test data. Train has 121405 (80%) amounts of data. The attack category is categorized in 15 features and is has specific binary label for attack and normal feature. Feature extraction technique is also a crucial part of data processing. It can vary the accuracy of the model. Machine learning experts often work with data that has many features, regardless of the number of rows in the data. These unique IDS may contain symbols and numbers. Language is frequently used in education to distribute information to improve user comfort and understanding. One way to convert a signal into text that a computer can read is to encode the tag. To create a pattern that represents a number, the symbol must be converted as part of the previously mentioned process. Programmers who design learning machines ultimately decide how the script will be used. Now an initial review should be performed on the given data, including any changes tracked.

3.3 Statistical Analysis

Transforming raw data into a suitable framework for evaluating and training models is the goal. The first step is to obtain the necessary features and labels from the database containing the malicious activities. After creating the file, the next step is to describe the dataset and the data types. Data types is also a very important element as it needed to be change before training the model. For this project I use DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, ExtraTreesClassifier and GaussianNB.

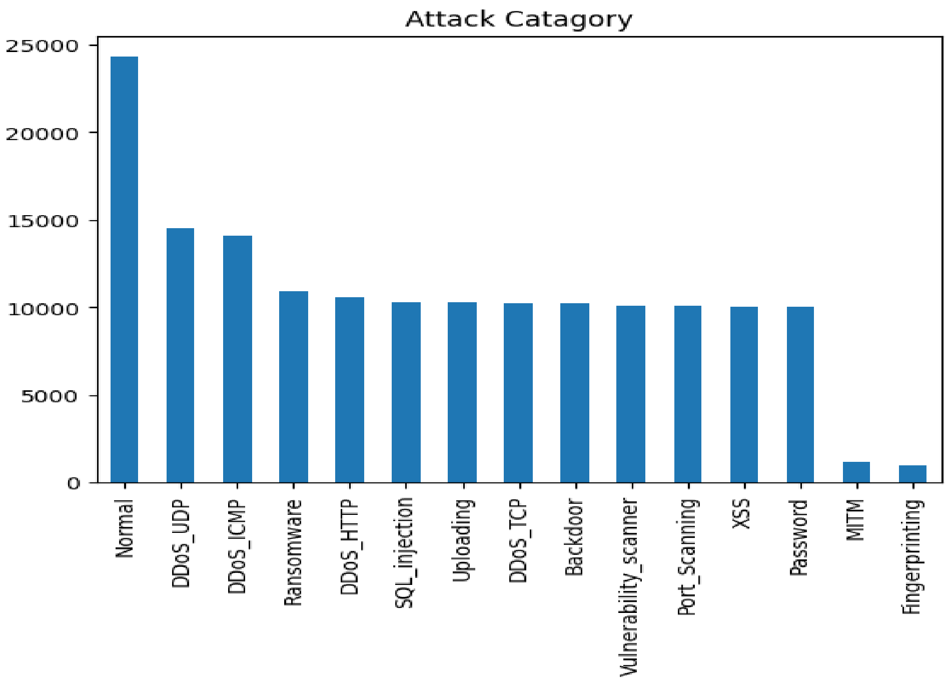


Figure 3.2: Attack category

In figure 3.2, the bar graph shows the attack category and the number of attacks. The chart provided shows the frequency of attacks by attack type. This group will have inappropriate attacks on other groups. Such events are fake events such as network connections that are not real attacks.

DDOS_UDP: This category is a denial-of-service attack that uses the UDP Data Protocol. UDP is a stateless protocol, meaning there is no connection between the attacker and the victim. It will be difficult to prevent DDOS_UDP attacks.

DDOS_ICMP: This category refers to distributed denial of service attacks that use the Internet Control Message Protocol. ICMP is used to exchange messages between network devices, including error messages and ping requests. Using ICMP, the attacker can flood the victim's network, making it inaccessible to legitimate users.

Ransomware: Ransomware is a type of malware that encrypts the victim's data and then demands a ransom to decrypt it. Ransomware attacks can cause significant disruptions and financial losses to businesses and organizations.

DDOS_HTTP: This category is a denial-of-service attack that uses Hypertext Transfer Protocol (HTTP). HTTP is the protocol used to communicate with web servers. An attacker could use HTTP to flood a website with requests, resulting in unauthorized users.

SQL_injection: SQL injection is a type of attack that allows a malicious actor to inject SQL code into a website or site. These codes can be used to steal data, modify data or control applications. Information downloads are generally referred to as attacks that involve sending malicious information to a website or website request. This information may be used to compromise the security of the website or website application.

DDOS_TCP: Its stands for a denial-of-service attack using Connection Protocol (TCP). TCP is a connection-oriented protocol; That is, before an attack can begin, a connection must be established between the attacker and the victim. This makes DDOS_TCP attacks more difficult to prevent than DDOS_UDP attacks.

Backdoor: A backdoor is a hidden entry point into a computer that an attacker can use to gain unauthorized access. Backdoors can be installed through a variety of methods, such as social engineering or malware.

Vulnerability_scanner: A vulnerability scanner is a tool used to identify vulnerabilities in a computer. Attackers can use vulnerability scanners to detect vulnerable systems.

Port_scanning: Port scanning is a process used to identify open ports on a computer. An open port is a port that is actively listening for communications. Attackers can use port scanning to find open ports that can be used to gain unauthorized access to the system.

XSS: Cross-site scripting (XSS) is a malicious technique that allows an attacker to inject malicious code into a website or web application. This injection can be used to steal sensitive data, modify existing data, or gain unauthorized control of the application.

Passwords: A password attack involves an attacker deliberately guessing or cracking a user's password. Attackers use a variety of methods, including dictionary attacks, brute force attacks, and the use of language rainbows, to capture passwords and gain unauthorized access.

MITM: A Man in the Middle (MITM) attack occurs when an attacker intercepts communication between two parties. By doing this, an attacker can monitor communications and even use them for his own malicious purposes.

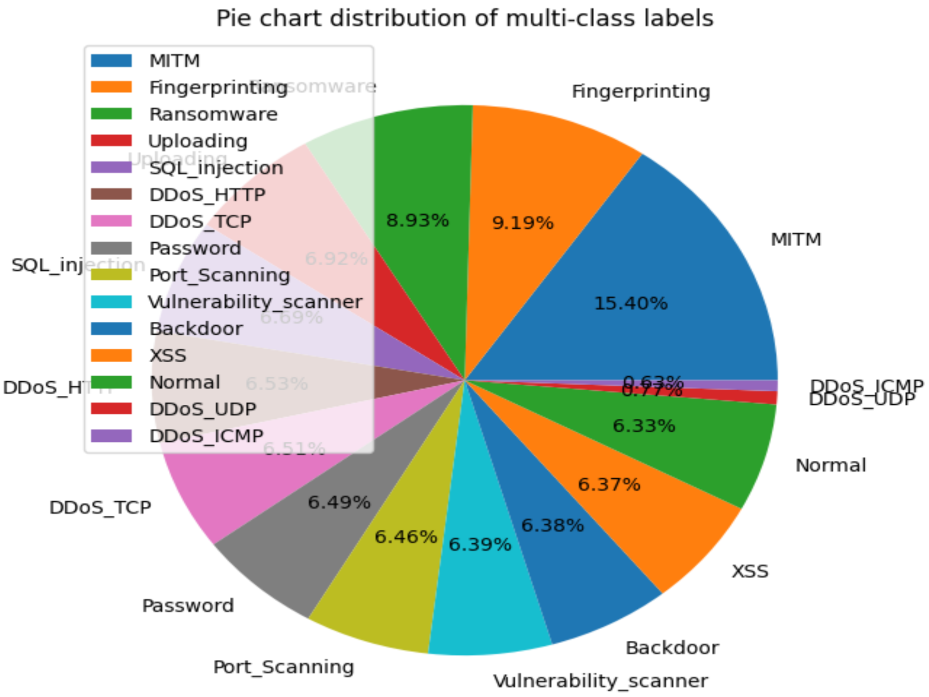


Figure 3.3: Attack label Pie chart

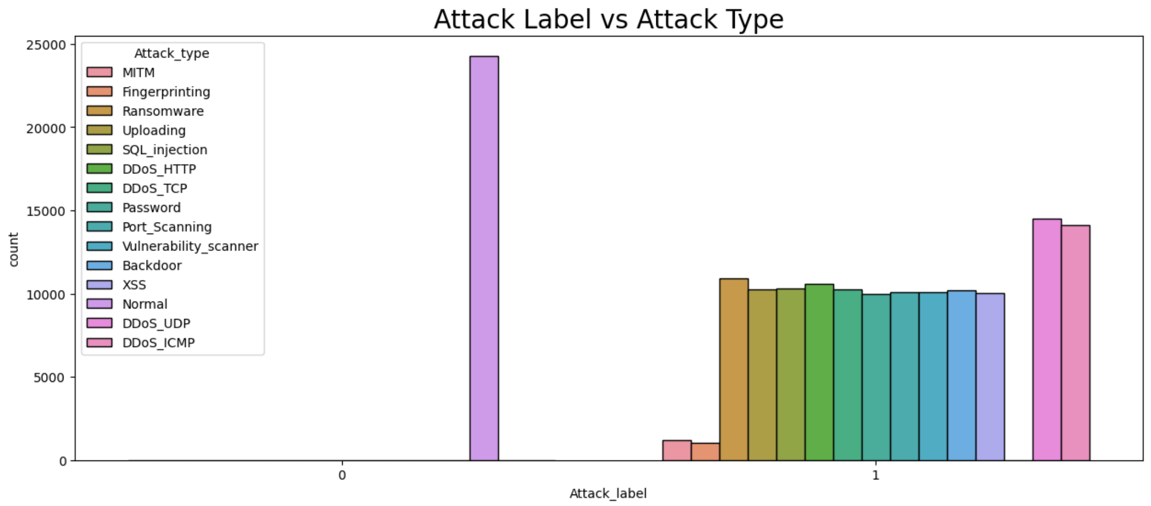


Figure 3.4: Attack label vs attack type

In figure 3.4, the distribution shows the attack type and attack label count.

3.4 Proposed Methodology

Various methods or techniques will be used to obtain results during this research. The experiments were carried out by selecting tasks, processing and collecting data, feature selection, and evaluating the efficiency using the results of the classifier.

- Data Collection: After collecting data from online, I preprocess it. Broad and complete data is lacking in this market due to the difficulty of collecting data on specific abuse devices and their distribution.
- Data Processing: Each piece of data is analyzed using various methods. Features selection, missing value handling, duplicate data drop is done during data preprocessing. I personally deal with the final stages of the selected data before it is used.
- Data Preparation: Organize and expand the dataset by content and labels. I prepare the data and present it to training. Only the minimum necessary work is required to prepare the split profile.
- Model Selection: To increase reliability, I select and specify a prediction method and evaluate it using our data. Machine learning involves the use of various filters.
- While various models were used to improve the effectiveness of the gadget layout and assist the ML model in classifying inaccurate data, only one device was finally selected to evaluate the accuracy of the data.
- Performance Evaluation: The next step is to complete the results. After training and testing these techniques give us confidence in both teaching methods. f1 measurements and statistical accuracy and confusion matrices are generated. This can help determine whether cyberbullying involves abusive language.

3.5 Implementation Requirements

The main purpose of an intrusion detection system (IDS) is to identify unauthorized or malicious behavior occurring in a network or system. To complete this task effectively, it is important to first consider various requirements**:**

Google Colab: Google Colab is an open-source publisher of the free Python programming language. I can work online via browser and Jupiter Notebook. But the main advantage of Google Collab is that it gives us free online virtual GPU access.

Matplotlib: Matplotlib's visualization toolkit includes Pyplot, which provides many tools for plotting, calculating, graphing, and graphing. It can be used to highlight the boundaries of the narrative or to visualize the perspective from which the story is told.

NumPy: The Python utility NumPy simplifies vector manipulation, especially in matrix calculations, Fourier transforms and index transformations. The NumPy module provides many tools and techniques for working with various types of matrices in Python. NumPy increases capability and realism in device design by focusing primarily on quantitative measurement.

Sklearn: Sklearn is an efficient and effective data analysis and forecasting tool. There are three Python programs: NumPy, SciPy and Matplotlib. These open-source tools can be customized by individual users.

Seaborn: The next version of this popular Python information visualization application can be used in conjunction with Matplotlib. Seaborn provides a user-friendly platform for creative data visualization.

Pandas: Pandas is a freeware toolkit specifically designed for analyzing and manipulating language-specific data. It provides essential data structures and statistical analysis methods to facilitate systematic data management, particularly for summary data.

Hardware Requirement:

1. Operating system (Windows, Mac)
2. Web browser (chrome, Firefox, safari)
3. Hard drive (at least 4 GB)
4. Memory (more than 4 GB)

Tools:

1. Python Environment
2. Anaconda
3. Spyder

CHAPTER 4
EXPERIMENTAL RESULT AND DISCUSSION

4.1Introduction

This section explains methods for classifying intrusion used in cybercrime. The entire process of building a model consists of just a few steps: selecting a model, collecting and evaluating data, feature extraction, model selection, and evaluating performance. I present the results of our experiments presented in the following section.

4.2Experimental Results & Analysis

Machine learning model can't produce perfect results. We can fix the bad model during training to improve accuracy. However, I found that the accuracy was very high using various techniques. This is a summary of my research activities. This image shows the following information: heatmap, recall, precision, f1 score and support. Depending on the strategy I use, I get different results. Using five machine learning algorithms and two features selection method, I was able to predict illegal activities associated with Intrusion. This method is used to determine the relationship between each part of the overall model and then use a series of verification techniques to reach a decision. Once all the data is selected, each model uses data that includes data from our own research as well as data published from online sources. The second is to use comparative data to determine whether the content should be classified as an Intrusion. Here I present a detailed analysis of various models using key performance indicators. Metrics such as total F1 score, accuracy, precision, and improvement provide a good understanding of the algorithm's performance.

I have used two different feature extraction method which cause different accuracy for each model. SelectKBest is a machine learning algorithm for feature selection that is part of the scikit-learn library in Python. The goal is to identify the top k features using specific metrics or scoring functions. In figure 3.5, the accuracy I got using SelectKBest method.

Table 3.1: Model Accuracy (SelectKBest)

Algorithm	**Accuracy**
DecisionTreeClassifier	94.67 %
RandomForestClassifier	96.12 %
KNeighborsClassifier	89.10 %
ExtraTreesClassifier	94.80%
GaussianNB	89.29 %

Correlation-based feature extraction is a method used in machine learning to select or extract features by considering the correlation between features and target variables. The goal is to identify and preserve features that show a strong relationship to the target, as they are expected to provide more valuable insight into the predictive model. For correlation-based feature extraction technique.

Table 3.2: Model Accuracy (correlation-based)

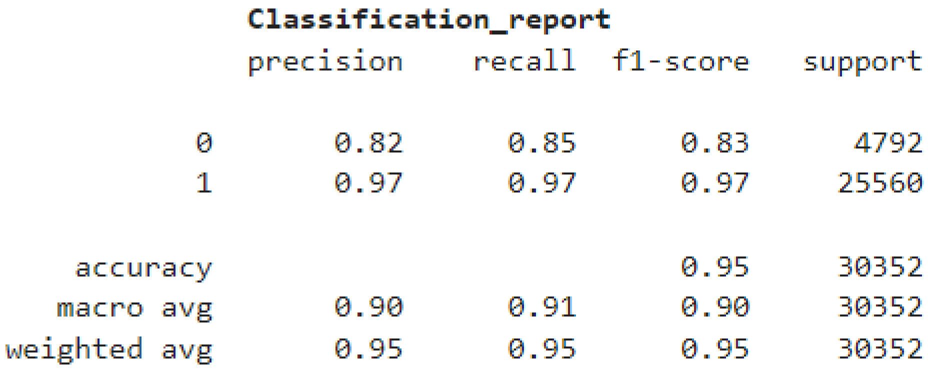
Algorithm	**Accuracy**
DecisionTreeClassifier	98.68 %
RandomForestClassifier	98.77 %
KNeighborsClassifier	96.30 %
ExtraTreesClassifier	98.75 %
GaussianNB	89.34 %

4.2.1 Algorithms

My analysis uses different types of studies to understand the malicious activities over a network. We examined DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, ExtraTreesClassifier and GaussianNB Machine learning techniques. I use multiple models to leverage our model's ability to represent structure and content. This positive research helps us achieve our goal of developing a better way to detect intrusion.

1. **Decision Tree:** Decision tree is a method of monitoring machine learning. This method can be used to solve classification and retrieval problems. It is a tree classifier, the distribution code is represented by branches, the quality of the dataset is shown in the nodes, and each leaf of the node shows the result. The decision tree has two nodes: the decision network and the leaf. In figure 3.5, Decision Tree achieve 94.67% for SelectKBest feature selection technique, Precision score 82%, Recall score of 85% and F1 score of 83%.

i. SelectKBest:



```
              Classification_report
              precision    recall  f1-score   support

           0       0.82      0.85      0.83      4792
           1       0.97      0.97      0.97     25560

    accuracy                           0.95     30352
   macro avg       0.90      0.91      0.90     30352
weighted avg       0.95      0.95      0.95     30352
```

Figure 3.5: Classification Report (Decision Tree Classifier)

In figure 3.6, the performance of Decision tree is given for SelectKBest which represent the TP, TN, FP and FN.

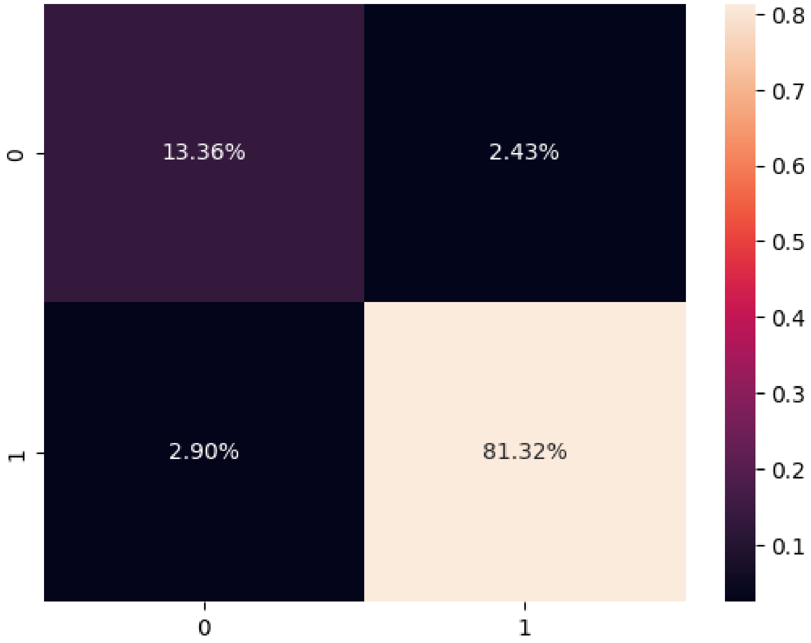
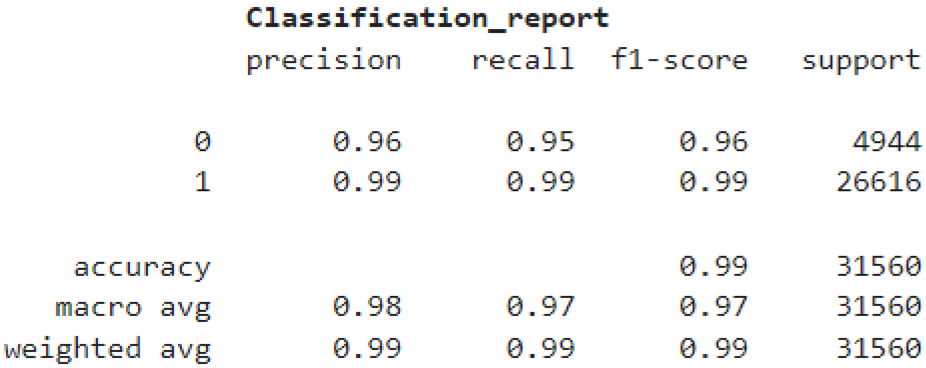


Figure 3.6: Confusion Matrix (Decision Tree Classifier)

ii. Correlation:

In figure 3.7, Decision Tree achieve 98.68% for correlation feature selection technique, Precision score 96%, Recall score of 95% and F1 score of 96%.



```
                 Classification_report
                 precision    recall  f1-score   support

           0       0.96      0.95      0.96      4944
           1       0.99      0.99      0.99     26616

    accuracy                           0.99     31560
   macro avg       0.98      0.97      0.97     31560
weighted avg       0.99      0.99      0.99     31560
```

Figure 3.7: Classification Report (Decision Tree Classifier)

In figure 3.8, the performance of Decision tree for correlation feature selection is given which represent the TP, TN, FP and FN.

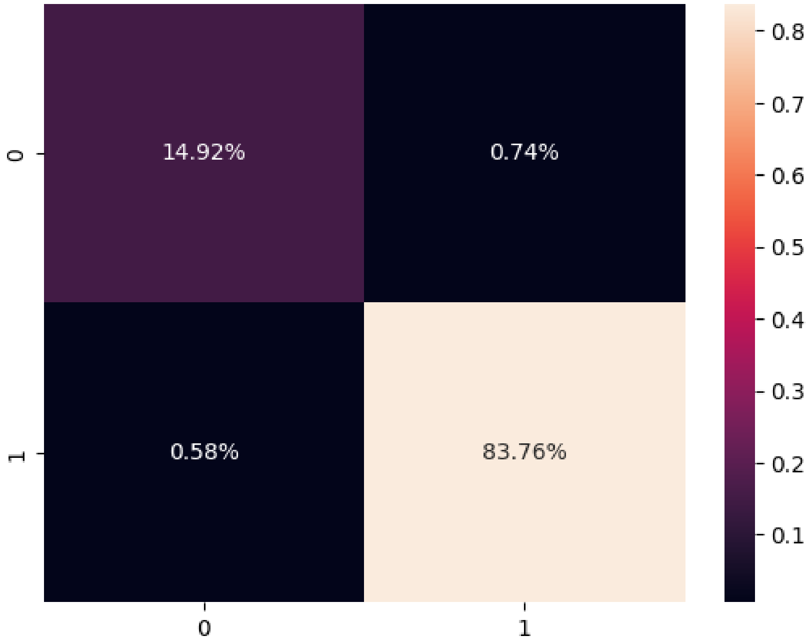
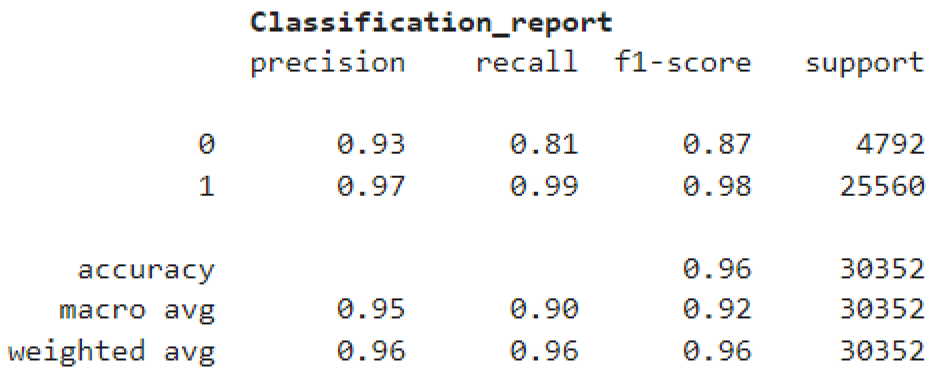


Figure 3.8: Confusion Matrix (Decision Tree Classifier)

2. **Random Forest:** A tree-based approach to RF classifiers that can be used for both regression and classification. It is a machine learning algorithm for creating hierarchical trees. AI uses this process to create hierarchical “decision trees.” When using the mixed method, the random forest model for classification results in multiple decision trees that are then averaged. This clearly demonstrates the problems with overfitting. In figure 3.9, Random forest achieve 96.12% for SelectKBest feature selection technique, Precision score 93%, Recall score of 81% and F1 score of 87%.

i. SelectKBest:



```
              Classification_report
              precision    recall  f1-score   support

           0       0.93      0.81      0.87      4792
           1       0.97      0.99      0.98     25560

    accuracy                           0.96     30352
   macro avg       0.95      0.90      0.92     30352
weighted avg       0.96      0.96      0.96     30352
```

Figure 3.9: Classification Report (Random Forest)

In figure 3.10, the performance of Random forest for SelectKBest feature selection is given which represent the TP, TN, FP and FN.

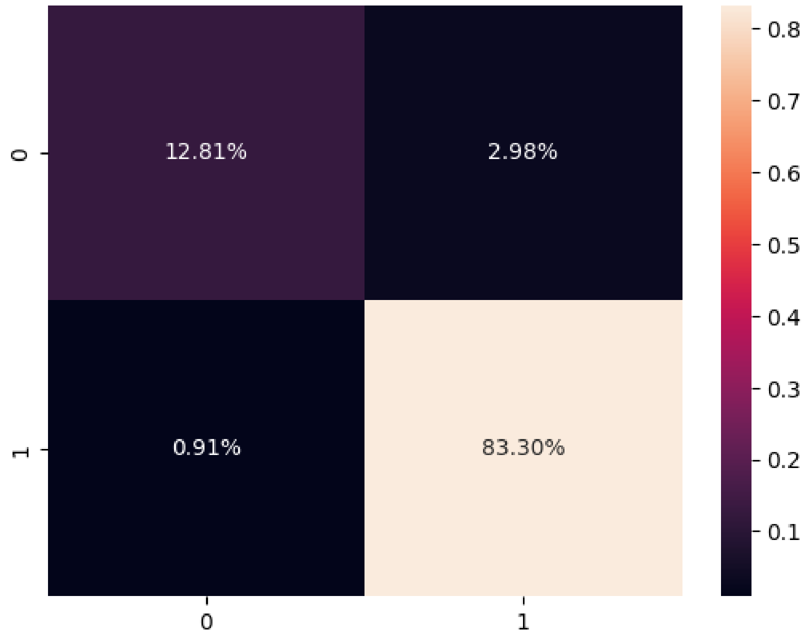
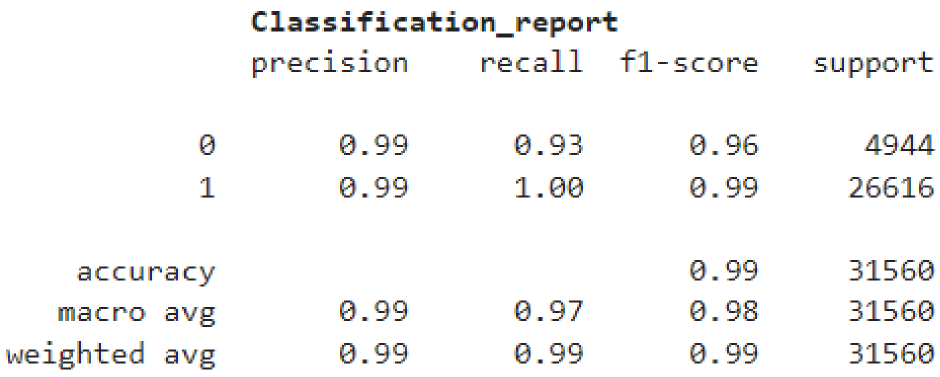


Figure 3.10: Confusion matrix (Random Forest)

In figure 3.11, Random forest achieve 98.77% for correlation feature selection technique, Precision score 99%, Recall score of 93% and F1 score of 96%.

ii. Correlation:



```
                  Classification_report
                  precision    recall  f1-score   support

              0       0.99      0.93      0.96      4944
              1       0.99      1.00      0.99     26616

       accuracy                           0.99     31560
      macro avg       0.99      0.97      0.98     31560
   weighted avg       0.99      0.99      0.99     31560
```

Figure 3.11: Classification Report (Random Forest)

In figure 3.12, the performance of Random forest for correlation feature selection is given which represent the TP, TN, FP and FN.

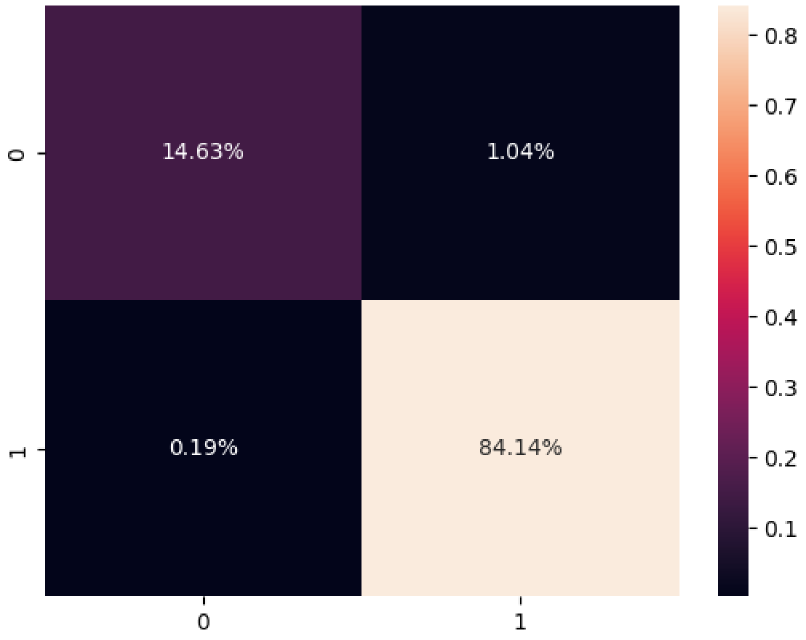
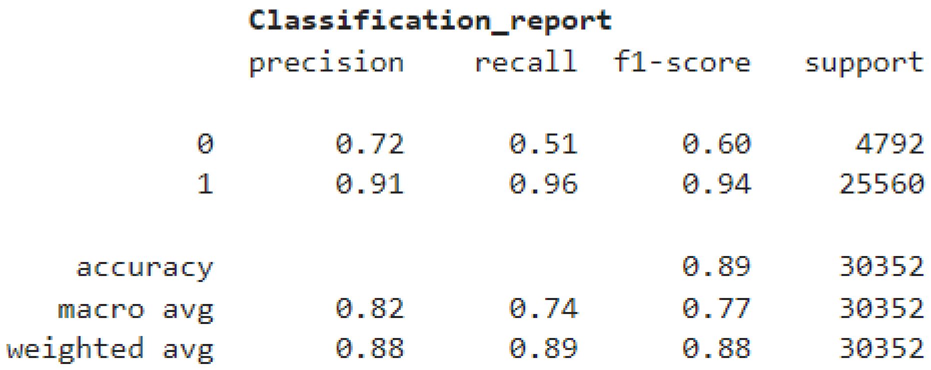


Figure 3.12: Confusion Matrix (Random Forest)

3. **K-Neighbor**: Nearest neighbor (k-NN) classifier is a simple and straightforward algorithm in machine learning for task classification. k-NN does not rely on statistical similarity, but predicts the ranking of a data point by considering the majority of neighbors close to a given location. In short, it assigns a category to the new information point according to the category of its neighbors. In figure 3.13, K-Neighbor achieve 89.10% for SelectKBest feature selection technique, Precision score 72%, Recall score of 51% and F1 score of 60%.

i. SelectKBest:



```
                Classification_report
                precision    recall  f1-score   support

           0       0.72      0.51      0.60      4792
           1       0.91      0.96      0.94     25560

    accuracy                           0.89     30352
   macro avg       0.82      0.74      0.77     30352
weighted avg       0.88      0.89      0.88     30352
```

Figure 3.13: Classification Report (K-Neighbor)

In figure 3.14, the performance of K-Neighbor for SelectKBest feature selection is given which represent the TP, TN, FP and FN.

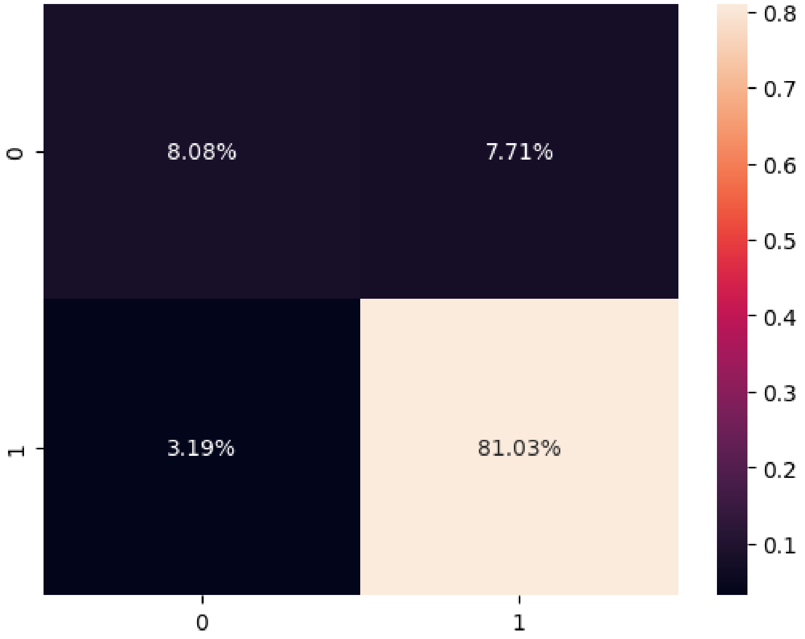
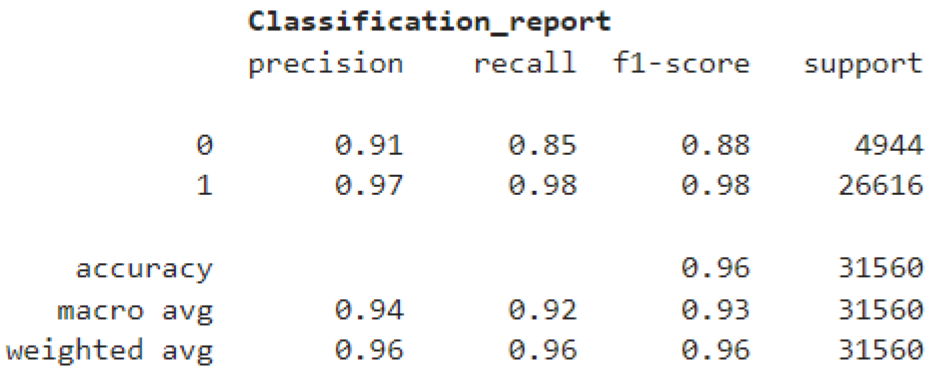


Figure 3.14: Confusion Matrix (K-Neighbor)

In figure 3.15, K-Neighbor achieve 96.30% for correlation feature selection technique, Precision score 91%, Recall score of 85% and F1 score of 88%.

ii. Correlation:



```
                Classification_report
                precision    recall  f1-score   support

           0         0.91      0.85      0.88      4944
           1         0.97      0.98      0.98     26616

    accuracy                           0.96     31560
   macro avg         0.94      0.92      0.93     31560
weighted avg         0.96      0.96      0.96     31560
```

Figure 3.15: Classification Report (K-Neighbor)

In figure 3.16, the performance of K-Neighbor for correlation feature selection is given which represent the TP, TN, FP and FN.

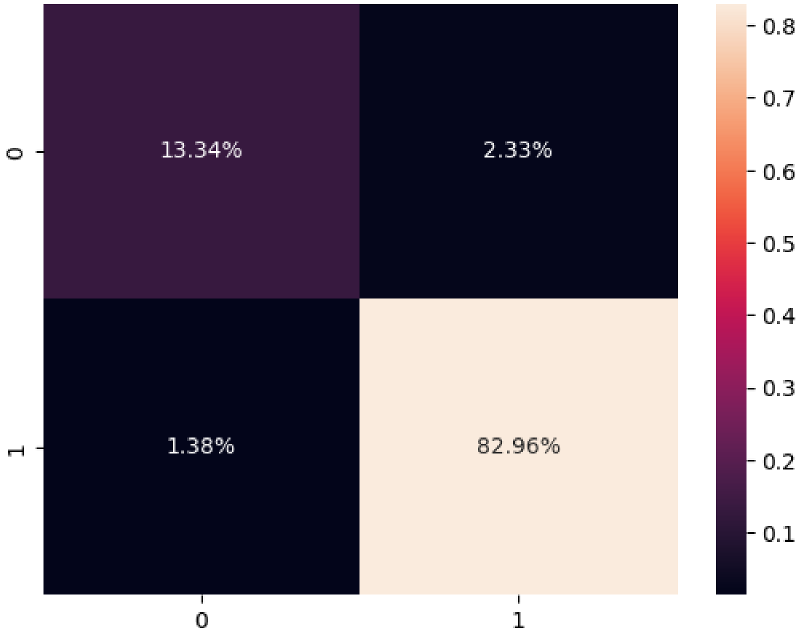
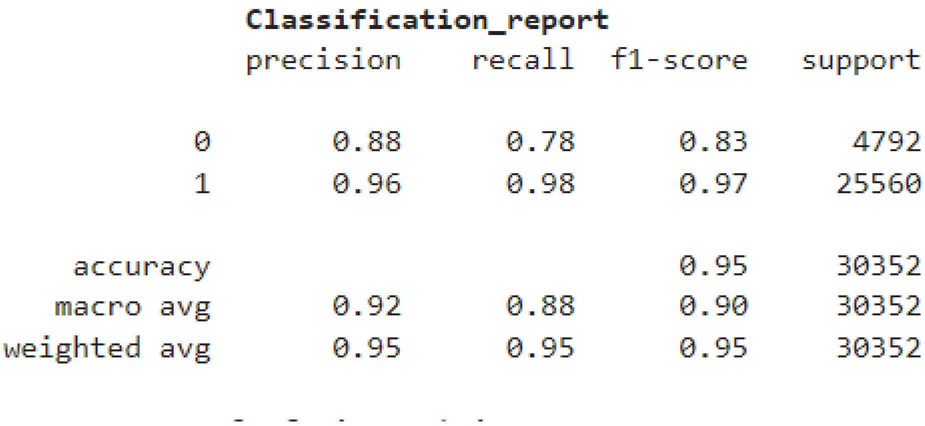


Figure 3.16: Confusion Matrix (K-Neighbor)

4. **Extra Tree:** Extra Trees classifier is a classification algorithm in machine learning. It falls under the umbrella of ensemble learning, which involves combining multiple models to increase the accuracy of predictions. Extra Tree has a unique way of creating decision trees from competitors. Unlike traditional decision trees, Complementary Tree randomly selects a set of features from each partition, reducing the risk of overfitting and improving the ability to process data. In figure 3.17, Extra Tree achieve 94.80% for SelectKBest feature selection technique, Precision score 88%, Recall score of 78% and F1 score of 83%.

i. SelectKBest:



```
           Classification_report
              precision    recall  f1-score   support

           0       0.88      0.78      0.83      4792
           1       0.96      0.98      0.97     25560

    accuracy                           0.95     30352
   macro avg       0.92      0.88      0.90     30352
weighted avg       0.95      0.95      0.95     30352
```

Figure 3.17: Classification Report (Extra Tree)

In figure 3.18, the performance of Extra Tree for SelectKBest feature selection is given which represent the TP, TN, FP and FN.

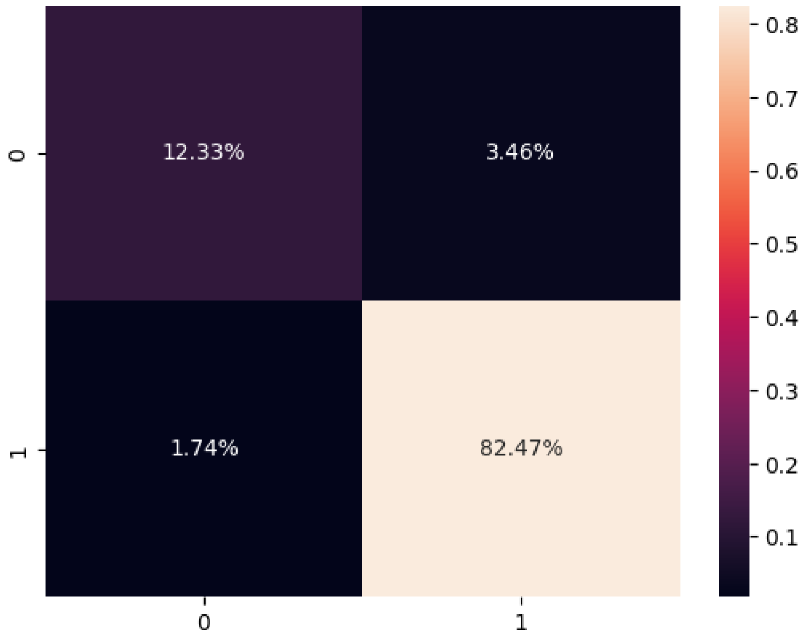
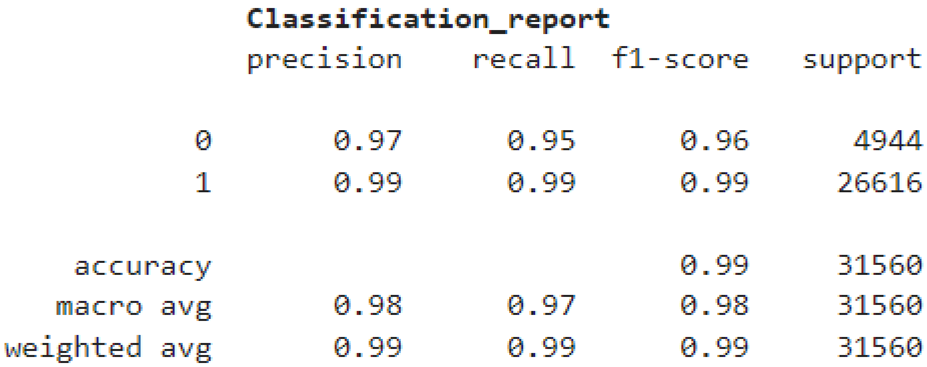


Figure 3.18: Confusion Matrix (Extra Tree)

In figure 3.19, Extra Tree achieve 98.75% for correlation feature selection technique, Precision score 97%, Recall score of 95% and F1 score of 96%.

ii. Correlation:



```
              Classification_report
              precision    recall  f1-score   support

           0       0.97      0.95      0.96      4944
           1       0.99      0.99      0.99     26616

    accuracy                           0.99     31560
   macro avg       0.98      0.97      0.98     31560
weighted avg       0.99      0.99      0.99     31560
```

Figure 3.19: Classification Report (Extra Tree)

In figure 3.20, the performance of Extra Tree for SelectKBest feature selection is given which represent the TP, TN, FP and FN.

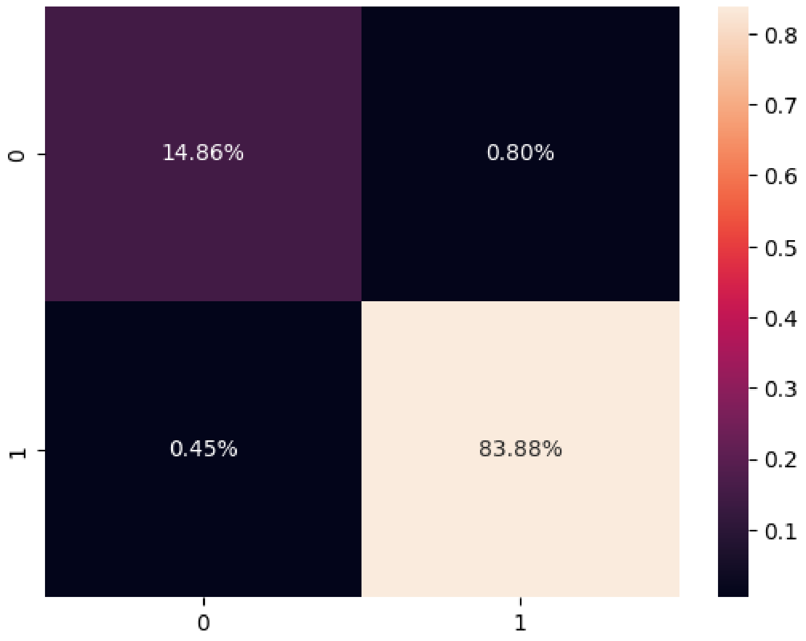
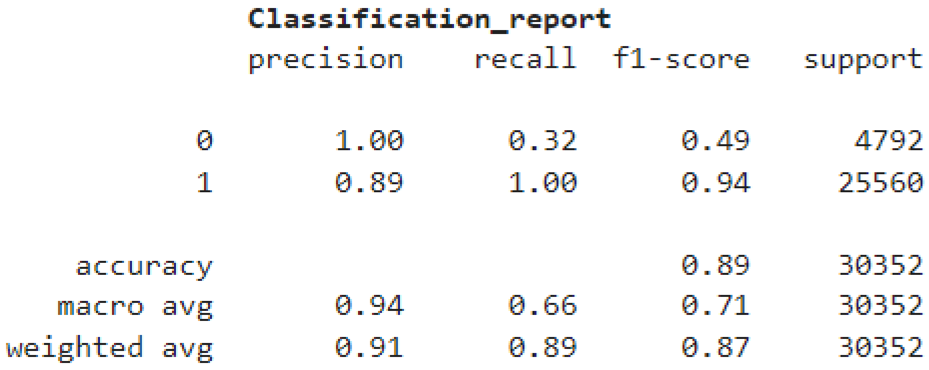


Figure 3.20: Confusion Matrix (Extra Tree)

5. **GaussianNB:** Gaussian Naive Bayes classifier is a popular machine learning algorithm used for classification purposes. It works using Bayes' Probability Principle and assumes that features are independent of their labels, hence the term "pure". Despite its simplicity and logic, Gaussian Naive Bayes has proven useful in many applications. It is especially useful when the features follow a Gaussian (normal) distribution. In figure 3.21, GaussianNB achieve 89.29% for SelectKBest feature selection technique, Precision score 1%, Recall score of 32% and F1 score of 49%.

i. SelectKBest:



```
                Classification_report
                precision    recall  f1-score   support

           0       1.00      0.32      0.49      4792
           1       0.89      1.00      0.94     25560

    accuracy                           0.89     30352
   macro avg       0.94      0.66      0.71     30352
weighted avg       0.91      0.89      0.87     30352
```

Figure 3.21: Classification Report (GaussianNB)

In figure 3.22, the performance of GaussianNB for SelectKBest feature selection is given which represent the TP, TN, FP and FN.

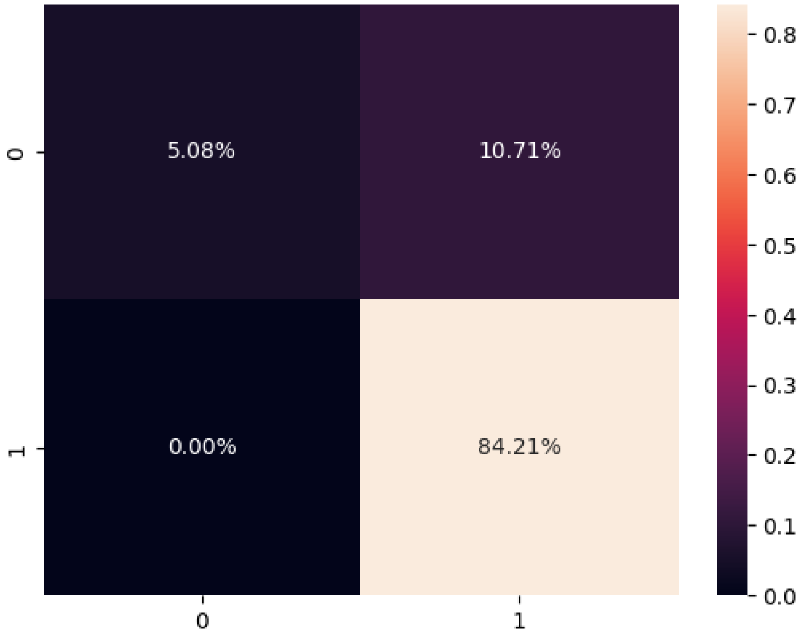
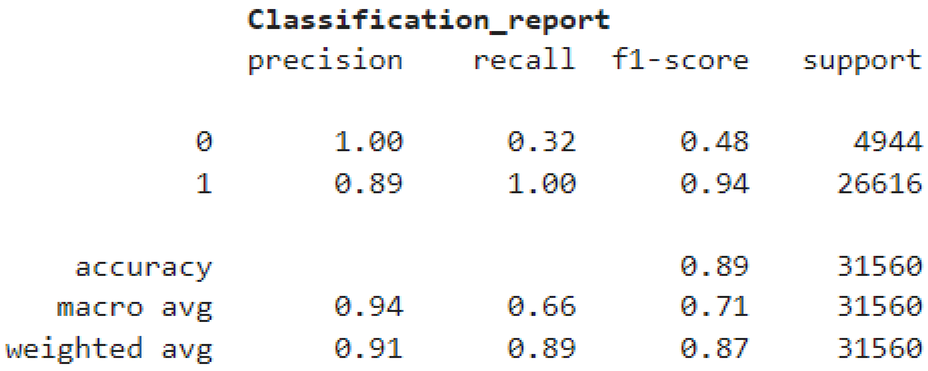


Figure 3.22: Confusion Matrix (GaussianNB)

In figure 3.23, GaussianNB achieve 89.34% for correlation feature selection technique, Precision score 1%, Recall score of 32% and F1 score of 48%.

ii. Correlation:



```
                Classification_report
                precision    recall  f1-score   support

           0       1.00      0.32      0.48      4944
           1       0.89      1.00      0.94     26616

    accuracy                           0.89     31560
   macro avg       0.94      0.66      0.71     31560
weighted avg       0.91      0.89      0.87     31560
```

Figure 3.23: Classification Report (GaussianNB)

In figure 3.24, the performance of GaussianNB for correlation feature selection is given which represent the TP, TN, FP and FN.

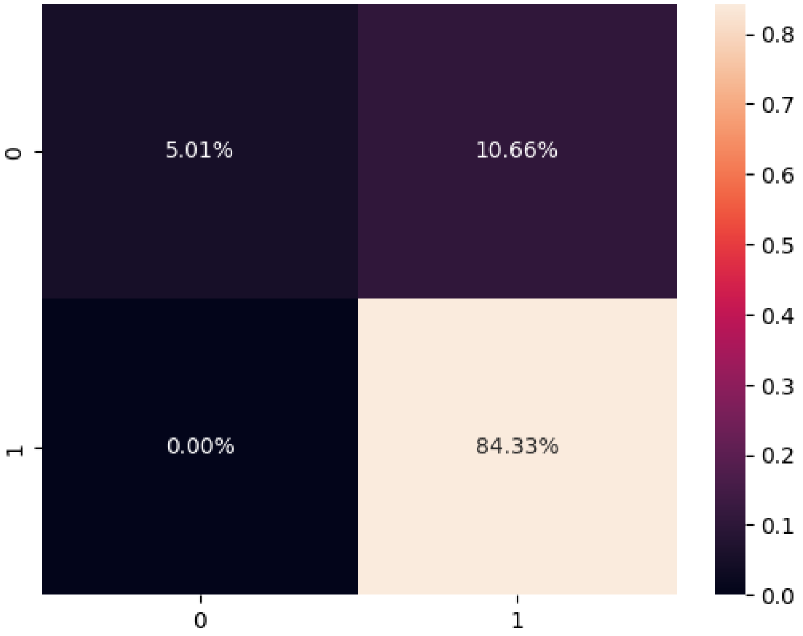


Figure 3.24: Confusion Matrix (GaussianNB)

By analyzing the result and classification report, I can say that Correlation feature selection technique perform better than SelectKBest. I choose Correlation based model for my detection project.

In figure 3.25, the graph demonstrates the difference between two features selection technique for 5 model I have used in this project. The correlation-based feature selection technique performs much better than SelectKBest in accuracy. Deciasion tree got 94.67% accuracy for SelectKBest and 98.68% for correlation. Similarly Random forest classifier

96.12% for SelectKBest and 98.77% for correlation, K-Neighbor 89.1% for SelectKBest and 96.3% for correlation, Extra Tree classifier 94.8% for SelectKBest and 98.75% for correlation and GaussianNB 89.29% for SelectKBest and 89.34% for correlation respectively. Comparing the result, I can come to a conclusion that correlation based feature selection technique is a better approach for feature selection method.

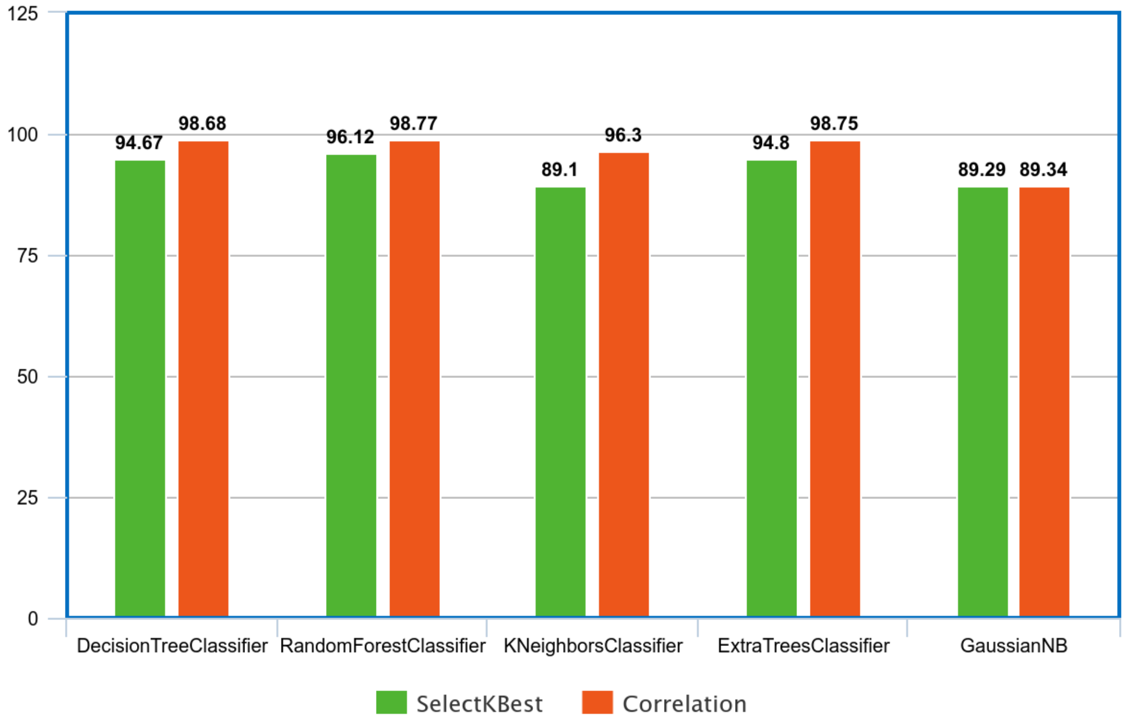


Figure 3.25: Performance Evaluation

4.3 Discussion

My research will use machine learning (ML) techniques to predict the detection of malicious activities that lead to cybercrime. Every attribute should play an important role in classifying research topics. The first goal of my research is to identify malicious signals. Data is one of the most important parts of any research. The results of the same test can vary greatly depending on the data provided. Because I am sharing data, I understand that results obtained by others using any of the previously published public data for this test may differ from my results when compared to actual data. I achieve my goals by using various machine learning techniques on reliability and average scores. I used a total of five different algorithms for this project. After choosing the algorithm, I start working on it. I then evaluate the accuracy of each algorithm. Meanwhile, I can achieve the Random Forest classifier (based on the classification of the following five models) with the highest accuracy of 98.77%. I have once again created a project to investigate Intrusion.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Machine learning-driven intrusion detection systems (IDS) have a huge impact on humans, both positive and negative. Unlike signature-based methods, machine learning-based IDS can detect new and unknown threats quickly and efficiently. These important processes, financial transactions, personal information, etc. is important for its protection. These systems can instantly analyze large amounts of data, allowing them to detect and respond to attacks faster, minimizing damage and protecting lives. Advanced algorithms used in these systems can distinguish malicious actions from network behavior, thus reducing security vulnerabilities and saving resources that would otherwise be wasted. Automation facilitated by machine learning allows security teams to manage large amounts of information and threats with fewer people, thereby reducing operational costs and allocating resources to other activities. Machine learning models continue to learn and adapt to new attack patterns, keeping defenses ahead of attackers and improving overall security. Privacy issues related to the collection and analysis of large amounts of data on the web are significant. It is important to balance security and privacy. Improper use of electronic equipment can lead to dangerous attacks, highlighting the need for ethical and protective measures. The introduction of machine learning simplifies the process and uncovers new attacks that need to be carefully evaluated and mitigated. Overall, machine learning has a positive impact on intrusion detection and greatly improves network security. However, negative issues need to be addressed through the development of accountability, deployment and ongoing monitoring to ensure that the device protects the person's safety and rights.

5.2 Impact on Environment

Despite the negative consequences of machine learning, the environmental impact of machine learning cannot be ignored when entering cybersecurity. The needs to train and run machine learning models will be significant, leading to significant energy needs and potentially higher carbon emissions. However, this problem can be solved by choosing efficient algorithms, taking advantage of advances in hardware and using fire products. Renewable energy. Continuous progress in the development of hardware and software required for advances in machine learning can lead to the creation of electronic products. To solve this problem, responsible recycling practice and the use of sustainable products are important. In addition, the production of specialized learning equipment for making materials often involves mining, which can have environmental impacts such as deforestation and depletion. In order to preserve these benefits, it is important to prioritize the sustainability of the economy and follow the circular economy model**.** Despite the negative effects of machine learning, the environmental impact of ML in the field of cybersecurity cannot be ignored. Training and running machine learning models require large amounts of energy and can lead to increased carbon emissions. However, this problem can be solved by choosing efficient algorithms, taking advantage of hardware development and using renewable energy sources. Continued advances in hardware and software critical to machine learning advancements can lead to wasted energy. To solve this problem, responsible recycling practice and the use of sustainable products are crucial. Additionally, the production of training equipment is often associated with mining, which can cause environmental impacts such as deforestation and depletion. In order to preserve these benefits, business sustainability must be prioritized and business circulation standards must be adhered to.

5.3 Ethical Aspects

Leveraging machine learning (ML) to access research raises ethical issues that require adequate attention. Monitoring the website or user's behavior may be a violation of privacy and therefore consent is required to monitor personal information to address privacy concerns. The presence of flaws and bad flaws in machine learning models can have serious consequences, affecting innocent people or leading to crime. The presence of bias in educational materials can lead to discrimination, highlighting the importance of quality control. Transparency plays an important role in ensuring accountability and trust by enabling machine learning standards to be defined. Obtaining consent from users is an ethical responsibility and data protection is essential to prevent unauthorized access or disclosure. Striking a balance between the benefits of automation and human care is important, including cultural and legal differences around the world. Continuously measuring impact and encouraging ongoing dialogue is critical to responding to changing technology and societal trends. Efforts to reduce the harm and reduce morale of engaging in search should also be an important part of their use.

5.4 Sustainability Plan

This research will help reduce the number of cyberattack incidents all over the internet. Our research has concluded that security in using the site will be sufficient. Based on this research, I can detect cases of intrusion using ML technology that detects anomaly activities. If you search on the internet, you will find the information more easily. Key elements such as training updates, resource efficiency, and defined standards have proven effective and efficient over time. Good works such as good financial means, seamless integration, teams with expertise and fair use to ensure good physical results. By evaluating and regularly updating these features, it can leverage the power of machine learning to securely adapt to changing threats while optimizing resources and managing operations.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This project has taught us a great aspect about this topic. Prediction of intrusion are still a complex topic. As a result, I used machine learning to categorize seven major risk: 'MITM', 'Fingerprinting', 'Ransomware', 'Uploading','SQL_injection', 'DDoS_HTTP', 'DDoS_TCP', 'Password','Port_Scanning', 'Vulnerability_scanner', 'Backdoor', 'XSS','Normal', 'DDoS_UDP', and 'DDoS_ICMP'. The use of the machine learning (ML) method has also been used to predict whether its and intrusion or not. As I have stated before, the study's goal is to learn as much as feasible about this topic. To do this, I collected recent data from online that combine 63 categories. With the help of the attribute of the dataset, I were able to analyze and train our machine learning models in precisely identifying the intrusion. The predictive model makes it easy to determine whether its intrusion or not. Some initial problems have been resolved. I met the need. For many students, different methods produce different results. This is covered in more detail in the next section

6.2 Conclusions

The recent increase in technologies has also led to an increase in cybercrime. To solve this problem, an intrusion detection system (IDS) is required to detect and report attacks. But finding a specific attack is difficult. Researchers around the world are interested in this topic and in particular the use of supervised learning algorithms for access control. This project presents an IDS model to compare the performance of different methods. Two features selection method is used to select feature. and then used to train different classes. In this project SelectKBest and Correlation based feature selection techniques is used. The reason for using a mixture of feature selection algorithms and classifiers is that each algorithm has its own advantages and disadvantages, making it difficult to choose one over the other for using an intrusion detection system. Experimental results show that machine learning can be effectively used to access search because all connections lead to truth. For this project, multiple machine learning model such as DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier, ExtraTreesClassifier, GaussianNB. Among the model, DecisionTreeClassifier shows the best performance, while Correlation based feature selection method outperforms other classifiers, while GaussianNB performs the worst. Therefore, this study concludes that the combination of Correlation feature selection and DecisionTreeClassifier can be used to create effective access detection.

6.3 Implication for Further Study

I've learnt a lot and I will keep digging further studies with this research. Intrusion became a big thread now a days as the technology evolving. There are many ways and method to detect and prevent intrusion. Further study can find more effective way to mitigate it. Deep learning method could be more effective for detection. More feature extraction method can improve the model accuracy.

References

[1] D. K. Anish Halimaa A, "MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM," in *Proceedings of the Third International Conference on Trends in Electronics and Informatics*, 2019.

[2] A. L. M. A. A. J. M. D. S. J. H. M. Akhil Krishna, "Intrusion Detection and Prevention System Using Deep Learning," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems*, 2020.

[3] M. C. A. A. M. K. Usman Shuaibu Musa, "Intrusion Detection System using Machine Learning Techniques: A Review," in *Proceedings of the International Conference on Smart Electronics and Communication*, 2020.

[4] J. A. Shikha Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science ELSEVIER*, vol. 60, pp. 708-713, 2015.

[5] A. R. O. M. A. K. H. M. R. F. M. S. a. D. M. F. N. F. Haq, "Application of machine learning approaches in intrusion detection system: A survey," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 2, pp. 9-18, 2015.

[6] Ahmed, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016.

[7] A. L. B. a. E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2015.

[8] K. S. V. A. &. H. T, "Machine learning classification model for network based intrusion detection system," in *11th International Conference for Internet Technology and Secured Transactions (ICITST)*, Barcelona, Spain, Dec 2016.

[9] D. &. M. Sreekesh, "A Two-tier Network based Intrusion Detection System Architecture using Machine Learning Approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 3-5 March 2016.

[10] K. M. R. a. M. A. R. Bhavani T. T, "Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques," in *International Conference on Sustainable Technologies for Computational Intelligence (ICSTCI)*, Springer, 2020.

[11] P. R. e. al, "Implementation of Machine Learning Algorithms for Detection of Network Intrusion," *International Journal of Computer Science Trends and Technology (IJCST)*, pp. 163-169, 2020.

[12] M. Z. a. C.-H. L, "Evaluation of Machine Learning Techniques for Network Intrusion Detection," *IEEE*, pp. 1-5, 2018.

[13] S. K. S. A. a. S. B. Verma P, "Network Intrusion Detection using Clustering and Gradient Boosting," in *International Conference on Computing, Communication and Networking Technologies (ICCCNT),*

2018.

[14] B. M. a. M. R. Kazi A., "Network Intrusion Detection using Supervised Machine Learning Technique with feature selection," in *International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2019.

[15] A. a. Almseidin, "Machine Learning Methods for Network Intrusions," in *International Confrernce on Computing, Communication (ICCCNT)*, 2018.

[16] B. M. a. M. R. Kazi A., "Network Intrusion Detection using Supervised Machine Learning Technique with feature selection," in *International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2019.

[17] I. a. Aftab, "A Feed-Forward ANN and Pattern Recognition ANN Model for Network Intrusion Detection," in *Researchgate* , International Journal of Computer Network and Information Security.

[18] V. Y. a. K. K, "Anomaly Based Network Intrusion Detection using Ensemble Machine Learning Technique," *International Journal of Research in Engineering, Science and Management. IJRESM,* pp. 290-296, 2020.

[19] K. M. R. a. M. A. R. Bhavani T. T, "Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques," in *International Conference on Sustainable Technologies for Computational Intelligence (ICSTCI)*, Springer, 2020.

[20] M. e. al., "Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches," *International Journal of Intelligent Engineering and Systems.INASS,* pp. 433-445, 2020.

[21] P. R. e. al, " Implementation of Machine Learning Algorithms for Detection of Network Intrusion," *International Journal of Computer Science Trends and Technology (IJCST),* 163-169.

[22] S. T. a. A. Mailewa, " The Role of Intrusion Detection/Prevension Systems in Modern Computer Networks: A Review," in *Midwest Instruction and Computing Symposium (MICS)*, Wisconsin, USA.

[23] B. –C.V, "Feature Selection and Classification in Multiple class datasets-An application to KDD Cup 99 dataset," in *https://doi.org/10.1016/j.eswa.2010.11.028* , 2012.

[24] S. T. a. A. Mailewa, " The Role of Intrusion Detection/Prevension Systems in Modern Computer Networks: A Review," in *Midwest Instruction and Computing Symposium (MICS).*, Wisconsin, USA.

A Machine learning approach to detect Intrusion

ORIGINALITY REPORT

19%	16%	5%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	4%
3	Preeti Mishra, Vijay Varadharajan, Uday Tupakula, Emmanuel S. Pilli. "A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection", IEEE Communications Surveys & Tutorials, 2018 Publication	1%
4	dokumen.pub Internet Source	1%
5	Submitted to University of North Carolina, Greensboro Student Paper	<1%
6	Submitted to TechKnowledge Student Paper	<1%
7	Submitted to Vels University Student Paper	<1%

8	fastercapital.com Internet Source	<1 %
9	Submitted to CSU, San Jose State University Student Paper	<1 %
10	Submitted to Coventry University Student Paper	<1 %
11	Submitted to Sheffield Hallam University Student Paper	<1 %
12	www.researchgate.net Internet Source	<1 %
13	Usman Shuaibu Musa, Megha Chhabra, Aniso Ali, Mandeep Kaur. "Intrusion Detection System using Machine Learning Techniques: A Review", 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020 Publication	<1 %
14	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
15	Submitted to University of Gloucestershire Student Paper	<1 %
16	Submitted to Royal Holloway and Bedford New College Student Paper	<1 %

17	www.coursehero.com Internet Source	<1 %
18	Submitted to Athlone Institute of Technology Student Paper	<1 %
19	www.ijraset.com Internet Source	<1 %
20	www.sersc.org Internet Source	<1 %
21	"First International Conference on Sustainable Technologies for Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1 %
22	Submitted to The University of the West of Scotland Student Paper	<1 %
23	Bhukya Madhu, M. Venu Gopala Chari, Ramdas Vankdothu, Arun Kumar Silivery, Veerender Aerranagula. "Intrusion detection models for IOT networks via deep learning approaches", Measurement: Sensors, 2023 Publication	<1 %
24	Submitted to University of West London Student Paper	<1 %
25	coek.info Internet Source	<1 %

26	pdfs.semanticscholar.org Internet Source	<1 %
27	Submitted to University of Westminster Student Paper	<1 %
28	www.livehacking.com Internet Source	<1 %
29	medium.com Internet Source	<1 %
30	www.slideshare.net Internet Source	<1 %
31	uwe-repository.worktribe.com Internet Source	<1 %
32	www.ijeat.org Internet Source	<1 %
33	www.mdpi.com Internet Source	<1 %
34	dr.ntu.edu.sg Internet Source	<1 %
35	ebin.pub Internet Source	<1 %
36	hal.archives-ouvertes.fr Internet Source	<1 %
37	link.springer.com Internet Source	<1 %

38	orbilu.uni.lu Internet Source	<1 %
39	researchonline.gcu.ac.uk Internet Source	<1 %
40	www.ndss-symposium.org Internet Source	<1 %
41	"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2021 Publication	<1 %
42	Kanwarpartap Singh Gill, Vatsala Anand, Rupesh Gupta. "Foetal Health Classification Using Input Dataset and Fine Tuning it using K-nearest Neighbour, Naïve Bayes and Decision Tree Classifier", 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI), 2023 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

A Machine learning approach to detect Intrusion

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51