MACHINE LEARNING ALGORITHMS TO DETECT RANSOMWARE ATTACK

by

NIKSHIPTHA PALLA

ST20208468

Master's degree in Data Science

Cardiff School of Technology

Cardiff Metropolitan University

ABSTRACT

The degree of sophistication of hacking strategies and the number of tools that are accessible to hackers are both increasing because of the ongoing development of technology in the contemporary world, which also brings with it an increase in the number of tools that are available to hackers. Because of this, the aggressor will feel encouraged to engage in acts of moral and financial blackmail because of the situation. As a direct result of this, it is essential to take prompt action against assaults of this kind, including ransomware. Additionally, it is essential to make use of technology to avoid ransomware by taking safeguards. We will use defences that are based on the idea of machine learning to protect ourselves against assaults of this kind. Our goal is to be able to forecast ransomware assaults, and the use of machine learning will help us do that. For the aim of attack prediction, the ransomware dataset may be used, and it is feasible to do so.

There are many methods of machine learning that have been brought up to this point and are going to be evaluated using performance measures like accuracy, precision, recall, and confusion matrix. These metrics, along with others, are going to be contrasted in our proposal. We will choose the machine learning algorithm combination that, when paired with a different types ofcombinations, delivers the highest number of results by using the outcomes of all performance measurements. This algorithm combination will be the one that we choose. Python is a computer language that may be used for a variety of applications, including the virtualisation of data, the analysis of data, and the prediction of ransomware based on the training and testing of a particular dataset.

ACKNOWLEDGEMENT

Working in this research has been one of the most rewarding experiences of my life. This research has assisted me in developing knowledge and diverse analytical skills. I would like to express my sincere thanks to my supervisor, Dr. Sandra Smart-Akande, for her continuous support and motivation throughout the journey of this research. The meetings and conversations were vital in inspiring me to think outside the box, from multiple perspectives to form a comprehensive and objective thesis.

Words cannot express my gratitude to thank my parents and my husband Lokesh, for their never-ending support and faith during the entire journey**.**

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
ML	Machine Learning
SVM	Support Vector Machine
ROC	Receiver Operating Characteristics
AUC	Area under curve
LSTM	Long term short memory
MD5	Message digestive
TP	True positive
TN	True negative
FP	False positive
FN	False-negative

CHAPTER1: INTRODUCTION

1.1 Title of Research

"Machine learning algorithms to detect Ransomware attack ".

1.2 Introduction

One of the most frightening sounds in "Cyber Security" is a Ransomware attack. A ransomware attack is one of the black hat attacks. Ransomware is a sort of virus that may limit or prevent a user from accessing their data, operating system, or device to demand money from the user. Ransomware can also encrypt the user's data. This is done with the intention to swindle money out of the user. Crypto-ransomware and locker ransomware are the forms of ransomware with which you are most likely to come into contact. The Locker ransomware presents its victims with a lock screen that prevents them from accessing their machines and asks for money in return. Crypto ransomware is a type of malicious software that, after encrypting sensitive data on a user's computer with complex encryption algorithms, then prompts the user to make a payment to decrypt the contents of an encrypted file. Ransomware has been around for a while, and throughout that time it has evolved into a form that is more pervasive, more complicated, and more devastating. "**Petya**, **Not Petya, Wanna cry, Golden Eye"** are some of the pretty well-known ransomware attacks in history. Reports that date back to the beginning of 2017 indicate that the total losses and profits that may be ascribed to ransomware are believed to have crossed the one-billion-dollar mark. These reports indicate that ransomware is responsible for these losses and earnings. (Groot 2018).

Machine learning is an efficient method for identifying malware in systems that are running Android as well as systems that are running Windows OS, as stated in the reference (Milosevic et al. 2016). An additional study on the capabilities of machine learning in the detection of malware is proposed as an alternative to the practice of using signatures in the article that serves as a reference (Anderson et al. 2011). This alternative is presented as an alternative to the practice of using signatures. This research presents evidence that detection approaches that depend on signatures are not as successful as those that rely on machine learning to identify threats. Because of their adaptability and strong ability to detect previously unseen samples of ransomware malware, it was decided to evaluate machine learning and deep learning approaches rather than other methods that were not based on machine learning. It has been

established via research that can be found in Reference (Zakaria et al. 2017) that methods for machine learning are more effective than approaches for static code analysis. With the ability to handle large and complex data sets and different types of data and fastness machine learning algorithms deal the ransomware attack in a pretty good way.

The rest of this paper is structured as follows:

Chapter2: Literature review and various existing methodologies and in-depth understanding of new terminology.

Chapter3: In this chapter, research methodologies will be explained along with the data set description.

Chapter4: Implementation of the proposed research methodology and results evaluation are included in this chapter.

Chapter5: It describes the conclusion based on the obtained results and future works described here.

1.3 Research Background

Cybercriminals and other malicious actors may use a broad array of attack tactics to generate new types of malwares in today's modern environment. In recent years, the kind of hack that has occurred the most often is known as ransomware attacks. In a manner like those of other security flaws in, it is very difficult for us to decrypt the data that has been encrypted by ransomware. There have been 290 malware attacks on companies in Europe, encompassing a broad range of sectors; of them, 50% are other ransomware attacks. The vast bulk of these assaults are taking place in European countries (virustotal.com, 2017). On the other hand, it is close that there have been 1.37 million new models submitted for the cyber-attacks. The period that must elapse before an attack and before its behaviour is seen is what differentiates ransomware from other forms of malicious software (malware). A simple malware attack will begin by hiding itself behind the programs next, it will infect the system; finally, it will damage the computer without first obtaining consent from the ransomware's creator.

All the remedies that are now accessible, including those that were proposed in previous study work, have not been effective in halting or blocking any malware attacks, and several obstacles must be overcome to find ransomware outbreaks. Ransomware must be differentiated from other kinds of attacks to safeguard users' devices from being compromised by ransomware-based attacks. The number of ransomware attack varieties is continually increasing year after

year. Even while ransomware and other types of attacks have certain similarities, each one also has its own set of traits that set it apart from the others.

To illustrate, ransomware will often carry out a variety of file-related operations in a condensed period to lock or encrypt data on the computer of a person who has been compromised by the infection. Signature-based malware detection methods make it difficult to identify zero-day ransomware and are not suitable for protecting users' data from harmful unknown ransomware attacks. These methods also make it difficult to identify the ransomware that was created in the past. It also becomes more difficult to identify the ransomware that was generated in the past because of these tactics. Because of this, ransomware requires a specialized security mechanism, and that mechanism must focus on ransomware-based operations to distinguish itself from other types of malicious software and harmful files.

1.4 Research Motivation

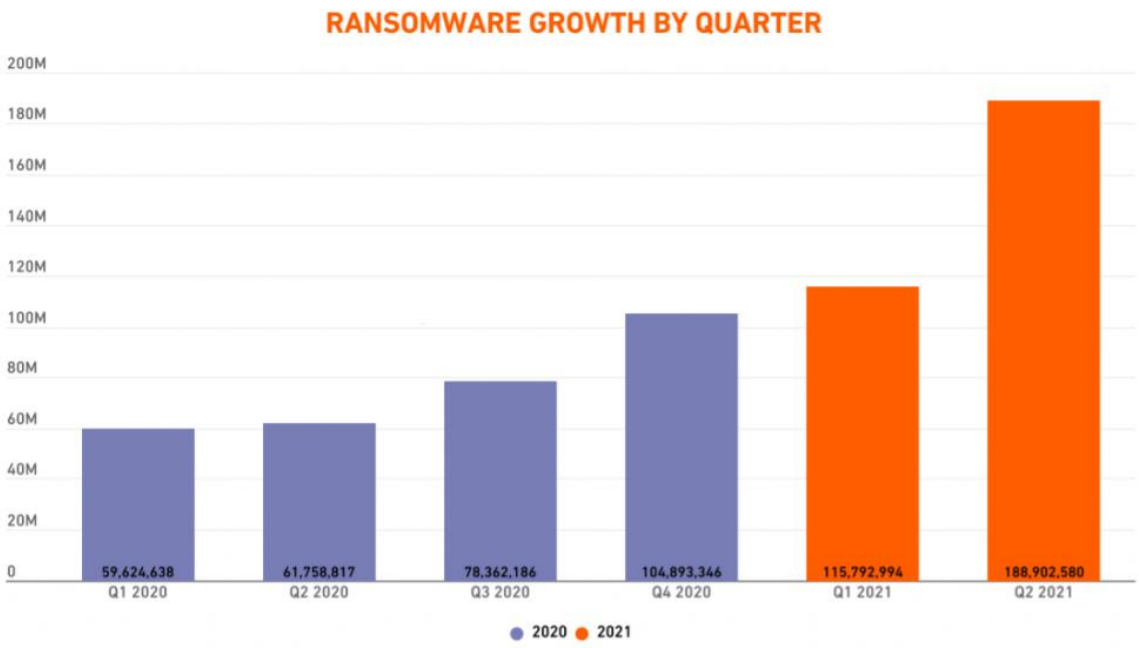

Figure 1:Ransomware attacks

(Source by threatpost.com, 2021)

The above figure shows the growth of the ransomware attacks for the years 2020-2022 by quarter, which is in an increasing manner (threatpost.com, 2021). It is a clear sign that, ransomware is one of the hazardous attacks which needs to deal with an ideal and accurate

solution. We never know the timing or target of an assault; therefore, we must forecast it in advance on the time based on several symptoms.

1.5Aim of the Research

The work aims to develop a machine learning model that can accurately anticipate ransomware attacks. In addition, one of our goals is to evaluate the performance metrics and outcomes of the many machine learning algorithms and choose the one that performs the best overall so that it may be combined with one another based on the requirements to get more accuracy in detection.

1.6 Objectives of the Research

- We may investigate the significance of machine learning in the age of information security.
- To identify the best algorithm for predicting ransomware attacks on encrypted data.
- Different machine learning algorithms and their implementation with the ransomware data.

1.7 Research Questions

- How can a machine learning algorithm deal with the most common attack in the era of information security?
- Which machine learning algorithm is suitable to detect ransomware attacks?
- To forecast ransomware assaults, is it possible to use various Machine learning algorithms in conjunction with one another?
- How should input properties be chosen such that assaults may be predicted with high accuracy?

CHAPTER 2: LITERATURE REVIEW

2.1 Ransomware Attack

Attacks using ransomware are becoming more common, each one exhibiting a unique set of behaviours and indicators. As a result, we were unable to identify Ransomware simply on its outward look. This is because Ransomware often modifies its behaviour, and the features of Ransomware are distinct. Now, we are unable to uncover ransomware assaults that can safeguard users' data since most of these attacks are based on unknown ransomware attacks. Therefore, the author creates a novel defence system to discover any ransomware assaults (Seong et al. 2019). The functioning of Ransomware and a few other forms of malware-based innocuous files was the primary focus of this study.

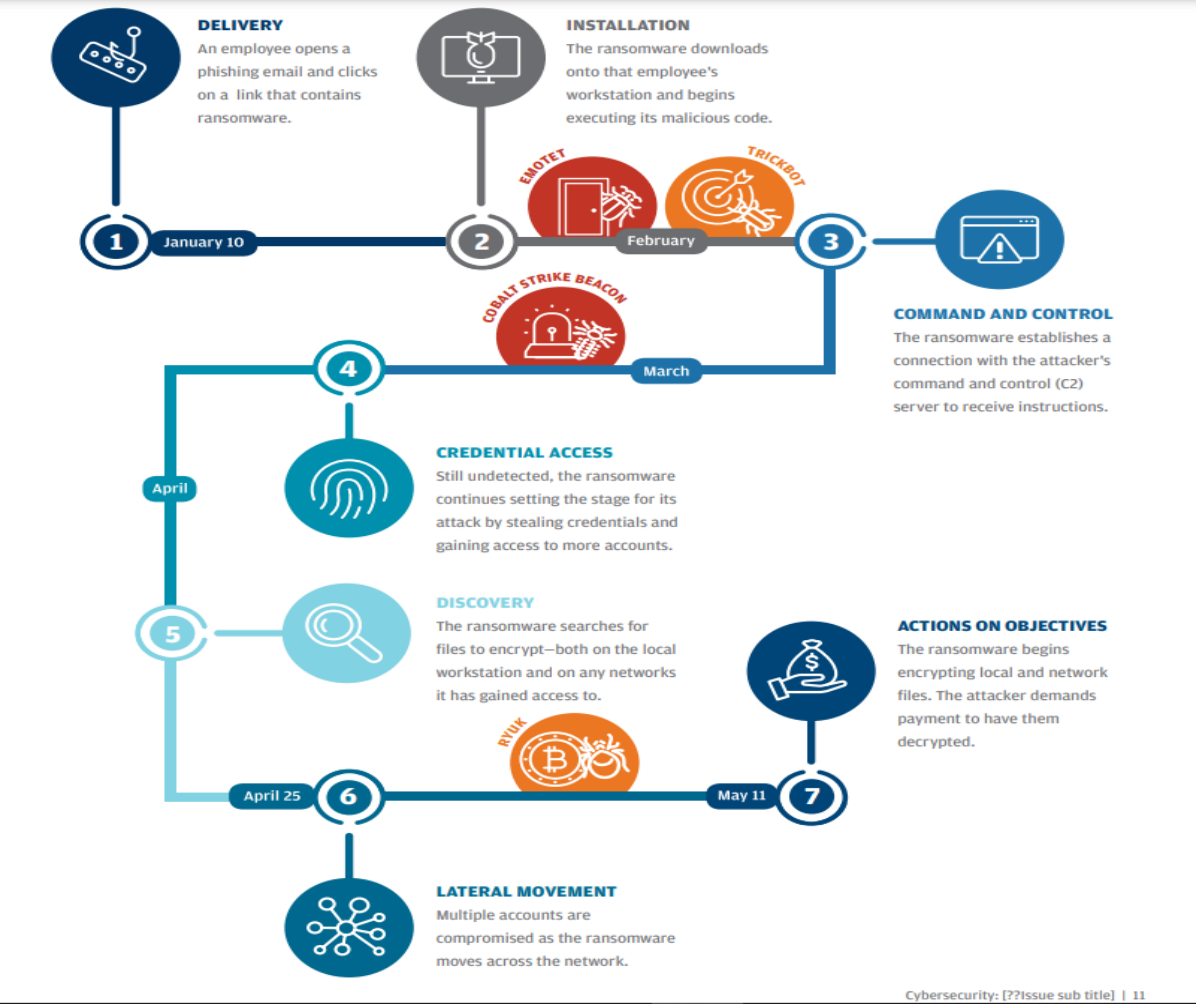

Figure 2: Seven stages of a ransomware attack

(The anatomy of a ransomware attack, n.d.)

The above figure explains the different stages of ransomware attacks right from opening an email to demanding the ransom amount from the victims. COVID-19 also contributed to the emergence of many ransomware attacks in recent times due to remote work and online services. There are different types of ransomware attacks are named based on their working strategy and agenda and nature. Ryuk, Maze, Lockbit, and DearCry are some of the popular ransomware variants. These ransomware attacks enter the system either by compromising the system credentials are network or through phishing emails or messages. Once the system is compromised, hackers demand Ransom amounts. RYUK is one of the most expensive ransom variants which demands an average of $ 1million.

Two popular varieties of ransomware: Locker, the first kind of ransomware, has as its main purpose to prevent victims from accessing the contents stored on their computers. The second kind of ransomware, known as Crypto Locker, is responsible for the encryption of data. It is the intention of Crypto Locker to encrypt personal data to prevent its victims from accessing those encrypted files (Shijo et al. 2015). If a piece of malware known as Lockers Ransomware infects a computer, the user will no longer be able to use either the infected computer in its whole or a particular piece of software that is kept on the computer that has been compromised and ransomware prevents the user from accessing the machine.

Best practices like giving cyber awareness education, regular software updates, multi-factor authentication, and data backup plans can stop the spreading of these ransomware attacks. Though many experts documented these practices, still we are encountering the attacks with devasting results. This leads to Machine learning algorithms to solve the problems in the era of security.

2.2 Existed detection techniques and challenges

2.2.1 Ransomware Analysis:

The ransomware attack has become very unknown after the “wannacry” attack though it was first discovered in 2013 by using RSA public key encryption. “Wannacry” happened in May 2017 year. It effected around 230,000 machines in 150 countries around the globe. Hackers demanded $300 bitcoins and increase it to $600 bitcoins within the time frame. (Kaspersky, 2019). Following this occurrence, several studies were conducted to identify and analyse ransomware assaults.

Ransomware analysis and detection methods have been classified into three main types

Static analysis:

Static analysis is an approach that does not execute directly on the original or target machine or code. Instead, Static features are those that can be observed without running the software, such as the size of the file.

Dynamic Analysis: Dynamic characteristics are those that can be observed while the software is running. For example, the size of the file is a static feature, but the file's size is a dynamic characteristic. Characteristics of how a piece of malicious software interacts with a computer system are referred to as the malware's characteristics. dynamical Interactions with a computer system may be caused by malware in several ways. (Grant et al. 2018).

Hybrid Analysis: Hybrid analysis is the combination of both static and dynamic analysis patterns. Various surveys and research have taken place in these three various varieties of analysis.

2.2.2 Existed Ransomware detection techniques

The dynamic analysis pattern has been used in most ransomware detection tools.

EldeRan is one of the frameworks that detect ransomware attacks by identifying the dynamic behaviour of the computer. During the assault, API requests and data strings were captured. It examined the distinct characteristics of a ransomware assault before injecting it into the system. Though this framework produces the best results with a high accuracy rate in identifying ransomware attacks, it has a few drawbacks in detecting ransomware that remains silent for some time and waits for the user to react. EldeRan did not extract the features properly in this instance. (Sgandurra et al., 2016).

HoneyPot techniques are also one of the tools which monitor and log the information and actions of unknown authorizations in the machine. It helps to analyse the pattern and kind and type of attack but anyhow, Honeypot only detects and monitors when hackers are working actively. Nowadays there are many ways hackers can enter the system using many passive ways (Moore, 2016).

LSTM-based malware detection, A long term short memory neural network with deep learning knowledge of the data set is proposed. This framework will also analyse behavioural patterns using API calls; the Cuckoo box works on API hooks to store API calls. It is often referred to as dynamic detection analysis. LSTM-based malware detection achieved 67.60

percent accuracy, although it can only operate for a limited time of fewer than 30 minutes (Maniath et al., 2017).

Crypto ransomware detection system, Locky ransomware was used as a case study in this ransomware detection. On the MCFP dataset, PCAP files were acquired and analysed. Many characteristics were collected from Locky's network operations utilizing various protocols such as TCP, HTTP, and DNS. It has been tested on two different levels: packet level and flow level. These retrieved attributes indicate hacking activity (Almashhadani et al., 2019).

DeepRan, an attention-based bi-directional LSTM detector, is used to identify system abnormalities and prevent network encryption. It pulls information from host logs obtained from bare metal servers using the TF-IDF method. DeepRan classifier is deployed on the aberrant data from these host logs in the second step.The below figure determines the clear working nature of the DeepRan detector and classifier.

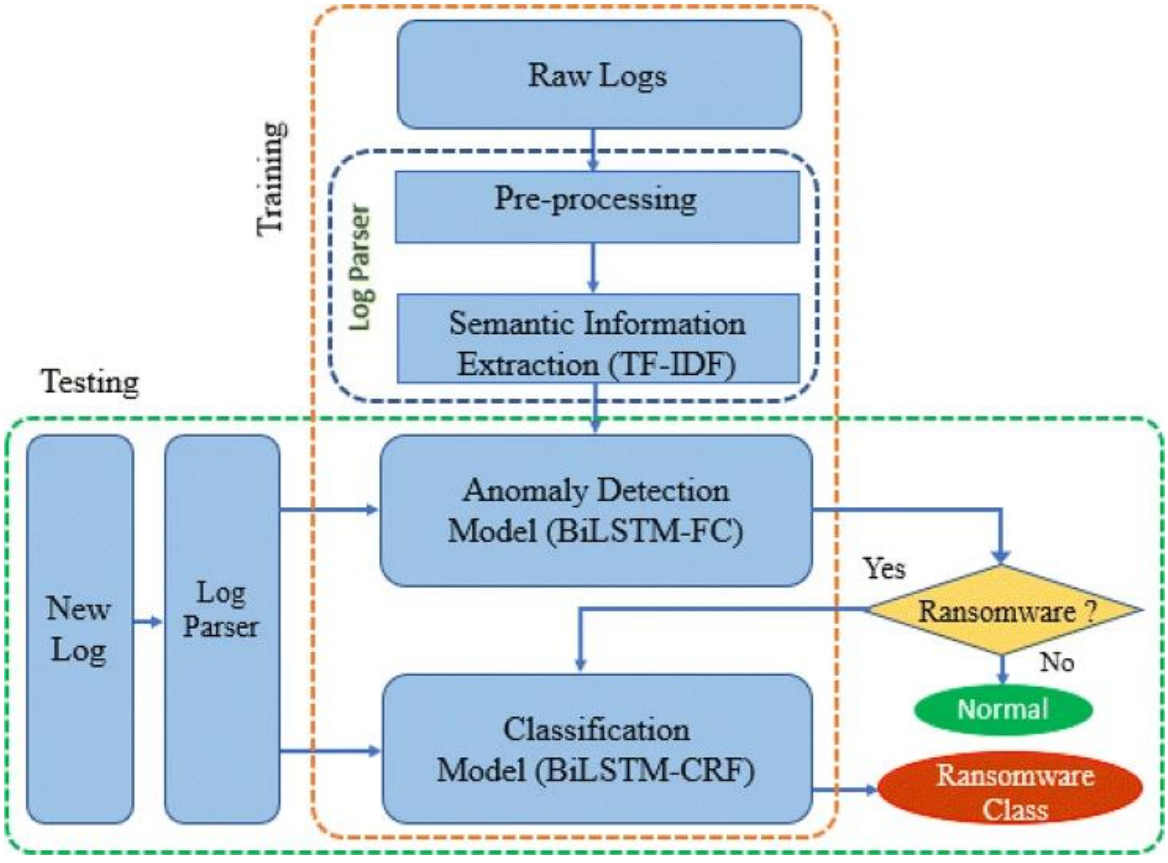

Figure 3:DeepRan framework

(Source by Roy and Chen, 2020)

Even though the DeepRan detector and classifier delivered 99.87 percent and 96.5 percent accuracy, respectively, they were unable to identify the zero-ransomware assault among the 17 ransomware families (Roy and Chen, 2020).

2.3 Machine Learning algorithms

Ransomware is an issue that impacts all aspects of society, including the government as well as private enterprises, according to the conclusions of the research that was carried out by the author (Usha et al. 2021). Because of this, they will probably grab the data and encrypt the user files, following which they will issue a demand for a ransom from the individuals who have been impacted. As a result, if you want to avoid problems of this kind, you should focus on building new approaches that are built on the machine learning algorithms that can deal with Ransomware attacks. According to the results of the research carried out on 2021 (Erik Larsen et al. 2021), the author includes a survey about machine learning techniques to predict the ransomware attack based on the input given data set, and then it has been applied to the exploratory data analysis, which is very useful to get an idea about the ransomware attacks.

A predictive model for ransomware attacks based on machine learning algorithms was proposed, and the authors (S. R et al. 2020) stated that the random forest algorithm is one of the best machine learning algorithms to perform ransomware attack detection because it provides the highest accuracy rate compared to other machine learning algorithms. This model has based on the authors' findings that the random forest algorithm is one of the best machine learning algorithms to perform ransomware attack detection. This approach searches for DLL properties such as code size and MD5 among other things using a variety of various sorts of features. These features are employed in this study. They were able to identify the ransomware assaults in advance of the actual operation.

The author of the study work done by (Khan et al. 2020) claimed that Ransomware has a huge number of categorization searches, including spyware, Ransomware, botnet Malware, rootkit, viruses, worms, and trojans. This information was obtained from the research work done by (Khan et al. 2020). Therefore, there are ransomware assaults that are depending on the activity of the attacker. Therefore, the author has concentrated their efforts on developing a new algorithm, which allows the machine learning algorithm to work in conjunction with the DNA Sequencing engine. (Ncube et al. 2020) DNA was analysed using an algorithm for machine learning. The ransomware attacks were based on continual surges, and they assessed the

performance metrics based on the suggested model. The attacks were triggered by the frequency factor. They have shown that they have been successful in classifying ransomware attacks based on the DNA sequence.

According to the findings of a research study that was carried out by (Ganta et al. 2020), attackers would make money because of ransomware attacks that are carried out on the data of users. This will influence the system, and when they have gained access to the assets, they will encrypt the user file that contains executable user files. (Harish et al. 2020) It is not expected that standard antivirus software would perform better. On The Ransomware Attacks, and it was unable to discover them earlier as mails. As a result, the method of machine learning has been the author's primary emphasis to identify the response attack for executable files. Additionally, it can anticipate ransomware assaults that are concealed inside the executable files.

Types of Machine learning algorithms:

There are four types of machine learning algorithms, which are supervised, semi-supervised, unsupervised, and reinforcement. There are two major terms involved in this subject, train data, and test data. The corresponding dataset will divide into a training dataset and a test dataset. The machine learning algorithm identifies the different patterns in train data. These considerations will apply to the test data for results. Supervised algorithms work with labelled data whereas unsupervised algorithms work with unlabelled data.

Classification and Regression fall under supervised algorithms, Classification algorithms are used to predict distinct values like True or False, Yes or No. Regression algorithms predict the information of real continuous values like salary, age, etc. For the below research, XGBoost and SVM and Logical Regression and Random Forest Algorithms were implemented for the given dataset.

Bagging

Although decision trees are one of the models that can be understood with the least amount of effort, their behaviours are very unpredictable. Consider a single training dataset that has been arbitrarily cut in half to create two separate datasets. Now, let's put each component through its decision tree training to get two different models.

When we put these two models through their paces, we would get quite different outcomes. Because of this tendency, it is argued that decision trees relate to a high level of variation.

Aggregation methods like bagging and boosting may assist limit the amount of variation that exists in every learner. Bagging is a learning strategy that uses many decision trees that are produced in tandem to build its basic learners. These learners get their instruction based on data that has been sampled with replacement. The ultimate prediction is the result obtained by averaging the output of all the learners.

Boosting:

When using boosting, the trees are constructed sequentially, and the goal of each successive tree is to minimize the mistakes that were produced by the tree that came before it. Each tree takes the knowledge it gained from its ancestors and uses it to correct any faults that remain. Because of this, the subsequent tree in the sequence will acquire knowledge using an improved version of the residuals.

The weak learners that are used in the boosting process are known as the base learners. These learners have a significant level of bias, and their predictive value is only slightly better than that of random guessing. The fact that each of these weak learners gives some essential information for prediction enables the boosting approach to efficiently combine these weak learners and build a strong learner because of this combination. The final strong learner lowers both the bias and the variance down to acceptable levels.

XGBoost algorithm:

XGBoost is a kind of learning technique called an ensemble. When faced with certain challenges, the outcomes of a single machine learning model may be insufficient on their own. The predictive potential of several learners may be combined via ensemble learning, which provides a methodical approach to this problem. The result is a single model that provides the output that is an aggregation of the results from numerous models.

The models that make up the ensemble, which is often referred to as base learners, might have been created using the same learning algorithm or they could have been created using an entirely different learning method. The two most common types of ensemble learners are known as bagging and boosting. Although these two methods may be used in a variety of statistical models, the application that has seen the greatest success yet has been with decision

trees.

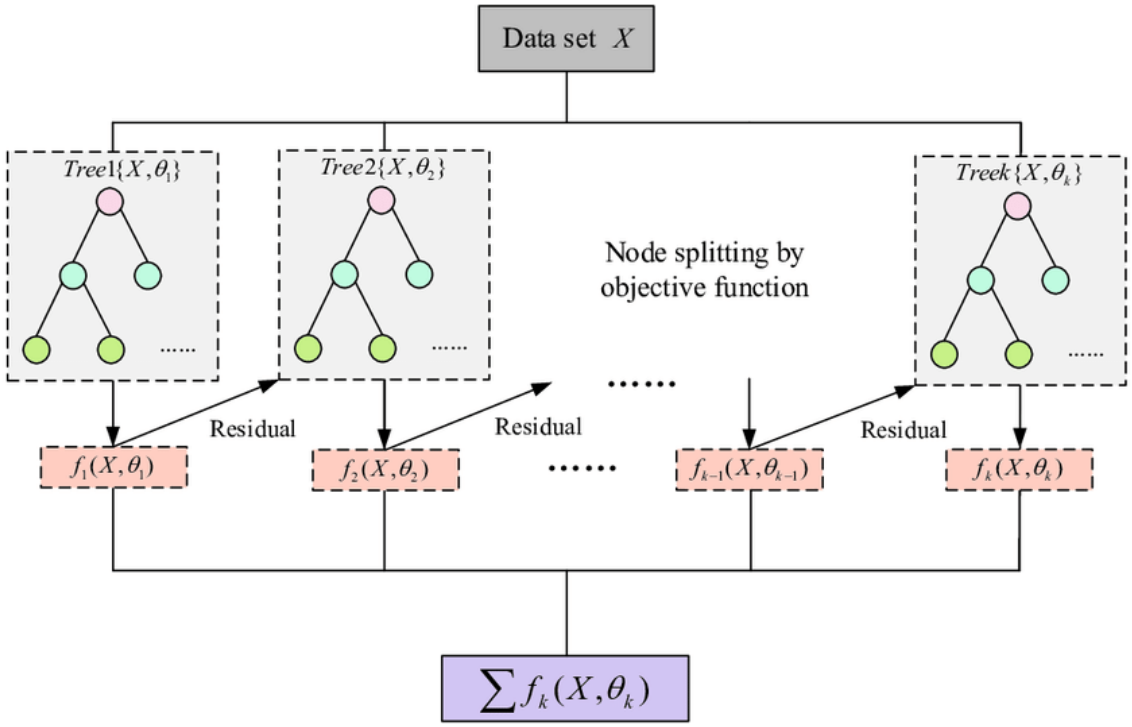

Figure 4:XGBoost framework

(Source by (GUO et al., 2020)

Support Vector Machine:

The Support Vector Machine, more often known by its acronym SVM, is one of the most widely used supervised learning algorithms. It may be used for issues involving classification as well as regression. However, its primary use is in the field of machine learning, namely for classification difficulties (javaTpoint.com 2022).

The purpose of the Support Vector Machine (SVM) technique is to generate the optimal line or decision boundary that can divide an n-dimensional space into classes. This will allow us to simply place any new data points in the appropriate category in the future. A hyperplane is a term used to describe this optimal decision boundary.

The extreme points and vectors that contribute to the creation of the hyperplane are selected using SVM. These exceptional circumstances are referred to as support vectors, which is how the method got its name: the Support Vector Machine. Look at the picture below, which shows how two distinct groups may be differentiated from one another with the use of a decision boundary or a hyperplane:

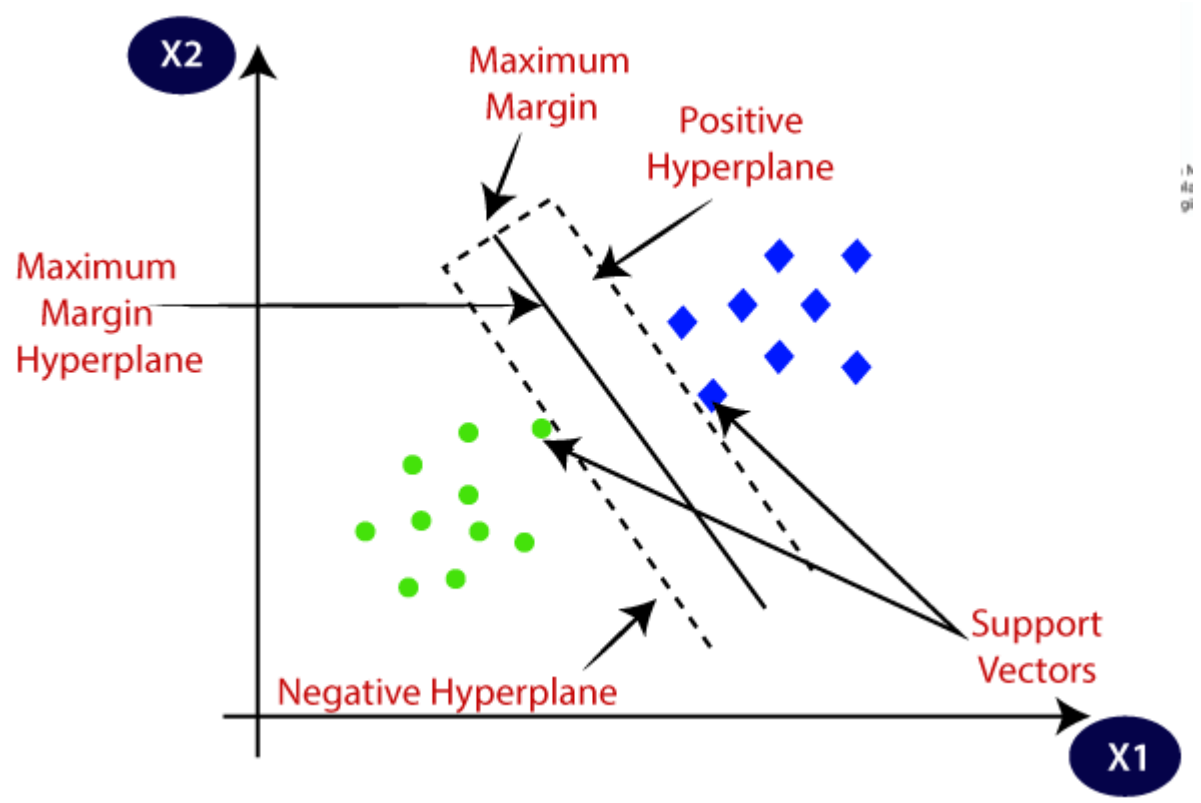

Figure 5: Support vector machine

Figure 5: Support vector machine

(Source by javaTpoint.com, 2022)

Logistic Regression Algorithm:

Logistic regression, which falls under the umbrella of the Supervised Learning methodology, is one of the most common and widely used machine learning algorithms. The categorical dependent variable may be predicted by utilizing it in conjunction with a predetermined group of independent factors. The outcome of a dependent variable that is categorical may be predicted using logistic regression. As a result, the output needs to be a value that is either categorical or discrete. It may either be yes or no, zero or one, the truth or a lie, etc. but rather than reporting the precise value as 0 or 1, it presents the probability values that fall between the two numbers. The Logistic Regression is quite like the Linear Regression, with the primary difference being how each method is used. When trying to solve regression difficulties, linear regression is the method of choice, but logistic regression is used when attempting to solve classification issues.

In logistic regression, rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values. This contrasts with traditional regression, which fits a straight line (0 or 1).

The curve that results from the logistic function shows the probability of anything happening, such as whether the cells are malignant, whether a mouse is a fat based on its weight, and so on. Logistic Regression is an important approach for machine learning since it can generate probabilities and classify new data by making use of continuous and discrete datasets. This makes it an extremely versatile algorithm.

Logistic regression is a method that can be used to categorize observations by using various kinds of data, and it can readily discover which factors are the most successful when it comes to classifying observations. The graphic that may be seen below illustrates the logistic function:

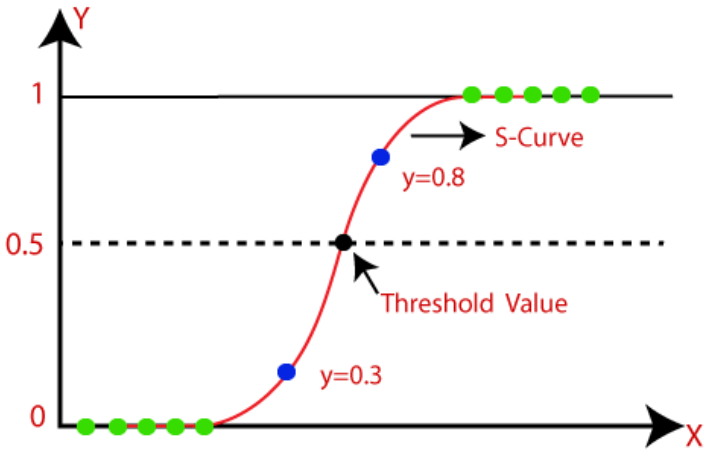

Figure 6: Logistic Regression

(Source by javaTpoint.com, 2022)

Random Forest algorithm:

The ensemble decision tree is known as a random forest. Random forest is the most popular machine learning algorithm which comes under in supervising technique. This algorithm approaches group learning based on the number of decision trees containing multiple decision trees in towards single random forest algorithm to improve the accuracy level for prediction. This will work better on both classification and regression problems. From this name itself, we can find that number of trees can be used to make a forest. The prediction of the random forest algorithm is based on the number of votes received from the multiple decision trees. It will take the consideration based on the maximum number of votes received from the decision trees. This performs well on the large data set for both classification and regression problems. It

provides maximum accuracy with the maximum number of decision trees. A map of the Random Forest is shown in the image below.

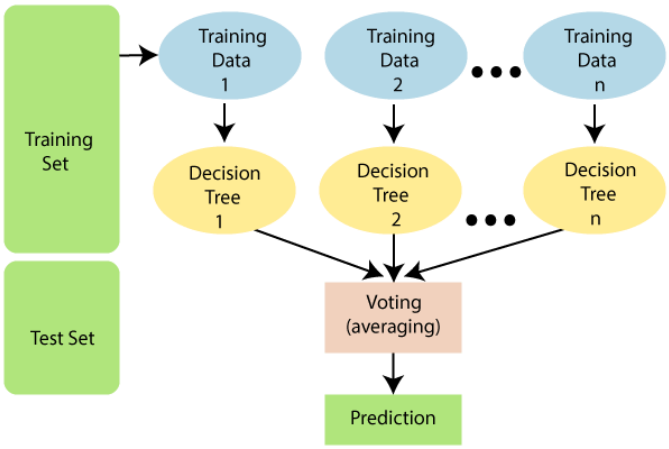

Figure 7: Random Forest Model

Figure 7: Random Forest Model. (Source by javaTpoint.com, 2022)

2.4 Performance metrics

Machine learning algorithms will be evaluated by performance metrics. For binary classification, the **confusion matrix** comprises two rows and two columns. A confusion matrix has the same number of rows and columns as classes. Columns represent expected classes, whereas rows represent actual classes. Four blocks of the confusion matrix are True positive, True negative, False positive, and False negative (Parte, 2020). According to this research TP, TN, FP, and FN values are based on good ware and malware files.

TP = Number of files were good ware files and model also classified the same.
TN= Number of files are ransomware files and model also classified the same.
FP= Number of files are malware files, but model classified as good ware files.
FN=Number of files are good ware files, but model classified as malicious files.

Accuracy is one of the most important metrics to define the performance of the algorithm. Accuracy is the ratio of some correct predictions to total predictions. The best fit will have an accuracy of more than 80% (Parte, 2020).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and **Recall** are other two metrics to evaluate the performance of ML algorithms. Precision is minimising the False positive predictions whereas Recall is used to reduce False negatives predictions. Both these metrics lie in the range from 0 to 1. F1-score is the harmonic mean of precision and recall (Parte, 2020).

$$\text{Precision} = \frac{TP}{Sum(TP+FP)}$$

$$\text{Recall} = \frac{TP}{Sum(TP+FN)}$$

Other visualisation metrics for machine learning algorithms include ROC and AUC Curves. ROC is an abbreviation for Receiver Operating Characteristics, whereas AUC is an abbreviation for Area Under Curve. The probability curve between the False positive rate and the True positive rate is known as the ROC. AUC is a measure of separability and the greater the AUC score, the better the classifier's performance. The AUC score can vary between 0 and 1(Parte, 2020).

CHAPTER 3. RESEARCH METHODOLOGY

3.1 Research Approach

The Research approach depends on the type and objective of the research. Research approach and design are interlinked with data analysis and data collection. Accordingly, research approaches are subdivided into three types:

- Deductive research approach
- Inductive research approach
- Abductive research approach

Deductive reasoning is based on previously established hypotheses or theories. Data is gathered to test the validity of the hypothesis. Finally, we can determine whether the hypothesis is correct.

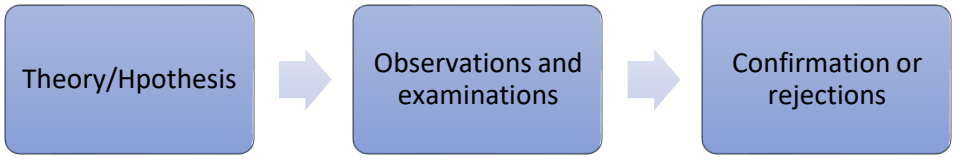

Figure 8: Deductive Approach

Source: (Dudovskiy, 2015)

Using the obtained data, the inductive technique will identify new phenomena and hypotheses. The technique will be explained based on the outcomes and patterns found.

Inductive research approaches include prediction-based research and working with a specific sample of data. The abductive technique works with existing theory, either developing a new theory from it or altering it (Dudovskiy, 2015).

Figure 9: Inductive Approach

Source: (Dudovskiy, 2015)

The research approach used for the below research is the inductive approach, where we have taken the sample of data and performed different algorithms, and identified and observed

predictions based on the result. From the prediction values like accuracy and recall, we can select the best model and method to detect ransomware attacks.

This research used static analysis for byte-level properties to identify ransomware. In this work, byte-level static analysis is used to address dynamic analysis's shortcomings. The characteristics are retrieved from executable raw bytes, and then pattern mining is employed. Ransomware and good ware files are classified using the Random Forest machine learning classifier.

3.2 Research Design

There are two sorts of research designs available in any study: quantitative research and qualitative research. Quantitative research is all about statistics and numbers, and it provides generalized facts. Qualitative research is concerned with meanings and words to comprehend a topic or idea. As we are dealing with calculations and accuracy measurement on different machine learning algorithms, Quantitative research has been used in the given research method (Streefkerk, 2019).

In this research, by using static analysis, we will implement different machine learning algorithms, and based on the accuracy level, we can select based on its performance metrics. In other words, the method is called a combined machine learning algorithm. The following is an outline of the logic phases that would be engaged in the proposed methodology:

- Collection of Data.
- Data Cleaning.
- Explorative Data Analysis.
- Data visualization.
- Train and test data split.
- Training and testing for SVM, Logistic regression, Random Forest, and XGBoost.
- Evaluating the performance metrics for the ML algorithms.

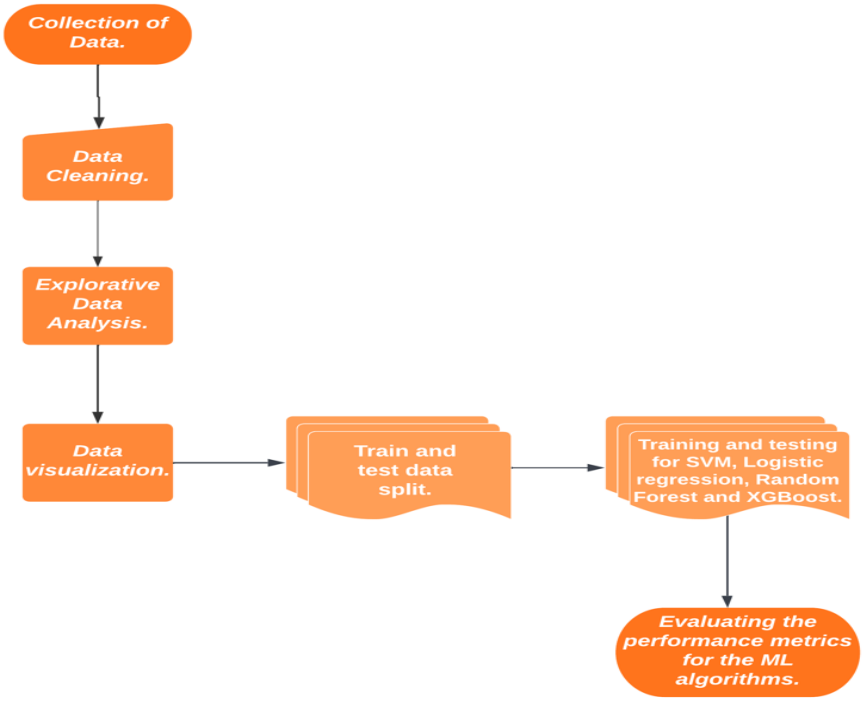

Figure 10: Design of research work

(Source by author)

3.3 Tools and Techniques

Machine algorithms play a key part in the field of "Information Security" in dealing with a wide range of issues ranging from data breaches to very dangerous assaults. It is due to its extensive data processing capability and other features. In the following study technique implementation, we will work with a dataset and employ several machine learning methods. The Jupyter Notebook application will use the Python programming language to execute coding and other tasks.

We are all aware that the Python programming language has extensive libraries for dealing with every type of operation. We imported many libraries to work with numbers and various machine learning methods. We divided the data into two groups based on the goal variable: train data and test data. These two datasets serve as input to many algorithms.

The "seaborn" and "matplotlib" libraries were used to visualise the graphs and plots.

The regression's "sklearn" library was utilized to deal with several classification approaches.

The XGBoost package was deployed in the background of the Anaconda program to ascertain the calculation metrics of the XGBoost algorithm. The Windows 11 operating system was used for this study and implementation.

3.4 Data Collection and Data Analysis

The dataset for ransomware attack detection is downloaded from rissgroup.org/ransomware-dataset in a CSV file format, which is a repository that is regarded as a secondary data source. The dataset comprises both good ware and ransomware files that were recovered from the Cuckoo sandbox in February of 2016. There are a total of 1524 samples, 582 of which are malicious files while the rest are excellent products. Ransomware is now divided into distinct families according to its characteristics. This dataset contains 11 different sorts of families, which are labelled with the "Type" column. Each family is assigned an ID, which is kept in the table below: (GitHub, 2022).

Family Name	**Variant ID**
Good ware	0
Critroni	1
CryptLocker	2
CryptoWall	3
Kollah	4
Kovter	5
Locker	6
Matsnu	7
PGPCoder	8
Reveton	9
TeslaCrypt	10
Trojan-Ransom	11

Table1: Ransomware families in the dataset

Source: (GitHub, 2022)

Along with the type of ransomware, there is another target variable which is a legitimate value that specifies whether the file is good ware or ransomware based on its value as mentioned in the below table.

Files	Value
Good ware	0
Ransomware	1

Table 2: Attack column value

The features of the datasets include the location, the year, the day, the length, the weight, and the count. Additionally, the neighbours, income, and label are all included. Every single one of the attributes, except for the location and the brand, makes use of the numeric data type. The address is thought of as belonging to the textual data type, whereas the title belongs to the categorical data type. This dataset takes into consideration social and ethical issues, and it does not include any personally identifying information about students or any other persons. Additionally, it does not include any information that may be used to identify the students. It is collected based on the talks that take place across the network. There are a lot of columns in our dataset, but the ones that stand out as the most significant are ID, Assault, and Type of assault. The following columns each include a time value that is expressed in milliseconds, as the unit of measurement.

CHAPTER 4: IMPLEMENTATION AND RESULTS EVALUATION

To identify ransomware attacks using a machine learning algorithm, detailed implementation steps have been taken, as detailed in the sections below.

4.1 Importing Libraries and dataset

The Python programming language has several libraries, which are sometimes known as modules. NumPy, Panda's, Matplotlib, and Sci-kit Learn are the main library packages required to develop predictive models for ransomware attack detection on ransomware data. Pandas are important in data cleansing, data exploration, and data visualization, among other things. The NumPy module, often known as Numerical Python, is used to perform numerical calculations.

Sci-kit learns package was used to deal with supervised and unsupervised algorithms, as well as to do classifications, regression, and clustering. It is a well-known open-source library.

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, confusion_matrix, roc_curve, roc_auc_score, classification_report, f1_scor
from sklearn.linear_model import LogisticRegression
```

Figure 11:Importing library packages

(Source by Author)

The dataset for ransomware attack detection is downloaded from rissgroup.org/ransomware-dataset in CSV file format. And then, the data set is extracted using Panda's libraries into a data frame. The ransomware data set has been uploaded and the sample data frame by using the head () function is shown in the below figure.

```
RansomeAttack = pd.read_csv(r"C:\Users\HP\OneDrive\Desktop\Project\Final dissertation\RansomwareData.csv")
```

```
RansomeAttack.head()
```

	10001	1	2	0	0.1	0.2	0.3	0.4	0.5	0.6	...	0.30925	0.30926	0.30927	0.30928	0.30929	0.30930	0.30931	0.30932	0.30933	0.30934
0	10002	1	3	1	0	1	0	1	0	1	...	0	0	0	0	0	0	0	0	0	0
1	10003	1	2	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	10005	1	5	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
3	10006	1	7	1	0	1	0	0	1	1	...	0	0	0	0	0	0	0	0	0	0
4	10007	1	5	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 12:ETL and Sample data frame

(Source by Author)

4.2 Exploratory data analysis

The attribute names for the data set are missing in this file, and we need to create attribute names for this data frame so we can generate attribute names for all the columns. For this, assigned the column names by using a string of column names. Exploratory data analysis has been performed for the ransomware attack data set, shown in the figure below.

```
columnames=['id','attack','type']
```

```
for i in range(30967):
    columnames.append(str(i))
```

```
RansomeAttack.columns=columnames #attribute name generation
```

```
RansomeAttack.columns
```

```
Index(['id', 'attack', 'type', '0', '1', '2', '3', '4', '5', '6',
       ...
       '30957', '30958', '30959', '30960', '30961', '30962', '30963', '30964',
       '30965', '30966'],
      dtype='object', length=30970)
```

Figure 13:Attribute name generation (Source by Author)

The statistical report of the data frame is shown below, which describes the mean-variance standard deviation minimum, maximum, and quartiles.

```
RansomeAttack.describe()
```

	id	attack	type	0	1	2	3	4	5	6	...	30957	3
count	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	1523.000000	...	1523.0	1523.00
mean	16806.216678	0.381484	2.029547	0.296126	0.003283	0.692712	0.001970	0.518713	0.027577	0.509521	...	0.0	0.00
std	4882.539498	0.485910	3.166189	0.456697	0.057222	0.461521	0.044353	0.499814	0.163812	0.500074	...	0.0	0.03
min	10002.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.00
25%	10807.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.00
50%	20232.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	0.000000	1.000000	...	0.0	0.00
75%	20754.000000	1.000000	3.000000	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	1.000000	...	0.0	0.00
max	21259.000000	1.000000	11.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	0.0	1.00

8 rows × 30970 columns

Figure 14: Statistical report (Source by Author)

We have checked for null values in the given dataset using the null () function, there are no null values observed. The data frame contains the ‘attack’ column, which is used for the label, and it includes the data about whether the data is regarding the attack on a normal one. It represents that data, 0 for standard and 1 for the attack. So, if we need to find, what is the

percentage of are a count of aggression that has been performed all over the data set. For that, we are filtering the data frame into another data frame, where the attack is equal to 1. This Ransom types of data frame contain the data about only attacked communication, which includes multiple types from 1 to 11. And we have created a count plot for the different types of ransomware attacks using the Matplot library, shown below figure.

```
Ransomtypes=RansomeAttack[RansomeAttack['attack']==1]
plt.figure(figsize=(12,8))
count_classes = pd.value_counts(Ransomtypes['type'], sort = False)
count_classes.plot (kind='bar')
plt.title("Count vs different types")
plt.xlabel("Type")
plt.ylabel("Count")
plt.show()
```

Figure 15: Filter and Plot attack data

Figure 15: Filter and Plot attack data

(Source by Author)

From the above histogram plot for different types of attacks, we can conclude that types 8 and 10 have the lowest count of occurrence, and type 2, 6, and 9 have the highest count of events where all other attack types are in the medium level. These types are different ransomware families which description available in the description of the dataset.

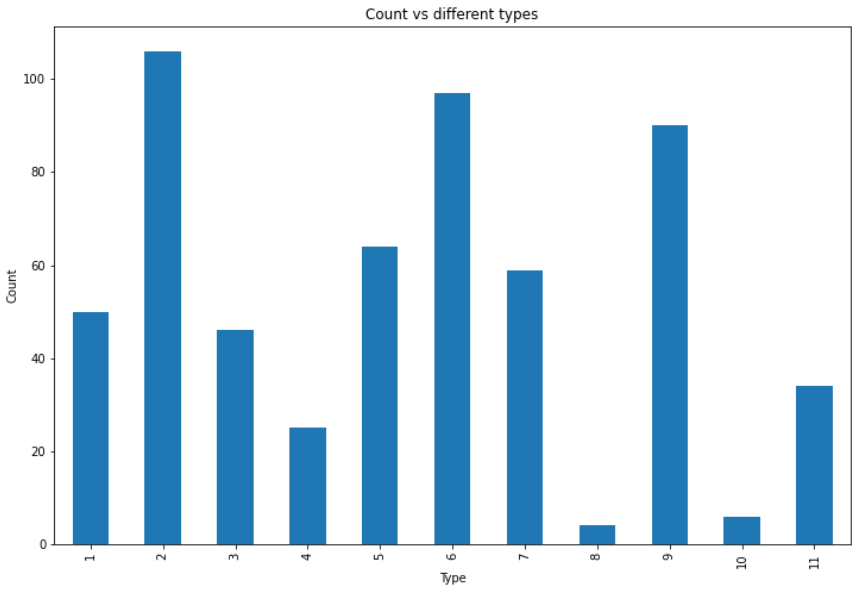

Figure 16: Histogram of different types of attacks

(Source by Author)

In the above bar plot, different type of ransomware attack families was compared with each other. From the visualisation, we reckon that crypto locker(type2) is the most happening type of ransomware in the given file. Type'8' has the lowest type of occurrence in the given file.

```
a=sum(RansomeAttack['attack'])
b=len(list(RansomeAttack['attack']))-a
fig = plt.figure(figsize = (10, 5))

# creating the bar plot
plt.bar(['No attack','Attack'], [b,a], color ='green',
        width = 0.4)

plt.title("Normal vs attack")
plt.xlabel("Ransomeware")
plt.ylabel("Count")
plt.show()
```

Figure 17: Plotting of Normal Vs Attack

(Source by Author)

We can create the plot for normal versus attack counts from the above figure. For this, first, we have obtained the count of the number of attacks and number of regular communications, then we have used a bar plot from the Mat plot library and the bar plot for frequent vs attacks as shown in the figure below.

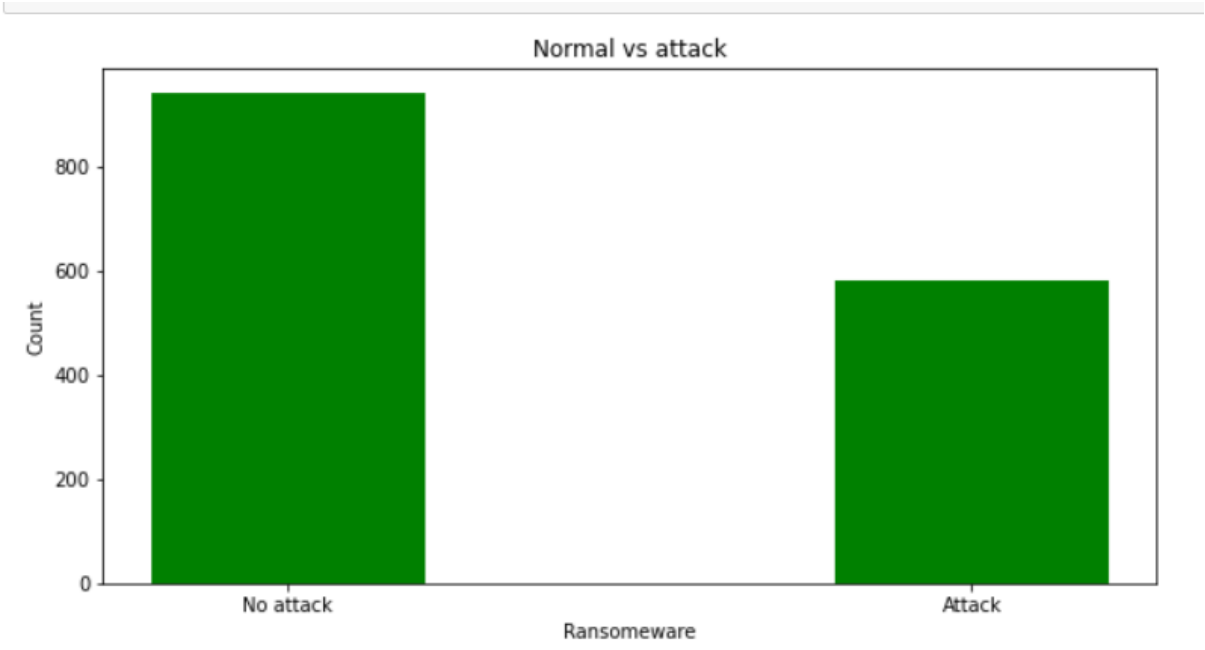

Figure 18: Normal Vs Attack bar plot (Source by Author

From the above plot, we know that more than 900 communications are routine, whereas more than 600 communications are regarding the attacks.

4.3 Classification based on an attribute of attack

The data set contains two types of output classification based on the attributes, ‘type’ and ‘attack’ are the two columns used to find whether the communication has an attack and the type of attack. In this proposed work, we have created two types of tours: label one is regarding the type of attribute, and another one is attack attributes, so we have developed machine learning models for SVM, Logistics Regression, XGBoost, and Random Forest. We have created two types of modelling based and the labels. In the first case, we used the target label's attack attributes.

In this classification, the ‘attack’ column is assigned to the 'y' variable, where all other columns related to input variables are stored in the 'X' variable. The input features and output targets are stored in the 'X' and 'y'. Then, these variable costs are applied to the train and test split, where the test size is 0.25. 75% of the data set is assigned to the training variable, and the remaining 25% of the data is given to the test variable. The train test split is shown in the below figure.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X,y,test_size = 0.25,random_state=100)
```

Figure 19: Train and test data split for attack classification

(Source by Author)

4.4 Machine algorithm application on attack classification

XGBoost Algorithm:

The predictive model for ransomware attack detection using XGBoost machine learning algorithm has been implemented as shown in the below figure, in which it has a creation of XGBoost model using XGBClassifier() and the training data, test data is classified using model predict function then predicted information is stored into a y_predicted. The classification report is a report which is used to evaluate the performances from the metric values called, Precision and Recall and F1-Score values.

The classification report is mentioned in the below list, and the confusion Matrix was the cause shown in the below figures.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.98	0.97	0.98	233
1	0.96	0.97	0.97	148
Accuracy			0.97	381
macro avg	0.97	0.97	0.97	381
weighted avg	0.97	0.97	0.97	381

Table 3: Performance metrics of XGBoost

(Source by Author)

From the above metric values, the XGboost algorithm has an accuracy of 09.7% and the F1-S core is near to 1 indicating that this is one of the best fit models. The below figure represents t he

confusion matrix of the XGBoost algorithm.

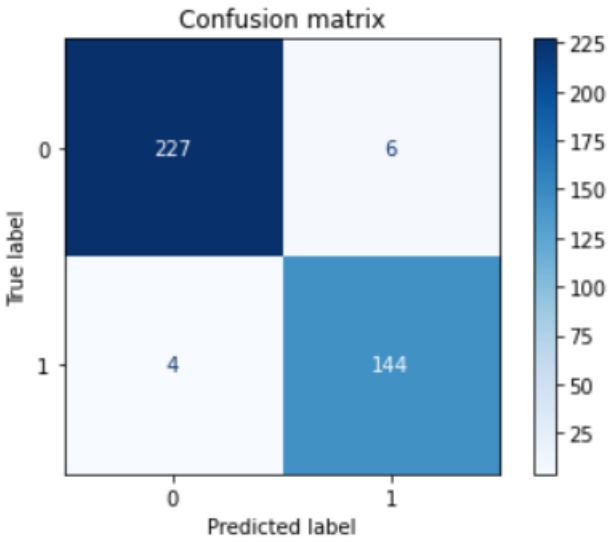

Figure 20: Confusion Matrix for XGBoost

(Source by Author)

The above Confusion matrix has four values which are True Positive, True negative, False positive, and False negative.

TP = 227 files were good ware files, and the model was also classified the same.

TN= 144 files are ransomware files, and the model is also classified the same.

FP= 6 files are malware files, but the model is classified as good ware files.

FN=4 files are good ware files, but the model is classified as malicious files.

In a nutshell, FP and FN values should be as possible as low the quality of a good machine learning algorithm.

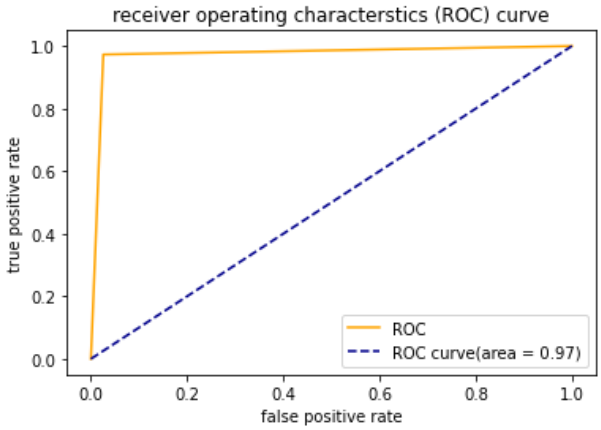

Figure 21: ROC curve for XGBoost

(Source by Author)

The AUC score for the XGBoost model is 0.973, and the ROC curve for this model is shown i n the above figure. The accuracy and recall are 0.97, where precision is 0.98.

Support Vector Machine algorithm:

The predictive model for ransomware attack detection using SVM machine learning algorithm has been implemented as shown in the below figure, in which it has a creation of SVM model and the training data, test data is classified using model predict function then predicted inform ation is stored into a y_predicted. The classification report is mentioned in the below list, and the confusion Matrix is shown in the below figures.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.99	0.97	0.98	233
1	0.95	0.98	0.97	148
Accuracy			0.97	381
macro avg	0.97	0.97	0.97	381
weighted avg	0.97	0.97	0.97	381

Table 4: Performance metrics of SVM (Source by Author)

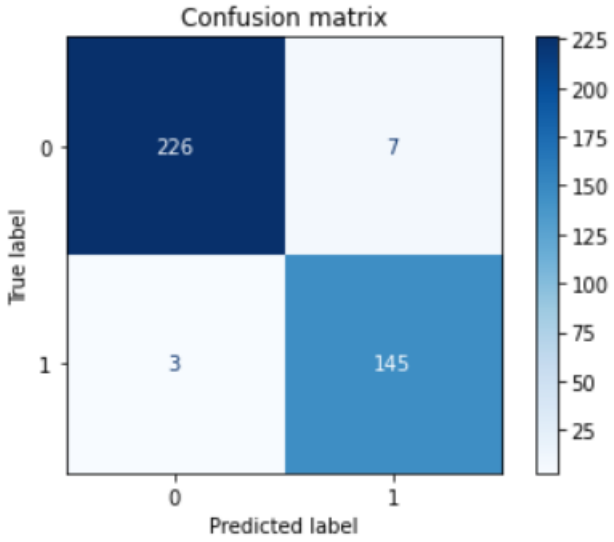

Figure 22: Confusion Matrix for SVM

(Source by Author)

From the above two interpretations, the accuracy of the SVM algorithm is 0.97% and FP and FN values are low when compared to the True positive and True negative values. The F1 Score value is also near 1 showing that SVM is also one of the good algorithms to predict ransomware attacks.

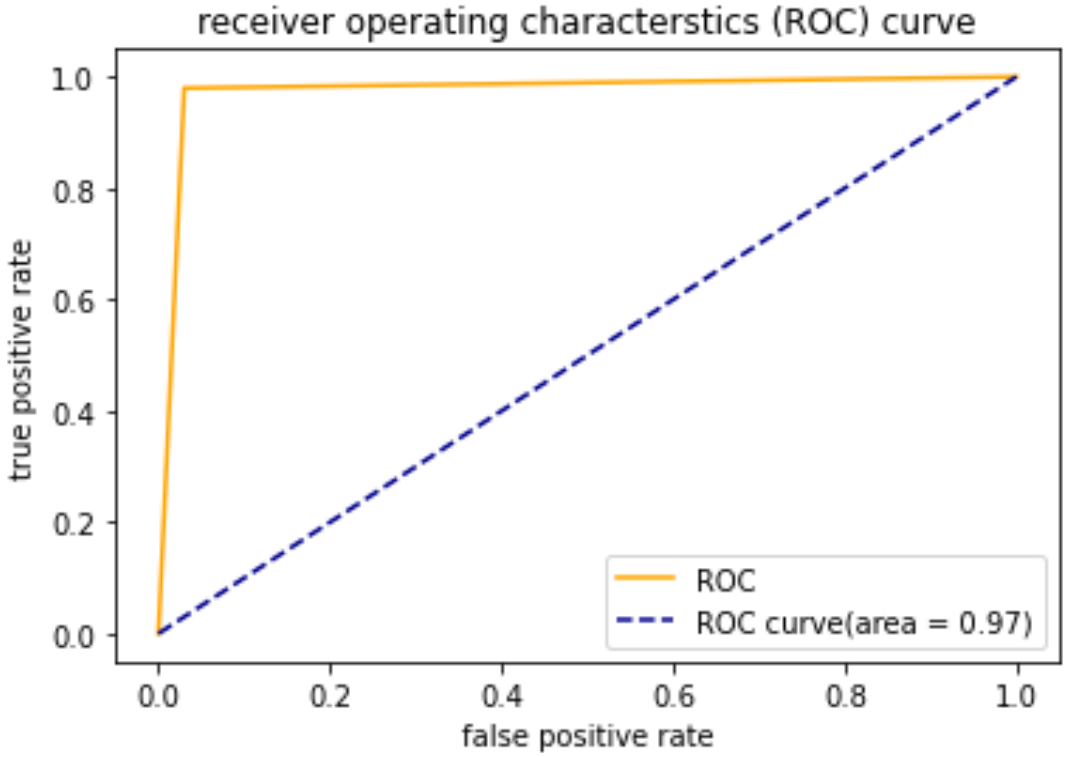

Figure 23: ROC curve for SVM

(Source by Author)

The AUC score for the SVM model is 0.974, and the ROC curve for this model is shown in the above figure. The accuracy and recall are 0.97, where precision is 0.99.

Logistic Regression machine learning algorithm:

The predictive model for ransomware attack detection using Logistic Regression machine learning algorithm has been implemented as shown in the below figure, in which it has a creation of Logistic Regression model and the training data, test data is classified using model predict function then predicted information is stored into a y_predicted. The classification report is mentioned in the below list, and the confusion Matrix is shown in the below figures.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.99	0.97	0.98	233
1	0.95	0.99	0.97	148
Accuracy			0.98	381
macro avg	0.97	0.98	0.98	381
weighted avg	0.98	0.98	0.98	381

Table 3: Performance metrics of Logistic Regression

(Source by Author)

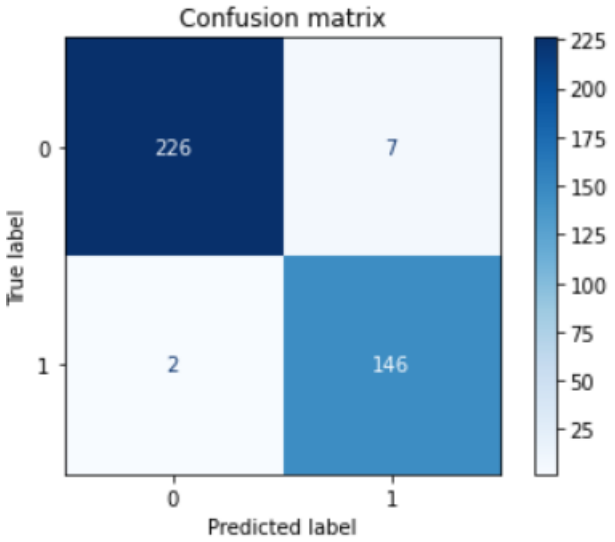

Figure 24: Confusion Matrix for Logistic Regression

(Source by Author)

From the above interpretation, Logistic Regression also has fewer False positive and False negative values when compared to True positive and True negative values.

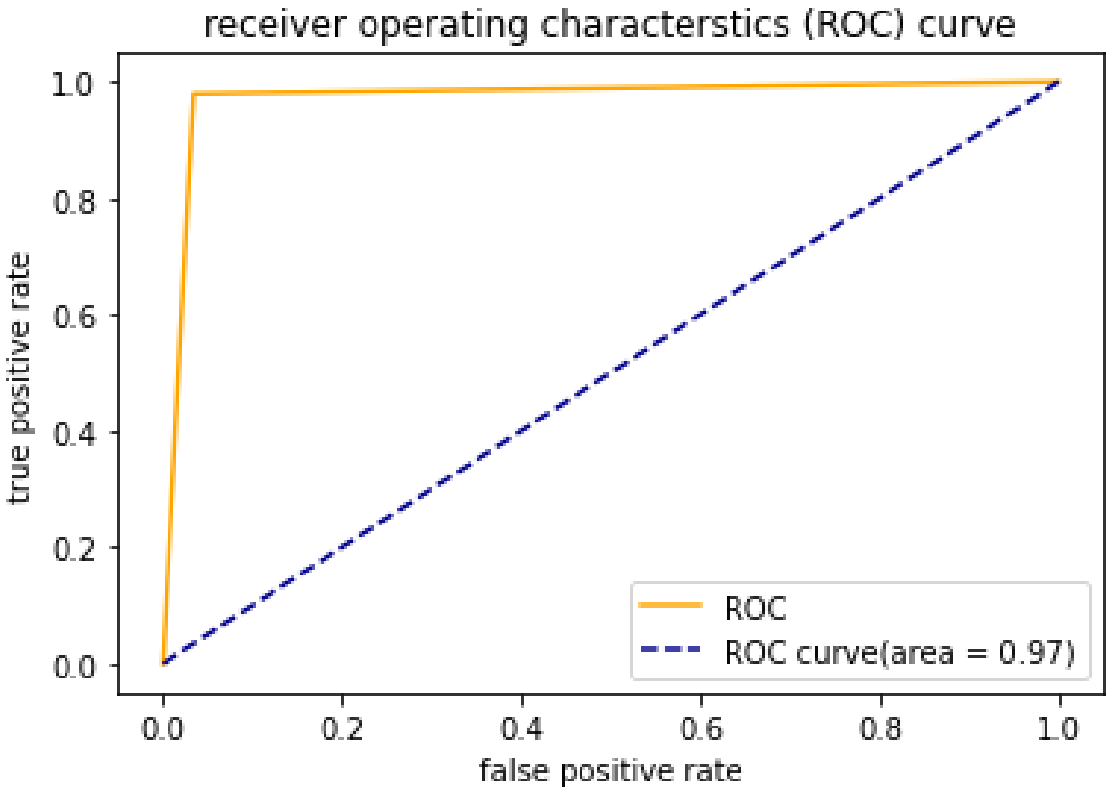

Figure 25: ROC curve for Logistic Regression

(Source by Author)

The AUC score for the Logistic Regression model is 0.974, and the ROC curve for this model is shown in the above figure. The accuracy and recall are 0.97, where precision is 0.98.

Random Forest machine learning algorithm:

The predictive model for ransomware attack detection using Random Forest machine learning algorithm has been implemented as shown in the below figure, in which it has a creation of Random Forest model and the training data, test data is classified using model predict function then predicted information is stored into a y_predicted. The classification report is mentioned in the below list, and the confusion Matrix is shown in the below figures.

	Precision	Recall	F1-Score	Support
0	**1.00**	**0.96**	**0.98**	**233**
1	**0.94**	**0.99**	**0.97**	**148**
Accuracy			**0.97**	**381**
macro avg	**0.97**	**0.98**	**0.97**	**381**
weighted avg	**0.97**	**0.97**	**0.97**	**381**

Table 4: Performance metrics of Random Forest (Source by Author)

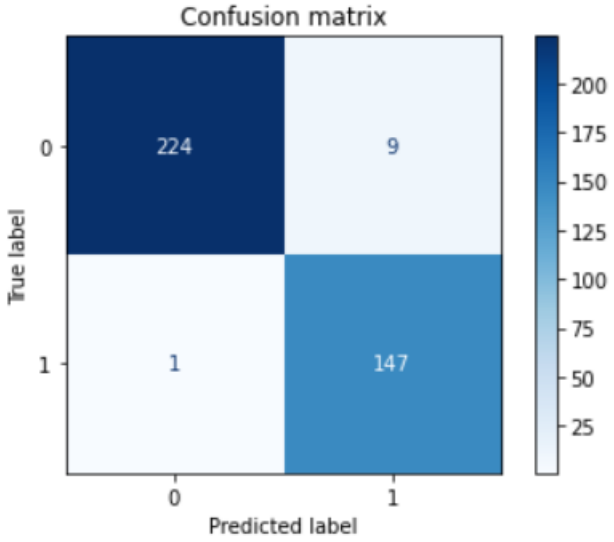

Figure 26: Confusion Matrix for Random Forest (Source by Author)

Random forest algorithm is also performed with the accuracy from 0.97% and 5 good ware files were classified as negative and 9 ransomware files fell under good ware files after classification.

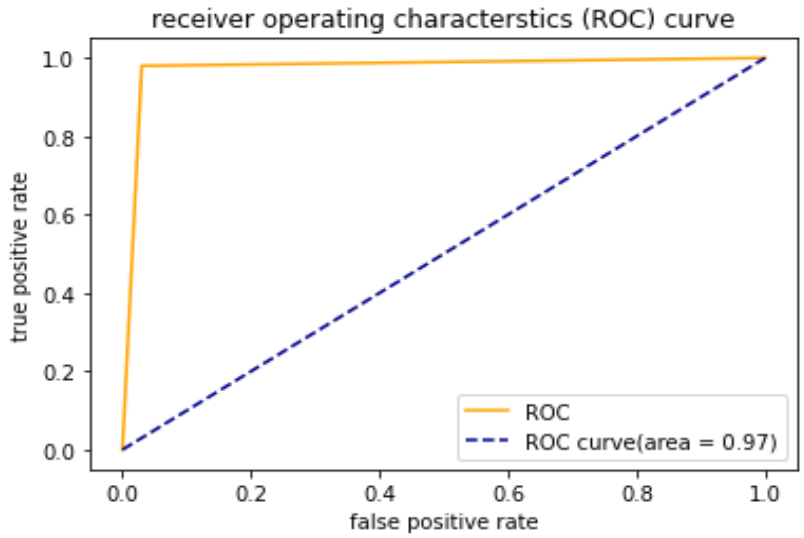

Figure 27: ROC curve for Random Forest **(Source by Author)**

The AUC score for the Random Forest model is 0.974, and the ROC curve for this model is s hown in the above figure. The accuracy and recall are 0.97, where precision is 0.99. From the above four models, we can conclude that all four models we have created work better for ransomware attack detection but Logistic Regression is doing better than the other three models with accuracy of 0.98%.

4.5 Classification based on the type of ransomware

The 'type' column is utilised for the target label in this classification and is compared for all four methods, as shown below. In this study, 11 ransomware families were involved in the type of ransomware attack. The 'type' column is allocated to the 'y1' variable, whereas the 'X' variable stores all other columns relating to input variables. The 'X' and 'y1' variables hold the input characteristics and output targets. The variable costs are then applied to the train and test split, with the test data set size set to 0.25, with 75 percent of the data set assigned to the training variable and the remaining 25 percent assigned to the test variable. The train test split is depicted in the image below.

```
from sklearn.model_selection import train_test_split
x1_train, x1_test, y1_train, y1_test = train_test_split(X,y1,test_size = 0.25,random_state=100)
```

Figure 28: Type Classification train and split data

Figure 28: Type classification train and split data

4.6 Machine algorithm application on type classification

Four machine learning methods were employed after partitioning the data into training and tes ting data sets, and their findings are detailed in the section below. Two performance measure ments were created for each method. The first is a confusion matrix, while the second is a clas sification report. Confusion matrix values are created for 11 different ransomware families in this form of categorization.

XGBoost model for type classification and its results are shown below;

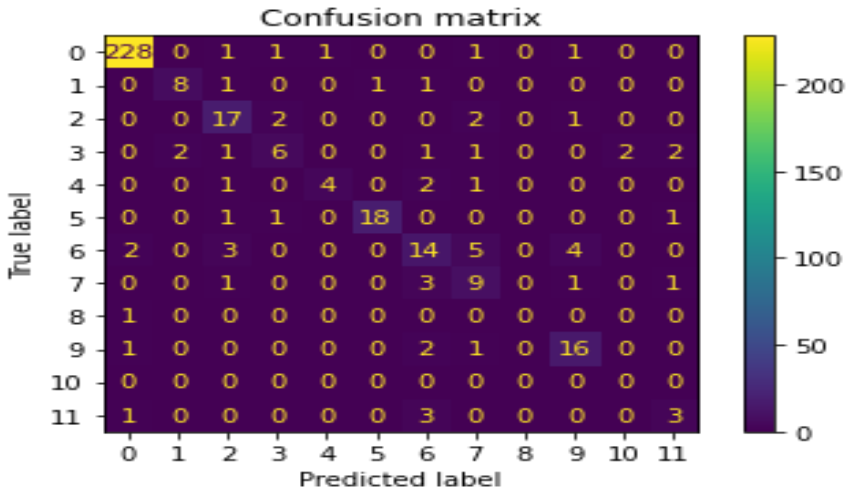

Figure 29: Confusion Matrix for XGBoost type classification

(Source by Author)

From the above matrix, all the diagonal values are True positive values of the class of ransomware family. Except for these values all the other values are minimal. From the classification report of the XGBoost algorithm, the accuracy obtained was 0.85%. The below picture depicts the performance metrics of the XGBoost algorithm.

accuracy			0.85	381
macro avg	0.57	0.55	0.55	381
ghted avg	0.85	0.85	0.85	381

Figure 30: Classification report of XGBoost type classification (Source by Author)

The second machine learning algorithm to detect ransomware attacks is the **Support Vector model**, is also one of the supervised learning models which are used for classification and regression and to deal with multi-dimensional space. Classification report and the confusion matrix for the SVM classifier is as follows:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.98	0.98	0.98	233
1	0.82	0.82	0.82	11
2	0.44	0.50	0.47	22
3	0.50	0.40	0.44	15
4	1.00	0.62	0.77	8

5	0.81	0.81	0.81	21
6	0.57	0.43	0.49	28
7	0.39	0.60	0.47	15
8	0.00	0.00	0.00	1
9	0.75	0.75	0.75	20
10	0.00	0.00	0.00	0
11	0.38	0.43	0.40	7
Accuracy			0.83	381
macro avg	0.55	0.53	0.53	381
weighted avg	0.84	0.83	0.83	381

Table 5: Performance metrics of SVM (Source by Author)

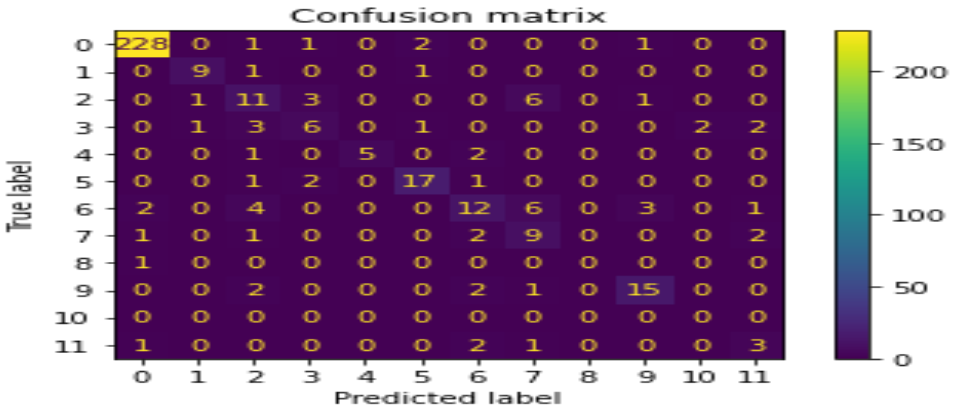

Figure 31: Confusion Matrix for SVM for type classification

(Source by Author)

The Logistic Regression Model is the third machine learning approach for detecting ransomware attacks. This model is also known as the logit model. Binary logistic regression and multinominal logistic regression are the two types of logistic regression models. We have 11 types of ransomware families in the dataset, and they are classified into 11 classes.

The Logistic Regression model's performance metrics are as follows:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.97	0.98	0.98	233
1	1.00	0.73	0.84	11
2	0.44	0.55	0.49	22
3	0.55	0.40	0.49	15

4	1.00	0.25	0.40	8
5	0.86	0.90	0.88	21
6	0.57	0.46	0.51	28
7	0.36	0.60	0.45	15
8	0.00	0.00	0.00	1
9	0.82	0.70	0.76	20
10	0.00	0.00	0.00	0
11	0.38	0.43	0.40	7
Accuracy			0.83	381
macro avg	0.58	0.50	0.51	381
weighted avg	0.84	0.83	0.83	381

Table 6: Performance metrics of Logistic Regression

(Source by Author)

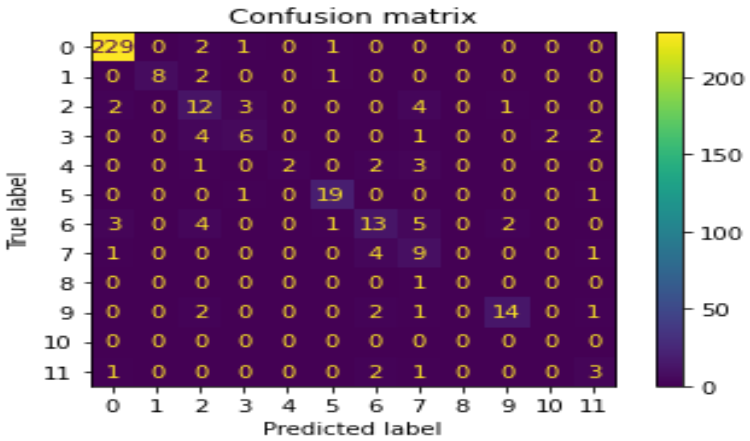

Figure 32: Confusion Matrix for Logistic Regression

(Source by Author)

The Random Forest model is also one of the supervised machine learning algorithms.

The workflow of the Random Forest algorithm goes from multiple decision trees to a single o utput. and its performance metric results are shown below.

Random Forest Classification Report : precision

0	0.92	1.00	0.96	233
1	1.00	0.73	0.84	11
2	0.53	0.41	0.46	22
3	0.58	0.47	0.52	15
4	1.00	0.62	0.77	8
5	0.83	0.71	0.77	21
6	0.64	0.50	0.56	28
7	0.47	0.60	0.53	15
8	0.00	0.00	0.00	1
9	0.83	0.75	0.79	20
10	0.00	0.00	0.00	0
11	0.38	0.43	0.40	7
accuracy			0.83	381
macro avg	0.60	0.52	0.55	381
weighted avg	0.83	0.83	0.83	381

Figure 33: Classification report of Random Forest for type classification

(Source by Author)

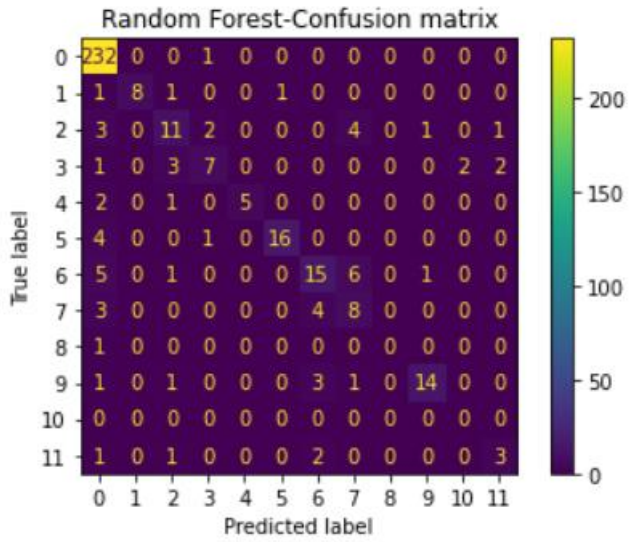

Figure 34: Confusion Matrix for Random Forest for type classification

(Source by Author)

In the above results, we can conclude that all four algorithms have a similar performance, where XGBoost has 85% accuracy for type classification, and other the three algorithms have 83% of accuracy.

4.7 Results Evaluation

To detect ransomware files from the good ware files using Machine learning classification algorithms, we have deployed four algorithms. To explore more predictions and results, two classifications were implemented in the above implementation. The first classification is based on the “attack” attribute and the later classification is on the type of ransomware.

In the attack-based classification, XGBoost obtained the accuracy of 0.97% and the values obtained from confusion matrix are [[227,06], [4,144]] whereas Support Vector Model produced the accuracy of 0.97% and confusion matrix values are [[226,7], [3,145]]. Logistic regression algorithm produced 0.98% of accuracy with confusion matrix values [[226,7], [2,146]] and last model Random Forest algorithm has given 0.97 percent of accuracy and [[224,9] [1,147]] as confusion matrix cell values. In comparison, the Logistic Regression algorithm is the best choice to detect ransomware files and all other algorithms are producing a pretty good percentage of accuracy. False positive and False negative values are less, they are in the range of 1 to 7 in between. It is one of the good signs that selected models are working to the best to find malicious files.

The "type" characteristic was utilized as the legal column for classification in the "type" classification, and the type is the family name of the Ransomware attack. The XGBoost algorithm has delivered 0.85% accuracy in this classification, and the False Negative and False Positive values are likewise smaller and in the range of 0 to 2. The Support Vector Model achieved 0.83% accuracy, while the Logistic Regression and Random Forest methods achieved 0.83% accuracy. All the models had false positive and false negative values between o and 2, indicating that the model is predicting well. If ransomware-type is a target variable, XGBoost is one of the best alternatives for malware prediction.

ROC and AUC Curve:

The ROC is the probability curve that connects the False positive rate with the True positive rate. The AUC is a measure of separability, the higher the AUC, the better the classifier's performance. The AUC score can range from 0 to 1 (Parte, 2020). Below picture depicts the evaluation of these two curves.

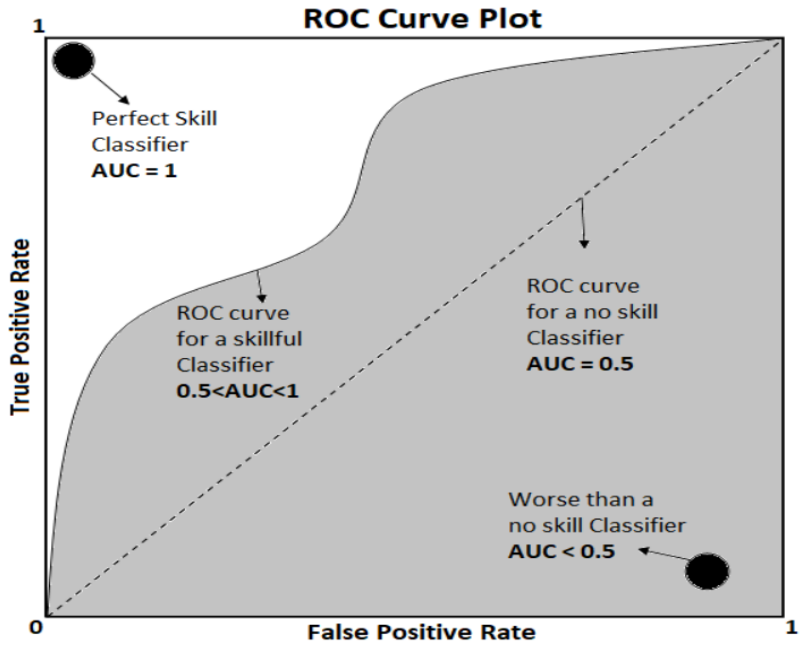

Figure 35: ROC and AUC curve

Source :(Parte, 2020)

The AUC curve for all the methods is listed below:

XGBoost algorithm = 0.973

Logistic Regression = 0.974

Support Vector model = 0.974

Random Forest algorithm = 0.977

All the above values are determining that, algorithms were performed well to predict the malicious attacks.

CHAPTER5: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Technology continues to advance in the contemporary world, increasing the level of sophistication of hacking strategies as well as the number of tools available to hackers. As a result of this, the aggressor will feel encouraged to engage in acts of moral and financial blackmail. As a result, it is critical to take prompt action against such assaults, including ransomware, and it is also crucial to employ technology to avoid ransomware by taking safeguards. To defend ourselves against such attacks, we shall employ defences based on the notion of machine learning. This study aimed to examine a machine learning approach for categorizing ransomware using random forest and characteristics collected from the raw bytes of the file. To build the most effective XGBoost classifier that can reliably identify ransomware, several different sizes of seeds and trees have been put through experimental testing.

It is feasible to utilize the ransomware dataset to forecast attacks. Machine learning is cut-edge technology that has several algorithms and four algorithms used in the recommended work. It is a static analysis for byte-level properties to identify ransomware. In addition, we will compare each algorithm to other machine learning approaches such as logistic regression, support vector machine, and Radom Forest, XGBoost. All the machine learning algorithms presented thus far will be evaluated using performance measures such as accuracy, precision, recall, and confusion matrix, which were compared in our proposition. All four algorithms have given their own best performance to detect bad files, however Logistic regression and XGBoost were given the best results in comparison with other algorithms. The best model can be selected based on the victim's platform and requirements and sample data.

5.2 Limitations

Though every research has its good outcomings, like every side of the coin, research has its own limitations. To begin, dataset accuracy ranks top on the list of limitations. Because the data is sensitive, organizations or victims are hesitant to release it after the assault or before it occurs. Second, a lack of knowledge of complicated material, particularly when dealing with ransomware files, is an issue.

Nowadays, hackers are becoming more mindful and intelligent to match scientists. Though many researchers are working to detect ransomware attacks, hackers coming up with counteractions in more depth ways to beat these new methods. There is no doubt that Machine Learning algorithms are more powerful to predict or classifying the data, but human errors can happen with one single digit change, and this impacts the train and test data. If the training and testing data has not been split correctly, there may be a chance of wrong predictions.

Another hurdle to forecasting the attack is data labelling; without suitable data naming conventions, it is difficult to distinguish between good and malicious files. Hackers with sophisticated expertise can sometimes use machine learning models to train their sample data for entry into the system. Even though many academics are working on detection methods, hackers continue to surprise with innovative defences.

5.3 Future Directions and Recommendations

These days, cybercriminals use cunning methods to develop new varieties of malware that are more lucrative. Ransomware is one of these attacks that has been spreading rapidly lately. In contrast to other security issues, ransomware cannot be removed and is very tough to eliminate. (Chittooparambil et al. 2019). In the future, we will focus on Ransomware detection in real-world data sets utilising both deep learning and a mix of machine learning methods. Data classification with a large variety of data possibilities such as network data characteristics and Opcode will be discussed. By testing and implementing in combinations, more algorithm combinations will be adhered to. The research focused on a Machine learning algorithm to detect this kind of attack. Though models have performed well, in the subject of ransomware detection, there is still potential for progress.

Rich Dataset: There is no real-world dataset that contains every imaginable Ransom attack pattern. If a dataset like the one stated exists, it will be useful to many researchers looking into this area.

Pre-encryption detection methods: While most detection methods operate with data after an attack has occurred, detection techniques that act at the right moment of attack or before the assault are always appreciated to prevent Ransomware attacks.

Techniques to work on distributed systems*:* One of the most dangerous habits of ransomware is spreading through all the systems in a particular group. Techniques to work in the distributed systems help to reduce the maximum percentage of loss to the victims (Urooj et al., 2021).

Working on fuzzy algorithms to tackle Ransomware is also an excellent place for researchers to start if they want to identify zero ransomware detections. Attacks in cyber security and various strategies to manage such attacks are a never-ending war; while researchers develop best practices to identify the assault, hackers develop a new sort of attack in a new way. It is our responsibility to protect our own data using new technologies.

REFERENCES

- Almashhadani, A.O., Kaiiali, M., Sezer, S. and O'Kane, P. (2019). A Multi-Classifier Network-Based Crypto Ransomware Detection System: A Case Study of Locky Ransomware. *IEEE Access*, 7, pp.47053–47067. doi:10.1109/access.2019.2907485.

- Anderson, B.; Quist, D.; Neil, J.; Storlie, C.; Lane, T. 2011, "Graph-based malware detection using dynamic analysis," *J. Comput. Virol. 2011, 7, 247–258.*

- Chittooparambil H.J. et al. 2018, "A review of ransomware families and detection methods." *International Conference of Reliable Information and Communication Technology, pp. 588-597.*

- De Groot, J. A 2017, "History of Ransomware Attack: The Biggest and Worst Ransomware Attack of All Time." *Available online:Https://digitalguardian.com/blog/history-ransomware-attacks-biggest-andworst-ransomware-attacks-all-time*.

- Dudovskiy, J. (2015). *Research Approach - Research-Methodology*. [online] Research-Methodology. Available at: https://research-methodology.net/research-methodology/research-approach/.

- Dudovskiy, J. (2015). *Research Approach - Research-Methodology*. [online] Research-Methodology. Available at: https://research-methodology.net/research-methodology/research-approach/.

- Erik Larsen, David Noever, Korey MacVittie 2021, "A Survey of Machine Learning Algorithms for Detecting Ransomware Encryption Activity," *In Cryptography and Security (cs.CR)*.

- E. Kolodenker, W. Koch, G. Stringhini and M. Egele 2017, "Pay-break: Defense against cryptographic ransomware," *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 599-611*.

- Erik Larsen, David Noever, Korey MacVittie 2021, "A Survey of Machine Learning Algorithms for Detecting Ransomware Encryption Activity," *In Cryptography and Security (cs.CR)*.

- F. Khan, C. Ncube, L. K. Ramasamy, S. Kadry and Y. Nam 2020, "A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning," *IEEE Access, vol. 8, pp. 119710-119719, 2020. IEEE.*

- G. Usha; P. Madhavan; Meenalosini Vimal Cruz; N A S Vinoth; Veena; Maria Nancy 2022, "Enhanced Ransomware Detection Techniques using Machine Learning Algorithms," *4th International Conference on Computing and Communications Technologies (ICCCT). IEEE. .*

- GitHub. (2022). *rissgrouphub/ransomwaredataset2016*. [online] Available at: https://github.com/rissgrouphub/ransomwaredataset2016 [Accessed 7 Aug. 2022].

- Grant, L. Parkinson, S 2018, "Identifying file interaction patterns in ransomware behaviour," *In: Guide to Vulnerability Analysis for Computer Networks and Systems, pp. 317–335. Springer.*

- Kaspersky (2019). *What is WannaCry ransomware?* [online] Kaspersky.co.uk. Available at: https://www.kaspersky.co.uk/resource-center/threats/ransomware-wannacry.

- Kharaz A. et al. 2016, "A large-scale, automated approach to detecting ransomware," *25th Security Symposium (Security 16), pp. 757-772*

.

- Maniath, S., Ashok, A., Poornachandran, P., Sujadevi, V., Sankar, A. and Jan, S. (2017). Deep learning LSTM based ransomware detection. *2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE)*. [online] doi:10.1109/RDCAPE.2017.8358312.

- Moore, C. (2016). Detecting Ransomware with Honeypot Techniques. 2016 Cybersecurity and Cyberforensics Conference (CCC). doi:10.1109/ccc.2016.14.

- N. Milosevic, A.D. Choo, K.K.R.: 2017, "Machine learning aided android malware classification," *Comput. Electr. Eng*.

- Parte, K. (2020). *Understanding Performance metrics for Machine Learning Algorithms*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/understanding-performance-metrics-for-machine-learning-algorithms-996dd7efde1e.

- Roy, K.C. and Chen, Q. (2020). DeepRan: Attention-based BiLSTM and CRF for Ransomware Early Detection and Classification. *Information Systems Frontiers*. doi:10.1007/s10796-020-10017-4.

- S. R, K. R and J. B 2021, "Implementation of Dynamic Scanner to Protect the Documents from Ransomware using Machine Learning Algorithms," *2021 International Conference on Computing, Electronics & Communications Engineering (access), 2021, pp. 65-70.IEEE.*

- Seong Bae, Gyu Bin Lee, Eul Gyu I'm. 2019, "June. Ransomware detection using machine learning algorithms." *In Concurrency and Computation Practice and Experience 32(3.20):e5422.*

- Sgandurra, D., Muñoz-González, L., Mohsen, R. and Lupu, E. (2016). *Automated Dynamic Analysis of Ransomware: Benefits, Limitations and use for Detection*. [online] Available at: https://arxiv.org/pdf/1609.03020.pdf.

- Sharma, A. Sahay, S.K. 2016, "An effective approach for classification of advanced malware with high accuracy," *arXiv preprint arXiv:1606.06897.*

- Shijo, P.V. Salim, A. 2015, "Integrated static and dynamic analysis for malware detection," *Procedia Comput. Sci. vol. 46, pp. 804–811*.
- Streefkerk, R. (2019). *Qualitative vs. Quantitative Research | Definitions, Differences & Methods*. [online] Scribbr. Available at: https://www.scribbr.com/methodology/qualitative-quantitative-research/.

- THE ANATOMY OF A RANSOMWARE ATTACK. (n.d.). [online] Available at: https://www.jpmorgan.com/content/dam/jpm/commercial-banking/documents/cybersecurity-fraud/2020SpringCyberMag_v5_RansomWare_ADA.pdf.

- Urooj, U., Al-rimy, B.A.S., Zainal, A., Ghaleb, F.A. and Rassam, M.A. (2021). Ransomware Detection Using the Dynamic Analysis and Machine Learning: A Survey and Research Directions. *Applied Sciences*, 12(1), p.172. doi:10.3390/app12010172.

- V. G. Ganta, G. V. Harish, V. P. Kumar and G. R. K. Rao 2020, "Ransomware Detection in Executable Files Using Machine Learning," *International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2020, pp. 282-286. IEEE.*

- V. G. Ganta, G. V. Harish, V. P. Kumar and G. R. K. Rao 2020, "Ransomware Detection in Executable Files Using Machine Learning," *International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2020, pp. 282-286. IEEE.*
- Weckstén M. et al. 2016, "A novel method for recovery from crypto-ransomware infections," *2nd IEEE International Conference on Computer and Communications (ICCC), pp. 1354-1358.*
- www.javatpoint.com. (2011). *Tutorials - Javatpoint*. [online] Available at: https://www.javatpoint.com/.
- Zakaria, W.Z.A.; Mohd, M.F.A.O.; Ariffin, A.F.M. 2017, "The Rise of Ransomware," *In Proceedings of the 2017 International Conference on Software and e-Business, ICSEB 2017, Hong Kong, 28–30 December 2017; pp. 66–70.*

APPENDIX

APPENDIX A :

A Jupyter notebook file was submitted with the name “RansomwareDetection.ipynb”.

APPENDIX B:

A CSV file (dataset) files was submitted with the name” RansomwareData”.