

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY (JKUAT)

DEPARTMENT OF COMPUTING

Course: **BSc. COMPUTER TECHNOLOGY**

ICS2406: COMPUTER SYSTEMS PROJECT

PROJECT PROPOSAL REPORT

REF: JKU/2/83/022

**TITLE: RESEARCH ON PHISHING DETECTION AND PREVENTION USING MACHINE LEARNING.**

***Author***

Name: **STANLEY NGUGI CHEGE** Reg. No: **SCT212-0065/2017**

Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Supervisor 1:***

Name: **Damaris Waema** Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Supervisor 1***

Name: **Petronilla Muriithi** Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Submission Date: **January 2021**

# Abstract

With the success of online businesses such as E-commerce platforms, cybercrimes have been on the rise. Often a fake website is publicly deployed to trick users into believing that the website is legitimate and safe to give away sensitive information such as passwords and credit card details. The perpetrators then use phishing techniques to attract vulnerable internet users. Users tend to overlook the URL of a website as the attacker may intelligently disguise themselves. This research will present a fraudulent URL detection mechanism to help internet users to stay protected from cyber criminals. This will be achieved by training a machine learning model to characterize the behavior of phishing attacks and therefore formulate a basis to predict the legitimacy of a URL. This will also provide suggest a response mechanism to alert an internet user in case of a fraudulent URL visit. This will enable online merchants and consumers to verify the authenticity of a website so that they can perform any business or data transactions with them without any hesitation.

Table of Contents

[Abstract 2](#_Toc83100146)

[Table of Contents 4](#_Toc83100147)

[1 Introduction 5](#_Toc83100148)

[1.1 Background 5](#_Toc83100149)

[1.2 Problem Statement 6](#_Toc83100150)

[1.3 Objectives 7](#_Toc83100151)

[1.4 Research Questions 7](#_Toc83100152)

[1.5 Justification 8](#_Toc83100153)

[1.6 Project Scope 8](#_Toc83100154)

[2 Literature Review 9](#_Toc83100155)

[2.1 Introduction 9](#_Toc83100156)

[2.2 Phishing 9](#_Toc83100157)

[2.3 Related Works 10](#_Toc83100158)

[2.3.1 Automated phishing detection 10](#_Toc83100159)

[2.4 Security Indicators 12](#_Toc83100160)

[2.5 Supervised Machine Learning 14](#_Toc83100161)

[2.5.1 Machine Learning Algorithms 15](#_Toc83100162)

[2.6 Features Extraction 18](#_Toc83100163)

[2.7 Why Phishing Still works 20](#_Toc83100164)

[2.8 Conclusion 22](#_Toc83100165)

[3 Research Methods and Design 23](#_Toc83100166)

[3.1 Data collection 23](#_Toc83100167)

[3.2 Data Analysis 23](#_Toc83100168)

[3.3 System implementation 23](#_Toc83100169)

[3.4 System Testing and Evaluation 24](#_Toc83100170)

[Conclusion 25](#_Toc83100171)

[References 27](#_Toc83100172)

[APPENDICES 28](#_Toc83100173)

[Appendix 1. Schedule 28](#_Toc83100174)

[Appendix 2. Budget 29](#_Toc83100175)

List of Figures

[Figure 4.1 Use case diagram 33](#_Toc86225172)

[Figure 5.1 Conceptual Design 34](#_Toc86225173)

[Figure 5.2 Database Schema 35](#_Toc86225174)

[Figure 5.3 Browser Extension alert screen 36](#_Toc86225175)

[Figure 5.4 Browser Extension feedback form 36](#_Toc86225176)

[Figure 5.5 Browser extension Intervention Screen 37](#_Toc86225177)

[Figure 5.6 Website Homepage 37](#_Toc86225178)

[Figure 5.7 Website samplye scan results 38](#_Toc86225179)

[Figure 5.8 Sequence Diagram 39](#_Toc86225180)

[Figure 5.9 Activity Diagram 39](#_Toc86225181)

# Introduction

## Background

Online businesses such as E-commerce platforms, auctioning websites and drop-shipping services are becoming the go-to model for most business today in order to keep their businesses afloat. Cyber Security threats have also been on the rise as we become more and more reliant on technology in our everyday lives. According to Sarno et al. (2019), often times such threats are a result of phishing scams in which individuals are tricked into clicking on malicious links attached to emails that are spoofed to look legitimate.

The number of phishing attacks has been growing considerably in recent years and is considered as one of the most dangerous modern internet crimes, which may lead individuals to lose confidence in e-commerce. Consequently, it has a tremendous negative effect on online commerce, marketing efforts, organizations’ incomes, relationships, customers, and overall business operations. (Ali, W. et.al., 2017). In a study conducted by Retruster in 2019, phishing is responsible for 90% of data breaches; Around 1.5M phishing sites are created every month; 30% of targets fall for phishing traps. Phishing attacks are not only increasing with time, but also evolving.

The Anti-Phishing Working Group (APWG) is an international consortium that brings together enterprises affected by phishing attacks. With support from 2300+ members and 1700 companies and agencies, The APWG tracks the number of unique phishing sites reported across the globe by the community. The following statistics were reported by APWG in the first quarter of 2021:

|  |  |  |  |
| --- | --- | --- | --- |
|  | January | February | March |
| Number of unique phishing Web sites detected | 245,771 | 158,898 | 207,208 |
| Number of brands targeted by phishing campaigns | 430 | 407 | 465 |

Phishing cases reported in January 2021 were the most reported cases ever reported to APWG. This is only a primary measure of reported phishing cases across the globe. There are clearly not enough prevention mechanisms in place to protect online consumers from attackers.

## Problem Statement

Internet fraudsters, an elite team of cyber-criminals, are individuals who make use of technology to commit malicious activities on the internet with the intention of personal gain such as making profits. The absence of laws that restrict the acquisition of domains for online businesses and the ease of acquiring them paves way for internet fraudsters who acquire domains with malicious intentions. They are able to disguise their identities and IP addresses through the use of Virtual Private Networks and proxies which makes it difficult to track them once their deception comes to light.

Most internet users are not computer security specialists and are therefore vulnerable to an attack. A phishing email would easily capture their attention and they will fall under the control of the phisher. The problem can be minimized by addressing it in two folds; developing more targeted anti-phishing detection and interventions techniques; implementing a round-the clock protection mechanism to notify internet users in case they come across fraudulent URLs.

## Objectives

The main objective of the study is to make use and evaluate the efficiency of Supervised Machine Learning in the detection and classification of websites as fraudulent or legitimate. To achieve this, the following objectives will be met:

1. To train a machine learning model using a suitable dataset to classify websites as legitimate or fraudulent.
2. To identify common distinguishing features between fraudulent and legitimate websites
3. To develop an application that uses the trained machine learning model to provide a round-the clock protection mechanism to keep the user aware of suspicious websites
4. To evaluate the ability of the application developed in 3 above in detecting fraudulent URLs

## Research Questions

1. How will the machine learning model be trained using a suitable dataset to classify websites as legitimate or fraudulent?
2. What properties of a website will be used to distinguish legitimate and fraudulent websites?
3. How will an application that uses the trained machine learning model be implemented to provide a round-the-clock protection mechanism to keep users aware of suspicious websites?
4. How will the application developed be evaluated in detecting fraudulent URLs?

## Justification

As phishers evolve and sharpen their skills, it becomes more difficult for novice users to detect or distinguish phishing websites from legitimate ones. Therefore, there is need to provide anti-phishing solutions that adopt Machine Learning which tends to be more practical and effective in combating phishing. Machine Learning anti-phishing techniques rely on website features to derive knowledge that can assist in identifying phishing websites. Phishing campaigns usually take significantly lower times in their attack (a few hours) such that relying on blacklists and whitelists does not guarantee their detection. Therefore, a binary classification of websites as either legitimate or fraudulent in real time will help internet users stay ahead of the game as they protect themselves from cyber-attacks.

## Project Scope

The scope of this project is limited to training a machine learning model to distinguish between potential phishing websites. This will be achieved by analyzing various properties of a website URL and training the model to detect potential phishing attacks based on the attributes. This will also include the development of an application to give users real time protection in order to keep them aware of suspicious websites. The application will then be evaluated to test the effectiveness of the application in detecting malicious URLs.

# Literature Review

## Introduction

This chapter consists of related literature on phishing detection and prevention mechanisms acquired from books, journals, academic papers and articles. This includes research on related approaches to phishing detection, machine learning algorithms that have been used to detect phishing, browser security indicators and features that can be used to flag websites as legitimate or fraudulent.

## Phishing

Phishing is a type of social engineering attack often used to trick users into revealing private or sensitive information by masquerading as a trusted entity. These attacks occur in large numbers and have caused billions of dollars in losses (R. Verma et al., 2018).

**How Phishing is conducted**

An attacker conducts a successful phishing in 5 stages (Anjum & Shabut et al., 2016):

Stage 1: **Planning and Setup**: the attackers identify the target, digs out the essential details regarding their target, then set up the attacks to redirect the victim to the fraudulent URL.

Stage 2: **Phishing**: The attackers disguise themselves as some reputable organization, attract victim(s) and request confidential information from them.

Stage 3: **Break-in/Infiltration**: The victim clicks on the malicious link and either a malware that allows the attacker to access the device automatically installs on his device or the victim is redirected to a URL.

Stage 4: **Data Collection**: As soon as the attackers gain access to the victim’s system, they extract the required data.

Stage 5: **Break-out/Ex-filtration**: Once the attacker has access and gained the required information, they remove all the evidence then track the degree of success of their attack to refine their future attacks

## Related Works

Research on phishing detection and prevention mainly explores four areas; automating phishing detection, providing user interface cues to help users detect phishing, educating users about protecting themselves and understanding user's vulnerability (Alsharnouby, Mohamed & Alaca, Furkan & Chiasson, Sonia. 2015). This research mainly explores automating phishing detection and providing user interface cues to help users to detect phishing.

### Automated phishing detection

Automatic phishing detectors exist at various levels such as web browser tools, internet service providers and mail servers and clients (Alsharnouby et. al. 2015). These tools restrict access to detected websites or request the website's internet service provider to take it down. Automatic email classification tools make use of machine learning algorithms, spam filter techniques and statistical classifiers to identify potential phishing messages. They have a varying degree of effectiveness and misclassifications are a common occurrence which affects the reliability of the service as users are likely to be intolerant to loosing legitimate messages.

Automated phishing detection techniques to detect phishing websites include the use of blacklists and whitelists, the use of heuristic methods, and use of machine learning principles. (Anjum & Shabut et al., 2016)

#### Blacklist and whitelists Approach

This method often maintains a list of URLs that are labelled as malicious or benign. Blacklists are essentially a database of URLs that have been confirmed to be malicious in the past. Whenever a new URL is visited, a database lookup is performed. The URL is checked whether it exists on the list and if it is found the label allocated to it is returned as output. A major problem with this method is the inability to maintain a list of all possible malicious URLs as new URLs can be easily generated daily, thus making it impossible for them to detect new threats. This is a critical concern when the attackers generate new URLs using algorithms, and can therefore bypass all blacklists as the URLs are dynamic (Doyen, Chenghao & Steven, 2019)

PhishTank is a popular blacklist launched in 2006 and has been in service ever since (Phishtank, n.d.). The blacklist is populated through crowdsourcing volunteers who submit potential phishing websites and vote on the legitimacy of websites. PhishTank is not protection. "PhishTank is an information clearinghouse, which helps to pour sunshine on some of the dark alleys of the Internet. PhishTank provides accurate, actionable information to anyone trying to identify bad actors, whether for themselves or for others" (Phishtank, n.d.). PhishTank offers a blacklist for use by other tools through an API. Popular organizations such as Kaspersky, MacAfee, APWG and Avira make use of Phishtank's dataset.

#### Heuristic Approach

Heuristic approaches are similar to blacklist methods since their basic idea is to create a blacklist of signatures. When common attacks are detected, a signature is assigned to the type of attack. The idea is to look out for a signature of malicious activity such as unusual process creation, repeated redirection etc. Intrusion detection systems are able to detect these behaviors and respond to them appropriately. These approaches are able to detect new threats but to a limited extent since new threats may be completely unrelated. Modern heuristic methods analyze the execution dynamics of webpages. They require visiting the actual URL which may initiate the attack. The techniques are resource intensive and require complete execution of the code (including the server-side scripts). The techniques may go undetected since the malware in place may not launch the attack immediately (Sahoo, Liu & Hoi, 2019).

#### Machine Learning Approach

These are intelligent heuristic-based methods which try to analyze the information of a URL and its corresponding webpages by extracting the features of URLs and training a prediction model using training data. In static analysis, the website is analyzed based on the features extracted from the URL string such as lexical features, information about the host, and sometimes the HTML and JavaScript content. The underlying assumption is that the distribution of these features is different for malicious and benign URLs. Using this distribution information, a prediction model can be built, which can make predictions on new URLs. Since no execution is required, they are safer than dynamic approaches which require complete execution of the URL. Dynamic methods monitor the behavior of the websites looking for anomalies. (Sahoo et. al., 2019)

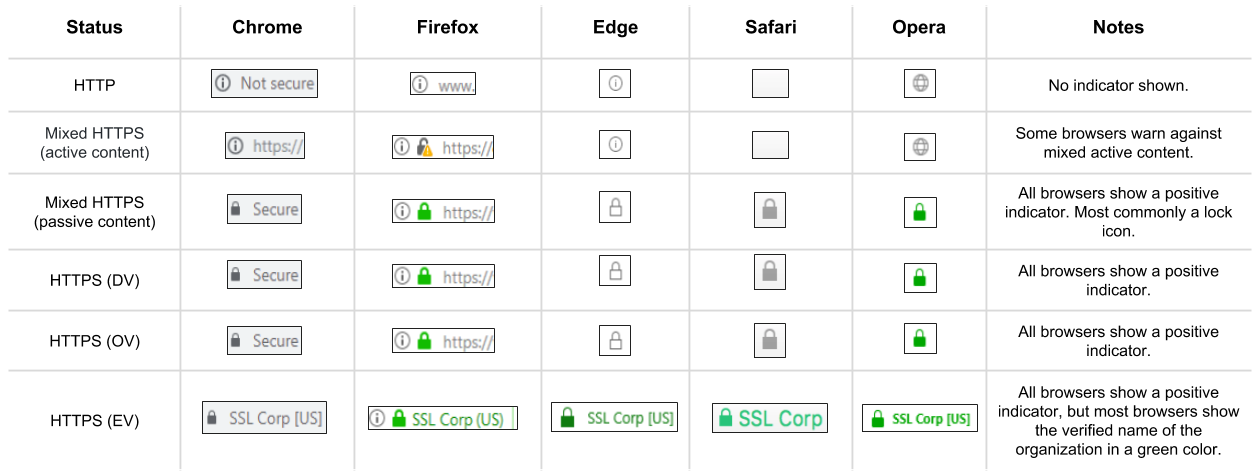
## Security Indicators

Phishing is primarily a problem because users are unable to verify the authenticity of a website. Security indicators are readily available every time a user launches the web browser. However, an average internet user rarely pays attention to them or may not understand their meaning. In general, most authors consider URL bar HTTPS indicators (SSL/TLS) and digital certiﬁcate indicators as the main cues (Jelovčan, Luka, Vrhovec, Simon, Mihelič, Anže. 2020)

Hypertext Transmission Protocol (HTTP), the most widely used protocol in the world, is the protocol that is used to view web pages. When you type an address, say www.example.com, HTTP is added automatically at the beginning of the address, ie http://www.example.com. HTTP sends and retrieves all data in clear text making it vulnerable to anyone who wants it, eg hackers. Secure Hypertext Transmission Protocol (HTTPS) is used to secure communications by encrypting the data exchanged between a person’s browser and the web site he or she is visiting. HTTPS is especially important on sites that offer online sales or password-protected accounts. Browsers indicate that a website uses HTTPS by use of a padlock (colored in some browsers). (Jelovčan et. al., 2020 & APWG 2021)

Secure Hypertext Transmission Protocol (HTTPS) uses Secure Sockets Layer (SSL), a protocol used to ensure security on the internet using public key encryption. When a computer connects to a website using SSL the computer's browser will ask the website to identify itself. The web server responds by sending the computer a copy of its SSL certificate, a small digital certificate used to authenticate the identity of a website. Once the browser establishes trust with the website, encrypted data can be transmitted to and from the website. (Jelovčan et. al., 2020)

The following table is a summary of the general state of security indicators in modern browsers. Starting with HTTP (which is not secure at all) each item further along the list is more secure than the previous ones. (Naziridis, 2018)



To help users stay safe on the internet, browsers require websites to use certificates from trusted organizations. This is because anyone can create a certificate (e.g., Using OpenSSL) claiming to be whatever website they claim to be. According to APWG (2021), PhishLabs, an active contributor to APWG, has been tracking the proportion of phishing sites that are protected by the HTTPS encryption protocol. Studying HTTP on phishing sites provides insight onto how phishers are fooling Internet users by turning an Internet security feature against them. 83% of the phishing attacks reported to APWG in the first quarter of 2019 used HTTPS protocol.

## Supervised Machine Learning

Machine learning concentrates on developing the computational algorithms that reason and induce patterns and rules from externally supplied instances and prior data in order to produce general models, which are able to make predictions about future instances. The machine learning is called supervised if known labels are given with instances in the training phase, whereas instances are unlabeled in unsupervised machine learning. (D. Sahoo et. al 2019.)

The phishing website can be detected based on some important characteristics like URL and Domain identity, and security and encryption criteria in the final phishing detection rate. According to APWG report for 2021, Phishers continue to use certain domain name registrars to obtain domains for their schemes. In a URL, lexical features can be extracted such as the URL string, information about the host, and sometimes even HTML and JavaScript content. Host-based features obtained from the hostname properties of the URL allow us to know the location, identity, and the management style and properties of malicious hosts. An underlying assumption is that there is an array of features to differentiate malicious and benign URLs. Based on this information, a prediction model can be built, which can make predictions on new URLs. This can be formalized as a binary classification task of a machine learning algorithm. (D. Sahoo et. al 2019.)

### Machine Learning Algorithms

Several machine learning algorithms have been implemented in real life applications. (Ali, Waleed, 2017). The identification of malicious and legitimate URLs can be regarded as a classification task and some of the popular machine learning algorithms applicable include the following:

1. **Back-Propagation Neural Network (BPNN)**

BPNN are popular algorithms in network models. They are particularly used in prediction and classification problems. They learn in two phases: the forward pass and the backward pass. In the forward pass, the input layer is presented with a training input pattern which is propagated from layer to layer until the desired output is produced. In the backward phase, the output is compared with the anticipated output in order to compute the error. The error is then propagated backward through the network from output to input layers and the weights are adjusted accordingly to minimize the error (Ali et. al., 2017).

1. **Radial Basis Function Network (RBFN)**

RBFN is a type of neural network that uses radial basis functions as activation functions. In the architecture of RBFN, there are three feedback networks: the input layer, the hidden layer and the output layer. In each hidden unit, a radial activation function is implemented while a weighted sum of outputs of hidden units is implemented for each output unit. Learning is conducted in two phases. The first stage involves clustering in order to determine the centers and widths of the hidden layer. In the next phase, the weights connecting the hidden layer with the output layer are optimized through the use of Least Mean Squared (LMS) or Singular Value Decomposition (SVD) algorithms. (Ali et. al., 2017).

1. **Support Vector Machine (SVM)**

SVM, very popular and robust machine earning techniques have been utilized effectively in many applications. They are based on maximizing the margin and thereby creating the largest possible distance between the hyperplane and the instances in order to reduce an upper bound on the anticipated generalization error. Support vectors close to the hyperplane provide the most useful information for classification. An appropriate kernel function is used to transform the data into a high-dimension to use linear discriminate functions (Ali et. al., 2017).

1. **Decision Tree and Random Forest**

In decision trees, a node corresponds to a feature of an instance being classified. The instances are classified through sorting based on feature values. Each branch represents a value that the node can predict. Random Forest is a popular decision tree that can be used for classification and regression. RF is a group of decision trees trained independently on selected training datasets. The classification is then determined by voting among all the trained decision trees. (Ali et. al., 2017).

The performances in terms of correct classification rate (CCR) of the above algorithms were compared together in a study (Ali et. al., 2017). The following table summarizes the results.

|  |  |
| --- | --- |
| **Classifier** | **Correct Classification Rate** |
| Back-Propagation Neural Network (BPNN) | 0.970 |
| Radial Basis Function Network (RBFN) | 0.928 |
| Support Vector Machine (SVM) | 0.963 |
| Random Forest (RF) | 0.971 |

*Fig: Performance measures of the machine learning classifiers*

The Random Forest and Back-Propagation Neural Network classifiers achieved the best correct classification rate while the Radial Basis Function Network attained the lowest.

## Features Extraction

Several features can be extracted from a website to distinguish phishing websites from legitimate ones. Feature selection is necessary in order to decrease computation time and to reduce noise and irrelevant features. The choice of extracted features is critical for the success of the detection mechanism in place. Once the selected features are selected, the machine learning model can be trained. (Ali et. al., 2017)

The following features can contribute to the effective prediction of the phishing websites: (Ali et. al., 2017).

|  |  |
| --- | --- |
| Feature Category | Feature Name |
| Address bar-based features | Using the IP Address  Long URL to Hide the Suspicious Part  Using URL Shortening Services “TinyURL”  URL’s having “@” Symbol  Redirecting using “//”  Adding Prefix or Suffix Separated by (-) to the Domain  Sub Domain and Multi Sub Domains  HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)  Domain Registration Length  Favicon  Using Non-Standard Ports  The Existence of “HTTPS” Token in the Domain Part of the URL |
| Abnormal-based features | URL of Anchor  Links in <Meta>, <Script> and <Link> tags  Server Form Handler (SFH)  Submitting Information to Email  Abnormal URL |
| HTML and JavaScript-based features | Website Forwarding  Status Bar Customization  Disabling Right Click  Using Pop-up Window  IFrame Redirection |
| Domain-based features | Age of Domain  DNS Record  Website Traffic  Page Rank  Google Index Number of Links  Pointing to Page  Statistical-Reports Based Feature |

## Why Phishing Still works

According to Retruster 2019, the problem lies in the detection and reporting of cybercrimes. It can take as long as 50 days from when a breach is discovered until the time when it is reported, a very huge risk for potential victims.

Users consider security as a secondary task. They are prone to concentrating on the real purpose of their interaction with their website making it unlikely for them to notice the security indicators displayed. Some security indicators are also only visible when visiting safe and secure websites (Alsharnouby, et. al., 2015). In a study conducted to assess whether browser security indicators and increased user awareness on phishing have led to users’ improved ability to protect themselves from phishing, a series of websites was presented to participants and they were asked to identify phishing websites. Participants were successfully able to detect only 53% of phishing websites even when forewarned to identify them. Using eye tracking, they found that two thirds of users looked at the SSL lock icon when prompted to be security-conscious but rarely used other cues on the browser-chrome (Alsharnouby, et. al., 2015). Users spend 85% of their time looking at the website content during a web interaction and only 6% of their time looking at security indicators. Even if users doubt the authenticity of the websites, they will still access it, primarily because they want the beneﬁts from it (Jelovčan et. al., 2020).

One major problem with cybercrime is establishing jurisdiction. Physical crimes are bound to a physical location. The crime is therefore considered territorial and its location determines the jurisdiction. Cybercrime activities are not bound to a physical location as the victim and the perpetrator can even be physically at different countries. To worsen the situation, some countries may not have extradition treaties and the law may therefore not be able to prosecute the perpetrators. Legislation of laws on cybercrime is still developing, and it may be challenging to bring justice to those who commit cybercrimes (Plachkinova, 2021).

## Conclusion

A-lot of research has been done in the attempt to eradicate phishing. The protection mechanisms in place such as HTTPS and the inbuilt browser indicators are insufficient in detecting threats and notifying the users. Therefore, developing a more effective browser indicator to alert users when they are in danger remains a crucial and unsolved problem in security. In the attempt to solve this problem, a browser extension bundled together with a machine learning model to evaluate the authenticity of a URL should be implemented.

A URL can be broken down into lexical elements which forms the basis of training a machine learning model. With a dataset containing URLs labelled as malicious or benign, the machine learning model can be trained and evaluated to classify new URLs. On this basis, the classification task involves using a multidimensional dataset since various attributes of the dataset are unevenly distributed. Logistic regression and Naïve Bayes algorithm, chosen for their simplicity and effectiveness in classification will be used in the proposed system to form a classification model for new URLS.

In summary, users being the weakest link to phishing campaigns, need to be educated on cybercrimes such as phishing, the potential consequences and ways in which they can protect themselves. Users should be highly discouraged from depending on their instincts as a protection mechanism. Users need to make use of the available automated cyber-crime protection tools.

# Research Methodology

This chapter covers the research methods and design used in this project. The preferred methodology in this research uses a qualitative approach, a popular and well proven methodology to implement machine learning models. This approach involves data collection, data pre-processing, model training, model testing and model evaluation and can be summarized as:

Data Collection

Model Testing

Model Evaluation

Model Training

Data Preprocessing

Model Deployment

## Data collection

This is a crucial and important step in conducting this research as it entails the acquisition of the data required in training and testing the model. In this particular research, the data required includes a dataset composed of legitimate and fraudulent websites acquired from a reliable source. The dataset in this case will be acquired from two sources, Kaggle and Phishtank. These will then be combined to form a single dataset. Phishtank updates its dataset hourly and is available in multiple formats making it easy to access and utilize (http://data.phishtank.com/data/online-valid.json). Kaggle also provides a reputable dataset containing a list of websites labelled as legitimate or fraudulent.

## Data pre-processing

This phase includes cleaning the data and preparing the data for model training. Data cleaning involves identifying and correcting errors in the dataset that may negatively impact the predictive model. There are many types of errors that exist in a dataset such as missing values and structural errors. Such malformed records as well as duplicated entries will be deleted to make the data more reliable and to produce a robust model. Since the data is acquired from a secondary data sources, the data needs to be structured in the format that can be used by the model.

## Model training

In this step, a model of the chosen algorithm will be developed and trained. This will involve extracting the various features of a URL through tokenization. Features extracted from various classes of URLs will then be used to set a classification basis for the model. This is necessary as the weights of the model must be initialized manually so as to enable the algorithm to learn to adjust to the weights accordingly. Consistent training will significantly improve the prediction rate of the model.

## Model testing

This phase will involve validating the model to ensure that it performs as expected, in line with the design objectives. This will make use of the data set reserved for the testing phase.

## Model evaluation

The model will then be evaluated to establish how accurate it is in terms of classifying new URLs. This is necessary in order to assess the accuracy of the model.

## Model deployment

The model will finally be expose as an online REST service, so that it can quickly be called upon to make prediction requests for new incoming URLs. This will make it possible to employ the model in applications through restful API calls.

The model development and evaluation steps can be summarized as:

DATA CLEANING

DATA INPUT

DATA PREPARATION

MODELLING

SPLIT DATA (TRAIN AND TEST)

FEATURE SELECTION

MODEL TRAINING

MODEL TESTING

CONCLUSION

MODEL DEPLOYMENT

## System Implementation

The machine learning algorithms that will be implemented in this research include Logistic Regression and Naive Bayes. The models will be built using Python programming language, the ideal language for machine learning modelling. This is because there are multiple libraries in Python which facilitate various aspects of the machine learning modelling process. Some of the tools and libraries used will include:

**Jupiter Notebook** - An open-source, interactive web tool used by researchers to combine software code, computational output, explanatory text and multimedia resources in a single document.

**Pandas** - A fast, powerful, flexible and easy to use open source data analysis and manipulation tool Python library for data analysis.

**Numpy** - A Python library that provides a simple yet powerful data structure: the n-dimensional array. This facilitates efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

**Matplot** - is a plotting library for Python that provides an object-oriented API for embedding plots into applications.

**SeaBorn** - An open-source Python library built on top of Matplot used for data visualization and exploratory data analysis. SeaBorn works easily with dataframes and the Pandas library.

**Fast API -** A modern, fast (high-performance), web framework for building APIs with Python.

# System Analysis

This chapter covers the feasibility study, requirements specification, system constraints and the use case diagrams for the proposed system.

## Feasibility Study

This section covers the technical, economic, operational and social feasibility aspects of the proposed solution. This is necessary in order to assess the viability, practicality and effectiveness of developing the proposed solution

### Technical Feasibility Study

Considering the hardware and software facilities required to develop and deploy the system, the platform is technically feasible. The hardware and software resources required to develop the system are immediately available and easily accessible. The infrastructure needed to host and run the system are available on cloud platforms.

The technologies required to develop the Machine Learning model are stable, well documented, and have a large community support to provide consultations in case of enquiries.

Use of cloud infrastructures facilitates the quick scaling, both up and down, of computing resources whenever needed to cater for the fluctuating number of user requests.

### Economic Feasibility Study

The use of cloud computing services to host and run the system reduces the maintenance costs substantially. The only limiting factor is the cost incurred to the cloud service provider in order to host the system on a cloud premises.

The softwares and technologies required to develop the system are freely available and accessible on the internet.

Due to the high automation in the system, it will require very little labor to run and maintain. This ensures that there are minimal costs in securing human resources.

### Operational Feasibility Study

With the use of cloud technology, the requirement to maintain the infrastructure upon which the system is deployed is outsourced to the cloud provider. This reduces the costs of maintenance of the system substantially. The system requires minimal human intervention in order to operate smoothly. Deploying the system on a cloud platform guarantees that the product will be available at any time over the internet. Cloud platforms also facilitate smooth transitions when upgrading to newer versions of the system.

### Social Feasibility Study

The sole purpose of this system is to automate phishing websites detection for end users with minimal user input. This system will alert users in the event of potential threats thereby providing real time protection on every URL visit. Users will easily be able to integrate the system onto the supported browsers therefore reducing the complexity of using the system. Users will also easily adapt to the system.

This solution ensures a viable, beneficial and acceptable solution to the users therefore guaranteeing that the system is socially feasible.

## Requirements Specification

### Functional Requirements

1. The system should be able to classify a URL as legitimate or fraudulent based on the training dataset.

2. The system should be able to train the model based on the given data set.

3. The system should alert users when they visit a URL that identifies as a potential phishing website on their browser.

4. The system should allow users to input URLS, scan for their authenticity and finally give the appropriate feedack.

5. The system should keep track of URL lookups performed by users.

6. The system should have a commenting sections on the online platforms encourage users to interact.

7. The system should facilitate user account management in order to keep track of user contributions to the platform.

8. The system should be indexed by google search.

### Non-Functional/Quality Requirements.

1. The system should be accessible at all times over the internet.

2. The system should be accessible across different devices as a website.

3. The application should be easily installable as a browser extension.

4. The system should be able to scale resources with increasing traffic.

5. The system should be secure to prevent stealing user information.

5. The application should not take not more than 10 seconds to alert users when a potential threat is detected.

6. The user interface should be able to scale across different screen sizes.

7. It should be easy to perfom URL scans on the website

8. The system should keep an audit trail of unsuccessful user requests.

### Hardware Requirements

The following are the hardware requirements for developing the proposed system:

\* Processor - 500MHz and above

\* RAM - 8GB

\* Hard Disk - 500GB (at least 1 GB reserved for this project)

\* Input Devices (Standard Mouse, Keyboard)

### Software Requiremnts

The following are the software requirements for developing the proposed system:

\* OS - Linux / Windows

\* Platform : Jupyter Notebook

\* Language : Python

\* IDE/Tool : Anaconda

## System Constraints

The system is limited to the following constraints:

\* In order to access the system, the end user is required to have a reliable internet connection to facilitate communication across the various interoperating services.

\* The browser extension made will be based on the chromium browser standardards hence it might not be compartible with every browser. Android users will be required to use a specific subset of browsers in order to integrate the browser extension to their smartphones/tablets

## Use Case Diagram

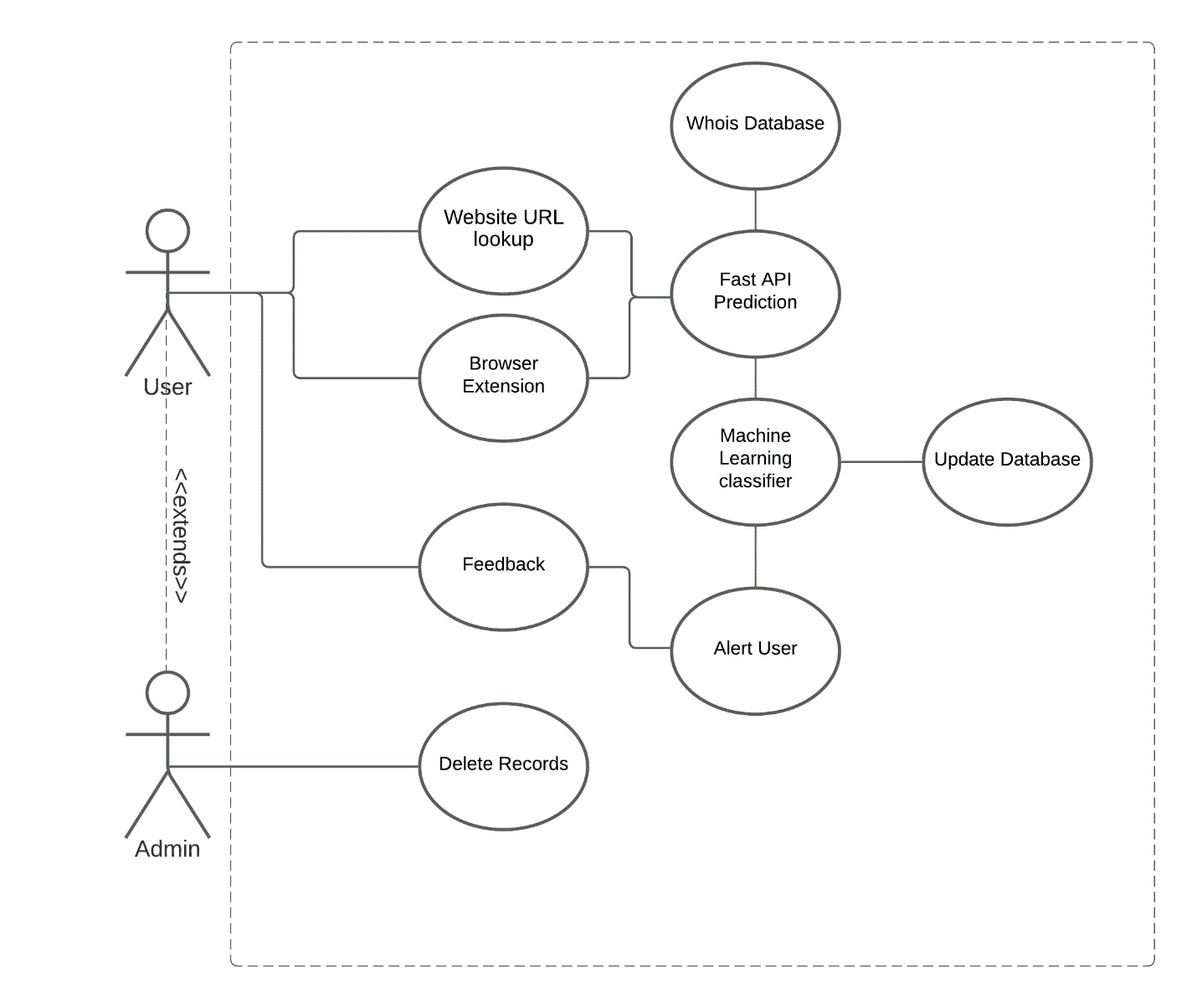


Figure 4.1 Use case diagram

# System Design

## Conceptual Design

The proposed algorithm follows a supervised learning model architecture. Labelled data (URLs) will first be loaded into the system. The favorable features will be extracted and processed based on the nature of the data used. The extracted features will then form the basis for the classifier model. The unlabeled data (new URLs) will then be loaded to the classifier model for classification. Finally, new data will be classified as either malicious or benign.

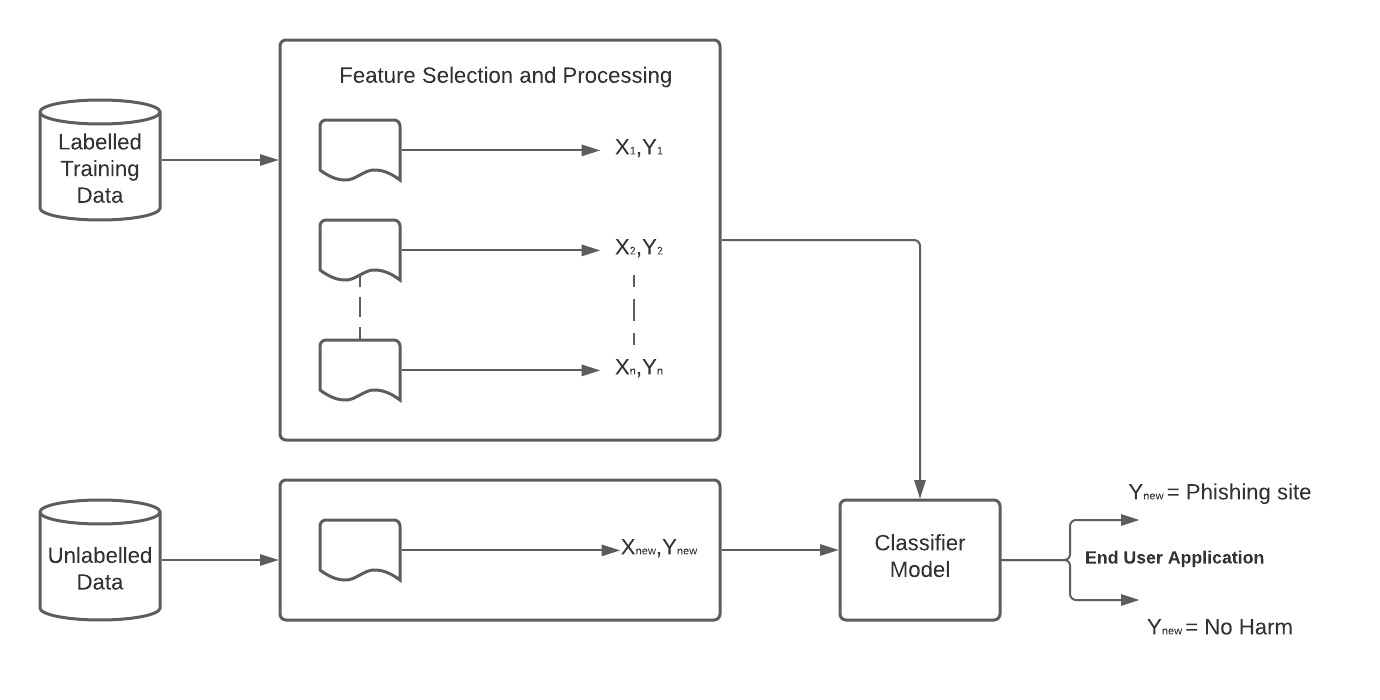


Figure 5.1 Conceptual Design

## Architectural Design and System Components

The proposed phishing detection and prevention system will be built upon a Microservices architecture. In Microservices architecture, a system is decomposed into loosely coupled services which can be developed, deployed and maintained individually. Each of these services is responsible for a discrete task and communicates with other services through simple APIs to solve a larger complex problem.

The proposed system will be composed of the following components.

* 1. Browser Extension – This module will add extra functionalities to supported browsers in order to inform users on the status of the URL that they intend to interact with. This component will be an installable package that will be publicly available for download from the website.
  2. Website – This module will enable users to perform URL scans without necessarily installing any extra packages on their devices. This will also give users information regarding the URLs being scanned.
  3. Machine learning model – This module will facilitate the binary classification of incoming URLS as previously trained and tested.
  4. Fast API Gateway `
  5. Relational Database – This module will serve as the central reservoir to store, maintain and manage data necessary for the various components.

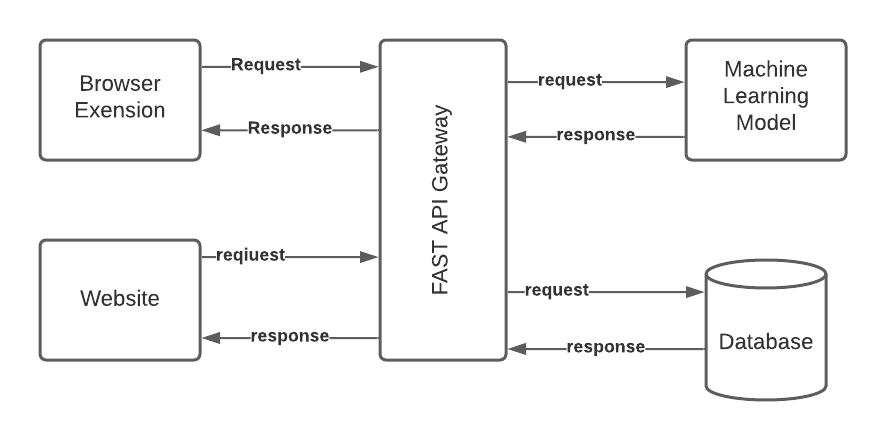


Figure 5.2 Architectural Design

## Database Design

**Schema**

Figure 5.3 Database Schema

## User Interface Design

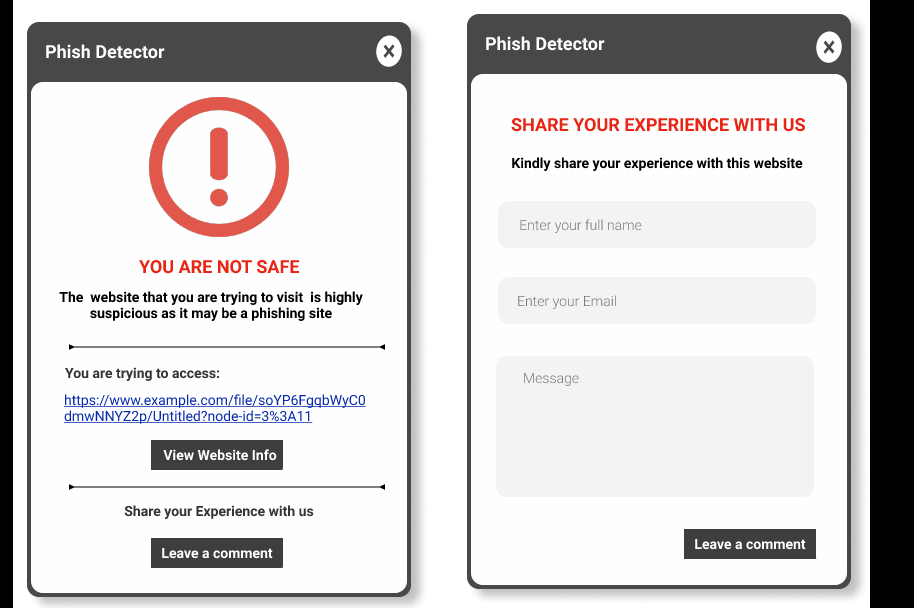


Figure 5.4 Browser Extension alert screen

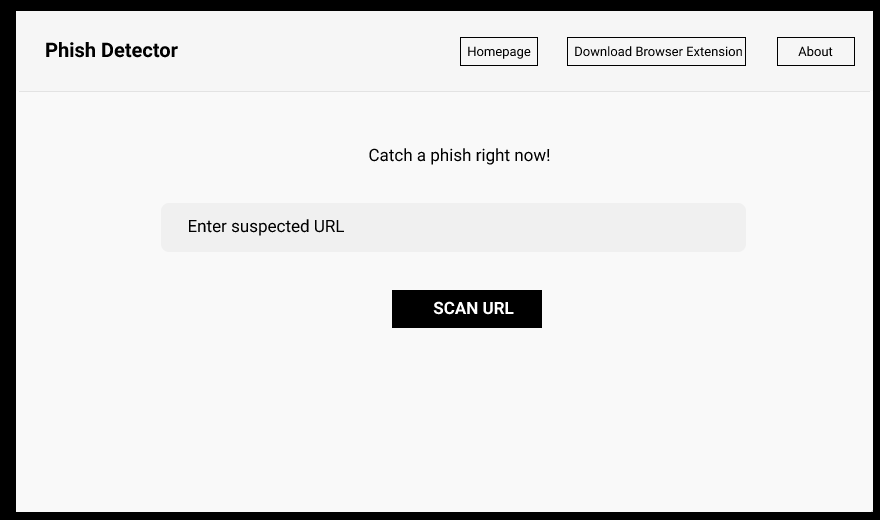


Figure 5.5 Website Homepage



Figure 5.6 Website sample scan results

### ACTIVITY DIAGRAM

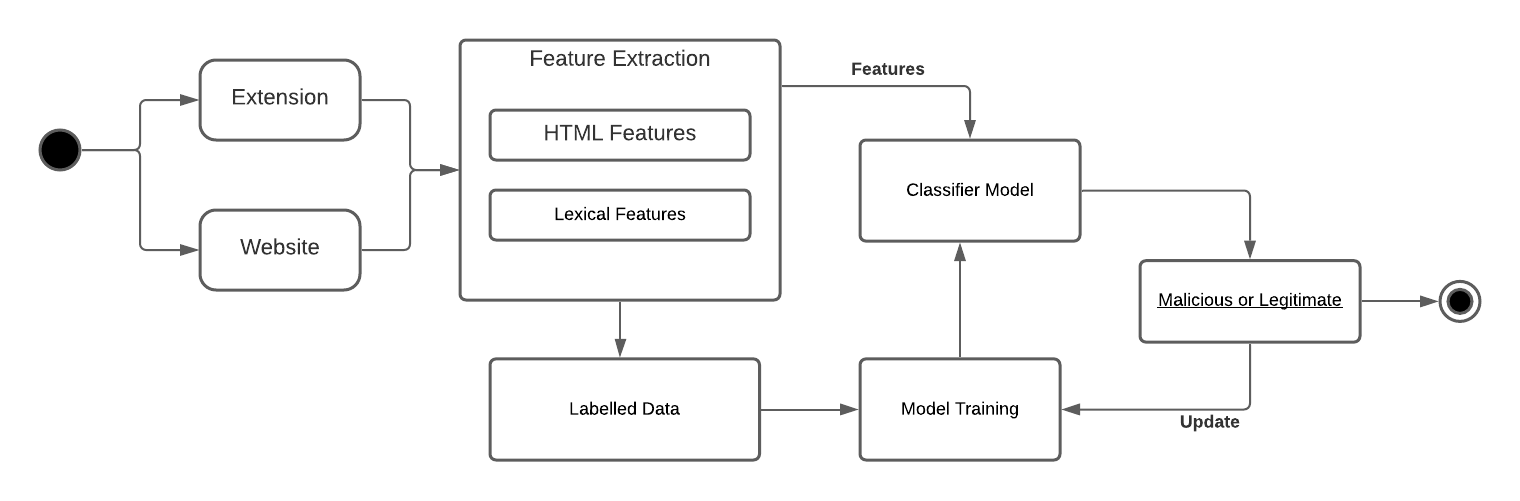


Figure 5.6 Activity Diagram

# System Implementation & Testing

## User Interface Design

Users of the system are expected to interact with the system through the chrome extension or the website.

### Website

The website is a collection of two webpages, the homepage and the scan results page. The website is a Laravel driven application that interacts with the database to hasten the acquisition of prediction results made by the model. This is achieved by storing URLS that have been predicted by the model which is availed by restful API calls to the Fast API gateway

**Home Page**

The homepage enables the user to insert a URL into a search box. This URL is sent over the Fast API gateway,

**[--SCREENSHOT--]**

**Scan Results Page**

This page gives the user the results of the prediction made by the model

**[--SCREENSHOT--]**

### Chrome Extension

This application displays as a popup window on the browser informing the user on the status of the webpage that they are currently visiting. This works by making restful API calls to the Fast API gateway.

**[--SCREENSHOT--]**

## System Testing and Evaluation

The system was evaluated through unit testing involves testing every individual module in the system.

### Prediction model evaluation

### User Interface Testing

# Conclusion

# Recommendations

# References

Rakesh V., Devin C., and Omprakash G., (2018). Phishing During and After Disaster: Hurricane Harvey. In 2018 Resilience Week (RWS). IEEE, 88–94

Anjum N., Antesar M. & M.A Hossain. (2016). A literature review on phishing crime, prevention review and investigation of gaps. (pp. 9-15, 10.1109/SKIMA.2016.7916190)

Phishtank (n.d.). Frequently Asked Questions: Phishtank, out of the net into the tank. Accessed August 1st 2021 from https://phishtank.org/faq.php

Retruster. 2019. Phishing Statistics and Email Fraud Statistics for 2019, Accessed 6 August 2020, (https://retruster.com/blog/2019-phishing-and-email-fraud-statistics.html)

Ali, Waleed. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. International Journal of Advanced Computer Science and Applications. Page 8, doi: 10.14569/IJACSA.2017.080910.

APWG (2021). Phishing Activity Trends Report: 1st Quarter 2021. Unifying the Global Response to Cybercrime. ([www.apwg.org](http://www.apwg.org)).

Sarno, D. M., Lewis, J. E., Bohil, C. J., & Neider, M. B. (2019). Which Phish Is on the Hook? Phishing Vulnerability for Older Versus Younger Adults. Human factors. Doi: 10.1177/0018720819855570

Mohammad R., Thabtah F., McCluskey L. (2015). Tutorial and critical analysis of phishing websites methods. Computer Science Review Journal. Volume 17, Pages 1–24.

Doyen S. and Chenghao L. and Steven C. H. Hoi, 2019. “Malicious URL Detection using Machine Learning: A Survey”. Eprint 1701.07179, arXiv, cs.LG

Jelovčan, Luka, Vrhovec, Simon, Mihelic, Anze. (2020). A literature survey of security indicators in web browsers. Elektrotehniski Vestnik/Electrotechnical Review. Doi 87. 31-38.

Naziridis N. (2018). A look at browser UI security indicators. Browser Indicators, HTTPS, SSL validations, The HTTPS-Only Standard, UI. (https://www.ssl.com/article/a-look-at-browser-ui-security-indicators/)

Plachkinova M. (2021). Exploring the Shift from Physical to Cybercrime at the Onset of the COVID-19 Pandemic, Journal of Cyber Forensics and Advanced Threat Investigations. [Doi: 10.46386/ijcfati.v2i1.29](https://doi.org/10.46386/ijcfati.v2i1.29).

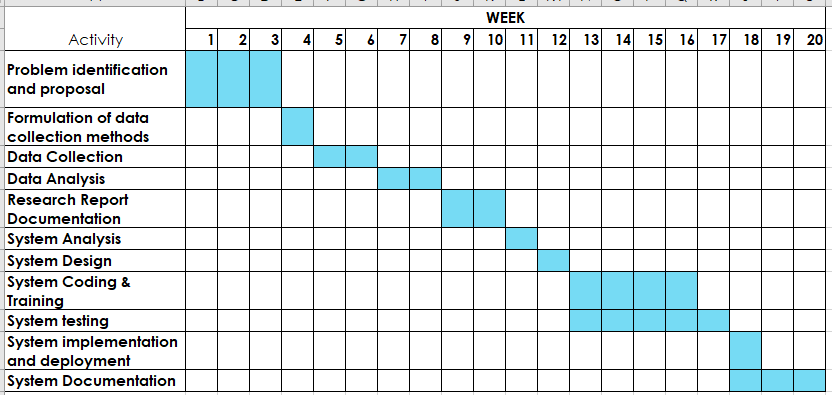
Alsharnouby, Mohamed & Alaca, Furkan & Chiasson, Sonia. (2015). Why phishing still works: User strategies for combating phishing attacks. International Journal of Human-Computer Studies. Doi: 2. 10.1016/j.ijhcs.2015.05.005.

# Appendices

## Appendix 1. Schedule

The aforementioned research with be carried out in phases during the following estimated timelines

|  |  |
| --- | --- |
| Task | Timeline |
| Problem identification and proposal | 3 Weeks |
| Formulation of data collection methods | 1 Week |
| Data Collection | 2 Weeks |
| Data Analysis | 2 Weeks |
| Research Report Documentation | 2 Weeks |
| System Analysis | 1 Week |
| System Design | 1 Week |
| System Coding | 5 Weeks |
| System testing | 1 Week |
| System implementation and deployment | 1 Week |
| System Documentation | 2 Weeks |



**Fig 1: *Gantt Chart – Project Schedule***

## Appendix 2. Budget

Below is an estimation of the expected expenses that will be incurred

|  |  |  |
| --- | --- | --- |
| ***Service*** | **Cost** | **Justification** |
| ***Hosting Space and Domain Name Registration*** | 5,000/- | The system will require a hosting space on the World Wide Web for testing and deployment |
| ***Stationery and Printing of material*** | 2,000/- | Printing documentation, binding reports, cost of notebooks and pens |
| *Total* | **7,000/-** | **-** |