Department of Computer Science,
University of Birmingham

An intelligent anti-phishing approach for Fraudulent URL detection using ML Model

Student ID - 2318246

UNIVERSITY OF BIRMINGHAM

Project Dissertation
submitted to the University Of Birmingham in fulfilment of degree of
MSc Advance Computer Science

Acknowledgement

This project is being undertaken by myself, Rajiv Kulkarni. Throughout this document, all of its content is work produced by myself, except otherwise stated. I want to give big thanks to Mr Rajesh Chitnis for being my supervisor and providing continuous feedback throughout the project. I want to thank Mr Paul Blain Levy for his constructive feedback on inspection.

Index

List of Abbreviation

1. URL - Uniform Resource Locator
2. DSN - Domain Name System

ABSTRACT

Phishing is a dishonest effort to obtain your personal information, mainly by email. Phishing is growing more harmful, and detection is critical. It is one of the social engineering strategies that acquire personal information from a corporation or an individual by posing as a trustworthy firm or organization using websites such as harmful websites and deceitful e-mail. Phishing often targets email by utilizing it as a vehicle and even delivering communications over email to users that represent a company or an entity that conducts business, such as a financial institution or a bank (2022). In order to get a better knowledge of the structure of URLs that promote phishing, we analyzed numerous data mining techniques for feature assessment. The fine-tuned parameters aid in the selection of the best machine learning method for distinguishing between phishing and benign sites. (RRPhish: Anti-phishing via mining brand resources request - researchr publication, 2022)

Keywords-*Phishing attack; security; anti-phishing*

Chapter 1

INTRODUCTION

OVERVIEW

What is Phishing?

First and foremost, the phisher must establish a phishing website in order to entice a victim who seems to be authentic [[(2022)]]. The site should then be hosted on the internet for the usage of victim confidential information. When a victim visits a phishing website, the victim is persuaded to input sensitive information. The phisher then obtains some inputted data, which he or she may later exploit.

Host-based analysis

Host-based characteristics describe "where" phishing sites are housed, "who" manages them, and "how" they are handled. [[(Gregory Paul and Gireesh Kumar, 2022)]] We employ these characteristics because phishing Web sites may be hosted in less trustworthy hosting facilities, on computers that are not typical Web hosts, or via shady registrars.

The following figure depicts the block design for the host-based analysis.

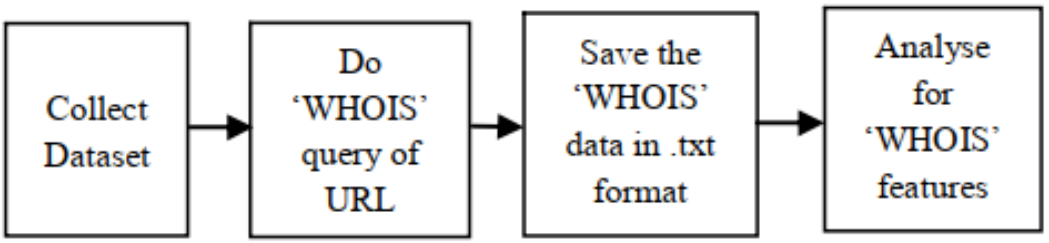


Figure: - Block diagram for host-based analysis

WHOIS

WHOIS features include information like as the date of registration, update, and expiration, as well as the registrar and registrant. If phishing sites are constantly pulled down, the registration dates will be fresher than those of authentic sites. Many phishing websites include an IP address in their hostname [(2022)]. So obtaining the data of such hostnames, which may be acquired through the Whois features, would be beneficial in attempts to point to phishing sites.

We want to exploit the WhoIs attributes of URLs to identify phishing websites. To identify phishing websites utilising URL and WhoIs information, we offer a unique approach, Phishing Detection Using Soft Computing and Machine Learning [(2022)]. The convolution Neural Network is used to train the network and ultimately determine whether or not the site is phishing. In a phishing attempt, attackers may obtain background information about the victim's personal and professional history, hobbies, and activities by using social engineering and other public information resources, such as social networks like LinkedIn, Facebook, and Twitter [(Gregory Paul and Gireesh Kumar, 2022)]. With this pre-discovery, attackers may learn the names, job titles, and email addresses of prospective victims, as well as the identities of important individuals in their colleagues' and employers' organizations.

Phishing may also be used to get a person's password or credit card information [[(Afroz and Greenstadt, 2022)], (2022)]. Computer users are driven to phony sites using e-mail that seem to be from a bank or official entity.

The following are examples of typical information stolen via a phishing attack:

- User IDs, Passwords, User Names, and Credit Card Details
- Specifics about Online Banking

AVOIDING PHISHING ATTACKS

In the context of phishing detection, the term "whitelist" refers to a simple list of websites that may be relied upon. More than simply, the address of the trustworthy website should be included on the list if URL detection is to function properly. Every item in the whitelist database has six strings: the URL of the trusted website, the domain of the website, the title of the website, the filename, and the URL content of the file. These strings are separated by commas. (2022)

1. The URL of the trusted site:

When the information in the database has to be updated at regular intervals, the URL of the reliable website is utilized. This is the Uniform Resource Locator (URL) of the website, such as "https:signin.ebay.com."

2. The domain of the site:

The trusted site's domain is the domain of the URL, such as "signin.ebay.com," and it is used to assess whether or not the current page shown in the browser is on the whitelist.

3. The title of the site:

A trustworthy site title, such as "Welcome to eBay," is the page title of the site and may be

used to speed up the matching potential of phishing site names with titles in the whitelist Database.

APPROACH

Because of the complexity of the phishing issue and the lack of a single, foolproof solution, it is common practice to use a combination of countermeasures. The suggested system uses a Convolutional Neural Network (CNN) algorithm to identify as many phishing attempts as possible.

PROBLEM DEFINITION AND OBJECTIVES

The term "phishing" refers to an assault that is more of a strategic operation than a direct one. Phishing Detection through Soft Computing and Machine Learning is a revolutionary approach we suggest for effectively identifying phishing websites using their URL and other attributes. Datasets of blacklisted and whitelisted URLs are utilized as features in machine learning techniques.

Objectives

• To create a pre-processing system capable of extracting a collection of characteristics representing all aspects of the phishing URLs and determining their merits in the detection process.
• Feature Extraction: Using the Whois system, features are extracted.
• Phish tank Dataset: This dataset contains instances of legitimate and phishing websites. In each example, the URL and the matching HTML page are supplied.
• Machine Learning: Using whois, features are retrieved, and then the CNN technique is utilized to identify the Phishing URL.

Project plan:

- Feasibility: The Phish tank dataset is open source available on the kaggle and it is big data with lots of URL data and can give better accuracy as well as detecting phishing/Non- Phishing urls. We also analyzed different techniques to detect phishing URLson state- of- Art.
- Resources: Dataset, Weka Tool for Analysis.
- Gantt Chart with tasks and milestones, reflecting the project objectives/deliverables.

Basic			
Goals		How would it be achieved	Outcome
1.	Study the type of approach I would use to identify a phishing attack	Research online on Related topics on the internet	√
2.	Build the website	Use CSS,HTML	√
3.	Build the CNN model	Study online about CNN and check similar approaches online	√
4.	Test the machine learning model	Run it on test data and on the phishy website list available online	√
5.	Analyze the results	Check the accuracy of the model	√

Chapter 2

LITERATURE SURVEY

A. Schemes based on heuristics: -

Heuristic-based systems rely heavily on characteristics retrieved from URL and HTML source code, which are subsequently validated using other approaches such as machine learning. (2022)

Schemes based on a blacklist —

Blacklist-based methods may identify phishing sites on the blacklist but not zero-day phishing attempts that have been active for days or even hours. New phishing sites may have already stolen user credentials or may have expired before being added to the blacklist.

B. Using an Artificial Neural Network to detect phishing URLs

It is a mechanism for categorizing Uniform Resource Locators (URLs) as Phishing URLs or Non-phishing URLs. Particle swarm optimization and classification training should be used to increase the performance of ANN. A single-layer artificial neural network is used in a dynamic strategy for identifying phishing attempts. This method is used in this study to determine the value of six heuristics in the first phase of the approach.

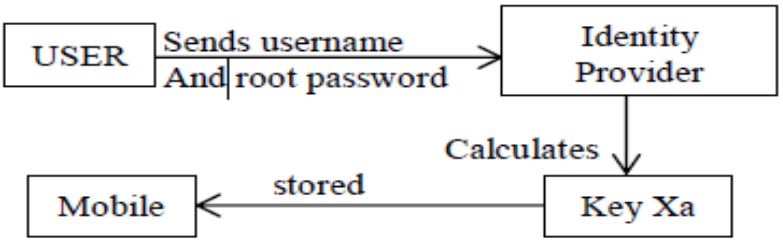
C. Anti-phishing single sign-on approach based on QR Code:

This solution tackles the issue of phishing on SSO authentication. Single sign-on is an authentication mechanism that allows users to access various apps or websites with a single login and password.

The technology use QR codes since they do not need mobile network data to be read and can hold a huge quantity of data. This method is divided into two stages;

- **User Registration Phase: -**

During the User Registration process, the user is given a secret key, which is subsequently utilized during the verification step to get access to the desired service.



• User Verification Phase:

During the verification phase, the user seeks service from the service provider, who transmits the user's identification to the identity provider.

D. Earth's Mover Distance (EMD)

EMD is a measure for comparing the similarity of two probability distributions across an area. The closer the two photos are, the lower the EMD value on employing image hash values in their work [(2022)]. The authors also computed the pHash of a screenshot and assessed the difference between two hash values using the hamming distance. They offer experimental data demonstrating that making a little alteration to the original picture resulted in a minor increase in hamming distance.

E. RRPHISH

Because browsers have a maximum concurrent connection restriction to the same domain, brand sites often host resource material on another domain, such as PayPal, which hosts CSS, JS, and image files on paypalobjects.com.

RRPhish can automatically expand the blacklist, which is a useful addition to the blacklisting approach. Include multiple algorithms of varying complexity for different application contexts, such as heuristic rules or machine learning algorithms. (RRPhish: Anti-phishing via mining brand resources request - researchr publication, 2022)

The summary of reviewed papers is included in following table —

Sr. No.	Title, Conference and Publication year	Description	Advantages and Dis-advantages
1	Mohammad Abu Qbeitah, and Mon-therAldwairi	Two labs are established to analyse samples. For genuine samples, a honeynet is employed. To analyse dynamic samples in a real-world setting.	**Advantage:** **1.** phishing emails can be analyzed in real-time, which is a huge advantage. 2. Experimental studies on the behaviour of potentially harmful samples are conducted. **Disadvantage:** The use of Remnux and similar tools for malware detection is a drawback. Linux distribution with a handy set of tools, known as Ubuntu.
2	Surbhi Gupta,Abhishek Singhal	A technique for categorizing Uniform Resource Locators (URLs) as phishing URLs or non-phishing URLs is provided. For appropriate authentication, an unknown website with an erroneous URL was utilized. Tab napping is a method used by attackers to reveal personal information. Email is one of the methods.	**Advantage:** For training ANN, particle swam op itemization was utilized. **Disadvantage :** The PSO method was utilised to solve the issue and get improved performance.

3	TianruiPeng,Ian G. Harris,YukiSaw a	Text analysis has been used in order to adapt phishing assaults via natural processing. A detection method has been developed that analyses each phrase one at a time and returns true if it contains an assault.	**Advantage:** To boost efficacy, a large number of e-mails are employed. **Disadvantage:** Malicious emails are recognized using an existing blacklist. To overcome the algorithm values, a multinomial function is utilized as a parameter.
4	Muhammet Baykara,Zahit Ziya	The attacker conducts cyber-attacks by sending malicious e-mails in the form of communications. For software reasons, an anti-phishing simulator is provided. Training defensive attacks is a parameter used for assessment purposes. In javascript, the URL from the address bar is processed.	**Disadvantage :** A limited collection of values are discovered using different spam filtering algorithms and various properties provided.
5	WeinaNiu, Xi-aosong Zhang, Guowu Yang, Zhiyuan Ma, ZhongliuZh uo	In the study of machine learning concepts, the Support Vector Machine (SVM) is more successful. CS- SVM collects 23 features and models a hybrid classifier based on them. Cuckoo Search (CS) and SVM have been used to optimise the Radial Basis Function (RBF)	**Advantage :** Cuckoo Search SVM is intended to increase accuracy.

Table 2.1: Literature Survey Table

Chapter 3

SOFTWARE REQUIREMENTS SPECIFICATION

SYSTEM FEATURE 1 (FUNCTIONAL REQUIREMENTS)

- URL Obfuscation phishing attacks
- Avoiding URL Obfuscation phishing attacks
- Avoiding phishing attacks

SYSTEM FEATURE 2 (FUNCTIONAL REQUIREMENTS)

User Module:

- Input URL

Avoiding phishing attacks Model

- The URL of the trusted site.
- The domain of the sites.
- The title of the site.

SYSTEM REQUIREMENTS

Database Requirements

MySQL Server is necessary to build and manage the database. Stored procedures are also used to retrieve and manipulate data.

Hardware Requirements

The minimum configuration required on the server platform.

1: Processor: Intel Core i5 or equivalent
2: RAM: 8 GB
3: Operating System: Microsoft Windows 10 Professional x64
4: Hard disk space: 500 GB internal storage drive

Software Interfaces

1: Eclipse Luna

2: Java JDK 1.8

3: Python

4: MySQL Database

Eclipse Luna

• Eclipse Luna is an open-source community whose initiatives include developing tools and frameworks for general-purpose applications. Eclipse Luna is most often used as a Java programming environment.

• Eclipse Luna is an open source community whose projects aim to provide an open development platform composed of extensible frameworks, tools, and runtimes for developing, delivering, and managing software throughout its lifespan.

• The Eclipse Luna Foundation is a non-profit, member-supported organization that hosts the Eclipse Luna projects and contributes to the development of an open source community as well as an ecosystem of associated goods and services.

• The independent non-profit business was formed in order to foster a vendor-neutral, open, and transparent community around Eclipse. The Eclipse community now includes people and companies from many sectors of the software industry.

Java 1.8 (JavaPipe, 2022)

- The Java Development Kit (JDK) is a software development environment that is used to create Java applications and applets. The Java Runtime Environment (JRE), an interpreter/loader (java), a compiler (javac), an archiver (jar), a documentation generator (Javadoc), and other tools required for Java development are all included.

- A Java virtual machine (JVM) is a computer abstraction that allows a computer to execute a Java application. The JVM has three concepts: specification, implementation, and instance.

- The specification is a document that officially outlines what a JVM implementation must perform. Having a single standard assures that all implementations are compatible.

- A JVM implementation is a computer programme that satisfies the JVM standard. A JVM instance is a process that runs a computer programme that has been compiled into Java bytecode.

MySQL

- A Java web application will need the storage of significant quantities of metadata as well as the organization of data. As a result, there was a need to host a Java web application using MySQL. Alternative advantages of utilizing MySQL rather than other database technologies for Java hosting include:
- Cutting-edge security: Because MySQL is widely regarded as the safest relational database currently in use, it is suitable for e-commerce sites that handle frequent online transactions and other sensitive data.
- Superior performance: It is designed to handle the most demanding websites with the highest traffic and is not slowed down by excessive use. MySQL retains its lightning-fast performance rates even when employed by traffic-heavy sites like Twitter and Facebook.
- Increased uptime: MySQL ensures 100% uptime, so you never have to worry about unexpected software failures.
- Simple upkeep: Because the software is open-source, it is continually being updated and debugged, which means you have less maintenance to worry about - all you have to worry about is your Java site or web application.
- It can be found everywhere: MySQL's ubiquity serves a dual purpose: since it is an industry-standard, it is interoperable with practically every operating system. The following are the fundamental stages of setting up a dedicated hosting server:
  1. Build your own dedicated server
  2. Install Apache Tomcat
  3. Install the latest version of MySQL (versions are available for Windows, Linux, and Mac)
  4. Configure and test your MySQL installation

ANALYSIS MODELS: SDLC MODEL TO BE APPLIED

Iterative SDLC Model

When using an iterative approach, such as the Software Development Life Cycle (SDLC), a comprehensive set of requirements is unnecessary upfront. Initial development begins with requirements for the core functionality that may be modified and enhanced as needed. The process is iterative, so new variants of the product may be made at each stage.

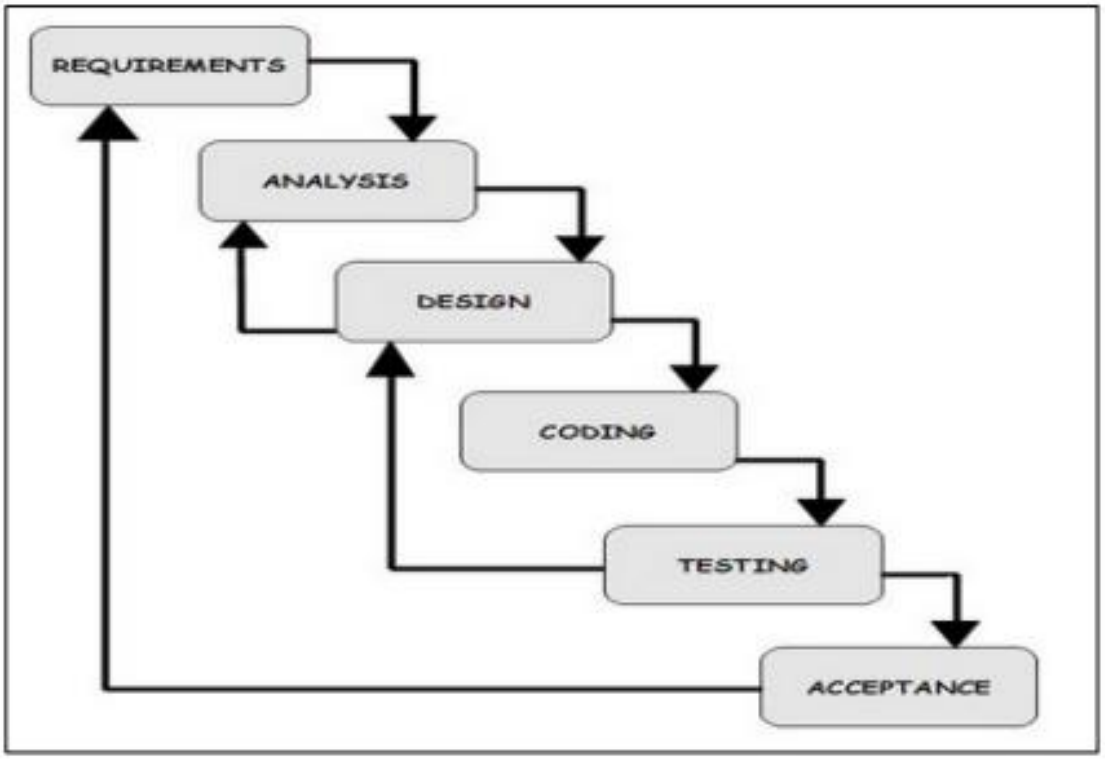


Figure: - Iterative SDLC Model

At the end of each cycle, the newly built functionality is integrated with the existing parts of the system.

The key phases of the software development life cycle are as follows:

• Requirement gathering:

All the project's needs, both practical and aesthetic, were mapped out. Users and other project stakeholders were interviewed to determine their needs, which ranged from the essential (such

as an audit trail and security settings) to the purely aesthetic (such as the design of the user interface). The various needs may be roughly classified as follows:

1. System features
2. Security parameters
3. User requirements4. Administrator requirements
5. User interface

• **Design**:

Creating a structure for the database came first. The whole database setup needed to carry out this project has been created. The next thing that was done was to create a plan for the project. Following a predetermined set of guidelines, the project was developed. There are three distinct parts to the framework:

a) There is a layer for "business entities," which lists everything that will be utilized in the project.

b) The business logic layer, which performs operations on the business entity to accomplish the objectives.

c) The data access layer, which mediates communication between the backend and the services.

- **Construction:**

This phase included the creation of all modules and the user interface. Java was used in the development process. MySQL was used for the database design and development.

• Integration and system testing:

Every component was joined with the others. The web-services-using modules were included in the UI, making for a seamless user experience. The flow of data was initiated by the MySQL database. The product was put through its paces, and any bugs were ironed out during testing. The project was tested from both the developers' and the users' perspectives, using a wide variety of custom-built test cases. Errors and exceptions were found via debugging and then fixed.

• Installation and maintenance:

All authorized administrators and users share a single computer running our system. Our system undergoes frequent maintenance to keep it running smoothly. As long as they do not interfere with the already implemented functionality, new requirements and features may be introduced as necessary.

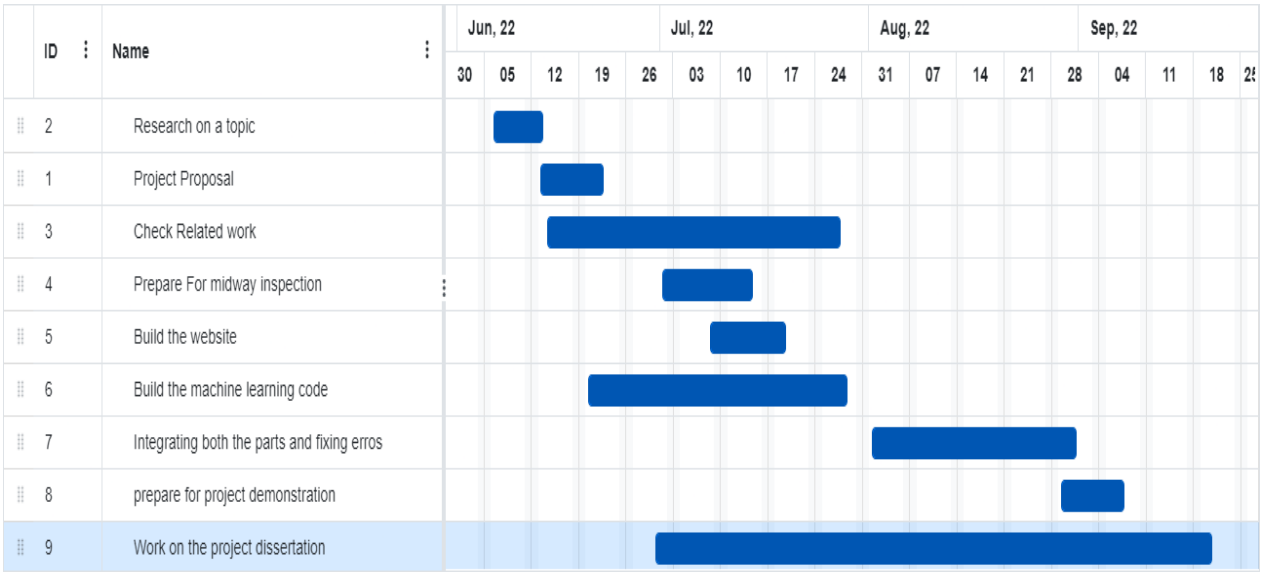
Timeline Chart

The use of schedules like Gantt charts to plan and then report progress within the context of a project is an important aspect of project management. From inception to completion, your project will be documented in a project management plan.

The scope and complexity of a project should determine the depth to which its analysis or prototype is developed. Software design often takes up 20% - 25% of development time.

1. Gathering of requirements
2. Review of current systems' literature
3. Modeling and training of requirements
4. Design of fake screens
5. Actual Application

The project's Gantt chart is shown below. The Gantt chart depicts project planning starting with the topic's selection.



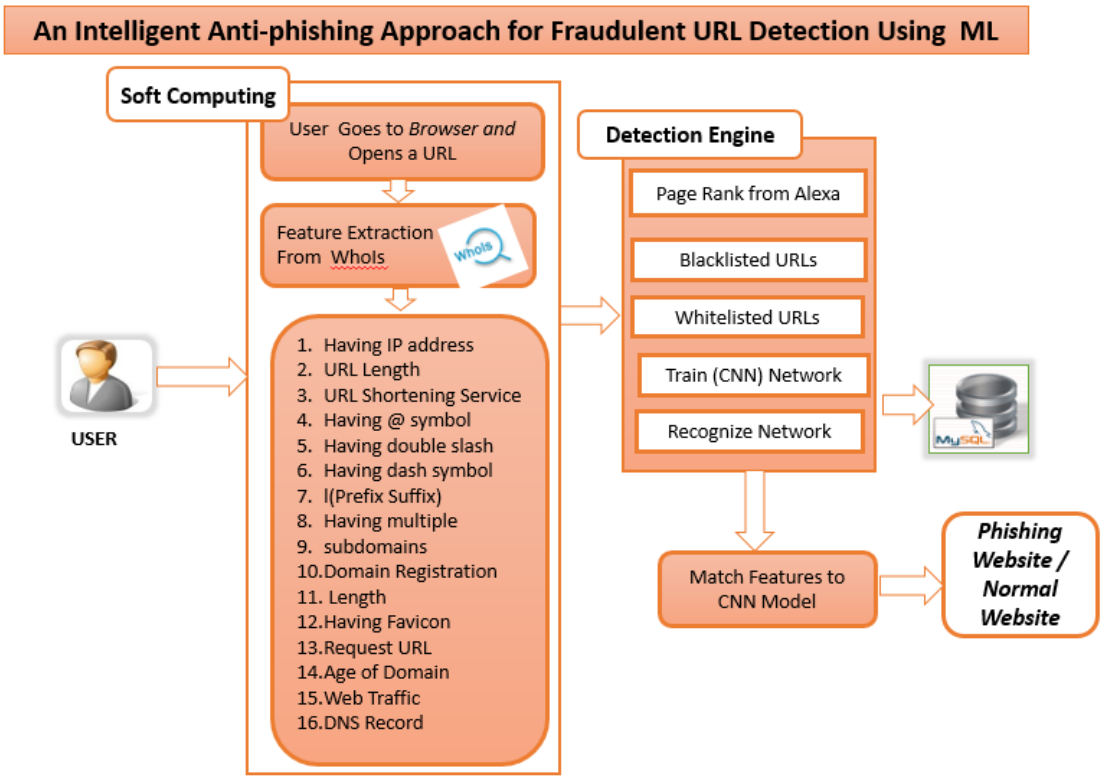
It represents the life cycle of software development (SDLC). Topic selection, requirements gathering, Software Requirements Specification, and Hardware Requirements Specification are all project stages.

Chapter 4

SYSTEM DESIGN

SYSTEM ARCHITECTURE

Phishing is a kind of cybercrime that involves copying a genuine company's website in order to collect personal information from unsuspecting internet users. There have recently been many new phishing detection algorithms developed, however, most of them still depend on pre-existing lists to determine if an email is fake or not. Detection algorithms thus struggle when a new phishing web page is released to accurately categorize it as phishing.



System design

1:Whois Domain:

A domain name, an IP address square, or a self-governing framework are all examples of Internet assets. WHOIS is a question-and-answer protocol that is used to query databases that include registered customers or trustees of these assets. A WHOIS search allows you to get

information about a domain's expiry date, registration details etc. in a public database. Whois can extract total of 13 key features some of them are listed as follows:

Sr.No	Features	Significance
1	Having IP Addr	A phishing website has an IP address in the domain name.
2.	URL_Length	Phishing sites have URLs longer than 75 characters.
3	Shortening_Servi	People get duped by the term known as ‘link sharpeners’
4	Having_At_Sym	Websites with a @ sign are fraudulent.
5	Double_slash_re	'/1' indicates a Phishing Website.
8	Links_in_tags	Some questionable websites may be accessed via specific tag
9	Abnormal_URL	The URL identifies genuine websites in the Who is Database.
10	Age_of_domain	Phishing websites have been up for more than a year.
11	Page_Rank	Phishing sites have few links, thus they rank poorly.

2. Developed a model for detection by using convolutional neural networks

The system recognizes the phishing website thanks to Convolution Neural Network (CNN) technology. CNN is used to determine whether a URL is malicious or not based on its analysis of a set of 13 characteristics used in the training process.

3. Anti-phishing measures

Whitelists are lists of trusted websites used to identify phishing.

1. The URL address of the reputable website

Database updates are carried out via the URL of a trusted website. "https://signin.ebay.com" is the URL of the website.

2. The website's URL

In order to establish whether the current page in the browser is on the whitelist, the domain of

the trusted site is checked against the URL, such as "signin.ebay.com."

3. The title of the site

It is possible to speed up the process of matching possible phishing site names with titles in the whitelist Database by using trusted site titles like "Welcome to eBay," for instance, which is an example of a site title that may be utilized.

4. Alexa Ranking

Changes in other sites' traffic affect your site's rating. Alexa calculates each site's three-month average daily traffic and page views every day. #1 is the site with the most visitors and page views in the previous three months. Because phishing sites are short-lived, Alexa may miss them (Alexa the Web Information Company., 1996). If a domain has no traffic or isn't in Alexa, it's branded "Phishing." It's "Suspicious" otherwise.

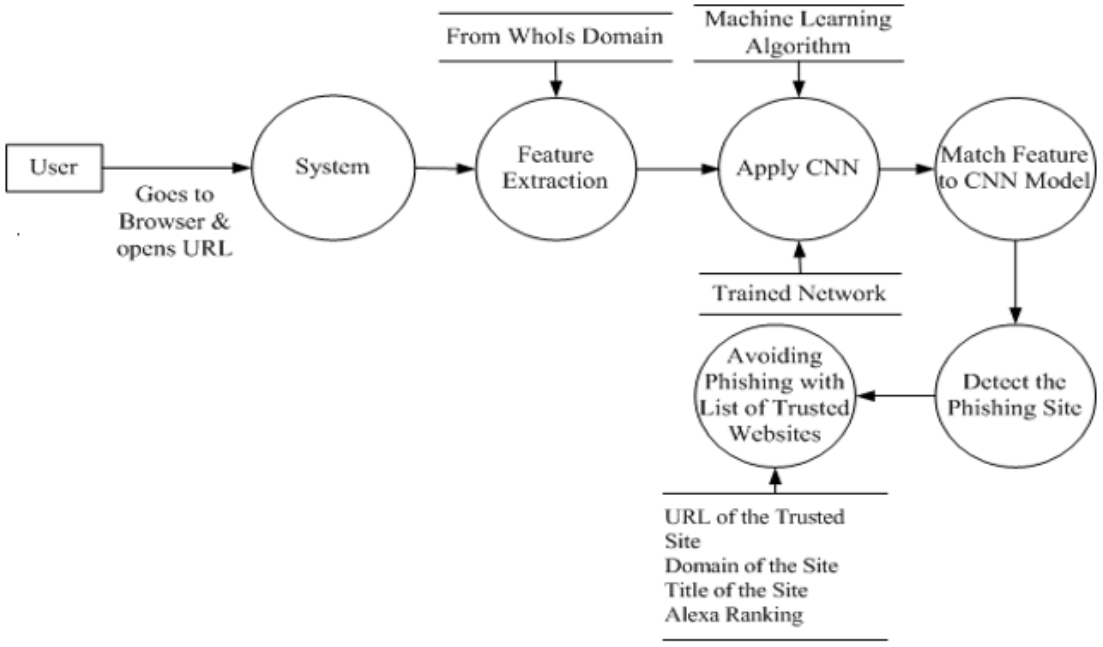


Figure: System Flow

Chapter 5

PROJECT MODULES

The system's graphical user interface was created using Java JSP. Java and JSP were the primary technologies used. The main programming was done in Eclipse Luna, and we utilised MY SQL GUI browser for the database.

The database is primarily used to store user information such as usernames, user identities, and keysets. MYSQL GUI Browser was used to perform database functions.

- **Data**

First, in terms of the dataset, we collect and test the system against a large number of example URLs, in which the phishing URLs are downloaded from phish tank (Kaggle). This increases the value of the performance review and analysis.

URL: https://rb.gy/zduhev

Dataset

- Total No of Instances: 11055
- Phishing: 6157
- Non-phishing: 4898

- **Analysis**

Second, we considerably boost the machine learning algorithm's richness and variety of features. In WEKA, we used the Nave Bayes, Random Forest, and SVM classification algorithms to assess the prepared URL feature dataset.

OVERVIEW OF PROJECT MODULES

To begin, the phisher must develop a phishing website that seems real to the target. The site should then be hosted on the internet for the usage of victim confidential information. When a victim visits a phishing website, the victim is persuaded to input sensitive information. The phisher then obtains some inputted data, which he or she may later exploit. (An Enhanced Phishing Email Detection Model Using Machine Learning Techniques (A REVIEW), 2022)

We wish to extract the WhoIs attributes of URLs to identify phishing websites. Phishing Detection Utilizing Soft Computing and Machine Learning is an innovative technique for detecting phishing web pages using URL and WhoIs characteristics. The convolution Neural Network is used to train the network and ultimately determine whether or not the site is phishing.

ALGORITHM DETAILS

Convolutional Neural Network

Neural Network Convolution Traditional feature learning approaches use picture semantic labels as supervision. They frequently think that the tags are equally exclusive and hence do not bring up the label issue. The learnt characteristics equip words with clear semantic relationships. We also create a new cross-modal feature that can represent both visual and textual material. CNN is a deep learning technique for image classification. In this step, we apply a single neural network to the whole picture. Convolution, subsampling, activation, and complete connectivity are the stages in CNN. (5 Layers of Convolutional Neural Network | upGrad blog, 2022)

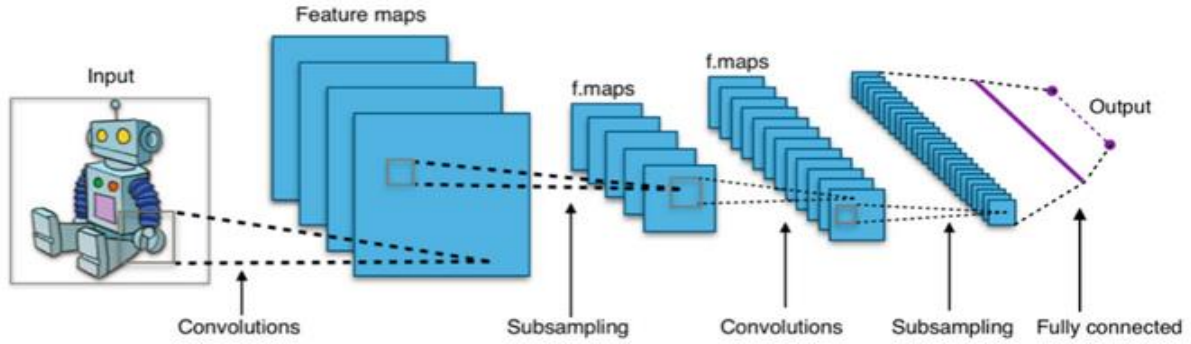
- Convolution is the first step. Convolution filters are the initial layers that take an input signal. Convolution is a method in which the network attempts to tag the input signal by referencing to what it has already learnt.

- Subsampling is the second step. Smoothing the convolution layer inputs reduces the susceptibility of the filters to noise and fluctuations. This smoothing method is known as

sub-sampling, and it is accomplished by averaging or considering the maximum across a sample of the signal.

- Activation is the third step. The activation layer controls signal flows from one layer to the next. Output signals that are firmly linked to previous references will activate more neurons, allowing signals to be transmitted more effectively for identification.

- Step 4: Completely linked the network's final levels are completely linked, with neurons from previous layers coupled to neurons from following ones. This simulates high-level thinking by measuring all possible paths from input to output.



CNN MODEL

I have split the data into categories using “to_categorical” and chosen a sequential model of activation layers, which include “sequential”, “Relu” and “Softmax”.

Categories helps convert normal data into categories, The sequential API allows you to create models layer-by-layer for most problems. It is limited in that it does not allow you to create models that share layers or have multiple inputs or outputs.

RELU: CNN model converts data into feature maps, and since they contain non-linearity therefore relu was used.

SOFTMAX: checks the strength of the signal and if it’s less then it passes back to the activation layer or else if its high then it pushes it forward.

For compiling of the model, I have chosen “adams” optimizer. Adams optimizer for the reason that, in CNN, speed reduces due to data processing for large dataset, therefore adams optimser is used to speed up the CNN model

This model is trained on a dataset and then tests are conducted on a different dataset.

“ServerSocket” feature is used for communication between CNN model (coded in python) and Website (coded in java).

Chapter 6
SOFTWARE TESTING

Software testing is an activity that evaluates a programmer or system's characteristic or capability and determines whether or not it achieves the desired outcomes. It entails more than just running a programmer to look for flaws. Every project is unique, with unique characteristics. No single yardstick may be appropriate in all situations. This is a distinct and vital field with distinct issues. Despite the fact that it is crucial to software quality and is commonly used by programmers and testers. Due to a lack of knowledge of software principles, software testing remains an art. The challenge derives from the software's complexity. Software testing may be done for quality assurance, verification and validation, or reliability estimate. Testing may also be used as a general metric. Software testing is a balance of cost, time, and quality.

For this project, to evaluate the effectiveness of the model, we ran it against test data and several other lists of phishing websites found on the internet. I ran additional test cases apart from my test data so that I could gain an accurate insight into the model. To further gain insights into the model’s effectiveness, we compared it with other models’ effectiveness on WEKA.

Test CASES & TEST RESULTS

A few random phishing websites found the internet on which the first trial of the model was done.

Statement	Did test cases pass	Did test cases fail
http://bfa.xxuz.com/	✓	
http://michaelelectronics.co.tz/	✓	
http://consulta-magazine-com.com/board.php?acessid=27570	✓	
http://shristifoundation.com/wp-includes/x3d/flying.htm	✓	
http://www.eraoshgnbdafihgn.com/about.html	✓	
http://conforteventos.com/	✓	
http://encuesta-covid19cuu-septiembre.000webhostapp.com/	✓	
https://aldara.co.id/wp-includes/js/tab/dab/conect/4d6ad	✓	
https://invalid-name-block.com/Login.php	✓	

Table: Test Cases

Comparison with other models

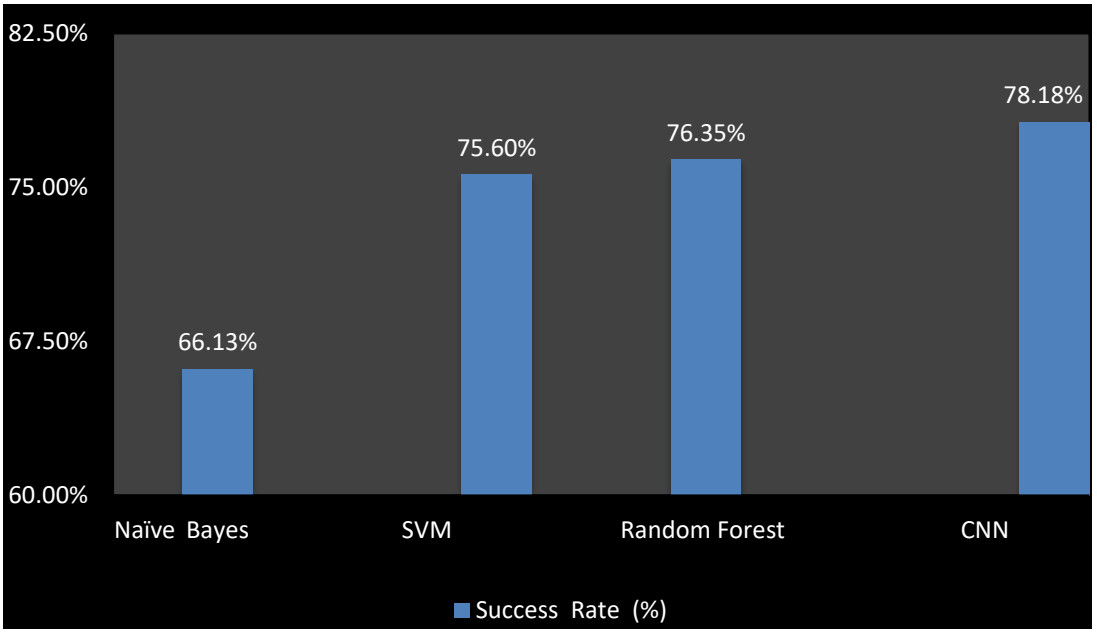
WEKA was used to evaluate the produced URL feature dataset using the Nave Bayes, Random Forest, and SVM classification algorithms. In this case, we compared the performance of other classifiers on the same dataset and then compare the results with ours.

The performance is then assessed using the Confusion Matrix, Detection Accuracy, True Positive Rate, and False Positive Rate. TABLE 1 summarizes the results. When compared to other chosen classifying algorithms in WEKA, SVM shows the highest accuracy rate on WEKA, but that is still lower than the CNN model built. In the below table (a referring to -1, means the URL is phishy and b referring to 1 means a normal website)

TABLE: Classifier Performance - WEKA

Classifier	**Confusion Matrix**	**Accuracy (%)**	**Error Rate (%)**
Naïve Bayes	**a b** 3471 1427 2317 3840	66.133%	33.867%
SVM	**a b** 4006 892 1805 4352	75.6038 %	24.3962%
Random Forest	**a b** 3605 1293 1322 4835	76.3455 %	23.6545%
CNN	a b 3682 1216 1196 4961	78.1818 %	21.8182%

Table: comparison of TP Rate, FP Rate and Detection Accuracy of different models



Graph showing Comparison between classifiers on weka vs CNN model

The Confusion Matrix results for CNN is as follows:-

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.47 seconds

=== Summary ===

Correctly Classified Instances        8643               78.1818 %
Incorrectly Classified Instances      2412               21.8182 %
Kappa statistic                          0.5577
Mean absolute error                      0.2827
Root mean squared error                  0.3745
Relative absolute error                 57.2894 %
Root relative squared error             75.3934 %
Total Number of Instances           11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.752    0.194    0.755      0.752    0.753      0.558    0.879     0.842     -1
                 0.806    0.248    0.803      0.806    0.804      0.558    0.879     0.906     1
Weighted Avg.    0.782    0.224    0.782      0.782    0.782      0.558    0.879     0.877

=== Confusion Matrix ===

    a    b   <-- classified as
 3682 1216 |    a = -1
 1196 4961 |    b = 1
```

Figure: - Confusion Matrix of CNN

Using a variety of data mining methods, many attributes are compared to one another. The findings demonstrate the effects that may be accomplished via the use of the Whois characteristics. We can attempt to detect phishing URLs by evaluating the host-based properties of the websites they lead to in order to prevent end users from accessing these sites. The fact that thieves are always coming up with new methods to get around our security protocols is a particularly difficult obstacle in this sector. For our algorithms to be successful in this competition, we need to be able to continuously adapt to new models and samples of phishing URLs. (RRPhish: Anti-phishing via mining brand resources request, 2022)

SCREEN SHOTS

Below are some screenshots of the website design.

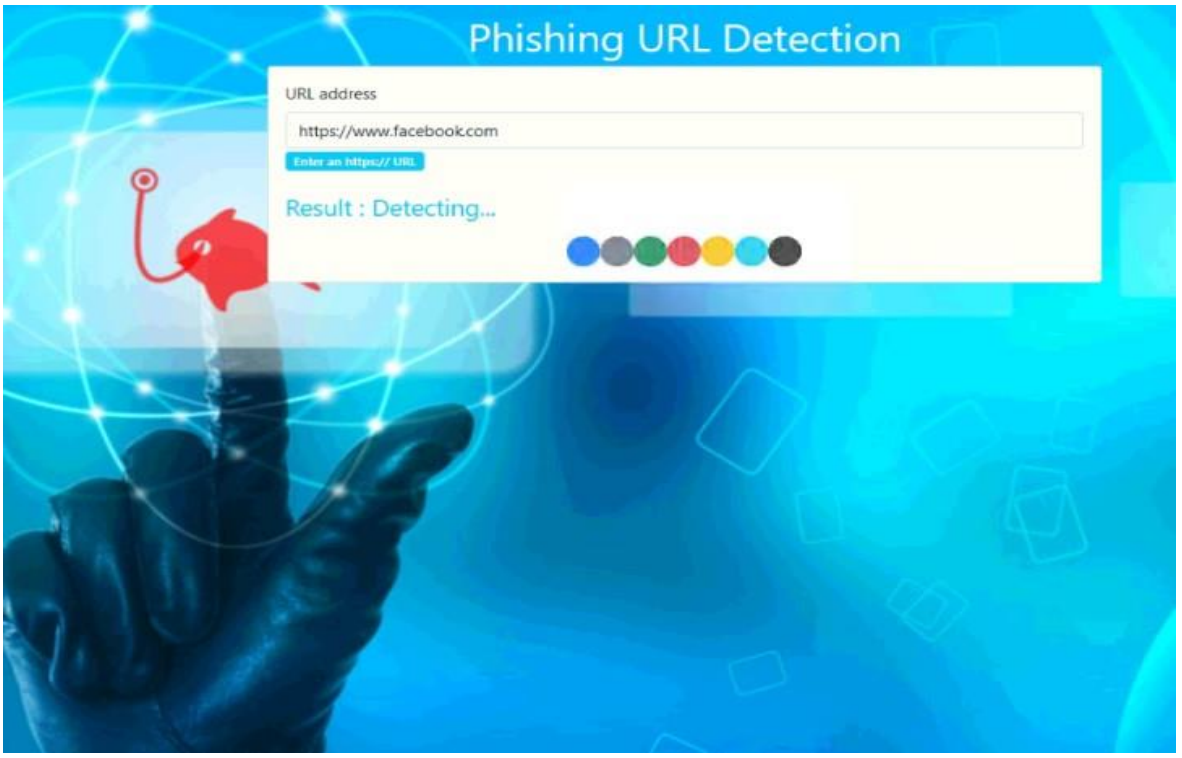


Figure : Phishing Detection

Figure : Phishing Site Detected

Figure : Non-Phishing (Normal) Site Detected

Chapter 7
RISK ANALYSIS

RISK	DESCRIPTION	LIKELIHOOD	SEVERITY	MITIGATION STRATEGY
Unable to implement the feature	The Risk associated with not being able to implement the planned features	2	5	Make sure to read past papers and practice everything that I have learned in ML and clear doubts with the supervisor if needed
The dissertation page limit crossed	More than the given limitation of pages passed	3	4	Make a clear structure to document in a draft format. Secondly, polish the contents to make them small And precise.
Unbalanced dataset/ Small dataset	If in case the dataset is unbalanced/small with its records, then the phishing detection will completely go wrong	3	5	There dataset balancing or searching for the new dataset will be the primary plan. Secondly, to analysis the Machine learning approach to training the dataset and visualize the best approach that gives better accuracy.

Personal risk

RISK	DESCRIPTION	LIKELIHOOD	SEVERITY	MITIGATION STRATEGY
Poor time management	If in case I have multiple submissions coming and I am unable to keep up with all of them	1	5	If in case there are multiple submissions due, make a schedule , divide tasks and start early
Difficulty in making the model	Unable to understand CNN	2	4	Start researching early onwards and ask for help if needed
Unforeseen illness	X	1	5	Inform the supervisor and discuss it with him

RISK	DESCRIPTION	LIKELIHOOD	SEVERITY	MITIGATION STRATEGY
Hardware failure	If anything happens to the device and data gets cleared	1	4	Upload the code on github and continuously save new changes
Requirements change	If my approach is not approved by the supervisor	2	2	Schedule a weekly meeting with the supervisor to provide updates and take regular feedback

Chapter 8

CONCLUSION

Phishing is a criminal plan that involves stealing the user's personal information as well as other credential information. It is a kind of fraud in which the attacker obtains the victim's sensitive information, such as a password, bank account information, credit card number, financial login and password, and so on, and then uses it.

To identify phishing websites using URL characteristics, we offer a unique technique, Phishing Detection Using Soft Computing and Machine Learning. For blacklisted and whitelisted URLs, features are retrieved and utilized as a dataset for machine learning algorithms. In this Project we have designed our own model, evaluated its effectiveness against test data and compared it accuracy with other models. The accuracy rate of our CNN model is 78%, which is higher in comparing to other models.

LIMITATIONS AND FUTURE WORK

This application is only applicable for phishing site detection through email only. Secondly, it is only after logging into the website and then running the CNN model on the website that ite we realize if it’s a phishing website or not, which not everyone will be inconvenient each and every time.

ONE MONTH PLAN

In the coming month, I will change the User interface to improve the user experience of the website. The second aspect would be to improve the model by training it on the additional dataset, more from the ones I have already done. This will help in improving the accuracy of the model furthermore.

LONG TERM PLAN

In the long term, for user convenience, instead of having him paste the URL on the website, the user can easily identify a phishy website by the use of AI that is integrated with the ML model.

References

[1] Internationaljournalssrg.org. 2022. [online] Available at: <https://www.internationaljournalssrg.org/IJCSE/2019/Volume6-Issue6/IJCSE-V6I6P106.pdf> [Accessed 19 September 2022].

[2] Researchr.org. 2022. *Phishing detection: A recent intelligent machine learning comparison based on models content and features - researchr publication bibtex*. [online] Available at: <https://researchr.org/publication/AbdelhamidTA17/bibtex> [Accessed 19 September 2022].

[3] 2022. [online] Available at: <https://www.researchgate.net/profile/Longfei-Wu-3/publication/286582584_MobiFish_A_lightweight_anti-phishing_scheme_for_mobile_phones/links/5b49d6ce0f7e9b4637d67cf1/MobiFish-A-lightweight-anti-phishing-scheme-for-mobile-phones.pdf> [Accessed 19 September 2022].

[4] Cis.temple.edu. 2022. [online] Available at: <https://cis.temple.edu/~jiewu/research/publications/Publication_files/Phishing_TVT_final.pdf> [Accessed 19 September 2022].

[5] Researchr.org. 2022. *RRPhish: Anti-phishing via mining brand resources request - researchr publication*. [online] Available at: <https://researchr.org/publication/GengYZJ18> [Accessed 19 September 2022].

[6] Afroz, S. and Greenstadt, R., 2022. *PhishZoo: Detecting Phishing Websites by Looking at Them*.

[7] 2022. [online] Available at: <https://www.semanticscholar.org/paper/Detection-of-phishing-attacks-Baykara-G%C3%BCrel/1b5ec7fa4ee1b549725d4541d6228e06f68d2a14> [Accessed 19 September 2022].

[8] Jetir.org. 2022. *An Enhanced Phishing Email Detection Model Using Machine Learning Techniques (A REVIEW)*. [online] Available at: <https://www.jetir.org/view?paper=JETIR1811068> [Accessed 19 September 2022].

[9] 2022. [online] Available at: <https://www.semanticscholar.org/paper/A-novel-approach-for-phishing-detection-using-Nguyen-To/ca7ea12ddc80d65eab9b2966c913164be007e9ba> [Accessed 19 September 2022].

[10] 2022. [online] Available at: <https://www.researchgate.net/publication/320651414_Detecting_spam_and_phishing_mails_using_SVM_and_obfuscation_URL_detection_algorithm> [Accessed 19 September 2022].

[11] 2022. [online] Available at: <https://www.researchgate.net/profile/Samuel-Marchal/publication/273169788_PhishStorm_Detecting_Phishing_With_Streaming_Analytics/links/582b2a0108ae102f07207f21/PhishStorm-Detecting-Phishing-With-Streaming-Analytics.pdf> [Accessed 19 September 2022].

[12] 2022. [online] Available at: <https://www.semanticscholar.org/paper/A-novel-approach-for-phishing-detection-using-Nguyen-To/ca7ea12ddc80d65eab9b2966c913164be007e9ba> [Accessed 19 September 2022].

[13] Jetir.org. 2022. [online] Available at: <https://www.jetir.org/papers/JETIR2103361.pdf> [Accessed 19 September 2022].

[14] Gregory Paul, T. and Gireesh Kumar, T., 2022. *A Framework for Dynamic Malware Analysis Based on Behavior Artifacts*.

[15] 2022. [online] Available at: <https://www.semanticscholar.org/paper/Predicting-phishing-websites-based-on-neural-Mohammad-Thabtah/2ba9e8879cc191f9bf68e733cd212b4f7edc1670> [Accessed 19 September 2022].

[16] 2022. [online] Available at: <https://www.researchgate.net/publication/320481930_Cuckoo_Search_Optimization-_A_Review> [Accessed 19 September 2022].

[17] upGrad blog. 2022. *Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network* | *upGrad blog*. [online] Available at: <https://www.upgrad.com/blog/basic-cnn-architecture/> [Accessed 19 September 2022].

Appendix

Github Link: https://github.com/Kulkarnirajiv14/AntiPhishing

(ML model, UI of website and the java(web app)-python(ML model) is designed by myself and rest is referenced and modified from open source)

How to RUN:-

1: turn XAMPP on and start port 3306

2: go to project/antiphishingweb/phishingpython/Pycharm and Click ‘run’ file

3: go to eclipse, make sure the location is of project file, Go to web content/pages/signin.jsp and run as/run on server

4: Then the website will open, copy the link and paste it on your browser.

Email: admin@gmail.com

Password: admin@123

You can also find the id and password from query browser