PHISHING WEBSITE DETECTION TECHNIQUE USING MACHINE LEARNING

NURUL AMIRA BINTI MOHD ZIN

BACHELOR OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING) WITH HONORS
UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	NURUL AMIRA BINTI MOHD ZIN
Date of Birth	:	
Title	:	PHISHING WEBSITE DETECTION TECHNIQUE USING MACHINE LEARNING
Academic Session	:	2022/2023

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997) *
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done) *
☒	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

_____	_____
(Student's Signature)	(Supervisor's Signature)
_____	MOHD FAIZAL BIN AB RAZAK
New IC/Passport Number	Name of Supervisor
Date: 8/6/2023	Date: 8/6/2023

NOTE: * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Perpustakaan Universiti Malaysia Pahang,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak,
26300, Gambang, Kuantan.

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's
Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

_____
(Supervisor's Signature)

Date:

Stamp:

Note: This letter should be written by the supervisor, and addressed to the Librarian, *Perpustakaan Universiti Malaysia Pahang* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I/We* hereby declare that I/We* have checked this thesis/project* and in my/our* opinion, this thesis/project* is adequate in terms of scope and quality for the award of the degree of *Doctor of Philosophy/ Master of Engineering/ Master of Science in ……..

_____

(Supervisor's Signature)

Full Name : DR MOHD FAIZAL BIN AB RAZAK

Position : SENIOR LECTURER

Date : 8/6/2023

_____

(Co-supervisor's Signature)

Full Name :

Position :

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

_____

(Student's Signature)

Full Name : NURUL AMIRA BINTI MOHD ZIN

ID Number : 000617060506

Date : 8/6/2023

PHISHING WEBSITE DETECTION TECHNIQUE USING MACHINE LEARNING

NURUL AMIRA BINTI MOHD ZIN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science
(Software Engineering) with Honors

Faculty of Computing
UNIVERSITI MALAYSIA PAHANG

NOVEMBER 2022

ACKNOWLEDGEMENTS

First, I would like to thank Allah for allowing me to complete this research. Secondly, I would like to offer my sincere gratitude to Dr. Mohd Faizal Bin Ab Razak, my supervisor, who not only helped me through the project but also provided helpful suggestions and direction for me to successfully complete the project.

In addition to assisting me in crafting a successful presentation of the project, he helped me comprehend the complex challenges that are involved in its creation. Without this intervention, these nuances would have been lost. His direction was the single most important factor in the accomplishment of my endeavor. Throughout the whole of the project work, Dr. Mohd Faizal Bin Ab Razak was a tremendous source of inspiration for me, supplying me with invaluable recommendations, inspirations, and ideas. I owe him a great debt of gratitude.

Then, I would want to express my gratitude to my parents, Mohd Zin Bin Iurat and Noraizah Binti Sapuan, for providing me with both emotional and monetary support throughout my life. In addition, I would want to express my gratitude to my fellow students and friends who have supported me throughout the process of completing this project by providing me with insightful comments and direction at key points along the way.

In closing, I would want to express my gratitude to everyone who has helped in any way, whether directly or indirectly, to bring this project to a successful conclusion.

ABSTRAK

Internet telah muncul sebagai alat yang sangat diperlukan dalam kedua-dua kehidupan peribadi dan profesional kita pada zaman moden kita. Internet adalah penting bukan sahaja untuk pengguna individu, tetapi juga untuk perniagaan, kerana perusahaan yang menyediakan perdagangan dalam talian mungkin mendapat kelebihan daya saing dengan memberi perkhidmatan kepada pelanggan di seluruh dunia. Ini menjadikan Internet penting untuk semua orang yang menggunakannya. Internet membolehkan syarikat menjalankan e-dagang yang berkesan dengan pelanggan yang berada di seluruh dunia tanpa mengambil kira kekangan geografi pasaran individu. Akibat langsung daripada ini, bilangan pelanggan yang membuat pembelian mereka melalui internet semakin meningkat dengan cepat. Setiap hari, transaksi berjumlah ratusan juta dolar dilakukan melalui Internet. Individu yang tidak jujur ini telah tergoda untuk mengambil bahagian dalam usaha penipuan mereka dengan kuantiti wang ini. Pengguna Internet mungkin terdedah kepada pelbagai jenis ancaman web akibat daripada fakta ini. Ancaman ini boleh mengakibatkan kerugian kewangan, penipuan penggunaan kad kredit, kehilangan data peribadi, potensi kerosakan pada reputasi jenama dan ketidakpercayaan pelanggan terhadap e-dagang dan perbankan dalam talian. Oleh sebab itu, melakukan transaksi kewangan melalui Internet penuh dengan potensi risiko. Pancingan data ialah sejenis ancaman siber yang boleh ditakrifkan sebagai amalan meniru tapak web tulen untuk tujuan mencuri maklumat sensitif seperti nama pengguna, kata laluan dan nombor kad kredit. Artikel ini akan menumpukan banyak ruang untuk membincangkan topik pancingan data. Di samping itu, kami menyediakan kemas kini mengenai penemuan terkini daripada penyelidikan yang dijalankan mengenai topik tersebut. Di samping itu, kami ingin menemui kemajuan terkini dalam pancingan data dan langkah pencegahan, serta menjalankan analisis dan semakan penuh penyelidikan ini, semuanya dengan matlamat untuk merapatkan jurang pengetahuan yang masih wujud dalam bidang tertentu ini. Penyelidikan ini memberi tumpuan kepada strategi untuk mengesan serangan pancingan data melalui internet dan bukannya cara untuk mengesan serangan melalui e-mel.

ABSTRACT

The Internet has emerged as an indispensable tool in both our personal and professional life in our modern day. The Internet is crucial not just for individual users, but also for businesses, since enterprises who provide online commerce may gain a competitive advantage by serving customers all over the globe. This makes the Internet essential for everyone who uses it. The Internet enables companies to conduct effective e-commerce with customers located all over the globe without regard to the geographical constraints of individual markets. As a direct consequence of this, the number of customers who make their purchases over the internet is quickly increasing. Daily, transactions totaling hundreds of millions of dollars are carried out through the Internet. These dishonest individuals were tempted to participate in their fraudulent endeavors by this quantity of money. Internet users may be vulnerable to a wide variety of web threats because of this fact. These threats may result in monetary loss, fraudulent use of credit cards, the loss of personal data, potential damage to the reputation of a brand, and customer mistrust in e-commerce and online banking. Because of this, doing financial transactions through the Internet is fraught with potential risks. Phishing is a sort of cyber threat that may be defined as the practice of imitating a genuine website for the purpose of stealing sensitive information such as usernames, passwords, and credit card numbers. This article will devote considerable space to discussing the topic of phishing. In addition, we provide an update on the most recent findings from research conducted on the topic. In addition, we want to discover recent advancements in phishing and preventative measures, as well as carry out a full analysis and review of this research, all with the goal of bridging the knowledge gap that still exists in this field. This research focuses on strategies for detecting phishing attacks through the internet rather than ways for detecting attacks via email.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

APWG	Anti-Phishing Working Group
URL	Uniform Resource Locator
IT	Information Technology
BEC	Business email compromise
PC	Personal Computer
DNS	Domain Name System
TF-IDF	Term frequency–inverse document frequency
ISP	Internet service provider
HTML	Hypertext Markup Language
LR	Logistic Regression
BART	Bayesian Additive Regression Trees
CART	Classification and Regression Trees
RF	Random Forests
NN	Neural Networks
MLBDMs	machine learning-based detection methods
NB	Naive Bayes
SDLC	Software Development Life Cycle

PD	phishing detection
PWD	Public Works Department
AI	Artificial intelligence
MLP	Multi-Layer Perceptron
KNN	K-Nearest Neighbors

CHAPTER 1

INTRODUCTION

This chapter addresses the fundamental theoretical underpinnings of the investigation. In order to provide readers an understandable synopsis of the research, this chapter has been broken up into six sections. The history of the research is covered in Section 1.1 of the paper. Section 1.2 provides a definition of the issue statements and places an emphasis on several subjects like application risk and phishing detection. The aim of the research as well as its goals are discussed in Section 1.3. The breadth of the investigation is discussed in Section 1.4. In Section 1.5, we go over the significance of research, and in Section 1.6, we demonstrate how the thesis should be structured.

1.1 Introduction

These days, phishing assaults are among the most common kinds of cyberattacks that may be carried out. Any mode of communication can be used to target an individual and trick them into revealing private data in a fake setting, which can later be used to harm the individual victim or even a whole corporation. The aim of the attacker and the type of data that is released are both factors that determine which mode of communication should be used.[1]

In addition to this, they sought a ransom and threatened to cancel the customer's account if they did not get it. Email spoofing is an additional kind of the fraudulent activity known as phishing. Customers are regularly tricked into giving sensitive information such as credit card numbers and passwords via the use of deceptive practices. Because of this, phishing is most often employed to get essential information like as login credentials for bank accounts and credit card details. Consumers and companies alike are losing trust in the legitimacy of online transactions as a result of the proliferation of this kind of fraudulent activity. Because of this, customers acquired a negative impression of the internet organization, and therefore, they lost faith in doing business online. Even while encryption software is being used to protect the data

that is being kept on computers, the machines themselves are still susceptible to being attacked.[2]

Phishing attacks are harmful, but they can be avoided if more people are aware of them and develop the habits of remaining vigilant, always being on the lookout while surfing the Internet, and only clicking links after first determining whether the source of the links is trustworthy and reliable in any way. There are further technologies,including as browser extensions, that may warn users whenever they enter their credentials on a fraudulent website, which might possibly transmit their credentials to a person who has criminal intentions. Other systems may enable networks to lock down everything while still allowing access to designated sites, which provides a higher level of protection but comes at the expense of user convenience.[1] In this work, machine learning was used to identify phishing.

1.2 Problem Statement

Everyone benefits greatly from the Internet through websites to interact with the world. Besides, several online activities may be performed utilizing the Internet, such as cloud storage, online banking, online shopping, and online communication, which are regrettably not safe due to phishing websites. Furthermore, while there are various contemporary ways of recognizing phishing websites, these systems are still incapable of detecting and blocking all kinds of phishing. However, when it comes to discriminating between phishing and legal websites, the current system still has very high false alarm rates to differentiate it. Phishing websites also target industries since many industries use websites their advertising their company services to people.[3]

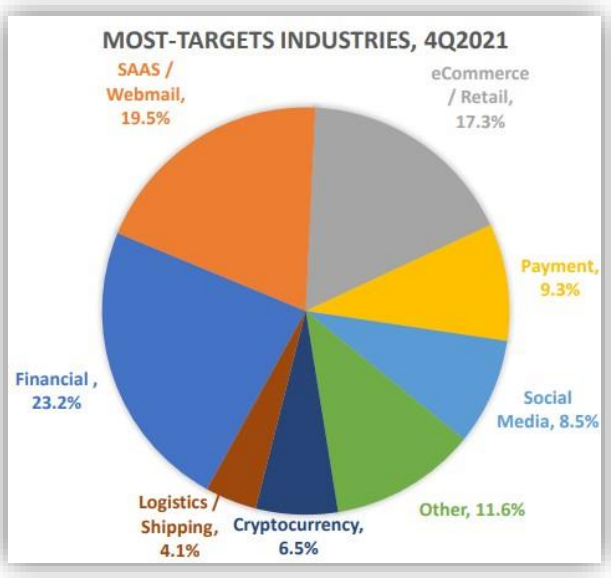

Figure 1 Most Targeted Industries

According to Figure 1, the Anti-Phishing Working Group (APWG) documented 316,747 attacks in the month of December 2021, making it the month with the greatest monthly total in the organization's history of reporting. Phishing scams have been more common since the beginning of the year 2020. In the fourth quarter of 2018, the financial industry was the industry that was the most often targeted by phishing, accounting for 23.2% of all attacks. The amount of cyberattacks directed against suppliers of SaaS and webmail has remained quite high. The percentage of attacks that were phishing scams aimed at cryptocurrency targets, such as bitcoin exchanges and wallet providers, rose to 6.5%. The number of companies that were discovered to be victims of ransomware jumped by 36% from the third quarter to the fourth quarter. 51.8% of emails identified as phishing attacks by business users were efforts to steal credentials, 38.6% of emails were response-based attacks (such as BEC, 419, and gift card scams), and 9.6% of emails were attempts to distribute malware.

1.3 Objective

The objectives of this research are:

i. To review the current phishing detection system issue.

ii. To develop a phishing detection technique system that analyses website applications using a Machine Learning approach.

iii. To evaluate the proposed detection technique system in terms of phishing detection accuracy.

1.4 Scope

The scope of this research:

i) Platform:
- This system is only for website applications.

ii) Development / Functionality:
- The system can only identify phishing websites but not remove them from the websites.
- The detection method applies to web-based only.

iii) User
- All computer users are students, operations and finance employees, and government employees.

1.5 Significant

This research will determine the importance of a phishing attack detection system. These are the advantages:

i. Able to prevent internet users from falling into scams and incurring financial losses.

ii. Able to secure the organization's website against phishing websites.

iii. Able to protect private information saved on websites.

iv. Able to provide security to Windows users from being hacked and reduce the threat.

1.6 Thesis Organization

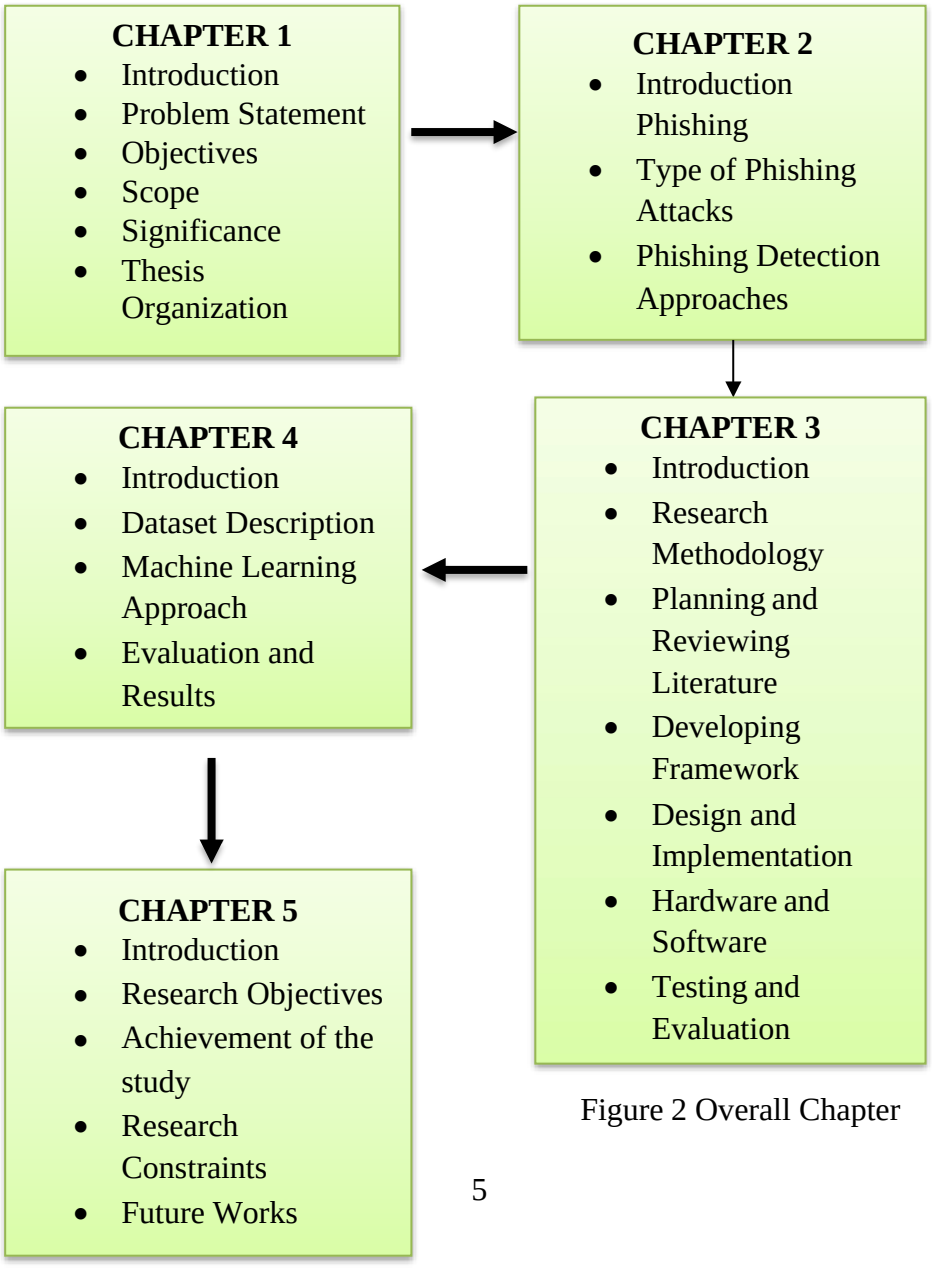

Figure 2 Overall Chapter

Figure 2 demonstrates three main chapters: Introduction, Literature Review, and Methodology.

This research will include five chapters. Chapter 1 consists of an introduction to this research as well as information on the present topic. This section also consists of the problem statement, aim, scope, and significance.

Next, Chapter 2 provides a debate topic regarding the literature view of existing research. The issue is a phishing attack, the type of phishing attack, and the comparison of current methods' solutions with prior relevant research.

Then, Chapter 3 shows a review of the methods applied in this research. The paper describes data collecting, data standardization, and the software utilized in this experiment on this topic.

Following that, Chapter 4 discusses project testing methods and project outcomes. It also includes a user manual and various appendices. The project's development must correspond to and achieve the project's goal.

The last chapter of this research will conclude the overall project based on the output that meets the objectives, the application of the methodology used for projects that need specifying software and hardware, as well as the constraint system and future work.

1.7 Conclusion

As may reasonably be deduced, this chapter constitutes one of the crucial parts of the whole investigation. This chapter contrasts prior solutions provided by another researcher with Machine Learning, which is the solution that is now being used. In addition, this chapter details the several methods that supported the researcher in arriving at their proposed malware detection technique and discusses how these methods worked. During the few decades that have just passed, several methods have been put up. However, these tactics need improvement so that in the future they can provide greater results. As a result, the current method to assist users of the internet in recognizing phishing websites is proposed in Chapter 3.

CHAPTER 2

LITERATURE REVIEW

This chapter gives an overview of the security component of the system that detects phishing websites and uses it as a springboard to explain the vulnerabilities that were discovered in the website application. The relevance of risk assessments and phishing detection for internet applications is brought to light in this chapter. The history of phishing on websites is discussed in order to acquire a better understanding of the challenges that are experienced by website users. Included in this discussion are the categories of website detection systems. These classes include analytic methods, detecting methodologies, and numerous additional deployments that may be used. Before the chapter is ended with a summary, this part of the chapter will talk about the dangers that users of websites confront.

2.1 Introduction

This research has been described in Chapter 1, introduction to analysis, which includes the problem statement, objective, importance, and scope. This research will examine the essential literature review in this chapter to understand the system approach and how the Phishing Website may be identified. As a result, existing Phishing Website Detection works will be expanded to justify the present work.

2.2 Phishing

Phishing is an attempt to collect sensitive data such as usernames, passwords, and credit card information by impersonating a trusted person in an electronic conversation with the goal of tricking the target into giving the information. This may be done in order to steal the information. This behavior is often motivated by evil intent. Phishing is often carried out via e-mail spoofing or instant messaging, and it commonly urges people to submit personal information at a bogus website that looks and feels like the authentic one. Phishing is illegal. The Uniform Resource Locator (also known as the

URL) of the website that is being phished is the only thing that is different between the two websites. Communication channels such as social networks, auction sites, banks, online payment processors, or IT (Information Technology) administrators are commonly misused in order to attract victims. This is done to get access to their personal information. Emails used in phishing scams may include links to websites that distribute viruses.

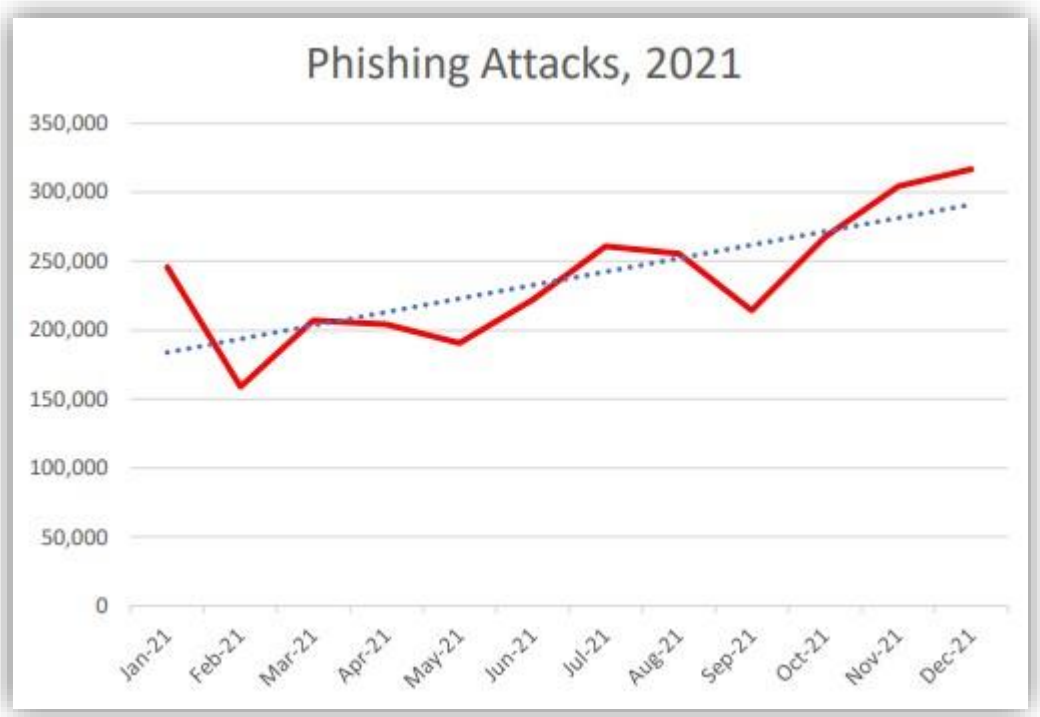

Figure 3 Phishing Sites Detected on 2021

According to Figure 3, the APWG documented 316,747 attacks in the month of December 2021. This is the greatest monthly total in the history of the APWG's reporting. Phishing scams have been more common since the beginning of the year 2020. In the fourth quarter of 2018, the financial industry was the industry that was the most often targeted by phishing, accounting for 23.2% of all attacks. The amount of cyberattacks directed against suppliers of SaaS and webmail has remained quite high. The percentage of attacks that were phishing scams aimed at cryptocurrency targets, such as bitcoin exchanges and wallet providers, rose to 6.5%. The number of companies that were

discovered to be victims of ransomware jumped by 36% from the third quarter to the fourth quarter.51.8% of emails reported by business users were credential theft phishing attacks, 38.6% were response-based attacks (such as Business email compromise (BEC), 419, and gift card scams), and 9.6% were.

2.3 Types of Phishing

There are different forms of phishing assaults available nowadays. It has been classified into three categories, as follows:

2.3.1 Content-Injection Phishing

Content-injection Phishing is the practice of injecting harmful material into a genuine website. The malicious material can drive users to other websites, install malware on their computers, or inject a frame of content that redirects data to a phishing server. [4]

2.3.2 Deceptive Phishing

Phishing attacks often take the form of misleading phishing, which is the most prevalent kind. Impersonating a reputable website and sending an email to the target that is designed to seem like it was sent from that website is required. The email would include a URL or link that led to a malicious website. It would provide the target the URL and instruct them to go there. The phishing website, after following the instructions, will capture the login credentials of the victim, along with any other sensitive information, and will transmit it to the attacker.[5]

2.3.3 Malware-based Phishing

Phishing attacks that are based on malware often include the installation of unnecessary software or applications on the target user's personal computer (PC). The virus makes use of a key logger as well as a screen logger in order to record the keystrokes that are made on the keyboard as well as the websites that are viewed on the internet. Key loggers, session hijacking, domain name system (DNS) phishing, content-injection

phishing, phone phishing, system reconfiguration, and link manipulation are the different types of attacks that fall under this category.[6]

2.4 Phishing Detection Approaches

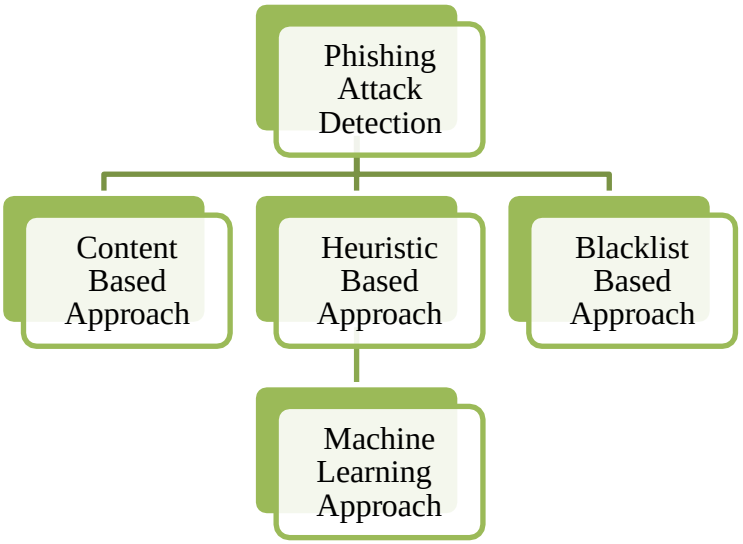

Figure 4 Phishing Website Detection Approaches

2.4.1 Content-based Approach

CANTINA is a content-based approach for identifying phishing websites that is based on the information retrieval algorithm term frequency-inverse document frequency (TF-IDF). CANTINA was developed by the University of California, Santa Barbara. CANTINA analyses the content of the page in order to determine whether the website in question is a phishing site. The recommended model by CANTINA included a total of eight features.[7]

The features are described as follows:

a) Age of Domain: With the use of this heuristic, one may assess whether a domain name is older than a year. The phishing site has a lifetime of 4.5 days when it first goes up. However, the heuristic does not consider phishing sites that are based on legitimate websites but where criminals have gained unauthorized access to the web server. It also does not consider phishing

sites that are hosted on otherwise legitimate domains, such as space provided by an Internet service provider (ISP) for personal homepages.[7]

b) Suspicious URL: In this heuristic, you should check to see whether the URL of the page includes the symbol '@' or '-'. The presence of the '@' symbol in the URL indicates that the text on its left side may be ignored; instead, you should only examine the right portion 59 of the string immediately after the sign. It is unusual to see a dash ('-') indicator on reputable websites.[7]

c) Suspicious Links: The purpose of this heuristic is to identify whether the links on the page fit the criterion stated before. If it fulfils the requirements, it will be marked as a potentially dangerous link.[7]

d) IP Address: It will figure out whether the URL that you provide uses an IP address as its domain.[7]

2.4.2 Heuristic-based Approach

The second tactic is known as heuristic techniques, and it involves collecting information from websites in order to determine whether they are authentic or phishing sites. In contrast to the blocklist method, the heuristic system can detect phishing websites as they are being constructed in real time. The success of heuristic methods is contingent on the selection of a set of discriminatory criteria that can assist in identifying the various kinds of websites. The heuristic approach makes use of a Hypertext Markup Language (HTML) or URL signature to identify websites that are engaged in phishing. Several studies are currently being carried out by utilizing this method.[8]

Spoof Guard is one of the solutions that uses a heuristic-based approach. It is a phishing-prevention add-on for your browser. For the purpose of determining the spoof value, this approach includes the assessment of stateless pages, evaluations of entire pages, and study of outgoing post-data. If the spoof index is found to be higher than a previously determined threshold value, the page in question is identified as phishing, and the user is given an appropriate warning.[8]

2.4.2.1 Machine Learning Approach

In addition to the approaches that have been described above, there is also a body of written material that evaluates the efficiency of machine learning and data mining algorithms. In terms of the accuracy of the predictions they make, Logistic Regression (LR), Bayesian Additive Regression Trees (BART), Classification and Regression Trees (CART), Random Forests (RF), and Neural Networks are compared (NN). In the experiments that compared the two systems, a dataset consisting of 1171 phishing emails and 1718 real emails was used. The process of training and testing the classifiers involved the utilization of a total of 43 functions. Experiments show that RF has the lowest error rate, which is 7.72%. This is followed by CART, which has an error rate of 08.13%, LR, which has an error rate of 08.58%, BART, which has an error rate of 09.69%, Support Vector Machines (SVM), which has an error rate of 09.90%, and NN, which has an error rate of 10.73%. The findings, on the other hand, indicate that there is no best classifier that can be applied to the task of predicting phishing sites.[9]

The effectiveness of machine learning-based detection methods (MLBDMs) such Bagging, AdaBoost, SVM, CART, NN, RF, LR, and Naive Bayes (NB), as well as BART, is analyzed and compared. A dataset consisting of 1,500 phishing websites and 1,500 trustworthy websites was used in the research activities. The evaluation performed by CANTINA is based on a total of eight factors.[9]

Before starting their experiments, a set of decisions made by authors as follows:

i. The Random Forest algorithm has been programmed to have a total of 300 trees.

ii. For all experiments that needed to be analyzed iteratively, the number of iterations was set to 500.

iii. The threshold value was set to 0 for some machine learning techniques, such as BART.

iv. The radial-based function was used in the support vector machine.

v. In the trials with the neural network, we used a value of 5 for the number of hidden neurons.

The results of the studies shown that seven MLBDMs, including AdaBoost, Bagging, LR, RF, NN, and NB, are more accurate than CANTINA.

A comparison of the accuracy of predictions made by several different machine learning algorithms, such as LR, CART, BART, SVM, RF, and NN. For the purpose of the comparative tests, a dataset consisting of 1171 phishing emails and 1718 real emails was used. A total of 43 attributes were used throughout the process of learning and testing the classifiers. The results of the trials show that RF has the lowest error rate, which is 7.72%, followed by CART, which has a rate of 08.13%, LR, which has a rate of 08.58%, BART, which has a rate of 09.69%, SVM, which has a rate of 09.90%, and NN, which has a rate of 10.73%. The data, on the other hand, indicate that there is no perfect classifier that can be used to forecast phishing websites. For instance, the FP rate for NN is 5.85%, however the FN rate is 21.72%. On the other hand, the FP rate for RF is 8.29%, although the FN rate is 11.12%. This demonstrates that NN outperforms RF in terms of FN, whilst RF exceeds NN in terms of FP. [10]

2.4.3 Blacklist-based Approach

The denylist strategy keeps a list of suspicious or harmful URLs gathered by various methods such as Google safe browsing, Phish Tank, and user voting. When a web page is launched, the browser searches the denylist for it and notifies the user if it is discovered. Finally, the blocklist can be kept locally or on a server. Blacklists are frequently used to determine if a website is harmful or legitimate. However, while these algorithms have low false-positive rates, they cannot classify freshly created dangerous URLs.[11]

The drawback of using this tactic is that blacklists are often unable to cover all phishing websites since the addition of a freshly constructed phoney website requires some amount of time. It is possible that phishers just need a short amount of time to accomplish their goals, since the time it takes to create a malicious website and add it to the list is not that long. Because of this, the technique for detecting phishing should be very fast and should typically begin as soon as the phishing page is uploaded and the user starts to enter his credentials.[12]

If the process for updating the blacklist takes too long, phishers who target websites may execute attacks without fear of being added to the list of those to avoid. The frequency with which blocklists are updated varies, but one study found that between 47 and 83 percent of phishing URLs are listed on blocklists within approximately 12 hours after their first release. This was nearly 12 hours after the initial launch. The same research also found that having zero hours of protection against the most well-known toolbars on blocklists resulted in a TP rate that ranged from 15% to 40%. Because of this, an effective blocklist must be kept up to date on a consistent basis in order to protect customers from phishing.[12]

Whenever a user browses the Internet, a little software programme known as Netcraft is launched automatically. The foundation of Netcraft is a blocklist of fraudulent websites that have been identified by Netcraft, as well as the URLs that have been provided by users and validated by Netcraft. This information is particularly helpful for visitors who are already acquainted with the web page's host server. Netcraft provides the location of the server.[13] The following are the primary factors that are considered by

Netcraft when determining the danger level of each site and whether to include iton the blocklist:

i. How long has the domain name been registered under which the website is housed.
ii. Domain names are not a part of the information included inside the Netcraft database.
iii. The fact that the same domain has been used to host phishing websites in the past.
iv. Replace any spaces in the URL with IP addresses or hostnames.
v. The history of the country and the role that the country's Internet service provider plays in the hosting of phishing websites
vi. An overview of the development of top-level domains for phishing websites.
vii. How well-known is the website among the community of people who use Netcraft Toolbar. [13]

The most significant drawback associated with Netcraft is the fact that the Netcraft server, and not the user's own computer, is the one that ultimately decides whether a website should be trusted. Consequently, if the user's connection to the server is severed for any reason, they will be subject to danger and exposed during this time.[13]

2.4.4 Comparison of Malware Detection Approaches

Types	Feature	Limitation
Content-based Approach	Determines whether the website is a phishing attempt by analyzing its content.	The extraction of keywords is disabled.
Heuristic-based Approach	Gather some information from the website to assess if it is phishing or not.	It is dependent on the selection of a collection of distinguishing traits that might aid in
Blacklist-based Approach	• A blocklist is a collection of malicious URLs. • When you visit a website, your browser checks the blocklist to determine whether the current URL is on it.	Because it takes such a long time for a recently created phoney website to be included to a blocklist, blocklists are often unable to cover all phishing websites.

Table 1 Comparison Phishing Detection Approaches

2.5 Conclusion

This chapter can be considered one of the most crucial aspects of this study. This chapter compares earlier methods suggested by other researchers and the present approach, Machine Learning. Furthermore, this chapter demonstrates the many processes that assisted the researcher in obtaining their proposed strategy for detecting phishing. There have been several strategies presented in recent years. However, to get a better outcome in the future, those procedures must be improved. As a result, the present approach is given in Chapter 3 to assist website users in identifying phishing on their websites.

CHAPTER 3

METHODOLOGY

In order to discuss the technique that was used for this research, this chapter is broken up into seven pieces. The first steps of the approach are outlined in Section 3.1. The research methodology, which may be a Software Development Life Cycle (SDLC) or some other one entirely, is discussed in Section 3.2. In Section 3.3, the need of developing and analyzing detection information for phishing websites is emphasized. Thedeveloping structure of the study is shown in Section 3.4. The conceptualization and execution of the study are both discussed in Section 3.5. The hardware and software that are used in the detection of phishing websites are detailed in Section 3.6, and the research, testing, and assessment processes are discussed in Section 3.7.

3.1 Introduction

The concept of phishing and its restriction capabilities were described in the previous chapter. Several previous studies that have been proposed to identify phishing have already been reviewed in Chapter 2. As a result, the specifics concerning the technique, strategy, and features employed during this study, as well as the methodology used in experimenting, will be presented in this chapter.

3.2 Research Methodology

This research approach is divided into four major phases: literature review, creation of a new framework, design, and execution, and testing and assessment. Because the phases may be evaluated and modified to achieve the best findings, this technique isappropriate and adaptable for this research project. This research technique varies from other system development lifecycles in that it focuses on thoroughly investigating the research issue.[14]

The review of the literature is the initial stage of this research process. At this point, existing papers on the study topic will be thoroughly evaluated and assessed. The aims and problem statements are then used to characterize the research definitions. The development of the frame is the following phase. The critical examination of prior studies will be considered in selecting an appropriate approach and algorithm to be employed in this research throughout this phase. The next step of research design and implementation will take place now that the framework of the research project has been defined. Language, hardware, and software requirements are provided during this step to set up the research experiment. The real implementation of the research project is applied to develop the detection model after the research needs are designed and prepared. Once the research experiment has been finished, it is tested and reviewed to establish the limitations of the research and the changes that may be made in future research.[14]

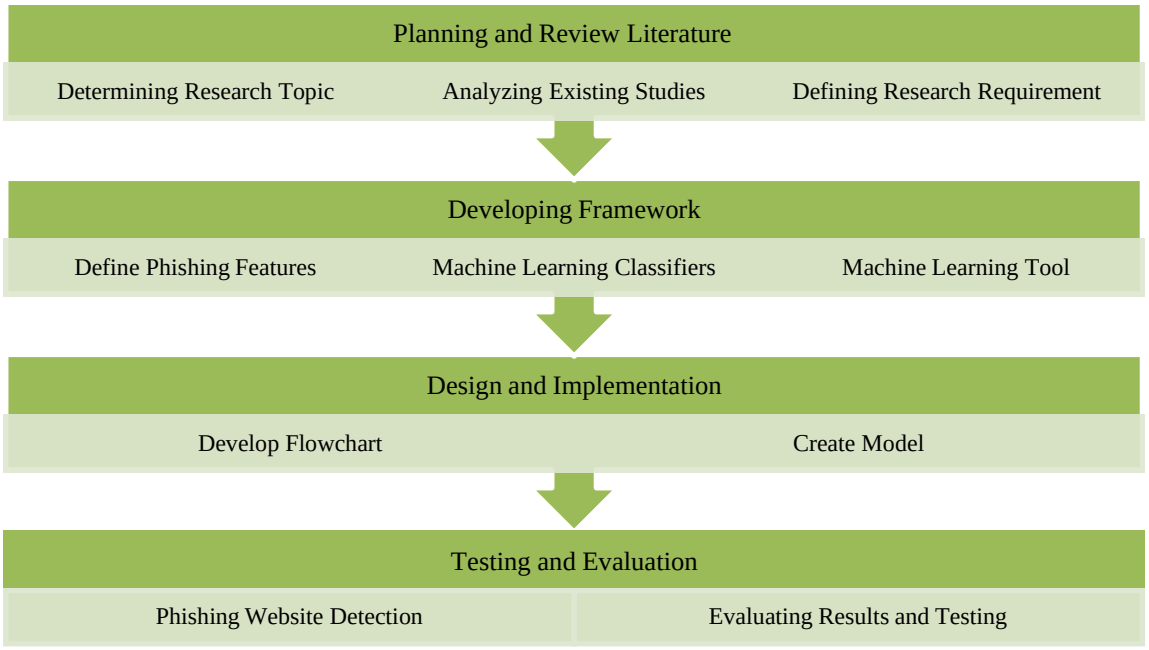

Figure 5 Main Stages for Research Methodology

Based on Figure 5 above, this research approach is being used in this project since it can be reverted to prior stages with low loss to apply fresh research advancements. Not only that, but this process enables adjustments to any step to be made to tackle problems that arise during the current stage. Finally, this research approach allows researchers to readily adjust to the study topic's demands.[14]

3.3 Planning and Reviewing Literature

The fundamental part of the research technique is study planning and a literature evaluation on the research subject. Before evaluating current studies, conceptualization is conducted to define the suitable study issue. When a research topic is chosen, related journals, publications, and studies are gathered to be studied. Existing studies might help in comprehending the research issue. This helps to specify the problem statement, the purpose, and the scope of the study.[14]

The materials were gathered from Internet journals, past student references, and online e-books. Existing scheme studies are thoroughly examined and filtered based on the topic's relevancy. The information gathered should be study-relevant to be used in developing this research.[14]

Based on the information acquired, many ways and techniques are learned to determine which approach and method are ideal for addressing the problem of the internet application, particularly phishing. Because their security issue is the primary worry in the website's application which focus on phishing detection (PD) for the website in this research. Existing PD research papers are thoroughly reviewed and categorized based on the location of the phishing code mechanism. Each suggested PD system is evaluated to discover its strengths and weaknesses. This data is critical in determining the methods followed by the researchers to conduct their experimental testing. As a result, in this study, research restrictions will be eliminated.[14]

3.4 Developing Framework

Considering our research into existing phishing code detection schemes, we decided to create a phishing code detection method that identifies characteristics using a machine learning technique. Figure 6 depicts the evolution of a Public Works Department (PWD) framework.

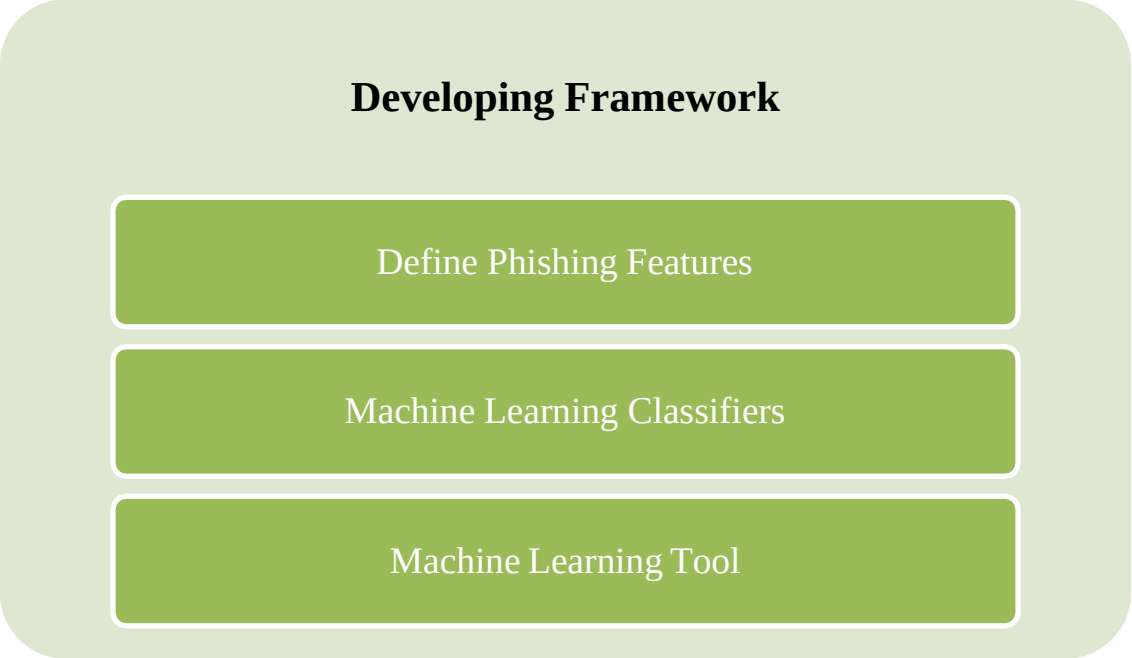

Figure 6 Development of PWD Framework

3.4.1 Define Phishing Features

In both academic literature and commercial applications, phishing detection may be accomplished using a variety of methodologies and data types. Phishing URLs and pages have several distinguishing traits that set them apart from malicious URLs. An adversary, for instance, might register long domain names that are difficult to decipherin order to hide the actual domain name.

In this investigation, our primary emphasis will be on traits based on URLs. When determining whether to phish a website, the very first thing that should be looked at is the URL. Certain distinguishing traits may be found in phishing domain

URLs. The URL is analyzed in order to get characteristics that are associated with these points. The following URL-based features are going to be studied as part of this project:

i. Address Bar-based Functionality

ii. Abnormal Base Characteristics

iii. HTML and JavaScript-based Functionality

iv. Domain-specific Functions

3.4.1.1 Dataset Description

The dataset gathering is the first step in the implementation process, crucial for accurate results. It provides insights into phishing and legitimate activities. After gathering, the dataset is analyzed for further research and used to anticipate phishing events. The dataset consists of 48 characteristics from 5000 authentic websites and 5000 fraudulent webpages. Unlike regular expression-based parsing, the browser automation framework offers a more accurate and reliable method for extracting features. The categorical values in the dataset are "Legitimate" and "Suspicious," which have been converted to numerical values. "Legitimate" is represented as "1" and "Suspicious" as "0." Additionally, the collected dataset contains other value [15]

In this research, the Correlation Attribute Evaluation technique was employed to assess the value of each characteristic by analyzing its relationship with the class. This technique is detailed in a referenced paper. By utilizing this approach, not only were the characteristics ranked in order of importance, but their respective rank numbers were also provided. Several characteristics attained the highest ranking due to their frequent utilization in the detection process.[16]

3.4.2 Machine Learning Classifiers

Machine learning is a kind of artificial intelligence (AI) that can acquire knowledge without being specifically programmed. It can predict future judgements and improve current ones when exposed to new evidence. The approach for making predictions is based on a search for patterns within the data set. This is referred to as learning. Classifier types influence both the learning process and the results of prediction. This method was often used to classify samples, particularly in intrusion detection systems (phishing and normal). Supervised machine learning and unsupervised machine learning are the most prevalent types. The approach for making predictions is based on a search for patterns within the data set. This is also referred to as learning. [17]

The different kinds of classifiers have an impact on both the learning method and the results of the predictions. This method was used rather extensively for classifying samples, particularly in the context of intrusion detection systems (malware and normal). Both supervised and unsupervised versions of machine learning are now in use.[17]

This research makes use of the supervised machine learning approach since the sample data set contains labels (phishing and normal). In addition, supervised machine learning produces useful results by minimizing the number of errors produced. This research makes use of five different classifiers in order to compare the conclusions obtained by using a variety of machine learning classifiers. Random Forest (RF), J48, Naïve Bayes, Logistics and K-Nearest Neighbors are the four classifiers that have been used (KNN).[17] They are as follows:

A common method of collective learning, known as Random Forest (RF), Random Forest may be used for supervised classification or regression. This method of machine learning is effective because it involves training a random number of decision trees and then determining the class mode (classification) or mean prediction of each individual tree (regression).[17]

3.4.3 Machine Learning Tool

Data analysis solutions that use machine learning provide functionality that automates the creation of the analysis model. This paradigm enables a system to learn from previous or ongoing data gathering by allowing predictions or judgements to be made while the process of learning is in progress. The analysis process may be made more straightforward and efficient by using a technology that is based on machine learning. Additionally, it may answer problems by automatically performing complicated mathematical computations, which does not need any machine learning approaches or prior expertise on the part of the user. Google Colab was chosen as the platform for the machine learning portion of this research.[18]

3.4.3.1 Google Colab

Google Colab is used for data set training because of its adaptability and cloud capabilities. It is useful in the context of machine learning using python. Optimizing performance by distributing GPU assets from Google servers to otherwise constrained hardware on the programmer end is critical to the memory-hogging machine learning algorithm. Google Storage provides a cloud drive infrastructure for storing this data set, which is then loaded and trained using the Colab online notebook. The trained model is then put into the Pi and validated using the collected data.[18]

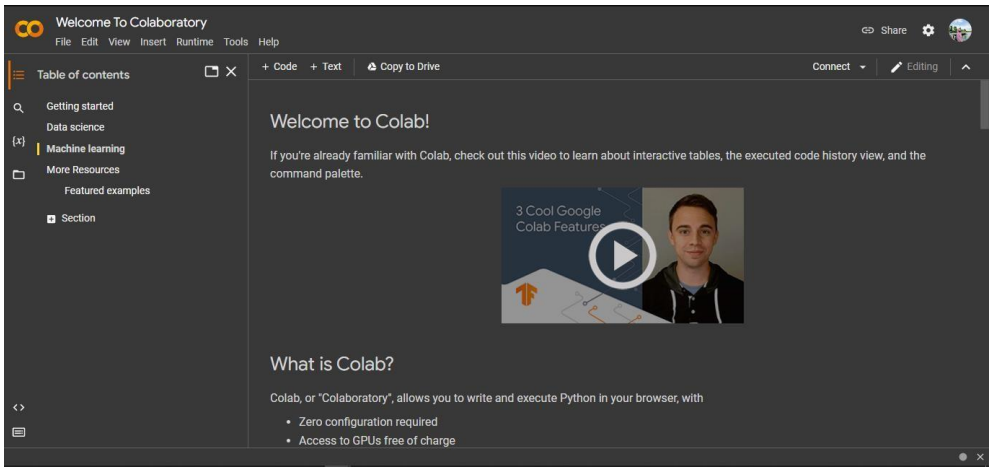

Figure 7 Google Colab before logging in

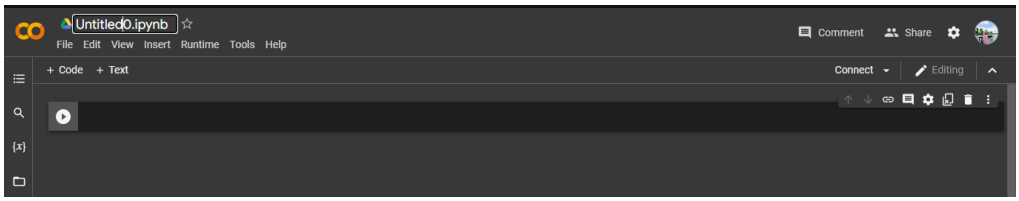

Figure 8 Google Colab before logging in

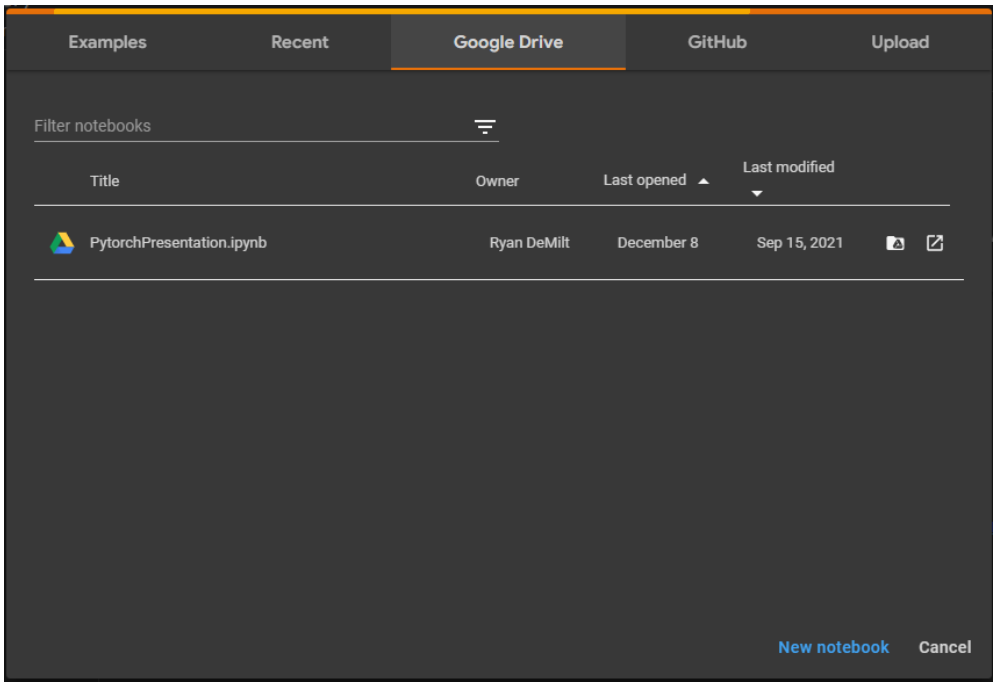

Figure 9 Google Colab blank code cell

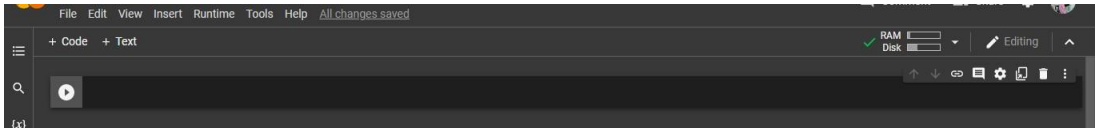

Figure 10 Google Colab executed code cell

Figures 7, 8, 9, and 10 [19] illustrate the Google Colab with a decent interface that even non-technologists may utilize. The program produces high-quality results, is simple to use, and requires only a Google account to get started, like any other Google product.[20]

3.5 Design and Implementation

After the framework has been developed, we will need to show whether the recommended framework is acceptable or unwanted. As a direct consequence of this, prior to the installation of the system, we developed a method to evaluate the precision of the abnormality detection strategy. Figure 11 depicts the design strategy that was developed to evaluate the idea prior to moving on to phishing website detection (PWD), which stands for "phishing website detection."

Collecting data, identifying key factors, testing the model, and finally comparing the findings are the four components that make up the design model. Each component will get a cursory examination in the next subtopic.

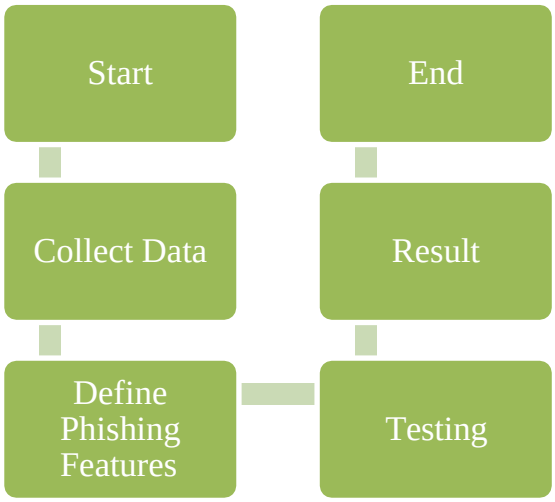

Figure 11 Procedures for Improving Detection Method

The implementation phase of the project is the very final stage in its progression. During this step, the design model will serve as a guide for putting the offered solution into action. The first thing that must be done in this phase is to install the necessary software for the project onto a personal computer or laptop, such as Google Colab. After then, the data set is gathered via the use of the internet or from persons who are ready to volunteer their data so that the project may be finished. After then, the project is carried out in accordance with the flow that was developed throughout the process of designing the project.

3.6 Hardware and Software

It is essential to explain all criteria required during the development of the project in order to successfully complete the project. In order to carry out the study experiment, we need to determine the software and hardware components that will be required to set up the inquiry. Because software and hardware are utilized to carry outthe research experiment and then test and analyses the results in the next phase, this phaseis an essential part of the research process.

3.6.1 Hardware Requirement

HARDWARE	PURPOSE
- **One unit of laptop** - **Processor: Intel(R) Core (TM) i3-7100U CPU @2.40GHz** - **Ram: 12 GB** - **System Type: 64-bit operating system, x64-based processor**	Utilized throughout the whole of the study project, including but not limited to the examination of available resources, implementation, testing, and documentation.

Table 2 Hardware Requirements and Purpose

3.6.2 Software Requirement

SOFTWARE	PURPOSE
Windows 10	The operating system used in this study
Microsoft Word	For documentation of this project
Microsoft Excel	To store the dataset/database
Google Chrome	To design a Gantt chart
Project Plan 365	To design a Gantt chart
Google Colab	To analyze and optimize the dataset using python

Table 3 Software Requirements and Purpose

3.7 Testing and Evaluation

The conduct of this study will end when this phase of testing and evaluation has been completed. The experiment will be assessed when each component has been included into the whole. In order to address the problem statement and assess whether the limitation of current journals is avoided, testing and evaluation are carried out. The primary objective of this examination is to demonstrate the mosteffective detection model that has been suggested in order to validate the validity of the outcomes and assertions made in this investigation. In addition, the testing and evaluation process enables the research experiment to identify flaws and restrictions, which in turn makes it possible for further adjustments to achieve the desired conclusion.

The research project, which was finally completed, served as a summary of the whole research process. In order to determine whether the goals have been reached, the results are also discussed and recorded. The next chapter will provide an in-depth look at the implementation step's explanation in more detail.

3.8 Conclusion

This chapter may be summed up by declaring that it is one of the topics that aided the researcher in deciding which model to apply to the inquiry. In addition, this chapter discusses the kind of approach and the tools that were used in order to accomplish the objectives of this thesis. This chapter also includes a comprehensive explanation of the methodology that will be used throughout the remainder of the research project. The researcher will require important instruments such as hardware and software to assist in the phishing website detection in order to accomplish the goal of this study and make it a success. The installation, testing, and evaluation processes will be covered in the next chapter.

CHAPTER 4

IMPLEMENTATION, RESULTS AND DISCUSSION

4.1 Introduction

The methodology, planning, analysis, and design that were prepared and drafted in Chapter 3 will be implemented in this chapter. The implementation step is important throughout the whole tool development process. This is since this step will depend on the process of identifying phishing websites applying technologies.

4.2 Dataset Description

The gathering of datasets is the initial step in the implementation process. In order to ensure that the results are accurate, the dataset phase is very important. The dataset will provide additional insight and explanation regarding phishing as well as legitimate activities. Following this step, the dataset is analyzed for additional research, and the findings are used to anticipate or forecast the events that will occur in phishing.

This dataset includes 48 characteristics that were taken from 5000 authentic websites and 5000 fraudulent webpages. When compared to the method of parsing that is based on regular expressions, the use of the browser automation framework allows for the utilization of an enhanced strategy for the extraction of features. This method is both more accurate and more reliable. The categorical values in the collected dataset are "Legitimate," and "Suspicious". These values have been converted to numerical values by substituting the values "1," and "0" for "Legitimate," and "Suspicious" in the appropriate places. The collected dataset also contains other values.[15]

4.3 Machine Learning Approach

The machine learning approach ensures that web browsers can optimize the phishing characteristics via the use of an approach called feature optimization. This method reduces the amount of time required for training and testing, hence making the phishing detection system easier to use. Methods of feature selection were used in order to locate and eradicate from the data any features that were deemed unnecessary or redundant and did not add to the precision of a prediction model.[21]

Following training, the elements of the phishing website were then categorized based on the key traits they had. The feature selection technique has been used in this research project in order to determine which characteristics are essential for reliable phishing website identification. There are various aspects that are utilized to ensuring that there is a distinct pattern emerging between the regular websites and the phishing websites. In Table 4, present a list of the characteristics of phishing websites that were evaluated in the research.

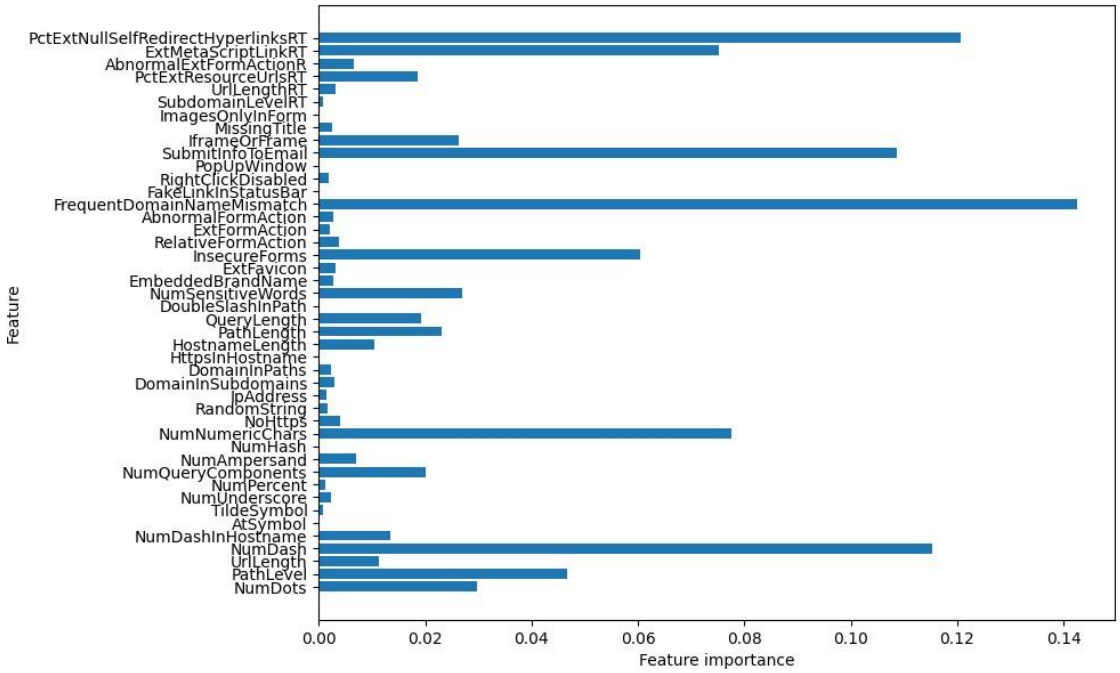

Figure 12 Features Ranking

Phishing Features	Description
Ip Address	An IP address is a unique identifier for a device connected to the Internet or a private network.
ExtMetaScriptLinkRT	Since META and SCRIPT are components of the HEAD tag, they are not allowed to appear anywhere else in the page than the header section: Additional tags used inside the META> header of the document This tag is used in a variety of contexts.
InsecureForms	It is mixed forms which are forms on HTTPS sites that do not submit on HTTPS, provide a security and privacy risk to users. The information supplied on these forms is exposed to eavesdroppers, enabling malevolent parties to see or modify sensitive data.
NumDots	This function verifies the number of dots in the hostname portion of a URL. Generally, a valid URL has two dots in the domain name, excluding 'www.' Using multiple dots in URLs, phishers add additional subdomains and the domain name of the original website as a subdomain to deceive users.[22]
NumSensitiveWords	Existence of a sensitive term 'Login,' 'Update,' 'Validate,' 'Activate,' 'Secure,' etc. are the tokens or words most typically used in phishing URLs. The use of these terms in a URL to project urgency and persuade visitors to immediately visit a phishing site to obtain sensitive information.[22]
UrlLength	The number of distinct characters that constitute a URL is what is referred to as its length. The programme being useddetermines the maximum number of characters that may

	be included in a URL. If the length of the URL is 2083 characters or less, it may be certain that it is secure.[23]
SubdomainLevel	Subdomains are extensions of the primary domain name. Subdomains assist organize and browse the main website. The main domain may have as many subdomains as needed to access all the website's pages.[24]
HostnameLength	Hostnames consist of a string of labels joined together with dots. "en.wikipedia.org" is an example of a hostname. Each label must include between 1 and 63 characters. The whole of the hostname, including the delimiter dots, may include no more than 253 American Standard Code for Information (ASCII) characters. [25]
NumDashInHostname	It is not possible to have several dashes inside a hostname at the same time. There are not allowed to be any spaces in a hostname, nor may it begin with a dash.[26]
NumUnderscore	An underscore is a punctuation mark that resembles a hyphen that has been stretched out. It is also often known as a low dash. In most cases, you will see underscores in things like domain names and email addresses.[16]

Table 4 List of Phishing Website Features Used

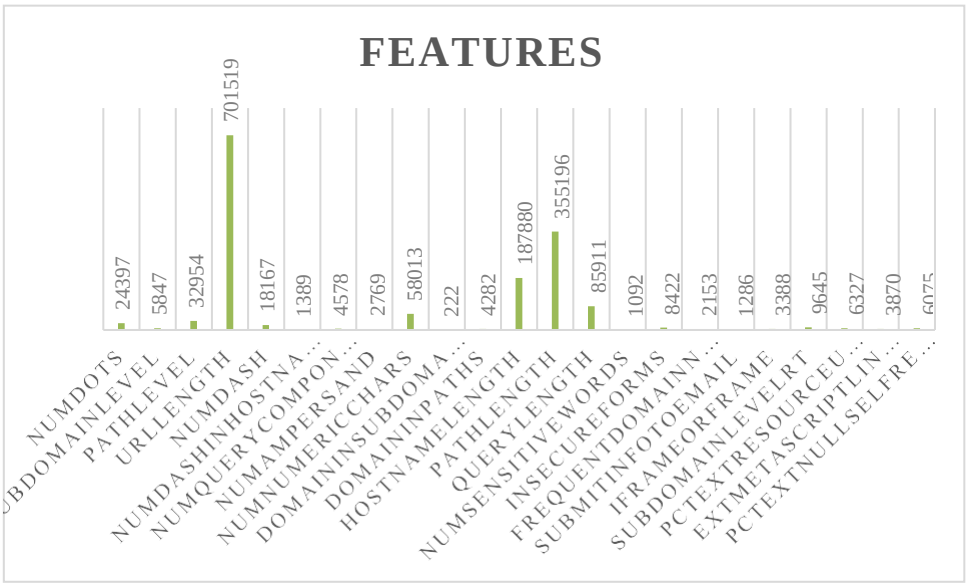

Figure 13 Chosen Features

This research used a technique known as Correlation Attribute Evaluation, which involves determining the value of a characteristic by analyzing the degree to which the attribute relates to the class. The approach may be found in this paper. Not only does it offer a ranking of the characteristics from best to worst, but it also shows the rank number for each quality.[16] According to Figure 13, some of the characteristics have the greatest ranking since they are used the most often throughout the detection process.

4.4 Evaluation and Results

The outcomes present the results attained using three different machine learning classifiers, namely random forest, J48, Naïve Bayes, KNN and Multilayer Perceptron. In addition, this investigation into the various measures made use of the metrics of accuracy, precision, and recall using python. In Table 5, the findings obtained from 25 phishing website features of the testing set that used five chosen classifiers are shown.

Classifiers	Accuracy	Precision	Recall	FPR	TPR
Random Forest	94.10%	0.978	0.904	0.021	0.904
J48	92.10%	0.917	0.926	0.084	0.926
Naïve Bayes	83.00%	0.921	0.771	0.071	0.771
KNN	92.21%	0.923	0.921	0.079	0.921
Logistic	89.50%	0.895	0.895	0.105	0.895

Table 5 Performance of Each Classifiers

According to the results, Random Forest classifiers had the greatest accuracy result, which was 94.10% percent, in comparison to Naïve Bayes classifiers, which only reached 83.00% percent accuracy. This result demonstrates that the Random Forest classifiers are more successful in identifying phishing websites than other chosen classifiers. It also demonstrates that the selection of the features to be used in the phishing website detection process is an extremely important one. The fact that the classifier provided more relevant results indicates that it also delivered accurate findings at a high accuracy rate.

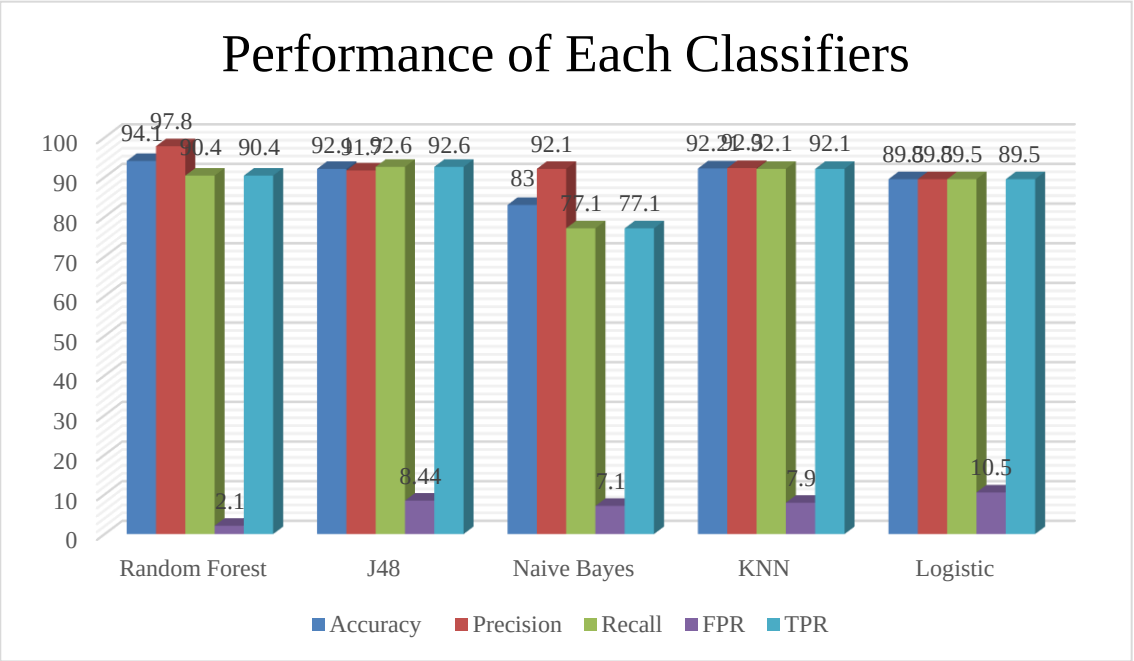

Figure 14 Percentage Accuracy

The percentage of correct classifications made by each of the five classifiers is shown in the Figure 14 that can be seen above. In comparison to other classifiers, the Random Forest classifier has the greatest percentage of accuracy, which comes in at 94.1%. The KNN classifier comes in second place with a score of 92.21%, third places are J48 with 92.1%, forth places are Logistic with 89.50% and Naïve Bayes in last place with 83.0%.

4.4.1 Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification model. The table shows two possible classes of prediction which are normal and phishing. For example, if a model predicts the presence of phishing activities, the result will show “phishing” and vice versa. Table 6 shows the performance of the five classifiers.

Classifiers	Actual	Predicted	
		Predicted Normal	Predicted Phishing
Random Forest	Actual Normal	967	21
	Actual Phishing	97	915
J48	Actual Normal	2278	210
	Actual Phishing	185	2327
Naïve Bayes	Actual Normal	1502	98
	Actual Phishing	463	1237
KNN	Actual Normal	2204	273
	Actual Phishing	130	2393
Logistic	Actual Normal	2191	286
	Actual Phishing	240	2283

Table 6 Confusion Matrix

The table above shows that the study produced corrected and magnificent results by predicting the unknown phishing with 2278 for the J48 classifiers. In the incorrectly predicted perspective, the Random Forest shows the most minimal value. Hence, the outcomes shows that J48 classifiers able to predict unknown phishing more accurately.

4.4.2 Receiver operating characteristics curve (ROC)

In this study, based on the phishing website features, the processes were classified as normal and phishing. Aside from using performance matrix, this study also calculated the receiver operating characteristics (ROC) curve for each of the machine learning classifiers. In this phase, the TPR was regarded as the detection rate which will correctly predict the phishing process and the FPR was selected as the detection rate which incorrectly predicted normal as phishing.

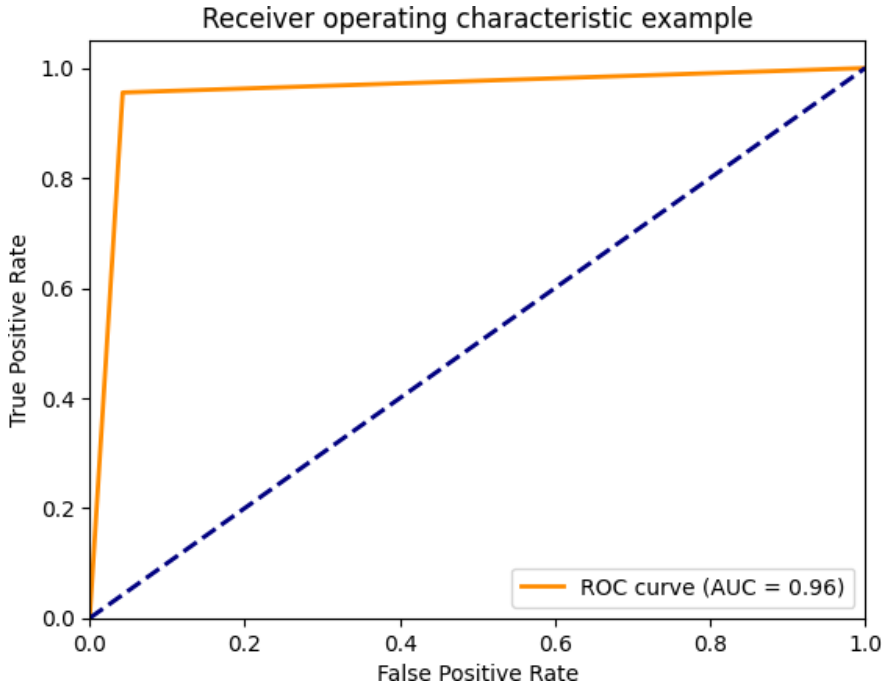

Figure 15 Receiver operating characteristics curve (ROC)

The horizontal axis in the above figure indicates the false positive rate meanwhile the vertical axis indicates the true positive rate. orange lines represent the individual ROC curve of the machine learning classifiers. The AUC results identified were able to measure whether the detection approach was good or bad. Table 7 shows the AUC performance.

Classifiers	AUC	Indicator
Random Forest	0.96	Perfect Prediction
J48	0.92	Perfect Prediction
Naïve Bayes	0.84	Perfect Prediction
KNN	0.95	Perfect Prediction
Logistic	0.91	Perfect Prediction

Table 7 AUC Performance

Table 7 shows that the random forest and Random Forest classifiers provide the best AUC values, with over 0.95. This signifies perfect prediction. Overall, the ROC and the AUC values confirmed that the most recent phishing experiments had provided compelling accurate results in the phishing website applications detection.

4.4.3 Threshold

The optimal threshold is the value that best separates the two detections that are related to the phishing and normal features. The threshold value is used to investigate whether the presence of behavior pattern indicator is normal (0) or phishing (1). The threshold value for random forest, J48, KNN, Naïve Bayes and Logistic are given in Table 8. As the threshold values were obtained based on the real behavior patterns of the normal and phishing applications, it can be said that the approach used in this study was able to detect phishing with more than 80 percent accuracy rate.

Classifiers	Accuracy	Threshold
Random Forest	0.960	0.941
J48	0.921	0.921
Naïve Bayes	0.848	0.830
KNN	0.921	0.901
Logistic	0.895	0.872

Table 8 Optimal Threshold

4.4.3 Robustness

Apart from evaluating effectiveness of the approach, the robustness of the approach for producing more dependable results were also tested. Robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset. In the other words, the robust algorithm is the one, the testing error of which is close to the training error. Table 9 shows the result of the classifiers' performance.

Classifiers	Accuracy	Precision	Recall	FPR	TPR	ROC
Random Forest	94.10%	97.8	90.4	0.021	90.4	96.0
J48	92.10%	91.7	92.6	0.084	92.6	92.0
Naïve Bayes	83.00%	92.1	77.1	0.071	77.1	84.0
KNN	92.21%	92.3	92.1	0.079	92.1	95.6
Logistic	87.20%	89.50	89.50	0.105	89.50	91.1

Table 9 Performance Result

The table 9 shows that the approach applied in this study was able to detect unknown phishing with over 80 percent accuracy rate.

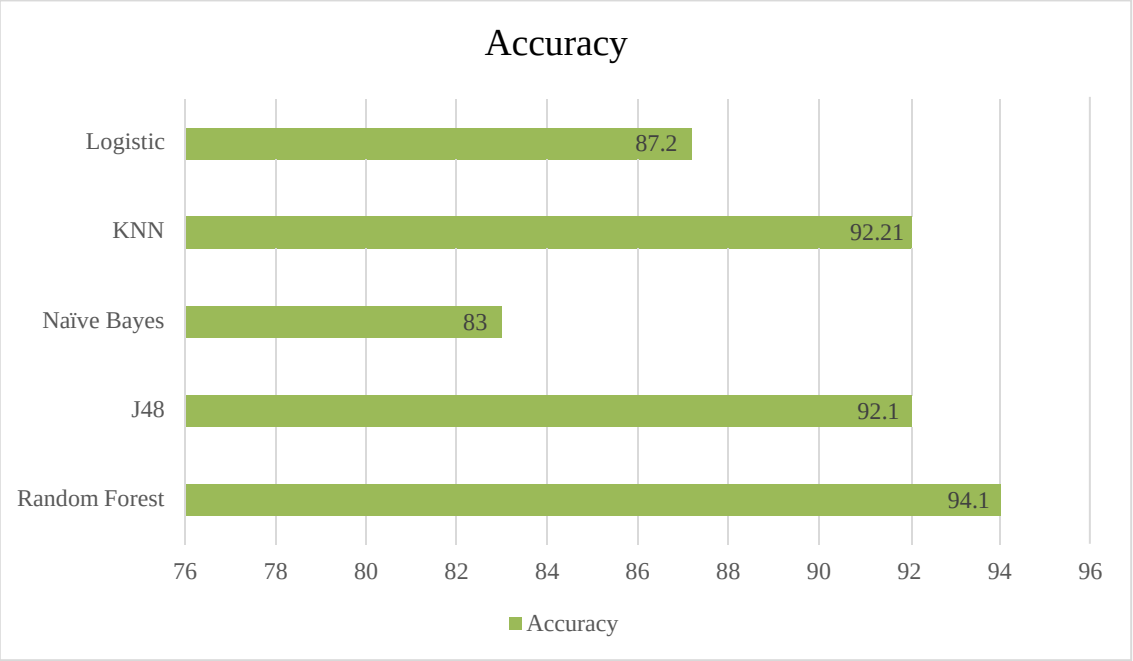

Figure 16 Percentages Accuracy

Figure 16 above shows percentage of accuracy of the detection based on the five classifiers. Random forest classifier shows the highest percentage of the accuracy by 94.1% compared to other classifiers. Second highest classifier is KNN with 92.21%, followed by J48 with 92.1%, then Logistic with 87.2% and the last is Naïve Bayes with 83%.

Classifiers	Accuracy	Source
Random Forest	94.10%	This research
	81.80%	[27]
J48	92.10%	This research
	73.90%	[27]
Naïve Bayes	83.00%	This research
	96.77%	[28]
K-Nearest Neighbor	92.21%	This research
	82.5%	[27]
Logistic	87.20%	This research
	90.32%	[28]

Table 10 The accuracy comparison of previous research papers

The accuracy results of the algorithms that were tested in this study are compared with the accuracy results of the algorithms that were tested in the previous research papers in Table 10. The recorded results are the highest results when compared to the results obtained by using other algorithms. According to the data shown in the table, the Naïve Bayes algorithm produced the most accurate results for the paper, with an accuracy rate of 96.77% when compared to both results. Then, when compared to the publications that came before it in the study, Random Forest achieved the greatest accuracy findings compared for J48, KNN, Naïve Bayes and Logistic.

Classifiers	Build Model
Random Forest	0.35
J48	0.15
Naïve Bayes	0.11
KNN	4.64
Logistic	0.27

Table 11 Time taken to build model (seconds)

The amount of time, in seconds, that it took to obtain the findings is shown in Table 11. According to the findings, Naïve Bayes has the simplest model complexity since it requires the least amount of time to construct the model.

4.5 Conclusion

From all the figures and tables in chapter 4, Random Forest algorithm produce the highest value for accuracy, precision, TPR, and ROC which is over 94%. J48 and KNN algorithms, also produce a consistence value for each test which is over 90% while the last are Naïve Baye and Logistics which produce below 90% for certain testing. From this observation, this thesis can conclude that, Random Forest is the best algorithm that can be used to detect the phishing attack.

CHAPTER 5

CONCLUSION

5.1 Introduction

Today, the internet has changed the way of life for humans. There is a wide range of activities from searching for information to entertainment, online shopping, financial services, and socializing. Frequent usage of the Internet makes people have come to trust the Internet to provide the gateway for office, home, and personal convenience.

Online transactions nowadays are becoming more relevant and provide the easiest and fastest way to manage and handle things. There is nothing impossible to be done quicker and simplest by having the Internet. Despite the advantages and benefits provided, it must be its disadvantages and that is security. Many people rarely realize these security issues which may bring harm to them.

This study provides an understanding of phishing. This study also aims to detect phishing websites by using machine learning. The dataset of phishing features is collected and they have been through a feature optimization approach. This approach makes the list of phishing features lesser and provides a smaller dataset. Then, it applies machine learning classifiers that are Random Forest, J48, Naïve Bayes, KNN, and Logistics. The parameters are taken account into to detect phishing websites effectively.

5.2 Research Objectives

The purpose of this study was to improve a phishing website detection system by using machine learning for website URL. Section 1.3 had described the three research objectives of this study.

Objective 1: To review the current phishing detection system issue.

The primary aim of the study was to thoroughly examine security vulnerabilities by conducting a comprehensive analysis of existing research on phishing website detection systems. This objective was accomplished through an extensive review of influential works published in reputable online scholarly journals. The findings of this review were presented in Chapter 2 of the study, which provided a detailed and in-depth overview of the phishing website detection system. Next, chapter 2 served as a valuable resource that synthesized the collective knowledge and insights from various studies in the field. It explored different aspects of the phishing website detection system, including its classification, the machine learning approaches employed, and the specific algorithms utilized for detecting phishing websites.

By delving into the classification of phishing website detection systems, the chapter shed light on the different methodologies and techniques employed to identify and combat phishing threats. It highlighted the importance of machine learning as a powerful approach in this domain and discussed the specific algorithms that have been utilized successfully in detecting phishing websites. Then, through the comprehensive review presented in Chapter 2, the study contributed to the existing body of knowledge by providing a cohesive and up-to-date understanding of phishing website detection systems. This information can serve as a foundation for further research and development in the field of cybersecurity, assisting in the creation of more robust and effective measures to combat phishing attacks.

Objective 2: To develop a phishing detection system that analyses website applications using a Machine Learning approach.

The second research objective aimed to assess the effectiveness of the phishing website detection system by employing a machine learning approach. The evaluation of the system was conducted using Python in Google Colab, which provided a suitable environment for implementing and executing the necessary experiments. In Chapter 4 of the study, a series of experiments were carefully designed and carried out to evaluate the system's performance. Then, the evaluation criteria utilized to measure the effectiveness of the system consisted of six key metrics: accuracy, False Positive Rate (FPR), True Positive Rate (TPR), precision, recall, and f-measure. These metrics were chosen because they provide a comprehensive understanding of the system's performance in terms of its ability to accurately detect and classify phishing websites.

Throughout Chapter 4, the experiments were conducted meticulously, following established methodologies and best practices in machine learning evaluation. The system was tested on a diverse set of data, comprising both legitimate and phishing websites, to ensure a realistic and representative evaluation. By analyzing the results obtained from the experiments using the defined evaluation criteria, the research successfully accomplished the objective of assessing the effectiveness of the phishing website detection system. The evaluation provided valuable insights into the system's performance, allowing for a robust assessment of its ability to accurately distinguish between legitimate and phishing websites. Overall, the second research objective was accomplished within Chapter 4, as the experiments were carried out, the evaluation criteria were measured, and the system's effectiveness in detecting phishing websites using a machine learning approach was thoroughly assessed.

Objective 3: To evaluate the proposed system in terms of phishing detection accuracy.

The third objective of the study focuses on evaluating the accuracy of the proposed system for detection. This evaluation involves testing the system's performance using five different classifiers: Random Forest, Logistic Regression, J48 (C4.5), Naive Bayes, and K-Nearest Neighbors (KNN). A dataset from Kaggle, consisting of examples of legitimate and phishing websites, was used for training, and evaluating the machine learning models. Before the evaluation, a data analysis step was performed to examine the correlation between the dataset's features. This analysis helped identify highly correlated features that might require preprocessing. The dataset was then divided into training and testing sets. The training set was used to train the machine learning models, while the testing set was used to assess their performance, ensuring an unbiased evaluation.

The five selected classifiers were trained on the training set and evaluated on the testing set. Performance was measured using various evaluation metrics, including accuracy, False Positive Rate (FPR), True Positive Rate (TPR), Precision, Recall, and Receiver Operating Characteristic (ROC) curve. These metrics provided insights into the models' ability to accurately detect phishing websites and distinguish them from legitimate ones. Moreover, the results from the evaluation showed that the random forest classifier achieved the highest accuracy in detecting phishing websites among the five tested classifiers. This indicates that the random forest algorithm outperformed the others in accurately identifying and classifying phishing threats. By incorporating these steps and evaluation metrics, the research provided a comprehensive assessment of the system's detection accuracy using machine learning classifiers. It highlighted the effectiveness of the random forest algorithm for detecting phishing websites.

5.3 Achievement of the study

The research commenced with an investigation into the evolution of phishing and an exploration of different types of phishing website detection systems. It thoroughly examined the challenges associated with detecting phishing websites and carefully considered the selection of pertinent features. Multiple machine learning classifiers were assessed, and their performance results were gathered for evaluation. In line with the study's objective, the obtained results were analyzed and various points of interest were identified, as outlined below.

5.3.1 A detection model for phishing

In this study, a model has been developed to detect phishing websites through static analysis. A machine learning approach was employed to create an adaptive detection model. The developed model exhibited strong performance in accurately identifying phishing websites using the provided dataset.

5.3.2 Issues in phishing website detection studies

Chapter 2 of this study provided an in-depth analysis of the various types of phishing website detection methods and their significance in combating phishing threats. By examining the strengths and weaknesses of these approaches, several strategies were identified to address their limitations. To enhance the efficiency of the phishing website detection system, extensive research was conducted to identify relevant features that could contribute to a more effective approach. The primary objective was to develop a more efficient and robust methodology for detecting phishing websites.

5.3.3 Issues in phishing website feature selection

This study has conducted a comprehensive analysis of various perspectives employed to tackle the significant challenges associated with feature selection. The primary objective was to enhance detection performance while minimizing complexity. By critically examining these perspectives, valuable insights were gained to inform the development of improved feature selection techniques.

5.4 Research Constraints

The discussions presented in the preceding chapters have effectively verified that this research has successfully accomplished its intended aims and objectives. Nonetheless, it is important to acknowledge and address the constraints and obstacles encountered during the study, which are outlined here for future reference and consideration.

5.4.1 Sample size

The utilization of a small sample size in this study posed challenges in identifying significant relationships within the data. It is important to acknowledge that the number of analytical samples employed has influenced the research, as statistical tests typically necessitate a larger sample size to ensure a representative distribution of the population.

5.4.2 The assessment of the study was carried out using a static detection model only

In this study, the collection of all input features solely through static analysis was employed. However, it is worth noting that in practical solutions, both static and dynamic analyses possess their respective advantages and disadvantages. Therefore, a comparative analysis of the results obtained from both approaches would provide greater utility and insight.

5.4.3 Time

The research time is limited by the fixed task deadline, constraining thorough investigation and measurement of change or stability. Time constraints limit in-depth exploration and capturing long-term trends. Strategic optimization is crucial within the finite timeframe to make accurate observations. Balancing comprehensive analysis with temporal constraints necessitates meticulous planning and efficient resource allocation for valuable insights despite limited time.

5.5 Future works

The following recommendations for future work outside the scope of this study were listed as follows:

5.5.1 Selection of relevant features

The more complex and extensive data becomes, the harder it becomes to choose relevant and suitable features to improve detection performance. The process requires further analysis to investigate the correlation between malware and benign applications. This will reduce false alarms, thus increase the detection accuracy.

5.5.2 Enhance false alarm rate

False alarm rate remains a problem as long as it exists in the detection module. False alarms refer to the statistical measurement of how well the sample dataset classifies the phishing website correctly. This means that the phishing data was incorrectly predicted as normal. This problem leads to incorrect detection of websites and even small amounts of false alarms can cause enormous impacts. A reliable and efficient detection module is therefore needed to solve this problem.

5.5.3 Dynamic analysis approach

This study also can be done by using Dynamic Analysis Approach. It can identify vulnerabilities in a runtime environment. This approach recognizes vulnerabilities that could have been false negatives in static code analysis.

5.6 Conclusion

The internet has transformed human life, providing convenience for various activities. However, security issues, particularly phishing, pose a threat that is often overlooked. This study focuses on improving phishing website detection through machine learning. It optimizes a dataset of phishing features, applies machine learning classifiers, and achieves high accuracy using the random forest classifier. The study develops a detection model for phishing websites, addresses issues in phishing detection studies and feature selection, and identifies constraints such as sample size, reliance on static analysis, and time limitations. Future work recommendations include selecting relevant features, enhancing the false alarm rate, and exploring dynamic analysis approaches.

REFERENCES

[1] S. Hossain, D. Sarma, and R. J. Chakma, “Machine Learning-Based Phishing Attack Detection,” 2020. [Online]. Available: www.ijacsa.thesai.org

[2] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. E. Ulfath, and S. Hossain, “Phishing attacks detection using machine learning approach,” in *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 1173–1179. doi: 10.1109/ICSSIT48917.2020.9214225.

[3] “Phishing E-mail Reports and Phishing Site Trends 4 Brand-Domain Pairs Measurement 5 Brands & Legitimate Entities Hijacked by E-mail Phishing Attacks 6 Use of Domain Names for Phishing 7-9 Phishing and Identity Theft in Brazil 10-11 Most Targeted Industry Sectors 12 APWG Phishing Trends Report Contributors 13 4 th PHISHING ACTIVITY TRENDS REPORT,” 2022. [Online]. Available: http://www.apwg.org,

[4] M. Badra, S. El-Sawda, and I. Hajjeh, “Phishing attacks and solutions,” in *MobiMedia 2007 - Proceedings of the 3rd International Conference on Mobile Multimedia Communications*, Association for Computing Machinery, Inc, Aug. 2007. doi: 10.4108/icst.mobimedia2007.1899.

[5] A. Shankar, R. Shetty, and B. Nath, “A Review on Phishing Attacks,” 2019. [Online]. Available: http://www.ripublication.com

[6] Sri Eshwar College of Engineering and Institute of Electrical and Electronics Engineers, *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS).*

[7] Anjali Gupta, Juili Joshi, Khyati Thakker, and Chitra bhole, “Content based approach for Detection of Phishing Sites,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 2, no. 1, Apr. 2015.

[8] N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, “Client-side defense against web-based identity theft.” [Online]. Available: www.ebaymode.com

[9] D. Miyamoto, H. Hazeyama, and Y. Kadobayashi, “An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites.”

[10] Saeed Abu-Nimeh, D. Nappa, Xinlei Wang, and S. Nair, “A comparison of machine learning techniques for phishing detection,” *APWG Symposium on Electronic Research*, p. 1, Oct. 2007.

[11] Institute of Electrical and Electronics Engineers, *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS) : proceedings : Amman, Jordan, 9-11 October 2019.*

[12] S. Sheng *et al.*, “An Empirical Analysis of Phishing Blacklists QR code phishing View project A Privacy Assistant for IoT View project An Empirical Analysis of Phishing Blacklists,” 2009. [Online]. Available: https://www.researchgate.net/publication/228932769

[13] Netcraft Ltd, “Netcraft,” *Netcraft Ltd*, 1995.

[14] “Research Methodology Methods and Techniques (PDFDrive)”.

[15] Choon Lin Tan, “Phishing Dataset for Machine Learning: Feature Evaluation,” *Mendeley Data*, Mar. 24, 2018.

[16] Betha Nurina Sari, “CorrelationAttributeEval,” *ResearchGate*, Apr. 25, 2017.

[17] F. Vanhoenshoven, G. Napoles, R. Falcon, K. Vanhoof, and M. Koppen, “Detecting malicious URLs using machine learning techniques,” in *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, Institute of Electrical and Electronics Engineers Inc., Feb. 2017. doi: 10.1109/SSCI.2016.7850079.

[18] Institute of Electrical and Electronics Engineers, IEEE Communications Society;, Denshi Jōhō Tsūshin Gakkai (Japan). Tsūshin Sosaieti, and Han'guk T'ongsin Hakhoe, *ICUFN 2019: the 11th International Conference on Ubiquitous and Future Networks: July 2 (Tue.)-July 5 (Fri.) 2019, Zagreb, Croatia*.

[19] M. Kuroki, "Using Python and Google Colab to teach undergraduate microeconomic theory," *International Review of Economics Education*, vol. 38, p. 100225, Nov. 2021, doi: 10.1016/J.IREE.2021.100225.

[20] Prabanjan Raja, "What is Google Colab?," *Scaler Topics*, Feb. 11, 2022.

[21] T.-S. Chou and J. Pickard, "Machine Learning based IP Network Traffic Classification using Feature Significance Analysis." [Online]. Available: www.highperformancebuildings.gov,http://www.highperformancebuildings.gov,ecu-cet-fra-01-

[22] S. das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," *Annals of Data Science*, 2022, doi: 10.1007/s40745-022-00379-8.

[23] SISTRIX, "URL length: how long can a URL be?," *SISTRIX*, Feb. 19, 2021.

[24] KRISTIN CRABB, "What is a Subdomain?," *DOMAIN.COM BLOG*, 2023.

[25] Wikipedia, "ASCII," *Wikipedia*, Jan. 12, 2023.

[26] No-IP.com, "What is a Valid Hostname?," *No-IP.com*, 2020.

[27] N. Mallios, E. Papageorgiou, M. Samarinas, and K. Skriapas, "Comparison of machine learning techniques using the WEKA environment for prostate cancer therapy plan," in *Proceedings of the 2011 20th IEEE International Workshops on Enabling Technologies:*

Infrastructure for Collaborative Enterprises, WETICE 2011, 2011, pp. 151–155. doi: 10.1109/WETICE.2011.28.

[28] B. Espinoza, J. Simba, W. Fuertes, E. Benavides, R. Andrade, and T. Toulkeridis, “Phishing attack detection: A solution based on the typical machine learning modeling cycle,” in *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019,* Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 202–207. doi: 10.1109/CSCI49370.2019.00041.