PHISHING ATTACK DETECTION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

RIYA VINCENT (TKM20MCA-2031)

to

The APJ Abdul Kalam Technological University

In partial fulfillment of the requirments for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

Thangal Kunju Musaliar College of Engineering
Kerala

DEPARTMENT OF COMPUTER APPLICATIONS

JULY 2022

DECLARATION

I undersigned hereby declare that the project report on P H I S H I N G A T T A C K D E T E C T I O N U S I N G M A C H I N E L E A R N I N G submitted for partial fulfillment of the requirements for the award of degree of Masterof Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Prof. NATHEERA BEEVI M. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources.I also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated anydata or idea or fact or source in our submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permissionhas not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Kollam

Date: 12-07-22

RIYA VINCENT

DEPARTMENT OF COMPUTER APPLICATIONS

TKM COLLEGE OF ENGINEERING

C E R T I F I C A T E

This is to certify that, the project report entitled **PHISHING ATTACK DETECTION USING MACHINE LEARNING** submitted by is **RIYA VINCENT(TKM20MCA-2031)** to the APJ Abdul Kalam Technological University in partial fulfillment of the M.C.A degree in Master of Computer Application is a bonafide record of the project work carried out by her under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor Head of the Department External Examiner

ACKNOWLEDGEMENT

First and foremost I thank GOD almighty and my parents for the success of this project. I owe sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.

I am extremely grateful to **Dr. FOUSIA M SHAMSUDEEN**, Head of the Department, for providing mewith best facilities.

I would like to thank my project guide, **Prof. NATHEERA BEEVI M** Department of Computer Applications, who motivated me throughout the work of my project.

I profusely thank all other faculty members in the department and all other members of TKM College of Engineering, for their guidance and inspirations throughout my course of study.

I owe my thanks to my friends and all others who have directly or indirectly helped me in the successful completion of this project.

Riya Vincent

ABSTRACT

PHISHING ATTACK DETECTION USING MACHINE LEARNING

About 30 000 online applications are hacked every day, and most of the time website owners or web designers aren't even aware of what's going on with their own websites. Phishing sites are used by web hackers to deceive users and collect their private and sensitive information. Different dangers can be used by online hackers to gain access to or compromise lawful web applications.

As a consequence, it's important to take the required safeguards to be aware of the threats and weaknesses that might affect the website and, as a result, the normal flow of business. The research also takes into account the creation of web application logs, which makes it easier to analyze the behavior of atypical users and identify instances where their actions are prohibited, inappropriate, or otherwise improper. The most prevalent web application dangerous threats are mitigated, and the web administrator is given detailed instructions on how to spot phishing links, a type of social engineering attack.

While a number of machine learning algorithms and deep learning techniques are employed for phishing link identification, secure coding approaches are used for mitigation. The testing process's outcomes demonstrated that the website had effectively countered these risky web application assaults. The application's component that recognises phishing URLs compares several algorithms to see which one performs best. The best model results in an accuracy of 85%.

Contents

List of figures

.

CHAPTER 1
INTRODUCTION

Attacks including phishing are increasing. As a result of the COVID pandemic, our methods of living, learning, and working have significantly changed, which has raised numerous new issues with cyber security. The 2020 Phishing Attack Landscape Report shows a dramatic rise in the number of phishing attack attempts. Secure coding approaches are used for mitigation, while a variety of machine learning algorithms and deep learning techniques are used for phishing link detection.

As the Internet has developed, people are choosing online commerce searching over traditional shopping. With a few special techniques, criminals now use this platform to look for their victims in the online computer community. Attackers demonstrate innovative techniques, such as phishing, to deceive victims by using bogus websites to obtain sensitive information, such as online account IDs, usernames, passwords, etc. The most common form of social engineering that targets the vulnerability found in contemporary methods as a result of internet users is phishing.

The phisher preys on online users by deceiving them into disclosing personal data with the purpose of using it fraudulently. Usually, multiple tactics are used to counteract distinct attacks.Data from the dataset will be gathered, and we'll classify it later. Afterward, dividing the data into test and training sets These sets will be split in an 80:20 ratio. The test set will be used to validate our model, and the train set will be used to train it. Applying the algorithms, or putting the ANN and XGBoost techniques into practice, is building the model. using the test set to train the model, which will train both the ANN and the XGBoost models.

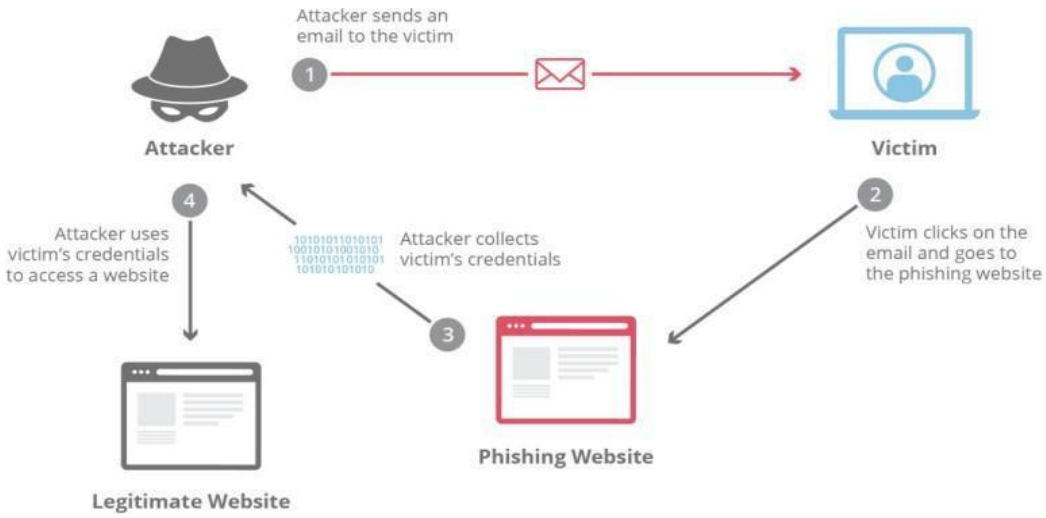

Fig 1 :Deceptive Phishing Attack Diagram

1.1 Problem Definition

Everyone uses the internet, which is expanding quickly since it can meet all of their demands while also offering a wide choice of services. The demand for ensuring secure and vital data exchanges rises as internet usage rapidly expands. This has made it vulnerable to a variety of security assaults, so it's critical to use web application security approaches while creating online web apps.Any modern online web application that isn't properly secured is open to hacker attacks. Hackers from all around the world can attack your website and steal confidential information or seriously harm it. When a user clicks on a link that contains malicious JavaScript code, many potential hacking scenarios could take place, including the theft of personal information or the hijacking of a web session. Detecting phishing links manually by examining every page of a website is ineffective, time-consuming, and necessitates a security system. Hackers will also frequently put phishing links into a website.

1.2 Objective

A well-known type of social engineering called "phishing" uses websites to impersonate legitimate web pages and uniform Resource locators (URLs). This job is specifically designed to train deep neural networks and device learning models to predict phishing websites using data collected from unique assertions. Phishing websites and URLs are compiled into a dataset, and functions based on URL and website content are needed from them.Artificial Neural Networks and XGBoost, two machine learning approaches, have been employed to solve this issue so that our computer can make decisions in real-time and identify websites as phishing sites or not in advance based on previously gathered data. The results of the testing process showed that the website had successfully mitigated these dangerous web application attacks. For the part of the application that detects phishing links, a comparison is made between various algorithms to find the best one, and the result of the best model gives 85 percent accuracy.

CHAPTER 2
LITERATURE SURVEY

In this section, several studies of Traffic sign and light detection techniques utilizing deep learning are discussed

2.1 Phishing Happens Beyond Technology:

Each step of the phishing process is impacted by human behaviors and demographics. Prior research has demonstrated that Internet users' actions and attitudes affect their chance of falling victim to phishing scams. In order to acquire the potential victim's trust and persuade them to execute the appropriate activities, many con artists create a step-by-step phishing strategy. Understanding which actions and attitudes can influence whether a victim follows the attacker through each stage of a phishing scam is crucial. This will make it possible for us to pinpoint the fundamental reasons of phishing, create tailored mitigation strategies for each stage of the technique, and boost preventive points. This study looks at how people's risk-taking and decision-making tendencies affect the likelihood of being a victim of phishing in three distinct steps.Here, the author performed a simulated phishing campaign to gauge participants' susceptibility to the three steps of phishing after asking them to play a risk-taking game and respond to questions about two psychological measures to examine their behaviors. We discover that the user's susceptibility to phishing in the various steps chosen can be predicted by gender and risk-taking attitude. However, there are more direct and indirect behavioral aspects that might be examined in future research. Starting from the underlying reasons of phishing attempts, a comprehensive framework to prevent their success can be built using the findings of this study and the model that was created.

2.2 Link Vulnerability leads to phishing attacks, Spear-Phishing electronic/UAV communication-scam targeted

This study examined current spear phishing attack methodologies and their defense mechanisms. The results astoundingly reveal that even just personality-targeted text can change the response to phishing assaults, and the article also outlines the processes to set up and manage successful phishing campaigns. Phishing Attack Emergence Threat and Machine Learning Models-Based Detection Methods - In this study, we employed a variety of classifiers to detect phishing urls, and because they had a higher degree of accuracy, we concentrated on timing when training the dataset. We discovered that the XGBoost classifier used less time and provided a higher accuracy of 94.44 percent.

2.3 A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites

Since a day has passed, identifying and discovering some phishing websites in real-time requires a number of traits and standards. Fuzzy logic algorithms may be crucial in identifying and testing phishing websites because of the uncertainties in the detection process. Compared to exact rules, fuzzy logic provides a more rational approach to addressing quality differences. A technique for fuzziness resolution as well as a robust and open phishing website detection model will be shown via the phishing website assessment. This tactic is based on machine learning algorithms that use logically sound definitions for a number of items on the phishing website. A total of 30 traits, features, and qualities of phishing websites may be employed for extremely precise phishing detection. A real-time phishing dataset is obtained from the UCI machine learning repository.

2.4 Systematic Literature Review on Phishing and Anti-Phishing Techniques

Phishing is the largest risk in the digital age. Phishing attacks have been around for a long, but as time goes on, they become worse as attackers find new and imaginative ways to get through defences. This paper discusses several phishing and anti-phishing techniques. To do this, the proposed study subjects are critically identified using the systematic literature review (SLR) approach. Initially, 80 pieces were taken from different sources. After then, these articles were filtered using the Tollgate Approach to reveal different phishing and anti-phishing techniques. According to a study analysis, spear phishing, email spoofing, email manipulation, and phone phishing are the most often used phishing techniques. In contrast, the SLR asserts that machine learning systems have the highest accuracy of all anti-phishing approaches for deterring and recognising phishing attempts.

2.5 Phishing Detection

The survey of phishing detection is part of this paper. This article reviews the research on phishing attack detection. Phishing attacks target holes in systems that exist because of the involvement of humans. Users are the weakest link in the security chain since many cyberattacks are spread via methods that take advantage of flaws in end users. Since there is no single, effective way to address all of the weaknesses in phishing, numerous strategies are frequently used to counteract particular attacks. Many of the recently proposed phishing mitigation strategies are surveyed in this study. In order to show how phishing detection strategies fit into the broader mitigation process, a high-level

overview of the many types of phishing mitigation approaches is also offered, including: detection, offensive defense, rectification, and prevention

2.6 Review of Detection Methodologies for Quick Response code Phishing Attacks

Attack vectors for phishing have been given a new dimension by the addition of quick response code. When an attacker lures a victim into giving over personal information like a password, personal identification number, username, or other credentials like online banking details using a quick response code, this is known as a rapid response code phishing attack. This attack is on the rise as more individuals embrace using mobile phones for easy purchases as well as communication. Because quick response codes are easy to make and use, both customers and companies that provide products and services have come to embrace them. This attack is semantic in nature since consumers have little chance of knowing what is hidden in the rapid response code before usage. This study looked at a variety of methods that earlier researchers have used to identify this semantic-based phishing effort. the benefits and drawbacks of every approach, along with any general research gaps.

2.7 The Emergence Threat of Phishing Attack and The Detection Techniques using Machine Learning Models

Cybercrime is growing rapidly along with the amount of people using the internet. Phishing has now been identified as the most effective and successful cyberattack method. Phishing attacks are the most prevalent and are utilized in a variety of methods to assault the intended user, according to our examination into them. Phishing attacks using phishing URLs, phishing emails, and phishing websites are fairly common.

However, attackers are now focusing their phishing attacks on social media and online gaming because of their growing popularity.

Both the users who want to avoid being attacked and the attackers themselves are entering a new era thanks to advances in machine learning. In the course of our work, we construct a machine-learning-based Twitter spear phishing bot. Phishing URLs, phishing emails, and phishing websites were all the subjects of experiments that we carried out. In order to detect phishing URLs, we employed a variety of classifiers, and in order to get a greater level of accuracy, we focused on the amount of time it took to train the dataset. We discovered that the XGBoost classifier has a higher accuracy rating of 94.44 percent and required significantly less time. We found that the naive bayes classifier was the most effective method for detecting phishing emails, with an accuracy of 95.15 percent. Among the several classifiers that we employed in our website identification procedures, we discovered that the Random Forest Classifier provides the highest accuracy, which was measured at 96.80 percent.

CHAPTER 3
METHODOLOGY

3.1 Proposed System

Any web application that is built for the internet today but does not have proper security measures in place is open to the possibility of being attacked by hackers. Hackers from all over the world have the capability of breaking into your website, which might result in the theft of confidential information or serious damage. When a user visits a link that contains malicious JavaScript code, it has the ability to steal personal information, hijack a web session, and trigger a wide variety of other forms of hacking problems. Phishing links are another common tactic utilised by hackers when attempting to compromise a website. In addition to being ineffective, manually identifying phishing links by inspecting each page of a website takes a significant amount of time and requires the skills of a security professional.

Because of this problem, we have implemented machine learning strategies such as Artificial Neural Networks and XGBoost so that our computer can make decisions in real time and can classify in advance whether a particular site is a phishing site or not based on data that has been gathered in the past. This allows our computer to make decisions in real time and can prevent us from falling victim to fraudulent websites. Because of this, the issue could finally be resolved.

3.2 System Architecture

The major purpose of this research is to identify and eliminate phishing frauds, which is why it was conducted in the first place. For the purposes of training, this system is capable of handling both phishing websites and websites that are not considered to be phishing. As a result of the training that the model underwent, the system is now able to recognise phishing websites. In addition to this, it is able to access the chrome and make an accurate prediction of the outcome. Figure 2 presents a diagram of the system's organisational structure.

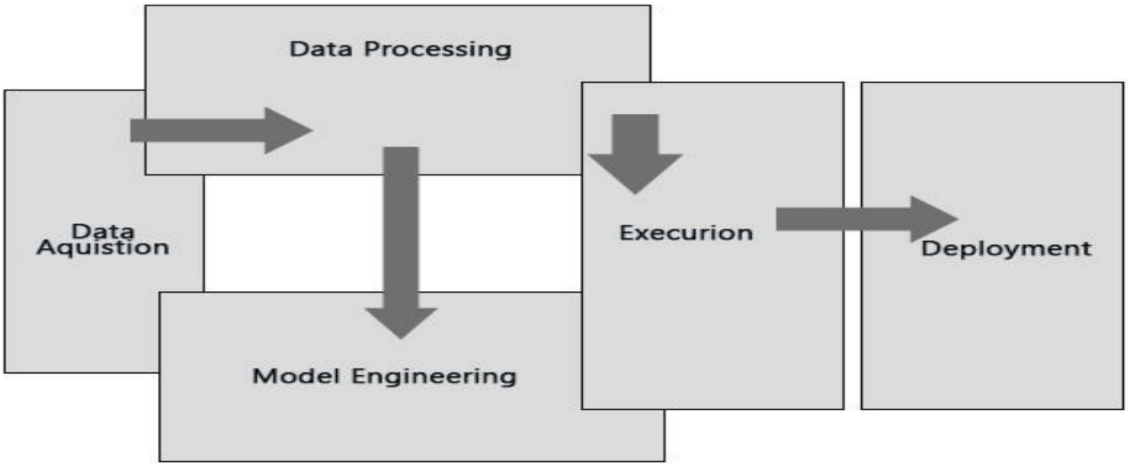

Fig :2 Block diagram

Data Acquisition : The availability of data is an important factor to consider while developing a knowledge-gathering and decision-making system for a mission device. To accomplish this, records must first be gathered, then the situation must be prepared and separated from the case in accordance with the positive characteristics involved in the decision-making process, and finally, the records must be sent to the processing unit so that they may be further categorized.

This stage of the procedure is sometimes referred to as the records preparation stage. The records version makes predictions that are accurate, prompt, and elastic, and those records can either be discrete or long-lasting. The information is then sent to a device that processes data continuously, and it is either permanently stored in a batch records warehouse or temporarily stored in a discrete records warehouse, depending on the kind of record, before it is sent on to the modeling or data processing phases.

Data Processing : After the data has been gathered inside the data acquisition layer, it is transferred to the data processing layer, where it is subjected to a more advanced level of integration and processing. This includes,

- Feature Extraction
- Normalization

- Training and Test Sets Preparation

Data Modeling : This level of architecture involves the selection of several algorithms that can adjust the machine to satisfy the challenge that the instruction is being processed, and these algorithms either change or are inherited by means of a set of libraries. In order to version the statistics in an appropriate manner, algorithms are utilized, and this prepares the machine for execution.

Execution : Experimentation, testing, and the development of improved versions of the device are all components of this stage of the device's research process. The overarching goal of improving the set of rules is to extract the given output from the device and maximize machine performance. Step output is a sophisticated solution that can provide the data that are required to make judgments.

Deployment : The results of machine learning should either be put to use or sent on to be processed further for additional research. The final outcome might be seen as a non-deterministic session, which is then supposed to be used inside the framework of the decision-making system. It is advised that the results of system learning be transferred immediately and seamlessly to production. This would permit the system to make decisions immediately primarily based on the results, which will reduce the need for further exploratory processes.

3.3 Dataset

A large number of invalid and authentic website URLs were collected from a variety of data sources. These website URLs cover a wide range of categories, including banking websites, online shopping websites, reservation websites, and others. The collection includes a significant number of URLs for real-world websites all across the world. The Aalto University's public databases are mined for the legitimate website URLs, which are then compiled into a list. This dataset is then used to generate random valid URLs, which are then used to educate machine learning models.

The collection of phishing URLs comes from an open-supply carrier that goes by the name PhishTank. This provider makes available an updated list of phishing URLs in a variety of formats, such as CSV

and JSON, and the list is refreshed every hour. The Phishing and Legitimate URLs Dataset that was obtained from Kaggle.com is another large and diverse collection.

3.4 Data Preparation and Preprocessing

It is imperative that we first preprocess the data before incorporating it into our model, since this step comes before the incorporation of our data into the model. This is owing to the fact that the quality of the data and the meaningful insights that can be gleaned from it have a direct influence on how well our model is able to learn from new experiences. In other words, the better the data, the better our model will be able to learn from new experiences. Because the substantial information that may be retrieved from data preparation has a direct influence on establishing the overall quality of the data, this phase in the process of machine learning is a vital one. As a consequence of this fact, this stage is an essential part of the procedure. "Data preparation" refers to the process of cleaning and arranging the raw data before it can be used with our model. This procedure ensures that the data can be used effectively. This particular stage of the process is referred to as "data preparation," which is the phrase that is used to describe it. This step of the process is referred to as "data preparation," which is a descriptive term for what it really entails. This phase, which is both the first and most crucial step in the process, is when the process of developing a model for machine learning, which is the process that was just explained, gets begun. This phase is also the most important stage in the process. In order for the model that is being utilised in the projects that involve machine learning to produce superior results, the data needs to be structured in the perfect approach. This is an absolutely required step in order to get everything done that has to be done. You cannot avoid taking it.

3.5 Data Cleaning

The information needs to be processed, and all of the relevant processes in training and screening must be finished. This might take a little bit of time. The URLs have requested that they be deleted, so if you

could just comply with that request. At this point, in addition to the URLs that have already been banned or whose validity has been exhausted, we delete any URLs that are unnecessary from the list. At this point, we will also delete any URLs whose attempts to block have already been successful. Because of this, the URLs of the websites whose access was being limited have been removed from the list where they had been previously included.

The information needs to be processed, and all of the relevant processes in training and screening must be finished. This might take a little bit of time. The URLs have requested that they be deleted, so if you could just comply with that request. At this point, in addition to the URLs that have already been banned or whose validity has been exhausted, we delete any URLs that are unnecessary from the list. At this point, we will also delete any URLs whose attempts to block have already been successful. Because of this, the URLs of the websites whose access was being limited have been removed from the list where they had been previously included.

3.6 Feature Extraction

Phishing websites are characterised by a variety of skills, which are outlined below. On the basis of these characteristics, several skills that have the potential to be highly useful in rating a website an 18 have been established. really useful for identifying potential phishing websites. These capabilities are referred to as HTML and JavaScript Features, Domain Features, and Address Bar Features respectively. There are three different categories of capabilities that may be derived from the URL by using our system.

- Address bar features
- Html and JavaScript features
- Domain features

3.6.1 Address Bar Features

There are a lot of different capabilities that may be retrieved, and some of them can be considered deal with bar base capabilities. The ones stated below are the ones that have been taken into consideration for this thesis out of them.

- IP Address in URL

These checks look for the existence of an IP address somewhere inside the URL. It's possible that URLs will have IP addresses rather than area calls. If an IP address is used in place of the area call inside the URL, it is highly likely that an individual is attempting to steal or get personal information through the usage of this URL. If the portion of the URL in question comprises an IP address, the value that is assigned to this option is 1, which denotes phishing; otherwise, it is 0. (legitimate).

- The "@" Character in a URL

In this instance, it examines the URL to determine whether or not it has an image of the symbol for the @ symbol. When the representation of an at symbol is contained within the URL of a website, it notifies the browser to disregard the entirety of the address. anterior to the image of a "@," but the real handle will almost always come after the image of a "@." If the URL contains an image of the @ symbol, then the price tag that goes along with this option is 1. (phishing). In any other scenario, there is no financial impact (legitimate).

- The length of the URL.

Specifies for how long the URL will continue to be accessible. Phishers frequently utilise long website URLs in order to obscure dubious components that are situated inside the address bar. This is done in order to steal sensitive information from victims. According to this thesis, a URL is regarded to be fraudulent when its length is larger than or equal to

fifty-four characters, whereas a legal URL has less than that number of characters. On the other hand, a URL is considered to be legitimate when its length is fewer than that number of characters. If the URL has a length of less than or equal to fifty-four characters, the cost associated with this option is set to zero; if the URL has a duration of more than fifty-four characters, the cost associated with this option is set to one (phishing) (legitimate).

- The breadth of URL

Identifies the extent to which the URL has been impacted. This function uses the forward slash character (/) as the primary factor in its computation in order to calculate the total number of subpages that are present inside the given URL. The value of the feature may be described as a number, and the URL has a completely decisive impact on what that number is.

- Redirection

If the slash character "//" appears anywhere in the URL other than immediately after the protocol, the price for this option is set at one (phishing), and if it does not, it is set at zero (legitimate).

- "http/https" in Domain name

This tests for the presence of "http/https" withinside the area a part of the URL. If the URL has "http/https" withinside the area part, the cost assigned to this selection is 1 (phishing) otherwise 0(legitimate)

- URL Shortening Services

If the website URL is the usage of Shortening Services, the output value assigned to this selection is 1 (phishing) in any other case 0 (legitimate).

- Prefix or Suffix "-" in Domain

Examining whether or not the given region contains the character '-' constitutes a component of the URL. Infrequently, the sprint picture might use URLs that were not valid. Phishers frequently add prefixes or suffixes to the domain name that are separated by dashes (-) in order to give consumers the impression that they are interacting with a legitimate website. This is done in the hopes that customers would provide sensitive information. If the URL of the Domain contains the symbol '-' inside the area section of the URL, the price allocated to this choice is 1 (phishing), but if it does not, the fee is 0. (valid).

3.6.2 Domain based Features

- DNS Record

When it comes to phishing websites, the stated identification isn't always determined by making use of the WHOIS, and there are no statistics based on the hostname. If the DNS report is empty or is no longer being monitored, then the cost associated with this option is set to 1 (phishing), and in any other circumstance, it is set to 0. (legitimate).

- Website Traffic

This function determines how popular the website is by determining the volume of traffic that it receives and the number of pages that users navigate through while they are there. On the other hand, due to the fact that phishing websites are only around for a short period of time, the Alexa database will no longer be able to identify them. If the website in question receives no visitors or isn't consistently identified as fraudulent by Alexa, the "Phishing" label is applied to it. The outcome of this option is 1 (phishing) if the rank of the domain is less than 100000; otherwise, it is 0. (legitimate).

- Domain Age

WHOIS may be used to find out information about this function. The majority of websites that are not real are only around for a short period of time. For the purposes of this thesis,

the minimum age of the legitimate region that is taken into consideration must be one year old. In this context, age does not refer to anything other than the span of time between a person's birth and their death. The outcome of this option is 1 (phishing) if the age of the domain is greater than one year; otherwise, it is 0. (valid).

- End Period of Domain

Calculating the closing area time for this feature requires finding the difference between the expiry time and the current time in order to get an accurate reading. For the purpose of this thesis, the stop period that is taken into consideration for the legitimate region is six months or significantly less. If the final term of the domain is less than six months, the cost of this pick is zero; if it is greater than six months, the cost is one (phishing) (legitimate).

3.6.3 HTML & JavaScript Features

- IFrame Redirection

If the iframe is empty or response isn't determined then, the fee assigned to this option is 1 (phishing) otherwise 0 (legitimate).

- Status Bar Customization

Phishers have another option available to them in the form of JavaScript, which allows them to display a sham URL to users by embedding it within the reputation bar. In order to extract this functionality, we will need to scavenge the web site supply code, specifically the "onMouseOver" event, and examine the reputation bar to see whether it undergoes any changes as a result of our efforts. If the answer is empty or an onmouseover location is found, then the price associated with this option is set to 1 (phishing), but otherwise it is set to 0. (legitimate)

- Disabling Right Click

Phishers employ JavaScript to suppress the right-click feature so that clients are unable to access and store the website's supply code. This prevents customers from falling for their scam. This method is

treated just as "Using onMouseOver to cover the Link," which you can find in the documentation. In any case, for the purpose of this operation, we are able to search for the event "event.button==2" inside the supplied code of the website in order to determine whether or not the appropriate click is disabled. The price that is assigned to this option is one (phishing) if the response is empty or onmouseover isn't always noticed; in any other circumstance, the price is zero (legitimate).

- Website Forwarding

The quality line that distinguishes phishing web sites from valid ones is how typically an internet site has been redirected. In our dataset, we discover that valid web sites were redirected one time max. On the other hand, phishing web sites containing this option were redirected as a minimum four times

3.7 Training the model

A neural network is a mathematical model that is stimulated with the assistance of using biological neural networks. An artificial neural network is another name for a neural network, however the two terms are used interchangeably (Ningxia Zhang, no date). The majority of the time, the apparatus in question is not one that can be modified in any way, shape, or form in order to accommodate the various stages of the training process.

A normalised dataset that is within the shape of 0s and 1s is presented in this section. After training with this dataset, many models perform better than they did before using this dataset. After we have finished training the models, we will go over the training dataset and eliminate any dataset values that the training models improperly categorised. In the future, we will be able to obtain a better degree of precision with the assistance of this. Utilizing models allows for the determination of the degree of precision;

- ➔ Decision Tree Classifier
- ➔ Logistic Regression
- ➔ Random Forest Classifiers
- ➔ XGBoost Classifier
- ➔ K Neighbors Classifier
- ➔ Multilayer Perceptron (MLPs): Deep Learning
- ➔ Autoencoder Neural Network
- ➔ Support Vector Machines
- ➔ **Decision Tree Classifier**

Although it can be applied to problems involving both classification and regression, the supervised learning method known as the decision tree is most frequently used to solve classification-related problems. It is a classifier that takes the shape of a tree, with internal nodes denoting a dataset's characteristics, branches denoting the decision-making processes, and leaf nodes denoting the classification's outcome. The two different types of nodes that can be found in a decision tree are the Decision Node and the Leaf Node. Choice nodes are used to make any decision and have many branches, whereas Leaf nodes are the output of those decisions and do not have any additional branches. A decision tree may contain leaf nodes. Based on the characteristics of the provided dataset, the judgments or test are conducted. It is a visual representation for discovering all feasible responses to a query or selecting a course of action based on the information provided. Because it starts with the root node and develops on subsequent branches to form a structure resembling a tree, it is called a decision tree. To build a tree, we employ a technique known as the CART algorithm, which stands for the Classification and Regression Tree algorithm. A decision tree begins with a single question to which the user can respond "Yes" or "No." The tree is then further divided into subtrees based on their responses.

➔ **Logistic Regression**

Logistic regression, which is a component of the approach known as Supervised Learning, is recognised as both one of the most well-known and commonly used machine learning algorithms. You will be able to make an accurate prediction of the categorical dependent variable by combining it with a predetermined group of independent variables. The outcome of a categorical dependent variable may be predicted by using the method of logistic regression. As a direct consequence of this, the value that is produced must either be discrete or categorical. It provides the probability values that lie between the two numbers as opposed to giving the exact value, which would be either 0 or 1. It might be either yes or no, 0 or 1, the truth or a falsehood, etc., depending on the circumstances. The manner in which each technique is implemented is the primary point of difference between the Logistic Regression and the Linear Regression. Linear regression is the method of choice when attempting to resolve issues pertaining to regression, whereas logistic regression is used when attempting to resolve issues pertaining to classification. As opposed to fitting a regression line, the fitting procedure for logistic regression involves fitting a logistic function in the form of a "S," which predicts two maximum values. The classic method of regression, on the other hand, almost always results in a straight line (0 or 1). The curve that results from evaluating the logistic function may be used to determine the chance of any occurrence, such as whether or not the cells are malignant, whether or not a mouse is overweight based on its weight, and so on. Logical regression is an essential machine learning strategy since it can generate probabilities and classify new data by making use of continuous and discrete datasets respectively. As a direct consequence of this, the method is very flexible. The term "logistic regression" refers to the process of classifying observations by using a wide variety of data.Regression, and it is capable of quickly identifying the factors that are most effective at doing so.

➔ **Random Forest Classifiers**

One of the more well-known machine learning algorithms, Random Forest, is included under the more general category of supervised learning. Classification and regression are two types of machine learning tasks that might benefit from its use. It is predicated on the idea of ensemble learning, which refers to the practise of integrating numerous classifiers in order to solve a difficult issue and to enhance the functionality of the model. "Random Forest" is a classifier that, as its name suggests, "contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Random Forest is a technique that "takes the average to improve the predictive accuracy of that dataset." The random forest model does not rely on a single decision tree; rather, it considers the forecast from each tree in the forest and determines the final output based on which tree's prediction received the majority of votes. The bigger the number of trees in the forest, the better the level of accuracy achieved, as well as the prevention of the issue of overfitting..

➔ **XGBoost Classifier**

The implementation of gradient-boosted decision trees known as XGBoost may be found here. The majority of Kaggle Competitions are completely controlled by XGBoost models. In this particular method, decision trees are constructed in a step-by-step fashion. The XGBoost algorithm places a significant emphasis on weights. Following the process of assigning weights to each of the independent variables and then feeding that information into the decision tree that predicts results, the process is complete. The variables whose results were incorrectly anticipated by the tree have their weights increased, and these results are subsequently provided to the second decision tree. After that, all of these separate classifiers and predictors are combined to produce a robust and more accurate model. It is able to solve issues involving regression, classification, ranking, as well as those that are user-defined.

➔ **K Neighbors Classifier**

The K-Nearest Neighbor algorithm, which uses the Supervised Learning method, is one of the most straightforward examples of machine learning. The K-Nearest Neighbors algorithm makes the assumption that the new case or data is similar to existing cases and places the new case into the category that is the most similar to the categories that are already available. The K-NN algorithm remembers all of the accessible data and determines how to categorise a new data point based on how similar it is to the stored data. This indicates that when new data becomes available, it can be easily classified into a well-suited category by making use of the K-NN algorithm. The K-Nearest Neighbors algorithm can be used for classification as well as regression, but the majority of the time, it is used for the classification problems. The K-Nearest Neighbors algorithm is non-parametric, which means that it does not make any assumptions about the data it is analysing. It is also known as a lazy learner algorithm because rather than immediately learning from the training set, it stores the dataset and then, when the time comes for classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

➔ Multilayer Perceptron (MLPs):

Deep Learning MLPs, also known as multilayer perceptrons, are a kind of artificial neural network that uses a feedforward architecture (ANN). The term "multi-layer perceptrons" (MLP) is sometimes used in a vague sense to refer to any feedforward artificial neural network (ANN), and sometimes it is used exclusively to refer to networks that are built of many layers of perceptrons (with threshold activation). In common parlance, multilayer perceptrons are commonly referred to as "vanilla" neural networks, particularly when they have only one hidden layer. This is especially true when the networks only have one hidden layer.

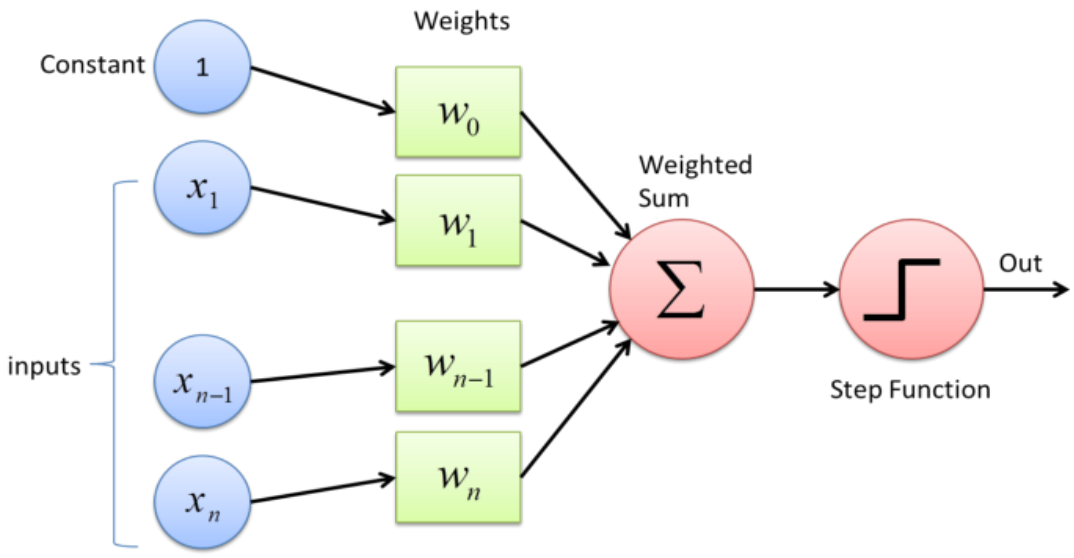

Fig 3: multilayer perceptron

➔ **Autoencoder Neural Network**

An auto-encoder is a type of common unsupervised artificial neural network that may acquire effective codings about the structure of the data by accepting unlabeled input and then applying those codings in a variety of different circumstances. The function that transforms data from the full input space to coordinates of a smaller dimension is called an auto-encoder. This function then generally returns the data to the same dimension as the input space with as little information lost as possible. For the purpose of improving the robustness of the model, auto-encoders may be utilised to extract features from the raw data for use in classification or regression tasks. An auto-encoder network may be put to a number of various uses thanks to its extensive list of potential applications.

➔ **Support Vector Machines**

Both classification and regression are performed with a technique to machine learning known as Support Vector Machine (SVM), which is supervised. In spite of the fact that we also refer to them as regression concerns, the classification system is where they fit in best. The purpose of the support vector machine (SVM) approach is to locate a hyperplane in an N-dimensional space that accurately classifies the input

points. The size of the hyperplane is determined by the total number of features. In the simplest case, where there are just two input characteristics, the hyperplane takes the form of a line. If there are three input characteristics, the hyperplane will transform into a two-dimensional plane. Trying to visualise anything gets more difficult whenever there are more than three aspects to consider.

3.8 Software Requirement

The software used for the project:

- Python
- Google Colab
- Juypter notebook

1.Python

Python is an interpreted programming language that may be used for a variety of purposes. Guido van Rossum was the creator of Python, and it was first made available to the public in the year 1991. Readability of the code is given a significant amount of weight in its design philosophy, and huge amounts of white space are effectively utilised throughout. Programmers will find that its language structures and object-oriented technique make it easier for them to write code that is clear and easy to understand, regardless of the size of the project they are working on. Garbage collection and dynamic typing are both features of the Python programming language. There is support for a wide variety of programming paradigms, including procedural, object-oriented, and functional programming. The programming language Python is sometimes referred to as a "batteries included" language because of its comprehensive standard library. The version 2.0 release of Python was made available on October 16, 2000, and it introduced a number of important new features. Two of them were a garbage collector that could identify cycles and support for the Unicode character set. The Python 3.0 release from December 12, 2008. The programming language

went through a considerable rewrite that is not fully compatible with earlier versions. Backports of many of Python's most essential features were made available in version series 2.6.x and 2.7.x of the programming language. The Python 3 versions include the utility known as 2 to 3, which, to some extent at least, automates the process of converting Python 2 code to Python 3 syntax. The programming language Python was developed with the goal of being easy to read. Its layout is uncluttered and easy on the eyes, and it commonly uses English terms in place of punctuation in other languages. Blocks are not separated from one another by using curly brackets, and the use of semicolons to terminate statements is not required. This is in contrast to the majority of other programming languages. When compared to Pascal or C, it has a much reduced number of syntactic exceptions and unique conditions. In 1999, Guido van Rossum defined his goals for the Python programming language:

Easy and intuitive language just as powerful as those of the major competitor

- Open source, so anyone can contribute to its development
- Code that is as understandable as plain English

The programming language Python was designed to be easy to read. When compared to other languages, it has a smaller number of syntactic structures, and it typically uses English terms in place of punctuation. Students and working professionals who wish to become exceptional software engineers should learn Python as soon as possible, especially if they work in the field of web development. Python is an absolute must for this. The following is a list of some of the primary advantages of learning Python:

Python is Interpreted : Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

- Python is an interactive programming language; you can really write your programme by interacting directly with the interpreter while working at a Python

prompt.

- Python is an object-oriented programming language, meaning that it supports the Object-Oriented programming style and method, which encloses code into objects.
- Python is a Great Language for Beginners: Python is a great language for the beginner level programmers and supports the development of a wide range of applications, ranging from simple text processing to browsers for the World Wide Web to games. It also supports the development of a wide range of games.

2.Google Colab

In order to facilitate the development of machine learning and deep learning models, Google Collab was developed to provide free access to computing resources such as GPUs and TPUs for anybody who has a need for such resources. Google Colab is an improved and more feature-rich version of Jupyter Notebook. Jupyter Notebook is an application that allows for the editing and execution of Notebook documents. It may be accessed through a web browser or an integrated development environment (IDE).

Google Colab provides tons of exciting features that any modern IDE offers, and much more. Some of the most exciting features are listed below.

- Machine learning and neural network instruction presented in the form of interactive lessons.
- You can still write and run Python 3 programmes even if you don't have a local setup..
- Using the "Execute terminal commands" option, you may write Python 3 code right in the Notebook and then immediately run it. Import datasets from other sources such as Kaggle while keeping your local setup intact.

3. Jupyter Notebook

Jupyter notebook is a web-based interface that Fernando Perez developed as a replacement for IPythonkernel. The Notebook project has been relocated to Project Jupyter, which provides a front interface for the programming environments Juila and R in addition to

Python, in an effort to provide an integrated interactive computing environment for various languages.

Rich text features, like as equations and diagrams, along with rich text that is formatted using HTML, are what make up a notebook page. Additionally, the notebook may be used as an executable document, which may include Python or code blocks from another compatible language. The programme Jupyter Notebook is a client-server application. The application initiates the installation of the server on a local computer and initiates the installation of the notebook interface in a web browser so that it may be changed and utilised. The notebook is kept on your computer as an ipynb file, and it is also capable of being exported as an html, pdf, and LaTex file.

CHAPTER 4
RESULT AND DISCUSSION

This section gives the experimental info of the skilled version type algorithms and the kinds of function extraction used are detailed. Then the outcomes of the comparative checks among those algorithms with associated traits are presented.

First part of this phase is to extract the features from the original dataset where it contains 0s (legitimate) and 1s (phishing). The extracted features are categorized into,

- Address Bar based Features
- Domain based Features
- HTML & JavaScript based Features

The status of the website is checked before the extraction process starts. Most of the websites were taken down by relevant authorities. Therefore, the dataset was filtered out with python functionality to check whether the website is still live or not. The dataset contains 5000 legit websites and 5000 phishing websites were extracted from the Original Dataset after performing the feature extraction. Dataset consists of 16 features.

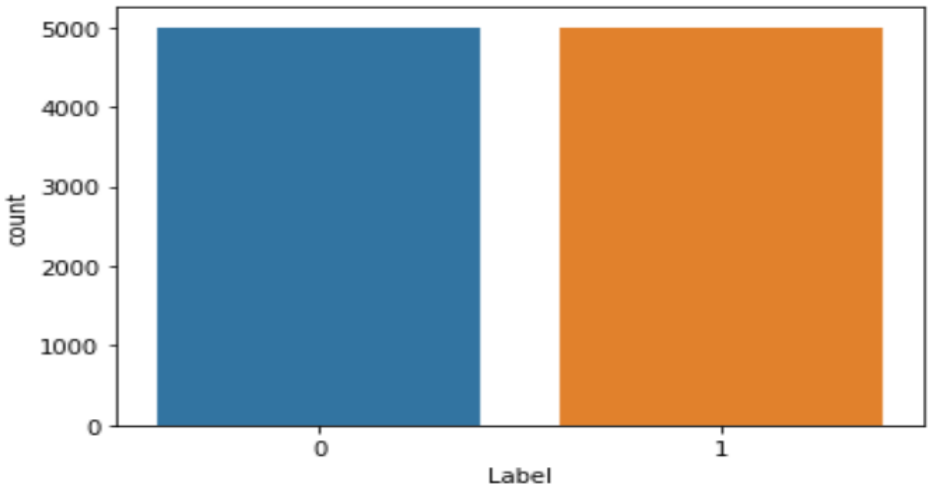

Fig 4 : Classes count of dataset

Since it took 5000 legitimate websites and 5000 phishing websites the class is well balanced and there are no null values in the dataset. Therefore, further preprocessing is not required.

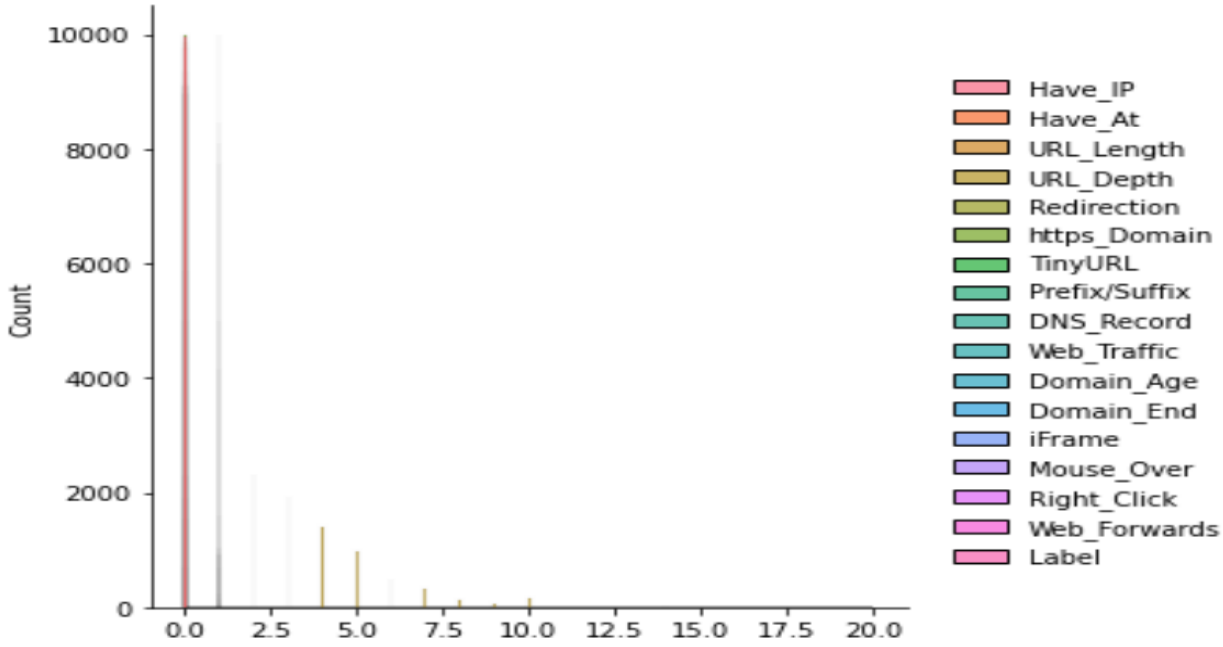

Fig 5: Overall distribution of continuous data variables

4.1 Classification report

Classification models that can be trained using this dataset are:

Decision Tree Classifier

- Logistic Regression
- Random Forest Classifiers
- XGBoost Classifier
- K Neighbors Classifier
- Multilayer Perceptron (MLPs): Deep Learning
- Autoencoder Neural Network
- Support Vector Machines

Four specific facts as sensitivity, f-measure, precision and accuracy are calculated to determine the usefulness and performance of the algorithms by the usage of the values in the confusion matrix. also are essential for creating a contrast among the examined system getting to know approaches.

- Precision

The capacity of a classifier to avoid incorrectly labelling as positive an occurrence that is, in fact, negative is what is meant by the term "precision." It is determined by comparing the number of true positives to the total number of true and false positives for each class.

$$Precision = TP /(TP + FP)$$

- Recall

The ability of a classifier to locate all positive examples is referred to as its recall. This metric is defined differently for each class and is expressed as a ratio of the number of true positives to the total number of true positives plus false negatives. To phrase it another way, "what percent of the cases that were genuinely positive were classified correctly?"

$$Sensitivity(Recall) = TP/(TP + FN)$$

- F1 score

The F1 score is a weighted harmonic mean of precision and recall, with a maximum score of 1.0 and a minimum score of 0.0. The best score possible is 1, while the lowest score possible is 0. Because F1 scores incorporate precision and recall into their computation, they often have a lower accuracy rating than other measures of accuracy. It is recommended that when comparing classifier models, the weighted average of F1 be utilised rather than global accuracy as the metric of choice. f1Score is equal to the percentage of correctly made positive forecasts.

$$F - Measure = 2 \times Precision \times Sensitivity /(Precision + Sensitivity)$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where TN means true negative, TP means the true positive, FP describes false positive, and FN implies the false negative rate of the classification algorithms.

- Decision Tree Classifier

The TP rate indicates the true positive rate for each class, which defines the numeric value of positive classifications by the model. In Decision Tree Classifier, Class 0 (legitimate) has the highest TP Rate. FN rate is indicating the number of false negatives classified by the model. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.70	0.99	0.82	932
1	0.98	0.64	0.77	1068

Fig : Classification report of Decision Tree Classifier

Precision is the proportion of true positive instances among the instances that the model classified as positive, and sensitivity (Recall) is the rate of the total number of positive instances classified as positive. In the Decision Tree Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Logistic Regression

In Logistic Regression Classifier, Class 0 (legitimate) has the highest TP Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.71	0.97	0.82	932
1	0.96	0.65	0.78	1068

Fig : Classification report of Logistic Regression

Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Random Forest

Classifiers In Random Forest Classifier, Class 0 (legitimate) has the highest TP Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.71	0.99	0.83	932
1	0.99	0.64	0.78	1068

Fig : Classification report of Random forest

In Random Forest Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate). This indicates a good performance.

- XGBoost Classifier

In XGBoost Classifier, Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.78	0.95	0.86	932
1	0.95	0.77	0.77	1068

Fig 6: Classification Report of XGBoost Classifier

Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate) in XGBoost Classifier. Comparatively Sensitivity and Precision values for all the classes are higher than other models in this. Therefore, it delivers a good performance.

- K Neighbors

Classifier Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.77	0.77	0.77	932
1	0. 80	0.80	0.80	1068

Fig : Classification Report of K Neighbours Classifier

Class 1 (phishing) has higher Sensitivity and Precision than Class 0 (legitimate) in K Neighbors Classifier

- Multilayer Perceptron (MLPs): Deep Learning

Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.79	0.93	0.86	932
1	0.93	0.78	0.85	1068

Fig : Classification Report of Multilayer Perceptron (MLPs): Deep Learning

In Multilayer Perceptron (MLPs): Deep Learning Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

- Support Vector Machines

Class 0 (legitimate) has the higher TP Rate and TN Rate. For all classes the TP and TN rate is higher value and the FP and FN rate is lower value.

	Precision	Recall	F1-Score	Support
0	0.69	0.98	0.80	932
1	0.97	0.61	0.75	1068

Fig : Classification Report of Support Vector Machines

In Support Vector Machines Classifier, Class 0 (legitimate) has higher Sensitivity than Class 1 (phishing) while Class 1 (phishing) has more Precision than Class 0 (legitimate).

URL Length has the highest rate of feature importance in the dataset. It has a broader range of values compared to other features. Secondly, the Prefix/Suffix feature has the highest rate feature importance in the dataset while other features have the same rate of feature importance.

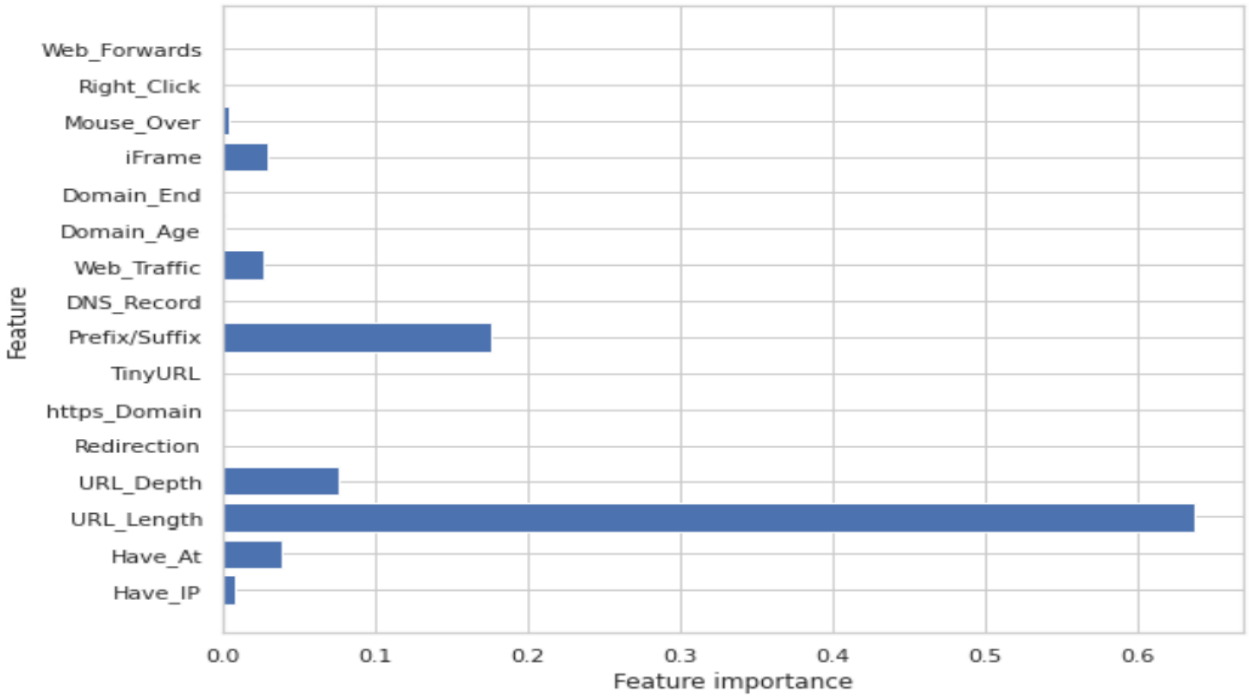

Fig 7: Importance of features

	ML Model	Accuracy
1	Logistic Regression	0.812
2	Decision Tree	0.820
3	Random forest	0.822
4	K Neighbors Classifier	0.826
5	XGBoost	0.872

	ML Model	Accuracy
6	SVM	0.811
7	Auto Encoder	0.858
8	Multilayer Perceptron	0.867

Fig : Model Accuracy before splitting on feature type

XGBoost algorithm has the best accuracy classification performance of 87.2% (table 8). The Multilayer Perceptron algorithm has the second highest accuracy of 86.7%.

Then the dataset was divided into three sub datasets based on the feature categorization. Each dataset was trained separately with above machine learning algorithms again. Domain Features and Html and JavaScript features have lower values compared to the previous results which was trained before spilling the dataset into three sub datasets. Here, the AutoEncoder algorithm has the best classification performance with 84.8% accuracy. Based on the experimental results, it is clear that the features based on the address bar do better than other features. This accuracy rating is interpreted as a great and appropriate end result for fraud detection. An exact ratio of 100% is impossible. Because while sysadmins are trying to use some new methods, attackers are trying to improve their attack methods in order to avoid mistakes. In a typical phishing attack, the website is designed as if it was a legitimate website, So attackers attempt to disguise it the use of a protracted URL the use of a few unique phrases to misinform customers as URLs can clutch extra quick clips made by users who has basic knowledge of phishing attacks. Finally, the tested model saved for API development and chrome extension was developed using that API. Once a user browses a website chrome extension automatically detects the website and sends it to a trained model via the API to check whether the website is legitimate or not. If the website is legitimate chrome extension will notify the user with green check mark if not danger mark

4.2 Screenshots

4.2.1 Checking flipkart website

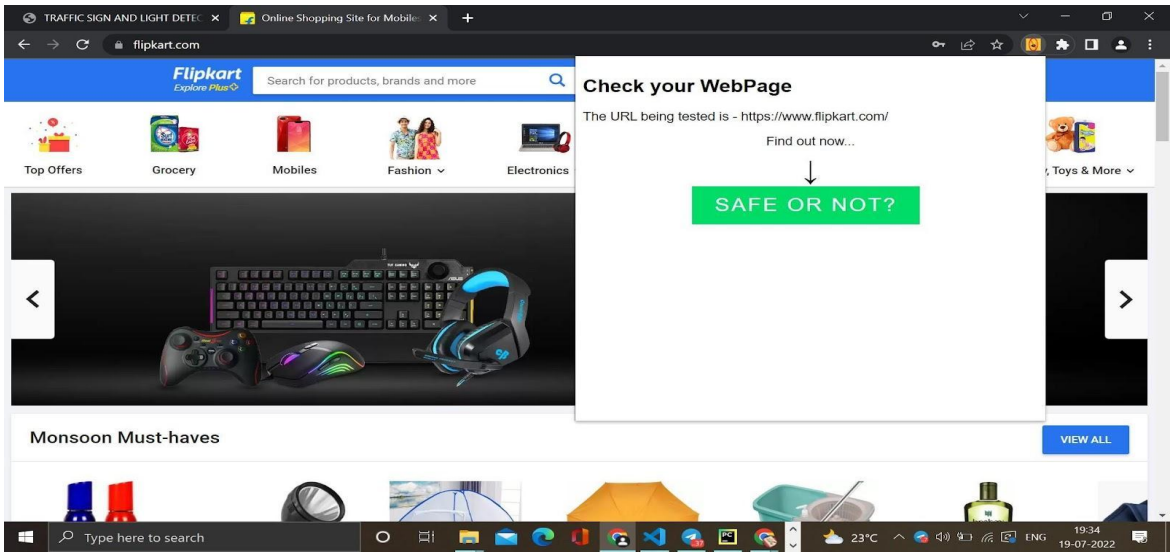

Fig 8: checking flipkart site

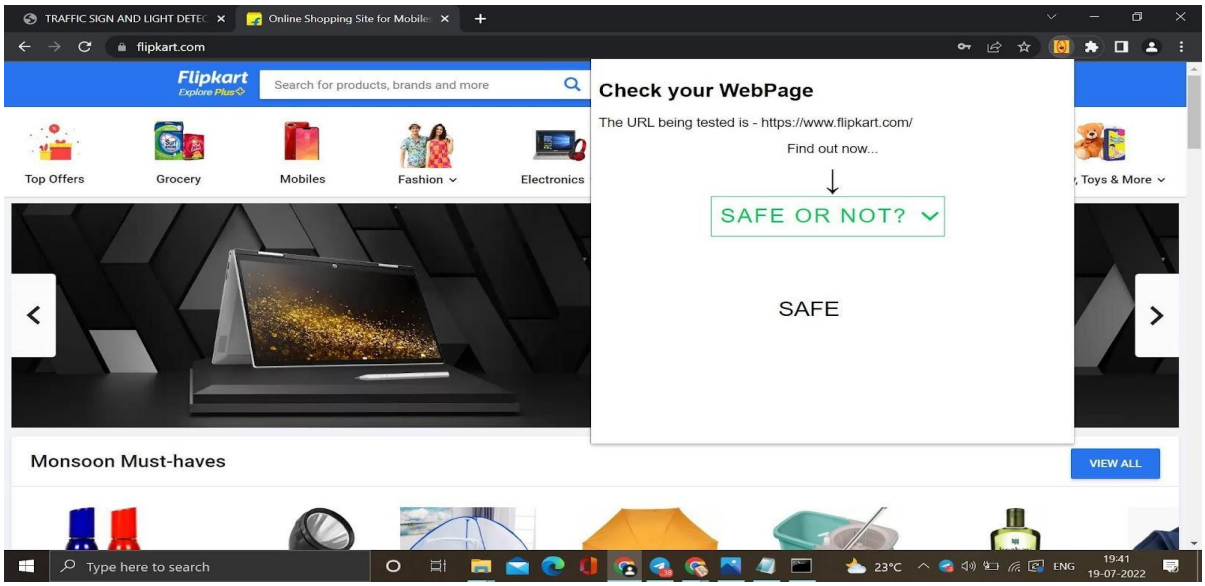

Fig 9: Result

4.2.2 Checking amazon prime website

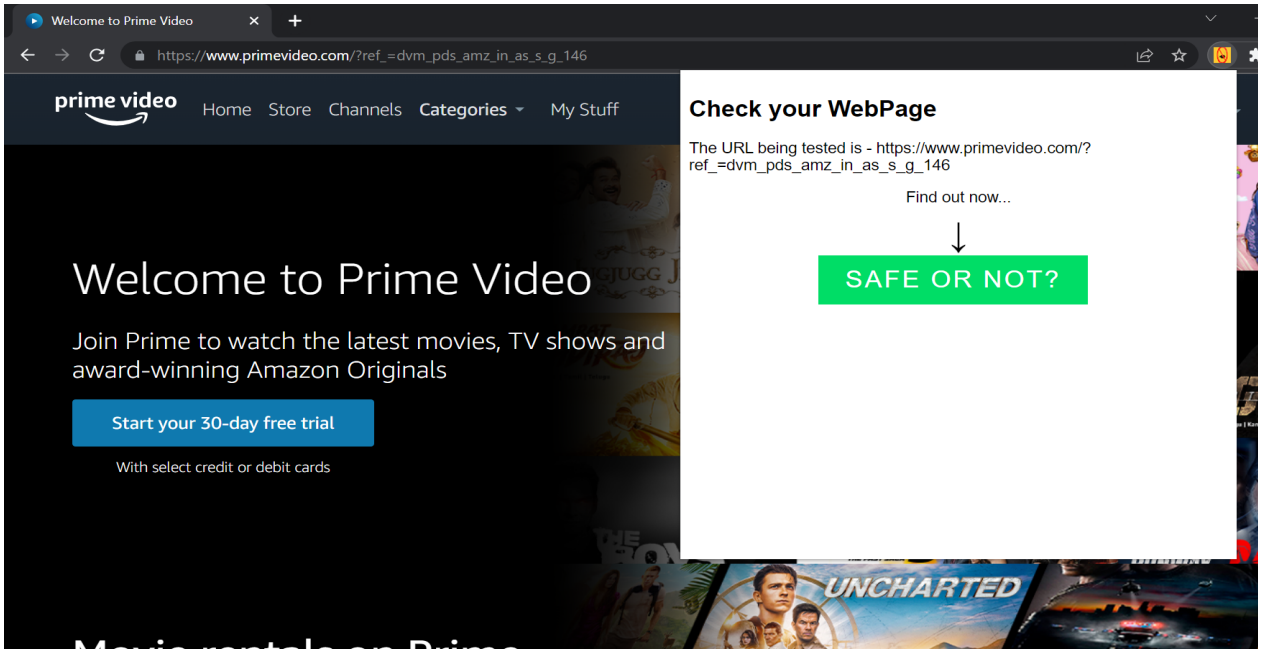

Fig 10: checking amazon prime site

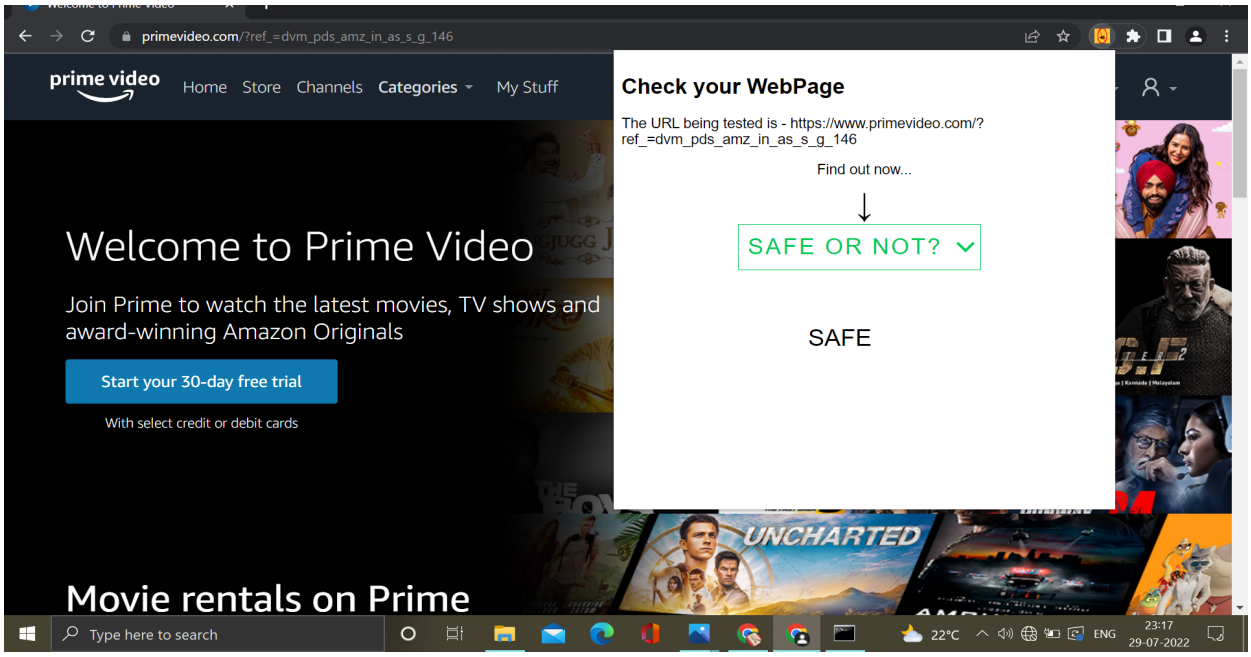

Fig 11: Result

CHAPTER 5
CONCLUSION

It has been documented that a reliable anti-phishing systems predicts phishing attacks over a period of time. Providing an honest anti-phishing remedy with an honest expiration date is also important for increasing the predicted number of phishing sites. Due to rapid growth of phishing scenarios, different techniques were used to fool the user who browse the internet. Sine it is very difficult to find a up to date dataset to perform the machine learning algorithms dataset was created by extracting features based on techniques which is used by phishers recently.

However, Feature extraction is a process that is very difficult and time consuming for this dataset. Not only for feature extraction but also for uptime of the websites were tested before doing the extraction. All these processes are time taking and good processing power is required. In this thesis, Phishing detection system has been implemented by using eight different machine learning algorithms, as XG Boost Classifier, Logistic Regression, Random Forest Classifiers, Decision Tree Classifier, K Neighbours Classifier, Multilayer Perceptron (MLPs): Deep Learning, Autoencoder Neural Network and Support Vector Machines. It is important to create an efficient list of functions to enhance the accuracy of the detection system. Therefore, feature list has been grouped in three different classes.

After training with above mentioned machine learning algorithms, XGBoost algorithm delivered the best classification performance (87.2% accuracy). The Multilayer Perceptron algorithm has the second highest accuracy of 86.7%. Then the dataset was divided into three sub datasets based on the feature categorization. Each dataset was trained separately with above machine learning algorithms again. By analyzing the result of machine learning algorithms' accuracy, it showed that there are some features performing low accuracy values and it might affect the entire 38 dataset training accuracy. Therefore, finding the right combination of features is a difficult task in this context. Attackers find new techniques

to perform the phishing attacks. Therefore, it became challenging to find out the latest techniques and have an up to date feature set. And also, the lifetime of the phishing website is short (less than a year). Websites should be live to perform all the feature extraction tasks.

CHAPTER 6
FUTURE WORKS

All datasets may be used for building the statistics base with the employment of this deep gaining knowledge of technology, having a huge collection of datasets will be good for training and accuracy. Therefore, a few multiprocessing strategies may be personalized to the system. Further, the system can be developed to get higher accuracy by trying the vectors of words that rely on using the words in the website URL without doing alternate operations. Even though finding and having a large dataset is challenging it will give more accuracy to the final output. One of the future improvements to the model is to add a methodology for evaluating the importance of features. The system can be integrated with a website where user can check a website is legitimate or not rather than having a chrome extension where it detects the website automatically

REFERENCE

1. Abrams A. Watch out: these unsubscribe emails only lead to further spam. 2021. https://www.bleepingcomputer.com/news/security/watch-out-these-unsubscribe-emails-only-lead-to-furth er-spam/.
2. Abroshan H, Devos J, Poels G, Laermans E. Phishing happens beyond technology: the eFfects of human behaviors and demographics on each step of a phishing process. IEEE. 2021. https://doi.org/10.1109/ACCESS.2021.3066383.
3. Alabdan R. Phishing attacks survey: types, vectors, and technical approaches. Future Internet. 2020;. https://doi.org/10.3390/ f12100168.
4. Al Khozaei M.G, Batarf OA. Phishing websites detection based on Phishing characteristics in the webpage source code. International Journal Information Communication Technology Research. 2011
5. Al Saadoon I, Ramadan RA, Khedr AY. Cultural impact on users' ability to protect themselves against phishing websites. International Journal Computer Science and Network Security. 2017
6. Bartoli A, De Lorenzo A, Medvet E, Tarlao F. How phishing pages look like?Cybern Inf Technol 2018. https://doi.org/10. 2478/cait-2018-0047.
7. Basit A, Zafar M, Liu X, Javed AR, Jalil Z, Kifayat K. A comprehensive survey of AI-enabled phishing susceptibility. Decis Support Syst. 2020. https://doi.org/10.1016/j.dss.2020.113287.
8. Fan W, Lwakatare K, Rong R. Social engineering: I-E based model of human weakness for attack and defense investigations. Int J Comput Netw Inf Secur. 2017. https://doi.org/10.5815/ijcnis..