

# **IBM NALAIYA THIRAN**

## **WEB PHISHING DETECTION**

**A PROJECT REPORT**

*Submitted by*

<b>NANDIKA L R</b>	<b>[711719104054]</b>
<b>SNEHA C I</b>	<b>[711719104094]</b>
<b>TIMOTHY JERALD XAVIER</b>	<b>[711719104102]</b>
<b>TREESA MARY GEORGE</b>	<b>[711719104103]</b>

*In partial fulfilment for the award of the degree of*

**BACHELOR OF ENGINEERING**  
**IN**  
**COMPUTER SCIENCE AND ENGINEERING**

**KGISL INSTITUTE OF TECHNOLOGY, SARAVANAMPATTI**

**ANNA UNIVERSITY: CHENNAI 600 025**

# **ABSTRACT**

Social engineering and cyber-attacks using phishing are the most popular. By deceiving unsuspecting internet users into disclosing private information, the phisher preys on their vulnerability and attempts to steal that information for fraudulent purposes. Users must be aware of phishing websites in order to prevent falling victim to them. Identify phishing websites and add them to a blacklist that only allows access after the website has been identified. Machine learning and deep neural network methods can be used to spot them in their early stages. The machine learning-based strategy, which outperforms the other two, has been shown to be the most efficient. Untrusted uniform resource locators (URLs) and web pages are mimicked by phishing websites, a popular social engineering technique. On the dataset produced to detect phishing websites, this project's goal is to train machine learning models and deep neural networks. The needed URL and website content-based attributes are retrieved from the dataset, which is made up of both phishing and benign URLs of websites. It is assessed and compared how well each model performs.

# TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
1.	<b>INTRODUCTION</b>	<b>1</b>
	1.1 PROJECT OVERVIEW	1
	1.2 PURPOSE	2
2.	<b>LITERATURE SURVEY</b>	<b>3</b>
	2.1 EXISTING PROBLEM	3
	2.2 REFERENCES	3
	2.3 PROBLEM STATEMENT DEFINITION	5
3.	<b>IDEATION AND PROPOSED SOLUTION</b>	<b>7</b>
	3.1 EMPATHY MAP CANVAS	7
	3.2 IDEATION AND BRAINSTORMING	8
	3.3 PROPOSED SOLUTION	12
	3.4 PROBLEM SOLUTION FIT	13
4.	<b>REQUIREMENT ANALYSIS</b>	<b>14</b>
	4.1 FUNCTIONAL REQUIREMENTS	14
	4.2 NON-FUNCTIONAL REQUIREMENTS	15
5.	<b>PROJECT DESIGN</b>	<b>16</b>
	5.1 DATA FLOW DIAGRAM	16
	5.2 SOLUTION & TECHNICAL ARCHITECTURE	17
	5.3 USER STORIES	18
6.	<b>PROJECT PLANNING AND SCHEDULING</b>	<b>19</b>
	6.1 SPRINT PLANNING AND ESTIMATION	19
	6.2 SPRINT DELIVERY SCHEDULE	20
7.	<b>CODING AND SOLUTIONING</b>	<b>21</b>
	7.1 FEATURE 1	21

<b>8.</b>	<b>TESTING</b>	
	8.1 TEST CASES	<b>23</b>
	8.2 USER ACCEPTANCE TESTING	<b>24</b>
	8.1.1 Test Case Analysis	<b>24</b>
<b>9.</b>	<b>RESULTS</b>	<b>25</b>
	9.1 PERFORMANCE METRICS	<b>25</b>
<b>10.</b>	<b>ADVANTAGES AND DISADVANTAGES</b>	<b>26</b>
<b>11.</b>	<b>CONCLUSION</b>	<b>27</b>
<b>12.</b>	<b>FUTURE SCOPE</b>	<b>28</b>
<b>13.</b>	<b>APPENDIX</b>	<b>29</b>
	13.1 SOURCE CODE	<b>29</b>
	13.2 GITHUB AND PROJECT DEMO	<b>37</b>
<b>14.</b>	<b>REFERENCES</b>	<b>38</b>

# **CHAPTER 1**

## **INTRODUCTION**

Phishing imitates the characteristics and alternatives of emails and makes it appear similar due to the fact the original one. It seems nearly like that of the legitimate supply. The consumer thinks that this e-mail has come back from a real employer or a corporation. This makes the consumer to forcefully visit the phishing internet site thru the hyperlinks given inside the phishing email. This phishing web sites region unit created to mock the seams of an ingenious website. The phishers force person to inventory up the nonpublic info via giving baleful messages or validate account messages etc. so that they inventory up the preferred data which might be utilized by them to misuse it. They come up with strategies that prevent users from constantly having an option but to visit their fake website.

The most dangerous criminal activity in the online world is phishing. Since the majority of users log on to access the services offered by governmental and financial institutions, phishing attempts have significantly increased over the past several years. Phishers started using this as a lucrative business to make money. The reason phishers do this crime is because it is incredibly trustworthy to do so, it doesn't cost anything, and it works. Phishing may be illegal. The phishing will actually get access to the email identity of someone, therefore it's extremely serious to hunt for out the email identification right now every day and send an email to everyone since it's widely available everywhere. These attackers have a terrible lack of resources, making it difficult for them to quickly and effectively advance vital knowledge.

Critical user information like a password, OTP, credit/debit card details, CVV, sensitive business knowledge, medical understanding, secret information, etc. is of interest to those cybercriminals. These crooks frequently also get information that will give them direct access to their emails and social media accounts. Many consumers utilize e-banking to pay for their online purchases of goods. There are e-banking websites that frequently request sensitive information from users—such as usernames, passwords, and credit card information—for malevolent purposes. Phishing websites are these kinds of e-banking websites. One of the most important online communications software services is the web service. Web phishing is one of several security risks to web services on the Internet.

### **1.1 PROJECT OVERVIEW**

A machine-learning technique is primarily used in this effort to identify phishing websites. We

suggested a clever, adaptable, and efficient solution based on classification algorithms to identify and forecast phishing websites. To extract the criteria from the phishing dataset and categorizes their authenticity, we used classification algorithms and methodologies. The final phishing detection rate may be used to determine the phishing website's identity based on several key factors, such as the URL and domain, as well as security and encryption standards. In order to determine if a website is a phishing website or not when a user enters it, our system uses a data mining algorithm.

## **1.2 PURPOSE**

Many consumers utilize e-banking to pay for their online purchases of goods. Some e-banking websites request users to submit private information, such as username, password, and credit card information, etc., frequently out of malice. Phishing websites are these kinds of e-banking websites. One of the most important Internet communications software services are web services. Web phishing is one of several security risks to web services on the Internet. Every hour, millions of incidents take place on the planet. These attacks cause unimaginable losses to people. Therefore, the only goal of our project is to defend users against such attacks. Phishing attacks are the simplest way to get sensitive information from unwitting users. Phishers want to get their hands on crucial information like usernames, passwords, and bank account numbers. Cyber security professionals are currently looking for effective and consistent ways to identify phishing websites. In order to identify phishing URLs, numerous properties of legitimate and malicious URLs are extracted and analyzed in this study. Support vector machines, decision trees, and random forests are among the algorithms used to detect phishing websites. The study aims to identify phishing URLs as well as to determine the best and efficient machine learning method by assessing each algorithm's accuracy thoroughly.

## **CHAPTER 2**

### **LITERATURE SURVEY**

#### **2.1 EXISTING PROBLEM**

Phishing has lately emerged as a major issue for security researchers because to how simple it is to develop a phone website that closely mimics a genuine one. Many individuals fall victim to phishing schemes because they are unable to recognize bogus websites as experts can, who can. Stolen bank account credentials are the attacker's main objective. Every year, customers falling for phishing schemes cost US businesses \$2 billion. In the third Microsoft Computing Safer Index Report, released in February 2014, it was calculated that the yearly worldwide effect of phishing might reach \$5 billion. Users are oblivious to phishing assaults, which is why they're succeeding more often.

Since phishing attacks prey on user weaknesses, it is extremely difficult to defend against them, yet it is imperative to develop phishing detection techniques. Adding Internet Protocol (IP) banned URLs to the antivirus database is a typical way for identifying phishing websites, sometimes known as the "blacklist" method. Attackers use deceptive strategies to trick users, such as altering the URL to make it appear genuine through obfuscation and many other simple techniques like fast-flux, in which proxies are automatically built to host the website, algorithmic generation of new URLs, etc. The main drawback of this approach is that it cannot detect phishing attacks that happen at 00:00.

Heuristic-based detection, which takes into account traits that have been noticed to exist in actual phishing attacks, can identify zero-hour phishing attacks. However, the presence of these traits is not always guaranteed in such attacks, and the false positive rate for detection is very high.

#### **2.2 REFERENCES**

##### **AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites**

*Yazan A. Al-Sariera, Victor Elijah Adeyemo, Abdullateef O. Balogun and Ammar K.*

Phishing is a type of social web-engineering attack in cyberspace where criminals steal valuable data or information from insensitive or uninformed users of the internet. Existing counter measures in the form of anti-phishing software and computational methods for detecting phishing activities have proven to be effective. However, new methods are deployed by hackers to thwart these countermeasures.

Due to the evolving nature of phishing attacks, the need for novel and efficient countermeasures becomes crucial as the effect of phishing attacks are often fatal and disastrous. Artificial Intelligence (AI) schemes have been the cornerstone of modern countermeasures used for mitigating phishing attacks. AI-based phishing countermeasures or methods possess their shortcomings particularly the high false alarm rate and the inability to interpret how most phishing methods perform their function. This study proposed four (4) meta-learner models (AdaBoost-Extra Tree (ABET), Bagging - Extra tree (BET), Rotation Forest - Extra Tree (RoFBET) and Logit Boost-Extra Tree (LBET)) developed using the extra-tree base classifier.

### **An Efficient Anti phishing Method to Secure e-Consumers**

*Guang-Gang Geng, Zhi-Wei Yan, Jong-Hyouk Lee, Xiao-Bo Jin and Dong-Jie Liu.*

In this paper, an antiphishing method, called resource request based phishing discovery (RRPD), has been discussed. By analyzing the resources request characteristics of phishing websites, this method can be used on both the client and server sides. On the client side, client RRPD can be used by the web browser for phishing sites detection. On the server side, server RRPD based on the domain name system dataflow can discover the suspicious phishing sites by analyzing a small amount of web content, which saves bandwidth and computing resources to the utmost. Experimental results demonstrate the effectiveness of the proposed methods.

### **Phishing website detection using novel machine learning fusion approach**

*Lakshmanarao, P.Surya Prabhakara Rao, M M Bala KrishnaRamez Elmasri*

The Phishing is a sort of social designing assault regularly used to take client information, including login accreditations and credit card numbers. With the enhancements in internet technology, websites are the major resource for the cyber-attacks. There are several counter measures available for avoiding phishing attacks, but phishers are changing their attacking methods from time to time. One of the most widely used techniques for solving cybersecurity issues is machine learning.

### **A Novel Machine Learning Approach to Detect Phishing Websites**

*Ishantn Tyagi, Jatin Shad, Shubham Sharma*

Phishing can be described as a way by which someone may try to steal some personal and important information like login id's, passwords, and details of credit/debit cards, for wrong reasons, by



appearing as a trusted body. Many websites, which look perfectly legitimate to us, can be phishing and could well be the reason for various online frauds. These phishing websites may try to obtain our important information through many ways, for example: phone calls, messages, and popup windows. So, the need of the hour is to secure information that is sent online and one concrete way of doing so is by countering these phishing attacks. This paper is focused on various Machine Learning algorithms aimed at predicting whether a website is phishing or legitimate.

## **Detection and Prevention of Phishing Websites Using Machine Learning Approach**

*Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, S. P. Godse*

Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning.

### **2.3 PROBLEM STATEMENT DEFINITION**

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet.

#### **Common threats of web phishing:**

- Stealing of one's personal information and illegal use of their details.
- It will lead to information disclosure and property damage.
- People may get trapped in different scams.

Our Project mainly focuses on applying a machine-learning algorithm along with data science concepts to detect Phishing websites and prevent them from being cornered.

In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking 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. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect if the e-banking website is a phishing website.

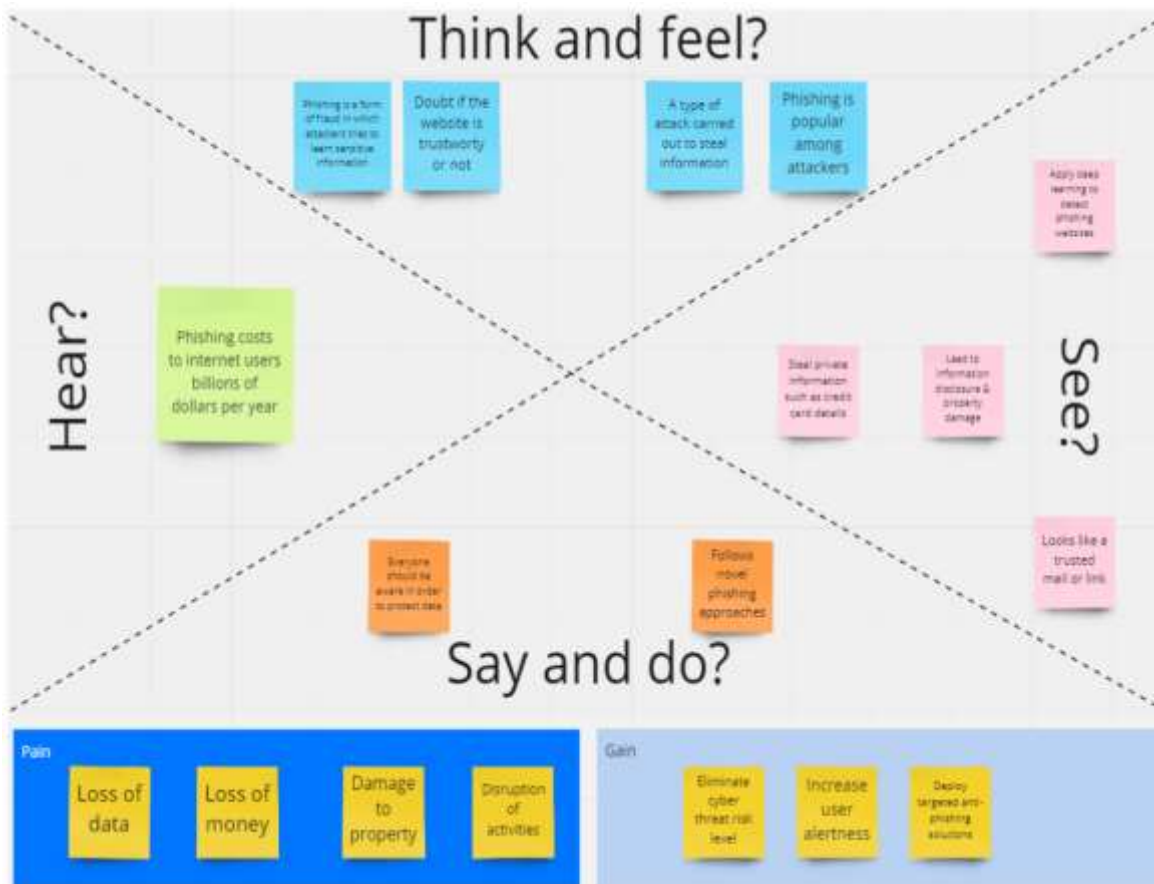
## **CHAPTER 3**

### **IDEATION AND PROPOSED SOLUTION**

#### **3.1 EMPATHY MAP CANVAS**

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user's behaviors and attitudes. It is a useful tool to help teams better understand their users. Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exercise of creating the map helps participants consider things from the user's perspective along with his or her goals and challenges.


Users think that these types of applications should have a very simple interface and need to be visually appealing for them to use, these can be also given to them as a mobile application for easy access. They feel that the application is very exciting as the digits are recognized but are confused whether they need these types of applications. They start trying to check the application by feeding various inputs get fascinated by the output produced by the application and then start recommending the application to their colleagues. Even though the application reduces the manual work and increases the efficiency of recognizing the digit, users think that there are few disadvantages also. Users feel that they may accidentally upload some sensitive files and taking photos of the digits is very annoying.



### 3.2 IDEATION AND BRAINSTORMING


Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich number of creative solutions


# Step-1: Team Gathering, Collaboration and Select the Problem Statement





## Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

 10 minutes to prepare


 1 hour to collaborate


 2-8 people recommended



### Before you collaborate


A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

 10 minutes




#### Team gathering

Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.



#### Set the goal


Think about the problem you'll be focusing on solving in the brainstorming session.




#### Learn how to use the facilitation tools

Use the Facilitation Superpowers to run a happy and productive session.


Open article






### Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.


 5 minutes


How might we (your problem statement)?





### Key rules of brainstorming


To run an smooth and productive session


 Stay in topic.

 Encourage wild ideas.

 Defer judgment.

 Listen to others.

 Go for volume.

 If possible, be visual.

9

## Step-2: Brainstorm, Idea Listing and Grouping

**2**  
**Brainstorm**  
Write down any ideas that come to mind that address your problem statement.  
⌚ 10 minutes

**TIP**  
You can sketch a story, flow and for the past, looking to identify what to start drawing!

**Nandika L R**

- Phishing Filter Mails
- Allow cookies in trusted websites
- Maintaining password in Alphanumeric & Special Characters
- Use of legitimate websites
- Training Users to be cautious

**Sneha C I**

- Maintaining of Authorization
- Block all spam calls
- Use indicators for phishing websites
- Not allowing cookies in unwanted websites
- Avoiding Public wifi when sharing sensitive information

**Treesa Mary George**

- Awareness to people about phishing
- Cautious about Fraudulent Websites
- Usage of trusted web browser
- To block all Phishing websites
- Cautious about websites asking for personal information

**Timothy Jerald Xavier**

- Cautious of sharing Financial information
- Use 2 or 3 step verification for mail
- Frequent change of password for websites
- Cautious about cyber attacks
- Avoid opening of Spam mails

## Step-3: Group Ideas

**3**  
**Group ideas**  
Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.  
⌚ 20 minutes

**TIP**  
Add customisable tags to sticky notes to make it easier to find, browse, organise and categorise important ideas as themes within your mural.

**Phishing Filter Mails**

**Block all spam calls**

**Awareness to people about phishing**

**Cautious about cyber attacks**

**Cautious about websites asking for personal information**

**Cautious about Fraudulent Websites**

**Use 2 or 3 step verification for mail**

**Frequent change of password for websites**

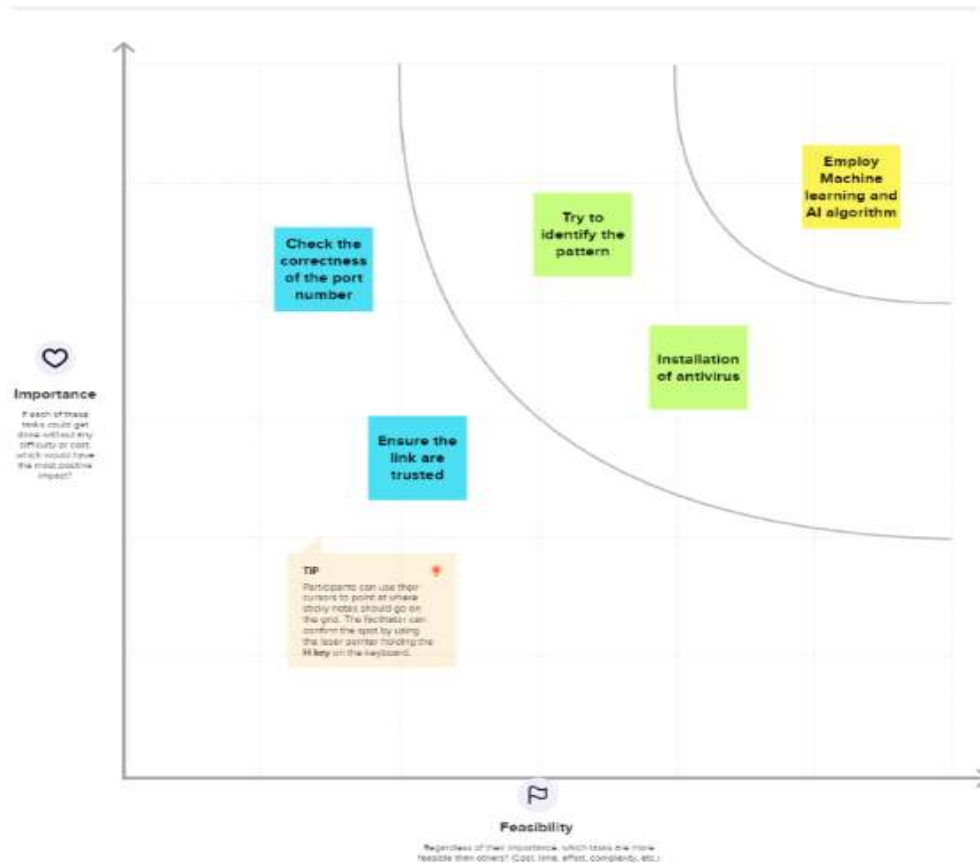
## Step-4: Idea Prioritization

4

### Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

🕒 20 minutes



### 3.3 PROPOSED SOLUTION

In order to detect and predict e-banking phishing websites, an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. Once a user makes a transaction online payment through an e-banking website our system will use a data mining algorithm to detect whether the website is a phishing website or not.

SNO	Parameter	Description
1.	<b>Problem Statement (problem to be solved)</b>	<ul style="list-style-type: none"><li>• There are a number of users who purchase products online and make payment through internet banking.</li><li>• Phishing is a fraudulent technique that is used over the internet to manipulate user to extract their personal information.</li></ul>
2.	<b>Idea / Solution</b>	<ul style="list-style-type: none"><li>• Websites, it must alert the user to verify whether the site is secured or not</li></ul>
3.	<b>Novelty / Uniqueness</b>	<ul style="list-style-type: none"><li>• The phishing websites can be detected based on some important characteristics like URL and domain identity, security and encryption.</li></ul>
4.	<b>Social Impact / Customer Satisfaction</b>	<ul style="list-style-type: none"><li>• There are more than 4505 increase in phishing websites from January to March 2021.</li><li>• As an impact of this model, people can find fraudulent sites of fake websites.</li></ul>
5.	<b>Business Model</b>	<ul style="list-style-type: none"><li>• It must be open source, so all the users can make use of it.</li><li>• The users may also buy annual subscription of advanced protection to protect their personal information.</li></ul>
6.	<b>Scalability of Solution</b>	<ul style="list-style-type: none"><li>• The Users can efficiently and effectively gain knowledge about the web phishing techniques and the way to eradicate them by detecting and reporting the sites.</li><li>• This System can also be integrated along with the future technologies.</li></ul>



### 3.4 PROBLEM SOLUTION FIT

The Problem-Solution Fit means that we have found a problem with our customer and that the solution we have realized for it actually solves the customer's problem. It helps entrepreneurs, marketers and corporate innovators identify behavioral patterns and recognize what would work and why. Few purposes of Problem-Solution Fit are:

- It can be used to solve complex problems in a way that fits the state of our customers
- Succeed faster and increase our solution adoption by tapping into existing mediums and channels of behavior.
- Sharpen our communication and marketing strategy with the right triggers and messaging

Problem-Solution fit canvas 2.0		Project Design Phase-I -Solution Fit Team ID: PNT2022TMID31654	
Define CS, fit into CC	<b>1. CUSTOMER SEGMENT(S)</b> <span>CS</span> Who is your customer? i.e. working parents of 0-5 y.o. kids  Internet users who utilizes the website for the purpose of e-commerce, e-shopping and internet banking.	<b>6. CUSTOMER CONSTRAINTS</b> <span>CC</span> What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.  phishing frequently results in the loss of user's personal information.	<b>5. AVAILABLE SOLUTIONS</b> <span>AS</span> Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital networking.  Provide awareness Web Phishing sites Proper safety while providing personal information.
	<b>2. JOBS-TO-BE-DONE / PROBLEMS</b> <span>J&amp;P</span> Which jobs to be done (or problems) do you address for your customers? There could be more than one; explore different sides.  The phishing websites must be detected before a user uses those websites  Otherwise the user may happen to lose all their personal information like credit card details etc..	<b>9. PROBLEM ROOT CAUSE</b> <span>RC</span> What is the real reason that this problem exists? What is the back story behind the need to do this job? i.e. customers have to do it because of the change in regulations.  Due to lack of awareness and carelessness of the user. Due to greedy scammers	<b>7. BEHAVIOUR</b> <span>BE</span> What does your customer do to address the problem and get the job done? i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)  Checking the link or the URL. Proper research about the website. Reporting the site Contacting the cybercrime department.
Focus on J&P, tap into BE, understand RC	<b>3. TRIGGERS</b> <span>TR</span> What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.  A warning can be popped before opening the website	<b>10. YOUR SOLUTION</b> <span>SL</span> If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality. If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.  The user can check the legitimacy of these kind of websites. To increase the awareness among the people. Pasting the url in phishing detection sites.	<b>8. CHANNELS of BEHAVIOUR</b> <span>CH</span> <b>8.1 ONLINE</b> What kind of actions do customers take online? Extract online channels from #7  The user's may tend to lose their personal data.  <b>8.2 OFFLINE</b> What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.  Aware the user to detect the phishing sites through books.
	<b>4. EMOTIONS: BEFORE / AFTER</b> <span>EM</span> How do customers feel when they face a problem or a job and afterwards? i.e. lost, insecure - confident, in control - use it in your communication strategy & design.  When an user's information is lost, they get panicked and insecure. And they never want to reuse any website.		<b>Extract online &amp; offline CH of BE</b>

## CHAPTER-4

### REQUIREMENT ANALYSIS

#### 4.1 FUNCTIONAL REQUIREMENTS

FR No	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Register through Form.
FR-2	User Authentication	Authentication via Password.
FR-3	User Input	User input an URL to check it is legal or phishing site.
FR-4	Website Comparison	Model comparing the entered URL with the help of Blacklist and Whitelist.
FR-5	Prediction	After comparing, if none found on comparison it extracts feature Using heuristic and visual similarity approach.
FR-6	Classifier	Model the displays whether the website is a legal or phishing site.
FR-7	Events	Model needs the capability of retrieving and displaying accurate result of website.

## 4.2 NON-FUNCTIONAL REQUIREMENTS

FR No	Non-Functional Requirement	Description
NFR-1	<b>Usability</b>	A set of specifications that describe the system's operation capabilities and constraints and attempt to improve its functionality.
NFR-2	<b>Security</b>	Assuring all data inside the system or its part will be protected against malware attacks or unauthorized access.
NFR-3	<b>Reliability</b>	This approach gives more accuracy than existing system.
NFR-4	<b>Performance</b>	Parameters for the proposed system gives accurate predicted value which is compared to the existing system.
NFR-5	<b>Availability</b>	The system is accessible by user at any time using web browser.
NFR-6	<b>Scalability</b>	The design will be suitable and performs with full efficiency according to the rising demands.

# CHAPTER 5

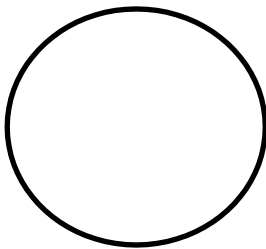
## PROJECT DESIGN

### 5.1 DATA FLOW DIAGRAM

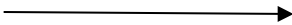
Data flow diagram is used to describe how the information is processed and stored and identifies how the information flows through the processes. Data flow diagram illustrates how the data is processed by a system in terms of inputs and outputs. The data flow diagram also depicts the flow of the process and it has various levels. The initial level is context level which describes the entire system functionality and the next level describes each and every sub module in the main system as a separate process or describes all the process involved in the system separately. Data flow diagram are made up of number of symbols,



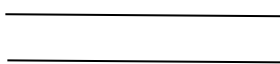
Square presenting external entities, which are resources or destination of data



Circle representing processes, which take data as input, do something to it and output it



Arrow representing the data flows, which can either be electronic data or physical items



Parallel lines representing data stores, including electronic stores such as databases or XML files and physical stores

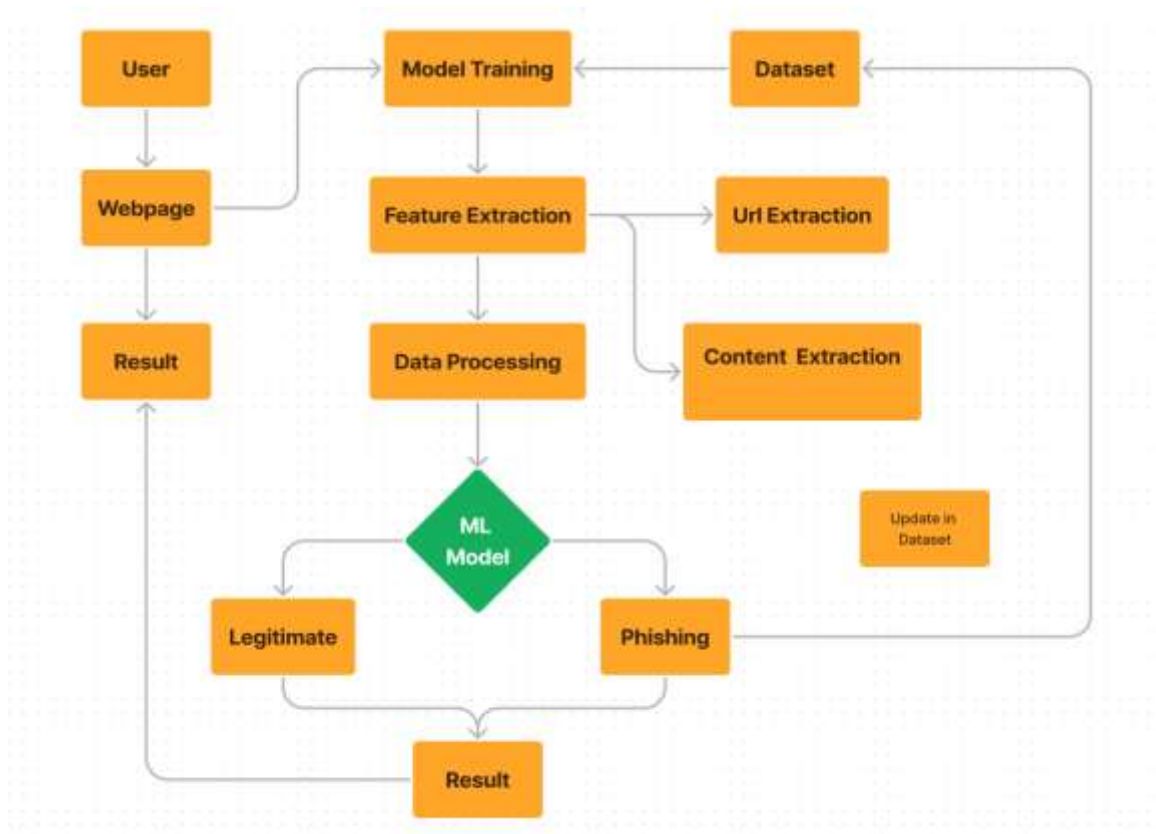


Fig:5.1

## 5.2 SOLUTION AND TECHNICAL ARCHITECTURE

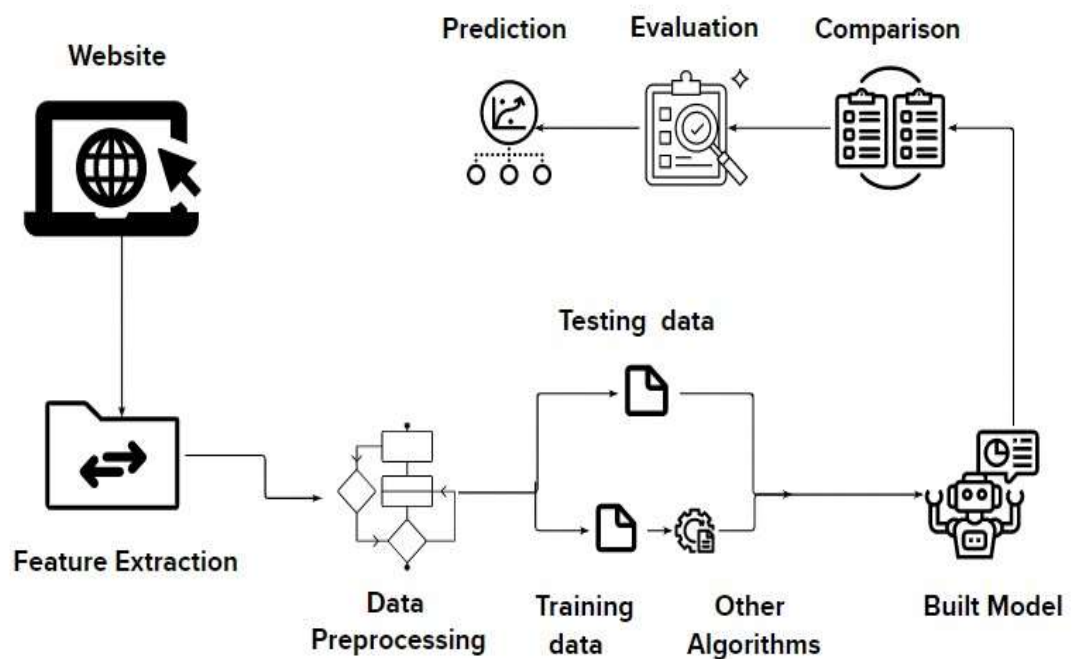


Fig:5.2

## 5.2 USER STORIES

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard					
Customer (Web user)	User Input	USN-1	As a user, I can enter the required URL in the box while awaiting validation.	I can access the website without any problem	High	Sprint-1
Customer Care Executive	Feature Extraction	USN-1	In the event that nothing is discovered during comparison, we can extract features using a heuristic and a visual similarity technique.	As a user I can have comparison between websites for security	High	Sprint-1
Administrator	Prediction	USN-1	The model will use machine learning algorithms like a logistics regression and KNN to forecast the URLs of the websites.	I can accurately forecast the specific algorithms in this way.	High	Sprint-1
	Classifier	USN-2	To create the final product, I will now feed all of the model output to classifier.	I'll use this to identify the appropriate classifier for generating the outcome.	Medium	Sprint-2

## CHAPTER 6

### PROJECT PLANNING AND SCHEDULING

#### 6.1 SPRINT PLANNING AND ESTIMATION

<b>Sprint</b>	<b>Functional Requirement (Epic)</b>	<b>User Story Number</b>	<b>User Story /Task</b>	<b>Story Points</b>	<b>Priority</b>	<b>Team Members</b>
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming mypassword.	2	High	Timothy Jerald Xavier
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	1	High	Timothy Jerald Xavier
Sprint-2		USN-3	As a user, I can register for the application through Facebook	2	Low	Nandika L R
Sprint-1		USN-4	As a user, I can register for the application through Gmail	2	Medium	Timothy Jerald Xavier
Sprint-1	Login	USN-5	As a user, I can log into the application by entering email & password	1	High	Timothy Jerald Xavier
Sprint-2	Dashboard	USN-6	As a user, I can easily navigate through dashboard and I can use the dashboard to get details about app and instruction to use the app.	1	Medium	Nandika L R
Sprint-2	Customer Care Executive (Login)	CCE1	As a CCE I can login to application using User id& Password and I can interact with user.	2	Medium	Nandika L R
Sprint-2	Customer Care Executive (Login)	CCE2	As a CCE I can access dashboard using User id and Password. I can see all user queries,explain app usage and rectify user queries	1	Low	Nandika L R

<b>Sprint</b>	<b>Functional Requirement (Epic)</b>	<b>User Story Number</b>	<b>User Story / Task</b>	<b>Story Points</b>	<b>Priority</b>	<b>Team Members</b>
Sprint-3	Administrator	A-1	As an administrator, I can login and access dashboard and manage and direct activities.	1	High	Sneha C I
Sprint-3	Model Building	M-1	As an User, I can enter the url and Predict it as a Phishing site or not.	2	High	Sneha C I
Sprint-4	Model Testing	MT-1	If the model Predict the URL as Phishing site or not with accuracy rate above 95%.	2	High	Treesa Mary George

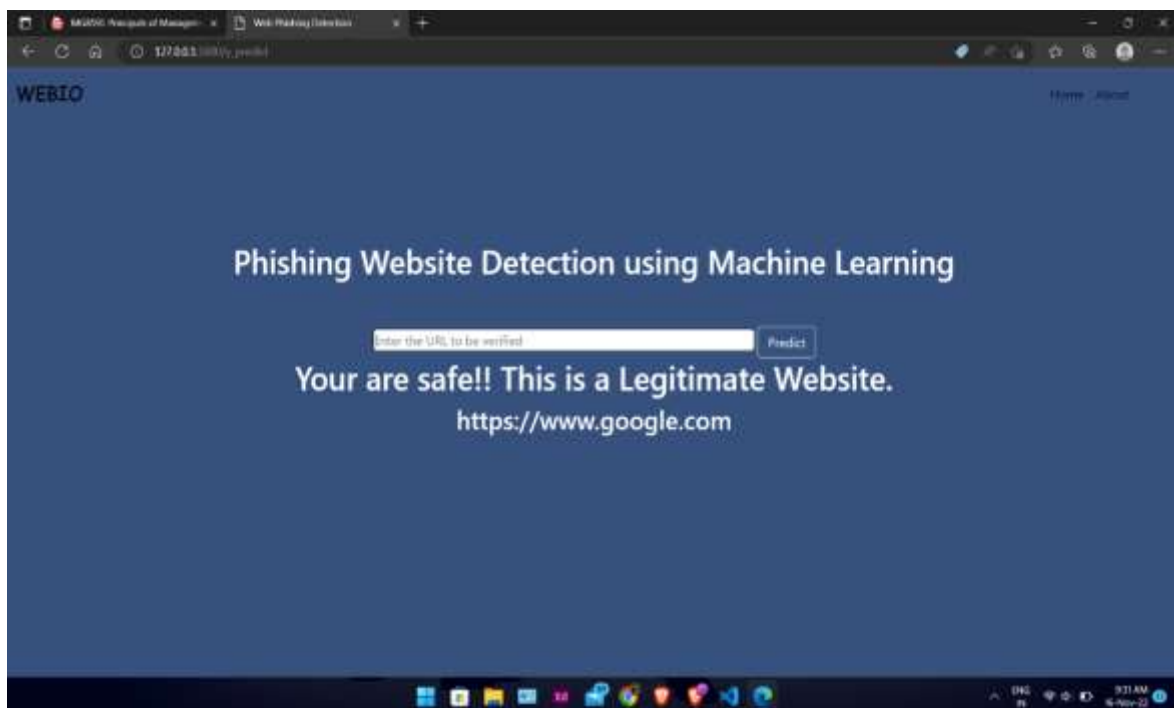
## 6.2 SPRINT DELIVERY SCHEDULE

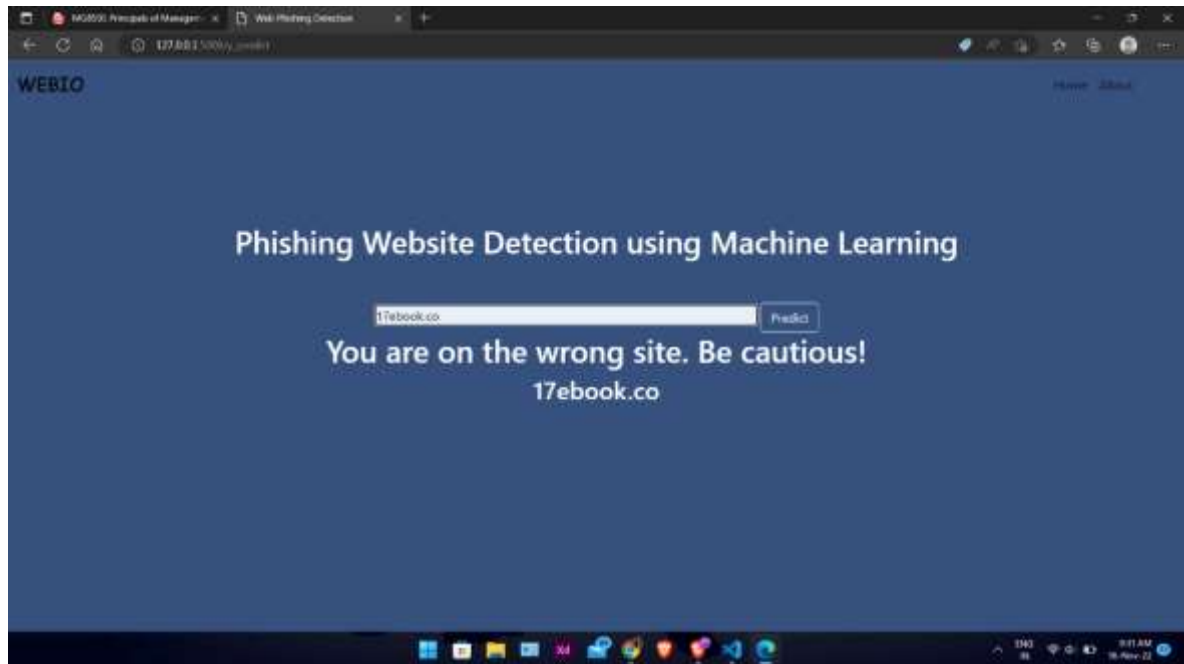
<b>Sprint</b>	<b>Total Story Points</b>	<b>Duration</b>	<b>Sprint Start Date</b>	<b>Sprint End Date (Planned)</b>	<b>Sprint Release Date (Actual)</b>
Sprint-1	20	6 Days	25 Oct 2022	29 Oct 2022	29 Oct 2022
Sprint-2	20	6 Days	01 Nov 2022	05 Nov 2022	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	13 Nov 2022	13 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	19 Nov 2022



# CHAPTER 7

## CODING AND SOLUTIONING





## CHAPTER 8

### TESTING

Test Case id	Feature type	Component	Test scenario	Expected result	Actual result	Status
TC_001	UI	Home Page	Verify UI elements in the Home Page	The Home Page must be displayed properly	Working as expected	PASS
TC_002	UI	Home Page	Verify whether the page is responsive	The Home Page must display in the same way in all devices	The UI is displayed correctly only on the desktop screens	FAIL
TC_003	Functional	Home Page	Check if user could navigate to the next page	The button in the Home Page is directing to next page	Working as expected	PASS
TC_004	Functional	Backend	Check if all the routes are working properly	All the routes should properly Work	Working as expected	PASS
TC_005	Functional	Model	Check if the model can handle various url	The model should accept various url and predict the results	Working as expected	PASS
TC_006	Functional	Model	Check if the model predicts the digit	The model should predict the number	Working as expected	PASS
TC_007	UI	Working	Check if the result is displayed properly	The result should be displayed properly	Working as expected	PASS

## 8.1 USER ACCEPTANCE TESTING

Acceptance Testing is a level of the software testing where a system is tested for acceptability. The purpose of this test is to evaluate the system's compliance with the business requirements and assess whether it is acceptable for delivery. Formal testing with respect to user needs, requirements, and business processes conducted to determine whether or not a system satisfies the acceptance criteria and to enable the user, customers or other authorized entity to determine whether or not to accept the system. In this application, the customer's acceptance is been monitored and it is been put into usage.

### 8.1.1 TEST CASE ANALYSIS

SECTION	TOTAL CASES	NOT TESTED	FAIL	PASS
Client Application	5	0	1	4
Security	1	0	0	1
Performance	3	0	1	2
Exception Reporting	1	0	0	1

# CHAPTER 9

## RESULTS

### 9.1 PERFORMANCE METRICS



## **CHAPTER 10**

### **ADVANTAGES AND DISADVANTAGES**

#### **10.1 ADVANTAGES**

- Reduces manual work.
- Reduces the loss of property and personal data in greater extend.
- Application is capable of handling a lot of data.
- Application can be used by every individual.
- They are highly compatible.

#### **10.2 DISADVANTAGES**

- All the URL must be in entered in correct format.
- Requires a high-performance server for faster predictions.
- Prone to occasional errors.
- Cannot handle complex data.

## **CHAPTER 11**

### **CONCLUSION**

This project demonstrates a web application build using HTML, CSS, JS, Flask and few other technologies. Each time the user enters an URL for recognizing, the link is pre-processed before feeding it into the model. After pre-processing, the model recognizes and verifies the URL by using random forest algorithm. In this model, the pre-processed URL passes into various layers and finally the link is validated. The output is being rendered into the web application and shown to the user. This application can be used in various devices for detecting the phishing sites.

## **CHAPTER 12**

### **FUTURE SCOPE**

In the future, application can be improved with following features:

- Add support to connect with our native browser for more effective usage.
- Add support to different languages to help users all over the world
- Add support to detect and inform automatically during search.
- Add support to process with the loss of personal data.



## CHAPTER 13

### APPENDIX

#### 13.1 SOURCE CODE

##### index.html

```
<!-- Home Page -->
<!DOCTYPE html>
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width,
    initial-scale=1.0">
  <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.0.2
    /dist/css/bootstrap.min.css" rel="stylesheet">
  <script src="https://cdn.jsdelivr.net/npm/bootstrap@5.0.2
    /dist/js/bootstrap.bundle.min.js">
  </script>
  <title>Phishing Website Detection</title>
  <link rel="stylesheet" href="style.css">
  <style>
    *{
      padding: 0%;
      margin: 0%;
    }
    .backg
    {
      background-color:#37517e;
      width: 100vw;
      height: 100vh;
      color: white;
    }
    .nav-bar
    {
      font-family: Arial, Helvetica, sans-serif;
      font-size: medium;
      position: fixed;
    }
    .comname
    {
      font-family:cursive;

      font-size: x-large;
    }
  </style>
</head>
<body>
  <div class="backg">
    <div class="nav-bar">
      <div class="comname">
        <h1>Phishing Website Detection</h1>
      </div>
    </div>
  </div>
</body>
</html>
```

```

.navc
{
    background-color: transparent;
    border:none;
}
.b2
{
    text-decoration: none;
    color: white;
}
.b2:hover
{
    color: white;
}
#abt
{
    background-color: white;
    color: #1b335d;
}
hr
{
    color: midnightblue;
}
hr:not([size])
{
    height: 10px;
    width: 30%;
    margin-left: 490px;
}
.webc
{
    background-color: #1b335d;
    width: 100%;
    height: auto;
}
.headi
{
    padding-top: 150px;
}
</style>
</head>

<body>

    <!-- Section 1 -->
    <div class="backg">

```

```

<!-- Navigation bar -->

```

```

<nav class="navbar navbar-expand-lg navbar-light
bg-blue">
<div class="container-fluid">
<a class="comname navbar-brand
fw-bold" href="#">WEBIO</a>
<div class="collapse navbar-collapse
justify-content-end pe-5">
<ul class="navbar-nav">
<li class="navc nav-item ">
<a class="text-center nav-link active"
aria-current="page" href="#">Home </a>
</li>
<li class="navc nav-item ">
<a class="nav-link
active" href="#abt">About</a>
</li>
</ul>
</div>
</div>
</nav>

```

```

<!-- Content -->
<div class="container d-flex
justify-content-center align-items-center">
<div class="text-right p-5 align-items-center">
<div class="container p-5">
<h1 class="fw-bolder fst-italic">
Solution to Detect<br>
Phishing Websites
</h1>
<p class="fs-5 text-wrap
justify-content-start">
Be aware of what's happening with your
confidential data
</p>
<button class="btn fw-bold btn-outline-light
rounded-pill" href="#abt">
GET STARTED</button>
<button class="btn fw-bold
btn-outline-light rounded-pill">
<a class="b2" href="https://youtu.be/Y7zNlEMDmI4">
WATCH VIDEO</a>
</button>
</div>
</div>
<div class="p-5">

```

```


</div>
</div>

<!-- About -->
<div class="p-5" id="abt">
<h1 class="text-center">ABOUT</h1>
<hr/>
<div class="con d-flex gap-5">
<div class="c1 justify-content-center">
<p class="text-wrap">Web service is one of the key communications
software services for the Internet. Web phishing is one of many
security threats to web services on the Internet. Web phishing aims
to steal private information, such as usernames, passwords, and
credit card details, by way of impersonating a legitimate entity.</p>
</div>
<div class="c2 justify-content-center">
<p class="text-wrap">The recipient is then tricked into clicking a
malicious link, which can lead to the installation of malware, the
freezing of the system as part of a ransomware attack or the
revealing of sensitive information. It will lead to information
disclosure and property damage.</p>
</div>
</div>
</div>

<!-- Check your website -->
<div class="webc d-flex justify-content-center align-items-center">
<div class="text-right p-5 align-items-center">
<div class="container p-5">
<h1 class="fw-bolder">
Check your Website
</h1>
<p class="fs-5 text-wrap justify-content-start">
Understanding if the website is a valid one or not is important
and<br>
plays a vital role in the securing the data.To know if the URL is
a<br>
valid one or you are information is at risk check your website.
</p>
</div>
</div>
<div class="p-5">

<button class="btn fw-bold btn-primary rounded-pill btn-lg"><a
href="/predict" class="b2" >Check Your Website</a></button>
</div>

```

```

</div>
</div>
</body>
</html>

```

## web.html

```

<!-- Prediction Page -->
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="viewport" content="width=device-width,
initial-scale=1.0">
<link href="https://cdn.jsdelivr.net/npm/
bootstrap@5.0.2/dist/css/bootstrap.min.css" rel="stylesheet">
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.0.2
/dist/js/bootstrap.bundle.min.js">
</script>

<link rel="stylesheet" href="style.css">
<title>Document</title>
<style>
*
{
padding: 0%;
margin: 0%;
}
.backg
{
background-color:#37517e;
width: 100vw;
height: 100vh;
color: white;
}
.nav-bar
{
font-family: Arial, Helvetica, sans-serif;
font-size: medium;
position: fixed;
}
.comname
{
font-family:cursive;
font-size: x-large;
}
.navc
{

```

```

background-color: transparent;
border:none;
}
.b2
{
text-decoration: none;
color: white;
}
.b2:hover
{
color: white;
}
#abt
{
background-color: white;
color: #1b335d;
}
hr
{
color: midnightblue;
}

hr:not([size])
{
height: 10px;
width: 30%;
margin-left: 490px;
}
.webc
{
background-color: #1b335d;
width: 100%;
height: auto;
}
.headi
{
padding-top: 150px;
}
</style>
</head>
<body>
<div class="backg">
<!-- Navigation bar -->
<nav class="navbar navbar-expand-lg navbar-light bg-blue">
<div class="container-fluid">
<a class="comname navbar-brand fw-bold" href="#">
WEBIO</a>
<div class="collapse navbar-collapse justify-content-
End pe-5">
<ul class="navbar-nav">

```

```

<li class="navc nav-item">
<a class="text-center nav-link active"
aria-current="page" href="index.html">Home
</a>
</li>
<li class="navc nav-item ">
<a class="nav-link active"
href="index.html">About
</a>
</li>
</ul>
</div>
</div>
</nav>

<div class="container text-center">
<h1 class="heading">Phishing Website Detection
using Machine Learning</h1>
<div class="p-5">
<form action="y_predict" method="post">
<input type="text" name="URL" id="URL"
placeholder="Enter the URL to be verified"
style="width: 500px;
border-radius:5px;">
<input type="submit" class="btn
btn-outline-light"
value="Predict"></input>
<h1>{{prediction_text}}</h1>
<h2><a>{{url}}</a></h2>
</form>
</div>
</div>
</div>
</body>
</html>

```

### **app.py**

```

import numpy as np from flask import Flask, request,
jsonify, render_template
import pickle
import feature
from sklearn import *
app = Flask(__name__)
model = pickle.load(open('Phishing_Website.pkl', 'rb'))
@app.route('/')

def predict1():
return render_template('index.html')
@app.route('/predict')

```

```

def predict():
    return render_template('web.html')
@app.route('/y_predict',methods=['POST'])
def y_predict():
    ...

For rendering results on HTML GUI
...

if request.method == 'POST':
    url = request.form['URL']
    checkprediction = feature.FeatureExtraction(url)
    prediction=model.predict(np.array(checkprediction.features).
    reshape(-1,30))
    print(prediction)
    output=prediction[0]
    if(output==1):
        pred="Your are safe!! This is a Legitimate Website."
    else:
        pred="You are on the wrong site. Be cautious!"
        return.render_template('web.html',
        prediction_text='{}'.format(pred),url=url)
        return.render_template('web.html',
        prediction_text='{}'.format(pred),url=url)
@app.route('/predict_api',methods=['POST'])
def predict_api():
    ...

For direct API calls trough request
...

data = request.get_json(force=True)
prediction = model.y_predict([np.array(list(data.values()))])

output = prediction[0]
return jsonify(output)

if __name__ == "__main__":
    app.run(debug=True)

if __name__ == '__main__':
    app.run(host='0.0.0.0', debug=True)

```



## **13.1 GITHUB AND PROJECT DEMO LINK**

### **13.1.1 GUTHUB LINK**

<https://github.com/IBM-EPBL/IBM-Project-46715-1660754625>

### **13.1.2 PROJECT DEMO LINK**

[https://github.com/Timothy025/project\\_recording](https://github.com/Timothy025/project_recording)

## CHAPTER 14

### REFERENCE

- [1] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol.484, pp. 153-166, 2019.
- [2] X. D. Hoang, "A website defacement detection method based on machine learning techniques," in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, 2018, pp. 443-448.
- [3] Odeh.A, Alarbi.A, Keshta.I & AbdelFattah.E (2020). Efficient predication of phishing prediction of phishing Website. *Journal of Theoretical and Applied Information Technology*, 98(16).
- [4] J. Singh and J. Singh, "A survey on machine learning-based malware detection in executable files," *Journal of Systems Architecture*, p. 101861, 2020
- [5] E. Zhu, Y. Ju, "DFOB-ANN: An Artificial Neural Network phishing detection model based on Decision Tree ," *Applied Soft Computing*, vol. 95, p. 106505, 2020
- [6] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN-MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites," *Neural Networks*, 2020.
- [7] R. Ravi, "A performance analysis of Software Defined Network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA)," *Computer Communications*, vol. 153, pp. 375-381, 2020.
- [8] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection URLs," *Expert Systems with Applications*, vol. 117, pp. 345-357, 2019.
- [9] P. Yi, Y. Guan, F. Zou, Y. Yao and T. Zhu, "Web phishing detection using a deep learning framework," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.
- [10] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using URL and HTML features for phishing webpage detection," *Future Generation Computer Systems*, vol. 94, pp. 27-39, 2019.