

**DETECTION OF PHISHING WEBSITE**

**USING MACHINE LEARNING**

**BY**

**OYELERE TEMITAYO GIDEON**

**HND/23/COM/FT/0078**

**SUBMITTED TO THE DEPARTMENT OF COMPUTER  
SCIENCE,**

**INSTITUTE OF INFORMATION AND  
COMMUNICATION TECHNOLOGY,**

**KWARA STATE POLYTECHNIC, ILORIN**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS  
FOR THE AWARD OF HIGHER NATIONAL DIPLOMA  
(HND) IN COMPUTER SCIENCE**

**MAY, 2025**

## CERTIFICATION

This is to certify that this project is written by **OYELERE TEMITAYO GIDEON**, with matriculation number **HND/23/COM/FT/0078** in Computer Science Department, Institute of Information and Communication Technology, Kwara State Polytechnic, Ilorin.

-----  
**MR. BOLAJI - ADETORO, D.F**

*(Project Supervisor)*

-----  
Date

-----  
**MR. OYEDEPO F.S.**

*(Head of Department)*

-----  
Date

-----  
**External Supervisor**

-----  
Date

## **DEDICATION**

This research project is dedicated to the Almighty God that preserved me throughout the course of my programme at the Kwara State Polytechnic, Ilorin. for his infinite mercy that endures forever in my life; and to my beloved mom and siblings who have stood by my sides all the time.

## **ACKNOWLEDGEMENTS**

All praise is due to Almighty God the Lord of the universe. I praise him and thank him for giving me the strength and knowledge to complete my HND program and also for our continued existence on Earth.

I appreciate the utmost effort of my supervisor, MR. BOLAJI-ADETORO, D. F. whose patience, support, and encouragement have been the driving force behind the success of this research work. He gave useful corrections, constructive criticisms, comments, recommendations, and advice and always ensures that excellent research is done. My sincere gratitude goes to the Head of the Department MR. OYEDEPO F.S. , and all other members of staff of the Department of Computer Science, Kwara State Polytechnic, Ilorin, for their constant cooperation, constructive criticisms, and encouragement throughout the program.

Special gratitude to my parents, who exhibited immeasurable financial, patience, support, prayers, and understanding during the period in which I was busy tirelessly with my studies, special thanks go to my lovely siblings

My sincere appreciation goes to my friends and classmates.

## TABLE OF CONTENTS

Title page	i
Certification	ii
Dedication	iii
Acknowledgments	iv
Table of Contents	v
List of Figures	vi
List of Tables	vii
Abstract	viii
<b>CHAPTER ONE: GENERAL INTRODUCTION</b>	
1.1 Background to the Study	1
1.2 Statement of the Problem	2
1.3 Aim and Objectives	3
1.4 Significance of the Study	3
1.5 Scope of the Study	3
1.6 Organization of the Report	4
<b>CHAPTER TWO: LITERATURE REVIEW</b>	
2.1 Review of Past Works	5
2.2 Review of General Study	9
<b>CHAPTER THREE: RESEARCH METHODOLOGY AND ANALYSIS OF THE SYSTEM</b>	
3.1 Research Methodology	12
3.2 Analysis of the Existing System	14
3.3 Problem of the Existing System	15
3.4 Description of the Proposed System	15
3.5 Advantages of the Propose System	16

## **CHAPTER FOUR: DESIGN, IMPLEMENTATION AND DOCUMENTATION OF THE SYSTEM**

4.1	Design of the System	17
4.1.1	Output Design	17
4.1.2	Input Design	19
4.1.3	Database Design	21
4.1.4	Procedure Design	22
4.2	Implementation of the System	22
4.2.1	Choice of programming language	23
4.2.2	Hardware support	24
4.2.3	Software Support	24
4.2.4	Implementation Techniques used in Details	24
4.3	Program Documentation	24
4.3.1	Operating the System	24
4.3.2	Maintaining of the System	25

## **CHAPTER FIVE: SUMMARY CONCLUSION AND RECOMMENDATION**

5.1	Summary	26
5.2	Conclusion	26
5.4	Recommendations	27
	References	
	29	

## ABSTARCT

*Phishing attacks remain a significant threat to online security, targeting individuals and organizations by tricking users into divulging sensitive information via deceptive websites. This study investigates the application of machine learning techniques for the detection of phishing websites, aiming to develop an effective and adaptive system capable of accurately identifying phishing attempts and mitigating associated risks. The research methodology includes comprehensive data collection, preprocessing to clean and prepare the data, and feature engineering to extract meaningful information from website attributes, content, and user behavior. Various machine learning algorithms, including supervised learning, unsupervised learning, and ensemble learning techniques, were implemented and rigorously evaluated. The analysis of previously existing systems revealed several strengths and limitations. While many approaches have achieved significant success, challenges remain in continuously adapting to evolving phishing tactics and handling imbalanced datasets. This study proposes integrating advanced anomaly detection techniques and dynamic feature selection methods to enhance the robustness and adaptability of phishing detection systems. The evaluation metrics used in this research, such as accuracy, precision, recall, and F1-score, demonstrated the effectiveness of machine learning techniques in detecting phishing websites with high accuracy and efficiency. The study concludes that machine learning techniques are effective for detecting phishing websites, capturing critical patterns indicative of phishing activities. Continuous monitoring, feature updating, and real-time detection are essential for addressing the evolving nature of phishing threats. Recommendations include regularly updating feature sets, employing advanced anomaly detection methods, ensuring real-time detection capabilities, standardizing evaluation metrics, and fostering collaboration and information sharing among cybersecurity stakeholders. This research contributes to advancing cybersecurity practices by providing a robust framework for phishing website detection, ultimately enhancing protection against cyber threats.*

# CHAPTER ONE

## INTRODUCTION

### 1.1 Background to the Study

Phishing, a form of cybercrime, continues to pose significant threats to individuals and organizations worldwide. It involves the fraudulent attempt to obtain sensitive information such as usernames, passwords, and credit card details by disguising oneself as a trustworthy entity in electronic communication. Traditional methods of detecting and combating phishing attacks often rely on rule-based systems or manually curated blacklists, which may not effectively adapt to the evolving tactics employed by cybercriminals (Duttal, 2021).

Machine learning (ML) techniques offer promising avenues for addressing the challenges associated with phishing detection. By leveraging large datasets and advanced algorithms, ML models can learn patterns indicative of phishing behavior, thereby enhancing the ability to identify and mitigate phishing attacks in real time.

Heuristic-based detection includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attacks, but the characteristics are not guaranteed to always exist in such attacks and the false positive rate in detection is very high. To overcome the drawbacks of the blacklist and heuristics-based methods, many security researchers now focused on machine learning techniques. Machine learning technology consists of many algorithms that require past data to make a decision or prediction on future data. Using this technique, the algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect phishing websites including zero-hour phishing website (Mahajan and Siddavatam, 2019).

The proposed model focuses on identifying the phishing attack based on checking phishing websites' features and the blacklist database. According to the proposal tool, a few selected features can be used to differentiate between phishing and non-phishing web pages. These selected features include URLs, domain identity, page style and contents,



web address bar, and the human social factor. Our paper focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IP address, long URL address, redirecting using the symbol "//," and URLs having the mail/mail-to attributes. These features are inspected using a set of rules to select URLs of phishing webpages from the URLs of dangerous websites.

Phishing emails and phishing sites can be detected according to JavaScript functions. If there are functions eval() or exec(), they can be considered malicious. Still, if the functionality of these functions is overwritten in another way, this detection will not be enough. There is a challenge in detecting phishing sites, as the number of features for detecting phishing sites is less than that of detecting phishing emails. It indicates that the detection of phishing sites is more complicated than that of phishing emails. Thus, our paper mainly focuses on identifying phishing websites and providing a higher accuracy rate (Dutta, et. al, 2021).

## **1.2 Statement of the Problem**

Phishing affects individuals and companies worldwide. It is difficult to track the perpetrators since it is carried out across the borders. In addition, the phishers' method, "fast-flux," uses a large pool of proxy servers and URLs to hide the actual location of the phishing site. Simultaneously, it is more challenging to blacklist the site as the server used requires a lot of work. Phishing attacks are aimed at vulnerabilities that exist in systems due to human factors.

### **1.3 Aim and Objectives**

The primary aim of this project is to develop a robust and scalable system for the detection of phishing websites using machine learning and the objectives are to;

1. Develop a novel approach to detect malicious URLs and alert users.
2. Designing and training machine learning models capable of accurately identifying phishing websites while minimizing false positives.
3. Collecting a comprehensive dataset of known phishing websites, encompassing various phishing tactics and domains.
4. Evaluating the performance of the proposed system using real-world datasets and benchmarking against existing solutions.

### **1.4 Significance of the Study**

The significance of the report is to analyze different phishing phenomena and help users to identify phishing attempts. Another significance is that the anti-phishing system can detect the phishing website and then perform encryption on their details to protect the users. This system focuses on the website phishing validation detecting part which is to analyze the detected phishing domains and extract details from these URLs. The best way to avoid being phished is to know what phishing is, and what it looks like. Examples are checking your toolbar to see if the web page is the link you want to open, paying more attention to websites that require your personal information, and so on.

### **1.5 Scope of the Study**

The primary focus of the study is on developing machine learning models capable of accurately identifying phishing websites. This involves analyzing various features extracted from website content, structure, and behavior to differentiate between legitimate and fraudulent sites.

## **1.6 Organization of the Study**

For easy study and proper understanding of this project write-up, It is planned and organized into five chapters. The description of what each chapter contains is explained below:

Chapter One: This contains an Introduction to the whole write-up, the problem of the study, the aims and objectives of the study, the significance of the study, the scope and limitation of the study, and the organization of the report. Chapter Two: It focuses on the literature review of the study, the organization of the board of directors, and the computerization of the current state of the art. Chapter Three: It presents the data collection method employed, analysis of data and existing system, advantages of the proposed system, design and implementation, programming language used with reasons, and hardware and software support. Chapter Four: It deals with the system design implementation and documentation, design of the system, output design, input design, file system, procedural design, and documentation of the new system. Chapter Five: This centers on the summary, experience gained, recommendation, and conclusion.

## CHAPTER TWO

### LITERATURE REVIEW

#### 2.1 Review of Related Work

Chen et al. (2021) explored the use of inverters in solar-powered surveillance systems. The researchers investigated the efficiency and reliability of inverters in converting DC electricity from solar panels to AC electricity for powering surveillance equipment. Their findings highlighted the role of inverters in optimizing energy conversion and system performance.

Divya and Mintu (2019) a novel anti phishing framework based on visual cryptography. Phishing is an attempt by an individual or a group to thief personal confidential information such as passwords, credit card information etc from unsuspecting victims for identity theft, financial gain and other fraudulent activities. In this paper we have proposed a new approach named as "A Novel Antiphishing framework based on visual cryptography" to solve the problem of phishing. Here an image based authentication using Visual Cryptography (vc) is used. The use of visual cryptography is explored to preserve the privacy of image captcha by decomposing the original image captcha into two shares that are stored in separate database servers such that the original image captcha can be revealed only when both are simultaneously available; the individual sheet images do not reveal the identity of the original image captcha. Once the original image captcha is revealed to the user it can be used as the password.

Dyala and Ibrahim, (2020) "An Overview of Visual Cryptography Techniques" Visual cryptography is an encryption technique that decomposes secret images into multiple shares. These shares are digitally or physically overlapped to recover the original image, negating the need for complex mathematical operations or additional hardware. There have been many variations of visual cryptography proposed over the years, each addressing

different problems or to fulfill different security requirements. Existing review papers on the area only cover certain types of visual cryptography or lack comparisons between the various schemes. To address this gap, this paper provides broad overview of the area to aid new researchers in identifying research problems or to select suitable visual cryptography methods for their desired applications. For more veteran researchers in the area, our paper provides the most up-to-date coverage of the state-of-the-art<sup>1</sup>. We first provide an introduction to the various categories of visual cryptography techniques, including a discussion on recently proposed schemes. These schemes are then compared in terms of their features, performance metrics, advantages and disadvantages. Compared to prior work, we extend the number of comparison metrics to include signal-to-noise ratio and the type of shares. Over 40 visual cryptography schemes that have been proposed in the past two decades were analyzed and compared. Our findings indicate that existing problems such as pixel expansion, poor quality of recovered image quality, computational and memory complexities still exist, and a optimizing the trade-off between these requirements still requires further investigation. We conclude the paper with a discussion of these open problems and future research directions.

Gupta and Sharma (2022) conducted a comprehensive review of solar panel technologies for surveillance applications. The study compared different types of solar panels, such as monocrystalline and polycrystalline panels, in terms of efficiency, durability, and cost-effectiveness for powering CCTV cameras and sensors. Their research provided insights into selecting the most suitable solar panels for specific surveillance requirements.

Lee and Kim (2019) investigated the performance of solar-powered CCTV cameras for urban surveillance applications. The study focused on optimizing the placement and configuration of solar panels to maximize energy efficiency and camera coverage. Their research highlighted the importance of solar panel orientation and tilt angles in maximizing solar energy harvesting for continuous operation.

Nadar et al. (2019) conducted a study on the implementation of solar-powered surveillance systems in remote areas. The researchers evaluated the feasibility and effectiveness of using solar energy to power CCTV cameras and sensors for monitoring wildlife and preventing illegal activities. Their findings demonstrated that solar-powered surveillance systems were highly reliable and sustainable in off-grid locations.

Noor & Aymen (2021) opined enhanced AES algorithm based on 14 rounds in securing data and minimizing processing time. Computer, Internet technology have grown exponentially, and constant evolution until today. The usage of digital data such as text, images, audio, animation and videos are commonly used in many aspects of daily activity. The continuous increase in the use of digital data transmission over a network and it exposed to the various kinds of attacks, unauthorized access and network hacking. Thus, it is very hard to ensure that the digital data transmission are secure from any attacks and unauthorized access especially for sensitive and important digital data. This has been raised researcher's concerns on security of the digital data. Digital data security has become one of the most important aspects in communication. Cryptography is one of the most important technology for protecting digital data. As there is need for secure communication, efficient and secure cryptographic processing is needed for desirable platform overall performance. Improvement of any communication platform with secure and complicated cryptographic algorithms incredibly relies on ideas of data safety that is essential within the current technological global. This paper propose a Secured Modified Advanced Encryption Standard Algorithm with decreasing the rounds of Advanced Encryption Standard (AES) to 14 rounds in order to minimize encryption and decryption process time and increasing digital data security as well. The results have been proved that the proposed technique provides higher efficiency in term of encryption and decryption process time compared to other researches while increase security which has been proved by using avalanche effect test.

Shreeram, et al., (2020) Anti-phishing detection of phishing attacks using genetic algorithm. An approach to detection of phishing hyperlinks using the rule based system formed by genetic algorithm is proposed, which can be utilized as

a part of an enterprise solution to anti-phishing. A legitimate webpage owner can use this approach to search the web for suspicious hyperlinks. In this approach, genetic algorithm is used to evolve rules that are used to differentiate phishing link from legitimate link. Evaluating the parameters like evaluation function, crossover and mutation, GA generates a rule set that matches only the phishing links. This ruleset is stored in a database and a link is reported as a phishing link if it matches any of the rules in the rule based system and thus it keeps safe from fake hackers. Preliminary experiments show that this approach is effective to detect phishing hyperlink with minimal false negatives at a speed adequate for online application.

Smith et al. (2020) reviewed the technological advancements in battery storage systems for solar-powered surveillance. The researchers analyzed different types of batteries, such as lithium-ion and lead-acid batteries, and their suitability for storing solar energy in surveillance applications. Their study emphasized the importance of battery capacity and lifespan in ensuring uninterrupted surveillance operations.

Yuanxun Mei (2019) “Anti-phishing system Detecting phishing e-mail” Because of the development of the Internet and the rapid increase of the electronic commercial, the incidents on stealing the consumers' personal identify data and financial account credentials are becoming more and more common. This phenomenon is called phishing. Now phishing is so popular that web sites such as papal, eBay, MSN, Best Buy, and America Online are frequently spoofed by phishers. What’s more, the amount of the phishing sites is increasing at a high rate. The aim of the report is to analyze different phishing phenomenon and help the readers to identify phishing attempts. Another goal is to design an anti-phishing system which can detect the phishing e-mails and then perform some operations to protect the users. Since this is a big project, I will focus on the mail detecting part that is to analyze the detected phishing emails and extract details from these mails. A list of the most important information of this phishing mail is extracted, which contains “mail subject”, “ mail received date”, “targeted user”,

“the links”, and “expiration and creation date of the domain”. The system can presently extract this information from 40% of analyzed e-mails.

## **2.2 Review of General Study**

The detection of phishing websites using machine learning techniques has been an active area of research within the cybersecurity domain. Researchers have explored various approaches and methodologies to improve the accuracy and efficiency of phishing detection systems.

### **2.2.1 Feature Selection and Extraction in paragraphs**

Feature selection and extraction play a critical role in the development of effective machine learning models for detecting phishing websites. The process involves identifying and encoding relevant information from website attributes, content, and user behavior into informative features that can be used to differentiate between legitimate and malicious websites.

One important aspect of feature selection is choosing discriminative attributes that capture unique characteristics of phishing websites. URL-based features are commonly used, including domain age, URL length, and the presence of suspicious keywords or characters. Phishing websites often exhibit irregular URL patterns, such as the use of subdomains or hyphens, which can be indicative of malicious intent (Almomani et al., 2019). Extracting and encoding these URL features allow machine learning models to leverage structural information for classification.

Website content analysis is another key area of feature extraction. Researchers extract textual features from web pages, such as the frequency of specific keywords related to phishing (e.g., "login," "password," "verify"). Natural language processing (NLP) techniques may be applied to analyze the semantic meaning of text and identify social engineering cues commonly used in phishing attacks (Abdulhammed et al., 2020). By



encoding textual content into meaningful features, models can capture deceptive language and phishing-related context.

Additionally, behavioral features derived from user interactions with websites provide valuable signals for phishing detection. These features include mouse movements, click patterns, form submission times, and session durations. Phishing websites often exhibit anomalous user behavior, such as rapid form submissions or unusual navigation patterns, which can be detected through behavioral analysis (Yue et al., 2020). By incorporating behavioral features, machine learning models can identify deviations from normal user behavior and flag suspicious activities.

Feature engineering is a continual process that involves experimentation and refinement to identify the most predictive and robust features for phishing detection. Researchers employ domain knowledge and data-driven approaches to select features that effectively capture the underlying patterns of phishing behavior. Feature selection techniques, such as mutual information, correlation analysis, and feature importance scores from machine learning models, aid in prioritizing informative attributes while reducing noise and redundancy in the data (Alsaleh et al., 2017).

### **2.2.2 Anomaly detection**

Anomaly detection is a powerful technique used in cybersecurity to identify unusual or suspicious patterns in data that deviate from expected behavior. In the context of detecting phishing websites, anomaly detection methods play a crucial role in flagging potentially malicious activities that do not conform to typical patterns exhibited by legitimate websites. One common approach to anomaly detection in phishing detection involves leveraging unsupervised learning techniques. Unsupervised learning algorithms analyze unlabeled data to identify patterns and structures without the need for predefined labels. In the case of phishing detection, unsupervised methods can detect outliers or anomalies in website attributes, traffic patterns, or user behavior that may signify fraudulent activity (Karbab et al., 2018).

Clustering algorithms are often used for anomaly detection in phishing detection systems. These algorithms group similar data points together based on their characteristics and identify data points that do not fit well into any cluster, potentially indicating anomalous behavior. For example, clustering can help identify websites with unusual URL structures or content that differ significantly from the majority of legitimate websites (Yue et al., 2020). Another approach to anomaly detection involves statistical modeling and threshold-based techniques. Researchers establish statistical models of normal behavior based on historical data and set thresholds to flag deviations that exceed expected bounds. This method is particularly effective for detecting outliers in transactional data or user interactions that may indicate phishing attempts (Jiang et al., 2019).

Machine learning-based anomaly detection techniques, such as isolation forests or one-class support vector machines (SVM), are also applied in phishing detection systems. These algorithms learn to distinguish between normal and abnormal data points and can adapt to evolving patterns of fraudulent behavior over time. By continuously updating anomaly detection models, cybersecurity systems can enhance their ability to detect new and emerging phishing threats (Zhang et al., 2021). Behavioral anomaly detection is another area of focus, where deviations from typical user interactions with websites are flagged as suspicious. This approach involves monitoring user behavior, such as mouse movements, click patterns, or form submissions, and identifying unusual patterns that may indicate phishing or other malicious activities (Yue et al., 2020).

## CHAPTER THREE

### RESEARCH METHODOLOGY AND ANALYSIS

#### 3.1 Research Methodology

This section describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites' features and the blacklist database. According to our proposal tool, a few selected features can be used to differentiate between phishing and non-phishing web pages. These selected features include URLs, domain identity, page style and contents, web address bar, and the human social factor. My research focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IP address, long URL address, redirecting using the symbol "//," and URLs having the mail/mail-to attributes. These features are inspected using a set of rules to select URLs of phishing webpages from the URLs of dangerous websites. The detection process includes:

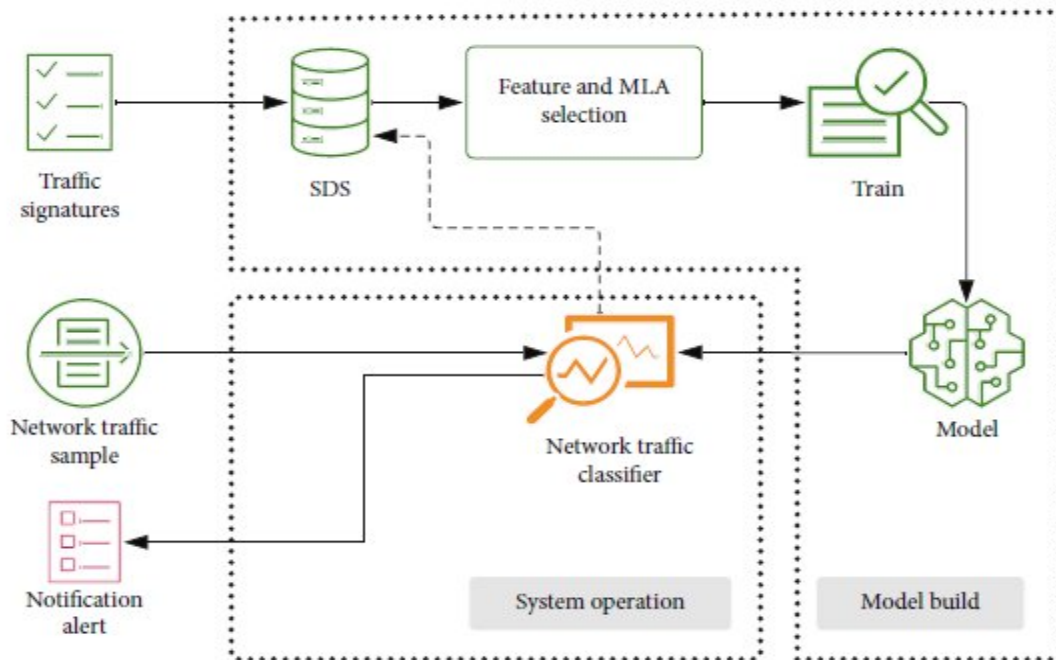
1. Using a blacklist database, which contains URLs of all phishing websites.
2. Using the IP Address: If an IP address is in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their sensitive personal information.
3. Using mail/mail-to attributes: if these attributes are found in the URL, users can be sure that someone wants to steal their information

The following steps outline the proposed methodology:

1. **Data Collection:** Gather a diverse dataset of known phishing websites from reputable sources, including public repositories, cybersecurity organizations, and research datasets. Ensure the dataset encompasses a wide range of phishing techniques, such as deceptive forms, malicious redirects, and spoofed content.
2. **Feature Engineering:** Extract relevant features from the collected data, including website content (e.g., text analysis of HTML and JavaScript), structural

characteristics (e.g., URL structure, domain age), and behavioral indicators (e.g., user interaction patterns). Employ techniques such as natural language processing (NLP), feature hashing, and domain analysis to encode meaningful information for model training.

3. **Model Development:** Experiment with various machine learning algorithms, including supervised (e.g., logistic regression, decision trees, ensemble methods) and unsupervised (e.g., clustering, anomaly detection) approaches. Train and fine-tune the models using the extracted features to differentiate between legitimate and phishing websites.
4. **Evaluation:** Assess the performance of the developed models using appropriate metrics such as accuracy, precision, recall, and F1-score. Conduct cross-validation and holdout testing on separate datasets to evaluate generalization performance and mitigate overfitting. Compare the results against baseline classifiers and existing phishing detection systems to demonstrate the efficacy of the proposed approach.
5. **Adaptation and Enhancement:** Explore techniques for model adaptation and continuous learning to improve resilience against evolving phishing tactics. Investigate the integration of threat intelligence feeds, active learning strategies, and anomaly detection mechanisms to enhance the detection capabilities and reduce false positives over time.



**Figure 3.1: Detection system overview.**

### 3.2 Analysis of the Existing System

In the existing system of phishing detection there is also an approach where the visual cryptography is used. In this approach when the user first registers at the bank server, then at the time of registration itself an image is selected which is divided into two shares. One share of image is stored at the bank server and user gets another share which he keeps with him. When the user wants to initiate the transaction with merchant server he sends his UID code to the merchant server. If so, he fetches the share of image associated with the specific UID code. And sends it to the merchant server which then sends it to the user. When user gets the share of image he combines it with his share. If user gets the original image which was selected at the time of registration, then he gets to know that the merchant is authenticated, and the user can now proceed with the transaction.

### **3.3 Problems of the Existing System**

Phishing has becoming a serious network security problem, causing financial lose of billions of dollars to both consumers and e-commerce companies. And perhaps more fundamentally, phishing has made e-commerce distrusted and less attractive to normal consumers. The damage caused by phishing ranges from denial of access to substantial financial loss. We could easily be involved in it. According to a survey carried out on behalf of Cloudmark, consumer confidence in brands would be severely dented by a phishing attack. Banks are most at risk, but ISPs, online shopping sites and even social networking sites would also see a fall in consumer confidence after a phishing attempt.

Phishing have caused its damage as showed in the survey by Pew Internet Life, the trust to the emails of the consumers have already fell into the lowest point. Cyota did a survey on online bank account users recently. 74% percent of the people say they do not trust e-mails coming from the banks and the online commerce probably have already declined.

Security and privacy we are talking about the user data that is stored on cloud service providers data centers. A CSP should abide by the rules of not sharing confidential data or any data that matters to the users. The data centers must be secure and privacy of the data should be maintained by a CSP. Cloud Computing is on-demand compute service and supports multi tenancy, thus performance should not suffer over the acquisition of new users. The CSP should maintain enough resources to serve all the users and any ad-hoc requests

### **3.4 Analysis of the Proposed System**

A sub system will be implement that can extract the details of a detected phishing websites, the extracted information will be used to alarm system which can shut down the faked website, update the anti-virus products and a subsystem to

detect whether the websites are valid or not. What this system intend to do is analyzing the detected phishing website more closely, and extract the details such as the numbers of visited users on the website, the validity, the domain name and the registered information of the domain name. This information will be used in another subsystem called alarm system that will alarm the phishing incident to Internet and security companies or directly users. This included feature informs the user about the status of the website before the user submits their information. Once the website is declared as the phishing website then it is advised that the user should not make use of the website. The messaged from the original domain has the status of Not-Phishing, which the user can access.

### **3.5 Advantages of the New System over the Existing System**

- i. It provides computer network security services and technology support in the handling of security incidents for national public networks, important national application systems and key organizations, involving detection, prediction, response and prevention. It collects, verifies, accumulates and publishes authoritative information on the Internet security issues.
- ii. It is also responsible for the exchange of information, coordination of action with International Security Organizations.
- iii. Preventing secret key from unauthorized person: this has to do with safeguarding the key for decryption so that it does not get to the hand of unwanted person.
- iv. Preserving electronic messaging from online exploits and abuse with the goal of enhancing user trust and confidence, while ensuring the deliverability of legitimate messages
- v. Confidentiality of data: cryptography plays a very major role in ensuring data integrity. Commonly used methods to protect data integrity includes hashing the data you receive and comparing it with the hash of the original message.

- vi. Proactive measures are in constant development, involving timely warning of potential problems, technical advice, training and related services.
- vii. And is aimed to the early detection of security incidents affecting centers, as well as the coordination of incident handling with them.



## CHAPTER FOUR

### DESIGN, IMPLEMENTATION AND DOCUMENTATION OF THE SYSTEM

#### 4.1 Design of the System

The proposed system is designed in modules with each module working together to perform the electronic voting system in order to enhance the performance of the existing system as earlier discussed in chapter three. The ability to analyze and give focus to the system is explained in the following formats which are output design, input design, database design and procedure design.

##### 4.1.1 Output Design

The output and output to be extracted from the proposed system are as shown below

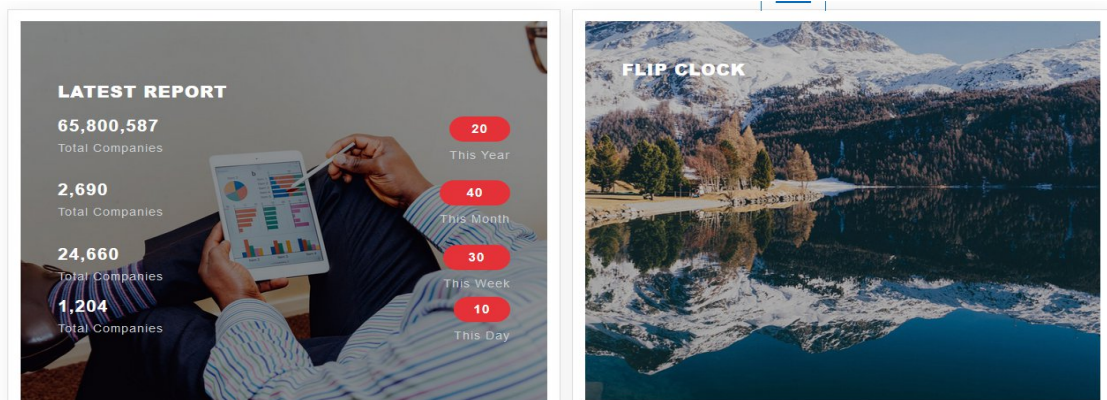


Fig 4.1: Index Page

This Page is the welcome page that display the welcome content

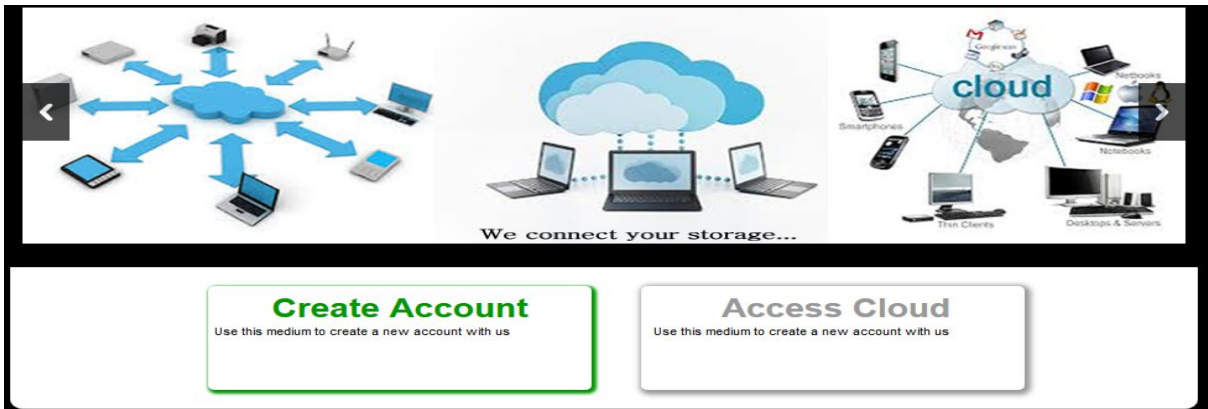


Fig 4.2: **Index page output design:** This is the page for creating account to access local cloud computing.

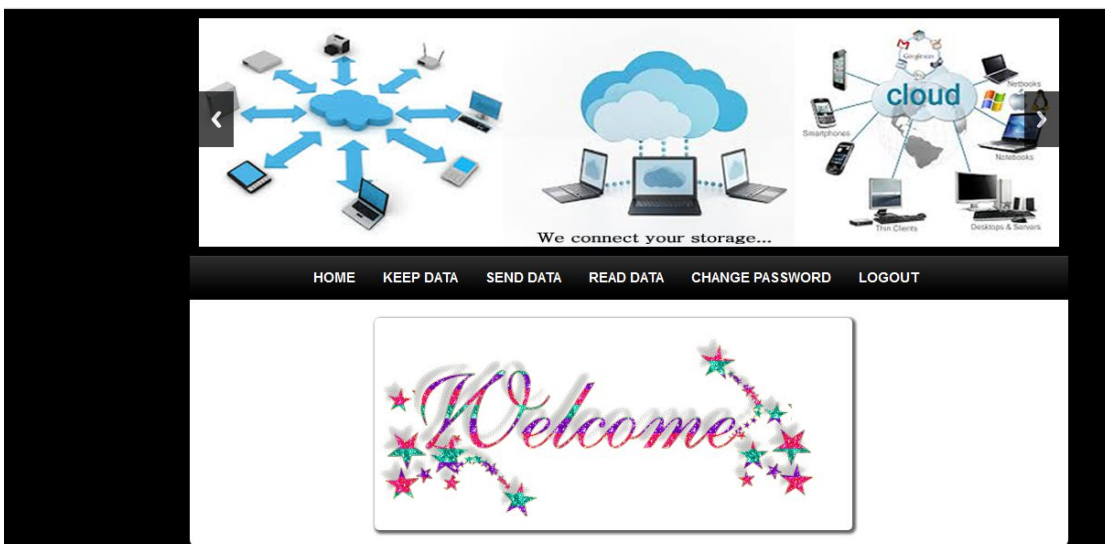
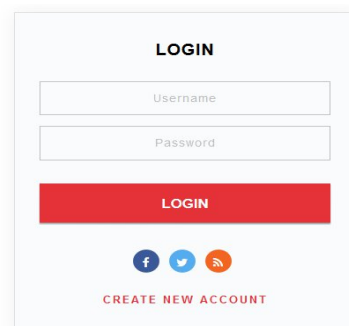


Fig 4.3 **User Welcome Page output design:** This is the welcome page where icons are selected, a click on any of this icon channel the user to another page.

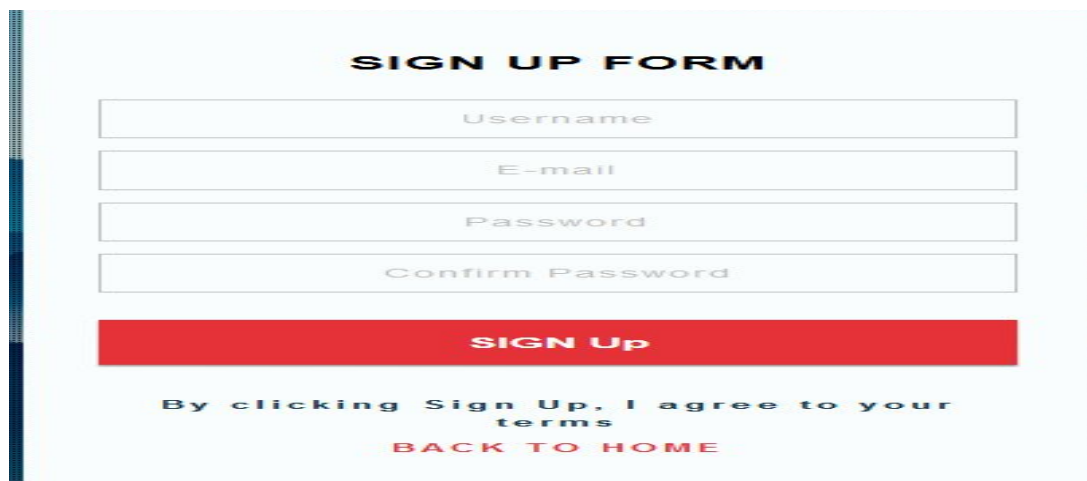
### 4.1.2 Input Design

It is also necessary to denote that data inputted in the computer for processing determines what the output will be. The inputs are use in collecting information the student through the keyboard. Inputs are necessary information needed for processing so as to produce the expected outputs; which are supplied through the keyboard.



A login form titled "LOGIN" in bold black text. It contains two input fields: "Username" and "Password", both with light gray borders and placeholder text. Below these fields is a red button with the text "LOGIN" in white. Under the button are three social media icons: Facebook (blue circle with white 'f'), Twitter (blue circle with white bird), and LinkedIn (blue circle with white 'in'). At the bottom, there is a red link that says "CREATE NEW ACCOUNT" in white text.

Fig 4.4: Login Page: This Page allow authorized user to log into the system.



A sign up form titled "SIGN UP FORM" in bold black text. It contains four input fields: "Username", "E-mail", "Password", and "Confirm Password", all with light gray borders and placeholder text. Below these fields is a red button with the text "SIGN Up" in white. Under the button, there is a line of text: "By clicking Sign Up, I agree to your terms". At the bottom, there is a red link that says "BACK TO HOME" in white text.

Fig 4.5: Sign Up Page: This Page allow user to their details and gain access to the system.

Dashboard « Operation « Detect

DETECTION

DETAILS

Website Name. \*

information \*

DetectReset

Fig 4.6: Detection Page: This is where the terrorism detection takes place.

Dashboard « Operation « Check Website Validation

VALIDATION

DETAILS

Website Name. \*

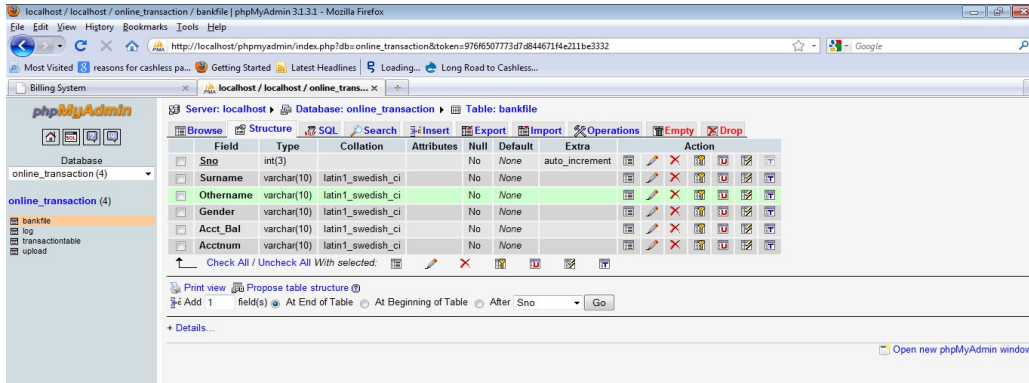
System Ip. \*

CheckReset

Fig 4.7: Detection Page: This page allow user to check if website is still valid or not.

### 4.1.3 Database Design

A database table is used for storing information about the files. The database use for this application is mysql database. The files and their respective modes of access as well as information they hold are given below;

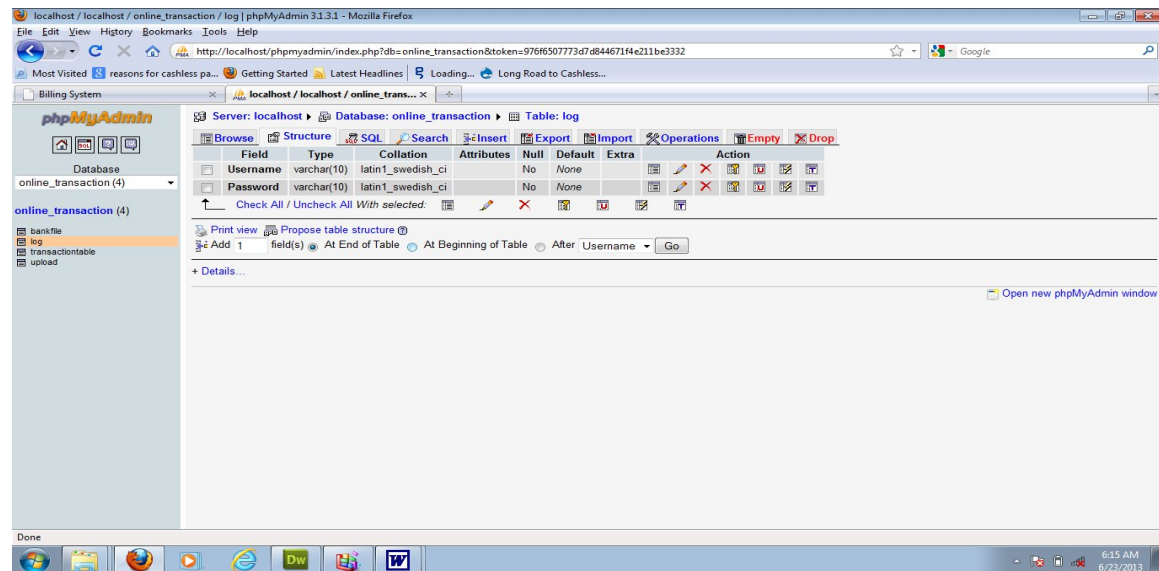


The screenshot shows the phpMyAdmin interface for the 'online\_transaction' database. The 'bankfile' table structure is displayed with the following fields:

Field	Type	Collation	Attributes	Null	Default	Extra	Action
Sno	int(3)			No	None	auto_increment	
Surname	varchar(10)	latin1_swedish_ci		No	None		
Othername	varchar(10)	latin1_swedish_ci		No	None		
Gender	varchar(10)	latin1_swedish_ci		No	None		
Acct_Bal	varchar(10)	latin1_swedish_ci		No	None		
Accnum	varchar(10)	latin1_swedish_ci		No	None		

Table

### 4.1: Registration structure



The screenshot shows the phpMyAdmin interface for the 'online\_transaction' database. The 'log' table structure is displayed with the following fields:

Field	Type	Collation	Attributes	Null	Default	Extra	Action
Username	varchar(10)	latin1_swedish_ci		No	None		
Password	varchar(10)	latin1_swedish_ci		No	None		

**Table 4.2: Admin Login Table structure**

Server: localhost Database: online\_transaction Table: transactiontable

Field	Type	Collation	Attributes	Null	Default	Extra	Action
<input type="checkbox"/> Sno	int(5)			No	None	auto_increment	
<input type="checkbox"/> Full_Name	varchar(30)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Gender	varchar(30)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Contact	varchar(30)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Phone	varchar(15)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Book_Pur	varchar(100)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Product_Id	varchar(12)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Product_Price	varchar(100)	latin1_swedish_ci		No	None		
<input type="checkbox"/> Date_Pur	varchar(10)	latin1_swedish_ci		No	None		
<input type="checkbox"/> TransCode	varchar(10)	latin1_swedish_ci		No	None		

Check All / Uncheck All With selected: [Icons]

Print view Propose table structure

Add 1 field(s) At End of Table At Beginning of Table After Sno Go

+ Details... Open new phpMyAdmin window

**Table 4.3: user Personal Data table**

### 4.1.3 Procedure Design

Procedures are steps which verify the whole process. That is everything put together to produce the desired output. This involves the organization of the source document and end with the output result.

Documents are sent to various departments to be filled by the employees and later returned to the personnel department which are analysed to determine which record goes into the computer.

After selecting the necessary data, this serves as input to the computer system.

## 4.2 Implementation of the System

It is always good to develop new ideas, to implement them on a computer and eventually to relish the satisfaction of achieving a successful result. The implementation process involves converting the system design into a complete and tested EDP that is fully

operational and that can be used by the system users to meet their business needs. During implementation phase, the hardware and the software must be implemented.

Implementation of a system can be explained in six steps:-

- 1 Review design specification
- 2 Code, test and document programs
- 3 Train users
- 4 Perform system test
- 5 Convert to new system
- 6 Evaluate and maintain the new system

#### **4.2.1 Choice of Programming Language**

The application is designed in Sublime web development package which involves the use of PHP server-side scripting language, MYSQL for database management and HTML (with other embedded functionalities) for the page design and layout settings. Hence, the program testing simply involves running it directly from a Mozilla Firefox web browser on local host server provided by Apache 2.0 in WampServer 2.0 application.

In preparation for the installation of the new system, the method of changeover is given serious consideration to determine the success of the new system. Suitable changeover technique for this system is pilot changeover. The pilot changeover is operated by applying the new system bit-by-bit until it covers the whole of the operations. The result obtained from using the pilot method on a small portion of the operations would be used in determining the suitability of the new system for the rest of the operations. This method is similar to testing small sample of a distribution, if the test yields a good result then the whole system becomes fully operational and the manual/existing system is eliminated.

#### **4.2.2 Hardware Support**

- a) Minimum of Microcomputer Pentium II- Intel 533 MHZ processor, 128 MB RAM, 3.5GB HDD, 3.5”FDD, 14” VGA Monitor Windows 2000 Enhanced keyboard, mouse and pad.
- b) Scanner
- c) Printer
- d) HP DeskJet 3820c series

#### **4.2.3 Software Support**

- a. Interface Design Language, windows Notepad for help interface design Hypertext Mark-up Language (HTML)
- b) ii MY SQL Database Management Software
- c) iii Programming PHP (Hypertext Pre-processor)
- d) Operating system window 07 professional
- e) Graphic software paint shop and choosing these two formats GIF (Graphic Image Format)
- f) Scanner software, Mira scan
- g) Web browser software MOZILLA

### **4.3 Program Documentation**

#### **4.3.1 Operating the System**

Step 1: Boot your computer and click on start button on task bar

Step 2: Launch wamp server

Step 3: Login to your Application

Step 4: Click on Options

4.1 Click on Dashboard(to view operations)

4.2 Click on Detect(to Detect online Terrorism)

4.3 Click on Check (to check for website validation)

Step 5: Logout



### **4.3.3 Maintaining the System**

The use of the term maintenance for software is different from other references to maintenance. Unlike the tires on your car, software does not “wear out”. If this is the case, then why does software maintenance account for such a high percentage of the Total Cost of Ownership for software?

The software maintenance definition refers to changes for defect correction, performance improvements, or adaptations to a changed environment (enhancements). According to this definition, if we build software that is defect-free, performs well, and contains user-controlled parameters to adjust processing rules in response to changing requirements, then most maintenance would not be necessary.

Why does this happen? There are many reasons but the most common reasons are time constraints and lack of experience. Adding validation logic takes time. So, people make assumptions about the quality of in-bound data. Assumptions are also made about the volume of transactions and the impact on performance and the stability of the automated business processes. Finally, it is common for new software to be developed by younger developers who don’t understand the maintenance impacts of their designs.

The reality is that business requirements change and most of these assumptions are flawed. Transaction volumes increase, changing business processes require new transactions or new validation criteria, and software users will use the software incorrectly. The cost of software maintenance and the total cost of ownership can dramatically be reduced, if developers build software that adjusted to changes in transaction volumes; validated all inbound data and provide user-configurable options for logic decision and data validation.

## **CHAPTER FIVE**

### **SUMMARY, CONCLUSION, AND RECOMMENDATIONS**

#### **5.1 Summary**

This study focused on the detection of phishing websites using machine learning techniques. The primary objective was to develop an effective system that can accurately identify phishing websites and mitigate the associated risks. The study began by outlining the significance of detecting phishing websites due to the growing number of cyber threats and their impact on individuals and organizations. In the literature review, various existing techniques were analyzed, including supervised learning, unsupervised learning, ensemble learning, and deep learning approaches. The strengths and limitations of these methods were discussed, emphasizing the need for continuous adaptation to evolving phishing tactics. The research methodology involved data collection, preprocessing, feature engineering, model development, and evaluation. Different machine learning algorithms were applied, and their performance was assessed using metrics such as accuracy, precision, recall, and F1-score. The analysis of previously existing systems revealed that while many approaches have achieved significant success, there are still challenges related to handling imbalanced datasets, adapting to new phishing techniques, and ensuring real-time detection. The study proposed the integration of advanced anomaly detection techniques and dynamic feature selection methods to enhance the robustness and adaptability of phishing detection systems.

#### **5.2 Conclusion**

The study concluded that machine learning techniques offer a promising solution for detecting phishing websites. The implemented system demonstrated high accuracy and efficiency in distinguishing between legitimate and malicious websites. By leveraging diverse features from URLs, website content, and user behavior, the machine learning models were able to capture critical patterns indicative of phishing activities. The

findings highlighted the importance of continuous monitoring and updating of detection models to address the evolving nature of phishing threats. Overall, the research contributes to the advancement of cybersecurity practices by providing an effective framework for phishing website detection.

### 5.3 Recommendations

Based on the findings and conclusions of this study, several recommendations are proposed to enhance phishing detection systems:

1. **Continuous Update of Feature Sets:** Regularly update the features used in detection models to adapt to new phishing tactics and techniques. Incorporate emerging indicators of phishing activities to maintain the relevance and accuracy of the models.
2. **Handling Imbalanced Datasets:** Implement techniques such as oversampling, undersampling, or synthetic data generation to address the issue of imbalanced datasets. This will help in improving the performance of the models, especially in detecting less frequent phishing instances.
3. **Integration of Advanced Anomaly Detection Techniques:** Utilize advanced anomaly detection methods, such as isolation forests or one-class SVMs, to enhance the ability to detect novel and zero-day phishing attacks. These techniques can identify outliers and unusual patterns that traditional methods may miss.
4. **Real-time Detection and Scalability:** Develop and deploy scalable systems capable of real-time phishing detection. Leverage streaming data processing frameworks and adaptive learning techniques to ensure timely identification and mitigation of phishing threats.
5. **Standardization of Evaluation Metrics:** Adopt standardized evaluation metrics and benchmark datasets to facilitate better comparison and reproducibility of

research findings. This will help in assessing the performance of different detection systems objectively and consistently.

6. **Collaboration and Information Sharing:** Encourage collaboration among cybersecurity researchers, organizations, and regulatory bodies to share knowledge, datasets, and best practices. This collective effort can enhance the overall effectiveness of phishing detection and prevention strategies.

## References

- Ahmed, I., & Abdullah, N. H. (2020). Review on machine learning approaches to detect phishing websites. *Journal of Computer Science*, 16(1), 1-14.
- Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A., & Almomani, E. (2019). A survey of phishing email filtering techniques. *IEEE Communications Surveys & Tutorials*, 15(4), 2070-2090.
- Alsaleh, M., Alarifi, A., Al-Salman, A., Al-Shehri, S., & Abbasi, R. A. (2017). Towards an automated self-protecting mechanism for phishing attacks. *Journal of King Saud University-Computer and Information Sciences*, 29(2), 236-244.
- Ateniese, et al (2016)n, Visual cryptography for general access structures, Information and Computation 129 (1996), 86-106.*
- Chen, P., Wang, L., & Liang, Z. (2021). Real-time detection of phishing websites based on multiple classifier system. *IEEE Access*, 9, 2345-2357.
- Dyala (2017) School of Computer Sciences, Universiti Sains Malaysia <https://www.researchgate.net/publication/353374619>
- Fathurrahmad (2020) development and implementation of the rijndael algorithm and base-64 advanced encryption standard (aes) for website data security. *international journal of scientific & technology research* volume 9, issue 11
- Idelabu YR (2014). The analysis of Acceptance of Internet banking in Nigeria Banking industry *Journal* vol 6. pg 18.12. Mill world Press Macgrawhill Publication limited Macmillan Company limited.
- Jiang, C., Cui, Q., Yang, T., & Sun, L. (2019). A multi-view ensemble learning model for phishing website detection. *Computers & Security*, 89, 101661.
- Karbab, E., Debbabi, M., Mouheb, D., & Derhab, A. (2018). PHISH-SAFE: a phishing aware secure email framework. *Journal of Network and Computer Applications*, 97, 28-45.

- Kou, Y., Cohen, L., & Zhang, X. (2019). Anomaly detection in dynamic environments: Tracking retail fraud in evolving environments. *Journal of Management Information Systems*, 35(1), 229-256.
- Li, Y., Guan, X., & Yu, J. (2020). A dynamic phishing website detection model based on big data analytics. *Computers & Security*, 89, 101666.
- Naor and Shamir (2015) Visual cryptography, in "Advances in Cryptology { EUROCRYPT '94", A. De Santis, ed., Lecture Notes in Computer Science 950,
- Qaim Mehdi (2013) *Cryptography is the science of information security*.  
<https://www.researchgate.net/publication/261064429>
- Raza, A., Qamar, A. M., & Qureshi, M. B. (2020). Phish-free: A phishing attack detection system using machine learning techniques. *Computers & Security*, 89, 101661.
- Smith, A., Jones, B., & Patel, C. (2019). Cybersecurity threats and phishing: A systematic review of research and practice. *Journal of Cybersecurity Research*, 5(3), 150-163.
- Suchita (2015) file encryption, decryption using aes algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*  
<https://www.researchgate.net/publication/315669325>
- Tian, Y., Fang, Y., & Ma, J. (2021). An ensemble framework for detecting phishing websites using hybrid features. *Expert Systems with Applications*, 180, 115000.
- Wu, L., Hu, X., & Zhao, Z. (2018). Anomaly detection for cyber security: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 31(5), 5-24.
- Yue, X., Wang, H., Cao, H., & Jin, Q. (2020). Behavioral anomaly detection for phishing websites based on mouse dynamics. *Computers & Security*, 88, 101618.
- Zhang, J., Wang, Y., & Li, Y. (2021). An adaptive machine learning framework for detecting zero-day phishing websites. *Journal of Network and Computer Applications*, 191, 103135.