

WEB PHISHING DETECTION

(TEAM ID: PNT2022TMID34830)

PROJECT REPORT

Submitted by

ANOJ ARUL DAS A.E. (962819104016)

DINESH S.K.M. (962819104032)

DIVYA LIFNA.G. (962819104034)

EVARISTUS ZEN.M. (962819104037)

KAYAL VIZHI.K. (962819104054)

in partial fulfillment for the award of a

degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



UNIVERSITY COLLEGE OF ENGINEERING, NAGERCOIL

ANNA UNIVERSITY: CHENNAI 600 025

NOVEMBER 2022

INDEX

CHAPTER NO	Title	Page No
1	INTRODUCTION	1
1.1	Project Overview	2
1.2	Purpose	2
2	LITERATURE SURVEY	3
2.1	Existing Problem	6
2.2	References	6
2.3	Problem Statement Definition	7
3	IDEATION & PROPOSED SOLUTION	8
3.1	Empathy Map Canvas	8
3.2	Ideation & Brainstorming	9
3.3	Proposed Solution	11
3.4	Problem Solution Fit	13
4	REQUIREMENT ANALYSIS	14
4.1	Functional Requirement	14
4.2	Non-Functional Requirement	15
5	PROJECT DESIGN	16
5.1	Data Flow Diagram	16
5.2	Solution & Technical Architecture	17
5.3	Components, Technologies, and Application Characteristics	18
5.4	User Stories	19
6	PROJECT PLANNING & SCHEDULING	20
6.1	Sprint Planning & Estimation	20
6.2	Sprint Delivery Schedule	21
6.3	Burndown chart	21
7	CODING & SOLUTIONING	22
7.1	Code	22
7.2	Output	26

S.NO	Title	Page No
8	ADVANTAGES & DISADVANTAGES	27
8.1	Advantages	27
8.2	Disadvantages	27
9	CONCLUSION	28
10	References	30

CHAPTER 1

INTRODUCTION

The Internet has become an indispensable infrastructure that brings great convenience to human society. However, the Internet is also characterized by some inevitable security problems, such as phishing, malicious software, and privacy disclosure, which have already brought serious threats to the economy of users. Phishing is a very popular method used in network attacks, leading to privacy leaks, identity theft, and property damage. Phishing is the most unsafe criminal exercise in cyberspace. Since most of these go online to access the services provided by the government and financial institutions, there has been a significant increase in phishing attacks for the past few years. Phishers started to earn and they are doing this as a successful business. Various methods are used by phishers to attack valuable users such as messaging, spoofed links, and counterfeit websites. It is very easy to create counterfeit websites, which looks like genuine website in terms of layout and content. Even the website content would be identical to legitimate websites. The reason for creating these websites is to get private data from users like account numbers, login IDs, passwords of debit and credit cards, etc. Phishing attacks can be prevented by detecting the websites and creating awareness among users to identify the phishing websites. Machine Learning algorithms have been one of the most powerful techniques in detecting phishing websites. Several websites provide information and help users in detect phishing websites but are reliable. The primary objective of this research is to develop a detection system to solve the problems users are facing while accessing internet services. This system will help the users to identify the phished websites and give awareness to users with the most possible accuracy. Multiple machine learning classification algorithms were evaluated to develop the system. This research suggests a new approach towards web phishing detection where the phished sites are requested to block by the server administrator and the original website is recommended to the user.

1.1 Project Overview

Internet users can be able to recognize the phished website and legitimate websites easily through this project. The ML-based phishing techniques depend on website functionalities to gather the information that can help classify websites for detecting phishing sites. The problem of phishing cannot be eradicated, nonetheless can be reduced by combating it in two ways, improving targeted anti-phishing procedures and techniques and informing the public on how fraudulent phishing websites can be detected and identified. In this project, the focus is on URLs and Domain name features. Features of URLs and domain names are checked using several criteria such as IP Address, long URL address, adding prefix or suffix, redirecting using the symbol “//”, and URLs having the symbol ‘@’. These features are inspected using a set of rules to distinguish URLs of phishing webpages from the URLs of legitimate websites.

1.2 Purpose

The main purpose of this project is to detect fake or phishing websites that are trying to get access to sensitive data or by creating fake websites and trying to access the user’s credentials. We are using machine learning algorithms to safeguard sensitive data and to detect phishing websites that are trying to gain access to sensitive data.

CHAPTER 2

LITERATURE SURVEY

1) Paper Name: A Methodical Overview on Detection, Identification, and Proactive Prevention of Phishing Websites.

Author Name: Bhagwat M. D., Dr. Patil P. H., Dr. T. S. Vishwanath

Journal Name: Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

LS Content: [Reference no.] Bhagwat M. D., Dr. Patil P. H. and Dr. T. S. Vishwanath suggest a new approach to detect phished websites. They selected a range of features that differentiate phished websites from genuine websites and arranged these websites according to their priority by machine learning algorithms. They evaluated the features of a URL based on fuzzy rule systems. Their prototype allowed the users to enter the genuine website but if it is a phished website then this prototype sends a notification about the phished website to the corresponding host server administrator and the host server administrator blocks that phished website.

2) Paper Name: Chawathe, Sudarshan. (2018). Improving Email Security with Fuzzy Rules.

Author Name: Sudarshan

Journal Name: Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4

LS Content: [Reference no.] The more extreme security risks are phishing and other malicious email messages. Automated or semi-automated malicious email detection is an effective tool for combating such email threats. For such purposes, Sudarshan reviews work on using fuzzy rules to identify communications. Experimental review of the usefulness of a fuzzy rule-based classification for other classifiers like those that rely on crisp rules and decision trees, real data sets, and an output comparison.

3) Paper Name: Demonstrating Different Phishing Attacks Using Fuzzy Logic.

Author Name: Ms. S. D. Shweta Dasharath Shirsat.

Journal Name: Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 57-61, Doi: 10.1109/ICICCT.2018.8473309.

LS Content: [Reference no.] Phishers create fake websites that look close to legitimate websites and enable the consumer to visit the malicious website. Therefore, to secure their confidential data, users must be aware of malicious websites. But, particularly for non-technical users, it is very hard to differentiate between legitimate and fake websites. Phishing sites, in addition, are increasing rapidly. Ms. Shweta Dasharath Shirsat's goal is to use fuzzy logic to demonstrate phishing detection and interpret results using distinct de-fuzzification methods.

4) Paper Name: Intelligent phishing detection system for e-banking using fuzzy data mining.

Author Name: Maher & Hossain, Mohammed & Dahal, Keshav & Thabtah, Fadi.

Journal Name: Expert Systems with Applications.37. 7913-7921. 10.1016/j.eswa.2010.04.044.

LS Content: [Reference no.] In evaluating the e-banking phishing website, Maher Aburrous introduced a new technique for fixing 'fuzziness' and proposed a smart, resilient, successful e-banking phishing website detection model. Their model is a mixture of flippant logic and data mining techniques to define the features of the phishing banking website, analyze its techniques by categorizing phishing forms, and define various parameters for attacking the structured e-banking phishing layer.

5) Paper Name: A Method for The Automated Detection of Phishing Websites Through Both Site Characteristics and Image Analysis.

Author Name: White, Joshua & Matthews, Jeanna & Stacy, John.

Journal Name: The International Society for Optical Engineering. 8408. 84080B 84080B.10.1117/12.918956.

LS Content: [Reference no.] Joshua S. White introduces a technique for rapid automated website detection and analysis of phishing. Our methodology relies on the aggregation and review of URLs posted on social media sites in near real-time. They fetch the pages that each URL points to and describe each page with a set of values that are easily measured, such as the number of images and links. As a form of visual comparison, they also take a screenshot of the rendered page image, compute a hash of the image and use the Hamming distance between these image hashes.

6) Paper Name: Detection of Phishing Websites and Secure Transactions Detection of Phishing Websites and Secure Transactions.

Author Name: Dhanalakshmi, R & Prabhu, C & Chellapan, C.

Journal Name: International Journal Communication & Network Security (IJCNS). 1.

LS Content: [Reference no.] The use of a mixture of techniques of social engineering and criminals spoofing the website is an automated extortion of online identity to trick a user to disclose sensitive data. It gathers personal identification details and financial credentials from the user. Most phishing attacks appear as spoofed e-mails that make users trust and reveal them by clicking on the links given in the e-mail. The spoofed emails appear as legitimate ones. To describe the website, the claimed title is combined with human experts and domain features. A variety of legal websites link to domain recognition services, while phishing generally covers domain names and suspicious domain names (fake identities). In addition to blacklists, in the state-of-the-art schemes, white lists, heuristics, and classifications used; R. Dhanalakshmi is proposing to consider the identity statements of websites. With MD5 hashing algorithms, password hashing has been done to allow secure transactions, which strengthens the authentication of web passwords. Often it is, it has been shown that getting the actual password from the hashed form is not an easy task due to adding the salt meaning. Get a session key through mobile if the user is legitimate, from which further access can be done.

7) Paper Name: PhishNet: predictive blacklisting to detect phishing attacks.

Author Name: Pawan Prakash, Manish Kumar, Ramana Rao Kompella, and Minaxi Gupta.

Journal Name: International Journal Communication & Network Security (IJCNS). 1.

LS Content: PhishNet is a predictive blacklisting scheme to detect phishing attacks. Traditional blacklist approaches (i.e., an exact match with the blacklisted entries) are easy for attackers to evade. Instead, PhishNet uses five heuristics (i.e., top-level domains, IP address, directory structure, query string, and brand name) to compute simple combinations of blacklisted sites to discover new phishing sites. Also, it proposes an approximate matching algorithm to determine whether a given URL is a phishing site or not. fishnet consists of two major components, namely, component I: predicting malicious URLs, and component II: approximate matching.

8) Paper Name: Large-scale automatic classification of phishing pages.

Author Name: Colin Whittaker, Brian Ryner, and Marria Nazif.

Journal Name: In NDSS, volume 10, 2010.

LS Content: Whittaker et. al. uses a logistic regression classifier to maintain Google's phishing blacklist automatically by examining the URL and the contents of a page. The proposed scheme correctly classifies more than 90% of phishing pages several weeks after training concludes. Marchal et. al. develops a phishing detection system that requires very little training data, which is language-independent, resilient to adaptive attacks, and implemented entirely on the client side. The proposed target identification algorithm is faster than previous works and can help reduce false positives. The proposed scheme achieves a 0.5% false positive rate and a 99% true positive rate.

2.1 Existing Problem

The current solutions of antivirus, firewall, and designated software do not fully prevent the web spoofing attack. The implementation of a Secure Socket Layer (SSL) and digital certificate (CA) also does not protect the web user against such attacks. In a web spoofing attack, the attacker diverts the requests to the fake web server. This project develops an anti-webphishing solution based on inspecting the URLs of fake web pages. This solution developed a series of steps to check the characteristics of the website's Uniform Resources Locators (URLs).

2.2 References

1. Anti-Phishing Working Group (APWG), https://docs.apwg.org/reports/apwg_trends_report_q4_2019. Pdf.
2. Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, https://doi.org/10.1007/978-981-10-8536-9_44.
3. Purbay M., Kumar D, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021, https://doi.org/10.1007/978-981-15-6840-4_40.
4. Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12.
5. Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.

6. Hong J., Kim T., Liu J., Park N., Kim SW, “Phishing URL Detection with Lexical Features and Blacklisted Domains”, Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.
7. J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Bindhumadhava, “Phishing Website Classification and Detection Using Machine Learning,” 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.
8. Hassan Y.A. and Abdelfettah B, “Using case-based reasoning for phishing detection”, *Procedia Computer Science*, vol. 109, 2017, pp. 281–288.
9. Rao RS, Pais AR. Jail-Phish: An improved search engine-based phishing detection system. *Computers & Security*. 2019 Jun 1; 83:246–67.
10. J. Anirudha and P. Tanuja, Phishing Attack Detection using Feature Selection.
11. Techniques “, Proceedings of International Conference on Communication and Information Processing (ICCIP), 2019, [http:// dx.doi.org/10.2139/ssrn.3418542](http://dx.doi.org/10.2139/ssrn.3418542).

2.3 Problem Statement Definition

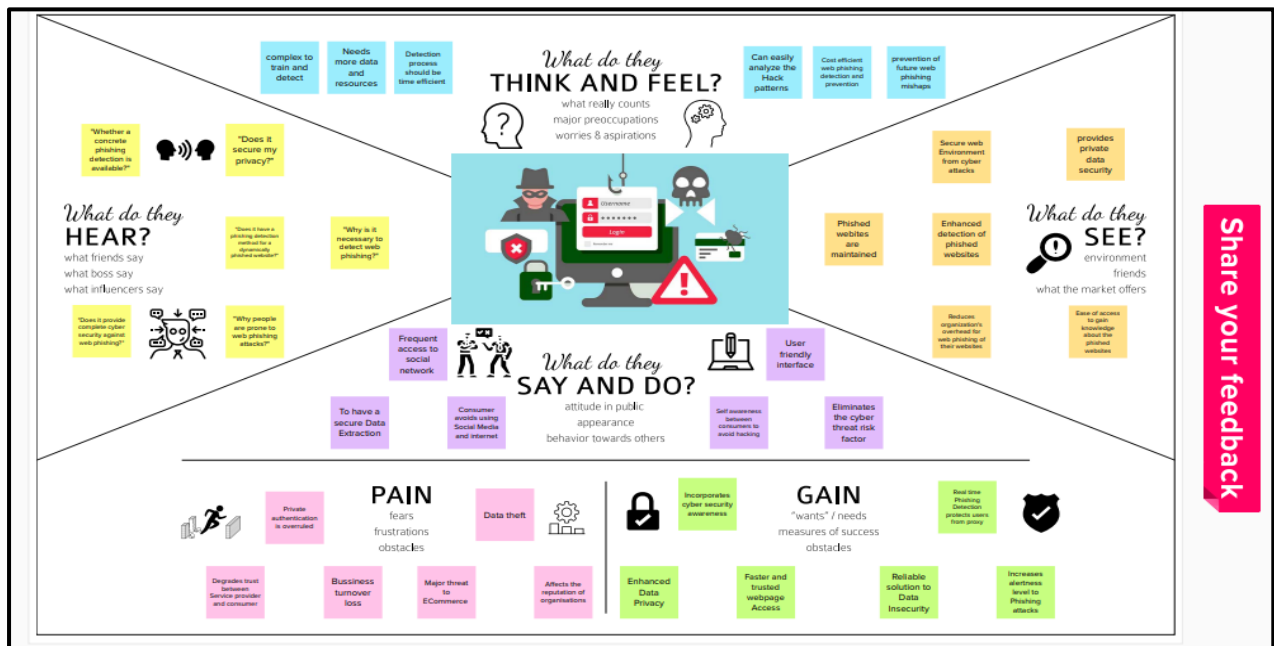
Phishing is one of the techniques which are used by intruders to get access to user credentials or to gain access to sensitive data. This type of access is done by creating a replica of the websites which look the same as the original websites which we use daily but when a user clicks on the link, he will see the website and think it is original and try to provide his credentials.

To overcome this problem, we are using some machine learning algorithms it will help us to identify phishing websites based on the features present in the algorithm. By using this algorithm, we can able to keep the user’s credentials or sensitive data safe from intruders.

CHAPTER 3

IDEATION & PROPOSED SOLUTION


3.1 Empathy Map Canvas



3.2 Ideation and Brainstorming

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

Template



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

Share template feedback

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 session →

1 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

Problem Statement

There are lot of cyber threats and crimes which allows the hackers to acquire sensible and valuable information of a user in a specific firm without their appropriate concern.

One such notorious cyber crime among them is Web Phishing through which a hacker creates a fake profile of a website and hacks the entire information from the user through the user themselves.

There are a lot of websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons.

This type of websites, resembling the original websites are known as a phished website and the process of creating such websites is called web Phishing.

Web services are one of the key communication software services for the Internet. Web phishing is one of many security threats to these web services on the Internet.

Step 2: Brainstorm, Idea Listing, and Grouping

2 Brainstorm
Write down any ideas that come to mind that address your problem statement.

10 minutes

ANOJ ARUL DHAS A E

- Develop a Content-rich website
- Send a report on cyber threats to the user
- Develop a Content-rich website
- Understand the common phishing threats
- Understand the common phishing threats
- Understand the common phishing threats

DINESH S K M

- Develop a Content-rich website
- Send a report on cyber threats to the user
- Develop a Content-rich website
- Understand the common phishing threats
- Understand the common phishing threats
- Understand the common phishing threats

EVARISTUS ZEN S

- Develop a Content-rich website
- Send a report on cyber threats to the user
- Develop a Content-rich website
- Understand the common phishing threats
- Understand the common phishing threats
- Understand the common phishing threats

KAYAL VIZHI K

- Develop a Content-rich website
- Send a report on cyber threats to the user
- Develop a Content-rich website
- Understand the common phishing threats
- Understand the common phishing threats
- Understand the common phishing threats

3 Group Ideas
Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, 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

About Web Phishing

By an average of over 30 billion websites are created by hackers every day.

Common threats are stealing private informations

US tops the table by hosting more phished web sites

Most targeted and many websites across the globe are targeted by hackers every day.

Detection Process

- Analyze the features of phishing websites
- Set the priority of each feature using Machine learning algorithms
- Assign the given url features using Decision tree algorithms
- Display whether the site is phished website or not

Step 3: 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

TIP

Participants can use their cursors to point at where sticky notes should go on the grid. The facilitator can confirm the spot by using the laser pointer holding the **H** key on the keyboard.



3.3 Proposed Solution

S. No	Parameter	Description
1.	Problem Statement	<ul style="list-style-type: none"> ➤ There are a lot of cyber threats and crimes which allow hackers to hack sensible and valuable information of a user in a specific firm without their appropriate concern. ➤ One such notorious cybercrime among them is Web Phishing through which a hacker creates a fake profile of a website and hacks the entire information from the user through the user. ➤ There are a lot of websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. ➤ This type of website resembling the original website is known as a phished website and the process of creating such websites is called web Phishing. ➤ Major web phishing attacks are held on E-commerce based websites, especially on banking websites. ➤ Web services are one of the key communication software services for the Internet. Web phishing is one of many security threats for web services on the Internet.
2.	Idea / Solution description	<ul style="list-style-type: none"> ➤ The solution for the phishing attack can be achieved by using a Machine Learning algorithm where two datasets are taken (Original Websites and Phished Websites) and trained. ➤ By detecting phishing attacks in the background user can easily identify cloned websites.
3.	Novelty / Uniqueness	<ul style="list-style-type: none"> ➤ Machine Learning Approach ➤ Pre-defined blacklisted website dataset ➤ Web address-based evaluation metric to achieve low-level phishing detection. ➤ Use of Heuristic rule-based detection techniques.

		<ul style="list-style-type: none"> ➤ The proposed idea suggests a new approach towards web phishing detection where the phished sites are requested to block by the server administrator and the original website is recommended to the user.
4.	Social Impact / Customer Satisfaction	<ul style="list-style-type: none"> ➤ By achieving efficient web phishing detection, the users are free from data theft. ➤ A huge barrier cross can be achieved in the case of E-banking websites. ➤ Secure users from proxies and scams.
5.	Business Model (Revenue Model)	<ul style="list-style-type: none"> ➤ Profitable to E-commerce and E-banking-based service providers. ➤ The government sector can be more digitalized and a secure web service experience can be achieved.
6.	Scalability of the Solution	<ul style="list-style-type: none"> ➤ Adapts to all sorts of web applications and ease of preventing users from scams. ➤ Apart from the E-banking sector, the idea proposed can be developed into platform independent model.

3.4 Proposed Solution Fit

Define CS, fit into CC	1. CUSTOMER SEGMENT(S) <ul style="list-style-type: none"> • Organizationbased on E-commerce • Banking and Insurance sectors 	6. CUSTOMER CONSTRAINTS <ul style="list-style-type: none"> • Lack of knowledge about web phishing • Lack of organized sectors against web phishing • Lack of law enforcement against web phishing 	5. AVAILABLE SOLUTION <ul style="list-style-type: none"> • Antivirus software • Firewall • web security gateway • Anti-phishing tools 	Explore AS, differentiate
Focus on JSP, tap into BE, understand RC	2. JOBS-TO-BE-DONE / PROBLEMS <ul style="list-style-type: none"> • Ensure safety of personal data • Ensure legit use of original website • Prevent user from suspicious malwares 	9. PROBLEM ROOT CAUSE <ul style="list-style-type: none"> • Increasing use of internet • Customers depending on comfortless • Lack of knowledge about web attacks 	7. BEHAVIOUR <ul style="list-style-type: none"> • Backup important files • Change login credentials often • Scan system for viruses 	Focus on JSP, tap into BE, understand RC

	3. TRIGGERS <ul style="list-style-type: none"> • Prevent important credentials of customers being stolen • Prevent financial losses and trust issues that occurs 4. EMOTIONS: BEFORE / AFTER <ul style="list-style-type: none"> • Before: Scared about data theft • After: Feeling safe that our data is secure now 	10. YOUR SOLUTION <ul style="list-style-type: none"> • Identify and block phish websites • Recommend original website • Validate website's identity regularly • Organize, group and secure original websites 	8. CHANNELS of BEHAVIOUR 8.1 ONLINE <ul style="list-style-type: none"> • Disconnect from internet • Check for viruses 8.2 OFFLINE <ul style="list-style-type: none"> • Call an expert 	
--	---	---	---	--

CHAPTER 4

REQUIREMENT ANALYSIS

Requirements analysis is critical to the success or failure of a systems or software project. The requirements should be documented, actionable, measurable, testable, traceable, related to identified business needs or opportunities, and defined to a level of detail sufficient for system design.

4.1 Functional Requirements:

Functional requirements explain what has to be done by identifying the necessary task, action, or activity that must be accomplished. Functional requirements analysis will be used as the top-level function for functional analysis.

- **Learning & Detection**

The samples and the topological structure of the machine learning Tensor Flow are built. The submitted URLs are tested against the samples in the database to perform classification

- **Testing& Alert**

URLs passed through the system are recorded in a database, thus each URL submitted by the user is tested to check or duplicate. If a phishing website is detected the popup message will alert the user. Give information about the malicious website with accurate results.

- **Deep Learning**

The phishing detection process could be done using the Recurrent Neural Network. The website could be detected.

4.2 Non-Functional Requirements

Non-functional requirements are requirements that specify criteria that can be used to judge the operation of a system, rather than specific behaviors.

- **Usability**

This system is used as it can able to detect phishing websites. By detecting malicious websites, our personal and professional data are confidential, secure, and accessible

- **Security**

Phishers spoof legitimate emails so that the victim trusts them. They send out massive numbers of fraudulent emails to catch a small percentage of recipients off guard. They create a sense of urgency so that the victim does not think twice before clicking the link or downloading the attachment. Lack of security awareness among employees is also one of the major reasons for the success of phishing. Organizations should be aware of how the benefits and purpose of security awareness training can secure their employees from falling victim to phishing attacks.

- **Reliability**

The performance of the system would be accurate. The probability of giving false information is very low. As the system is working based on the deep learning algorithm, it would easily predict and give the perfect information.

- **Performance**

The effectiveness of these methods relies on feature collection, training data, and classification algorithms and giving alerts when phished websites are detected. It must be processed and executed within a fraction of a second using the deep learning algorithm

- **Availability**

The availability of the solution is effective and it should be helpful in a great way to prevent our data to be exposed.

- **Scalability**

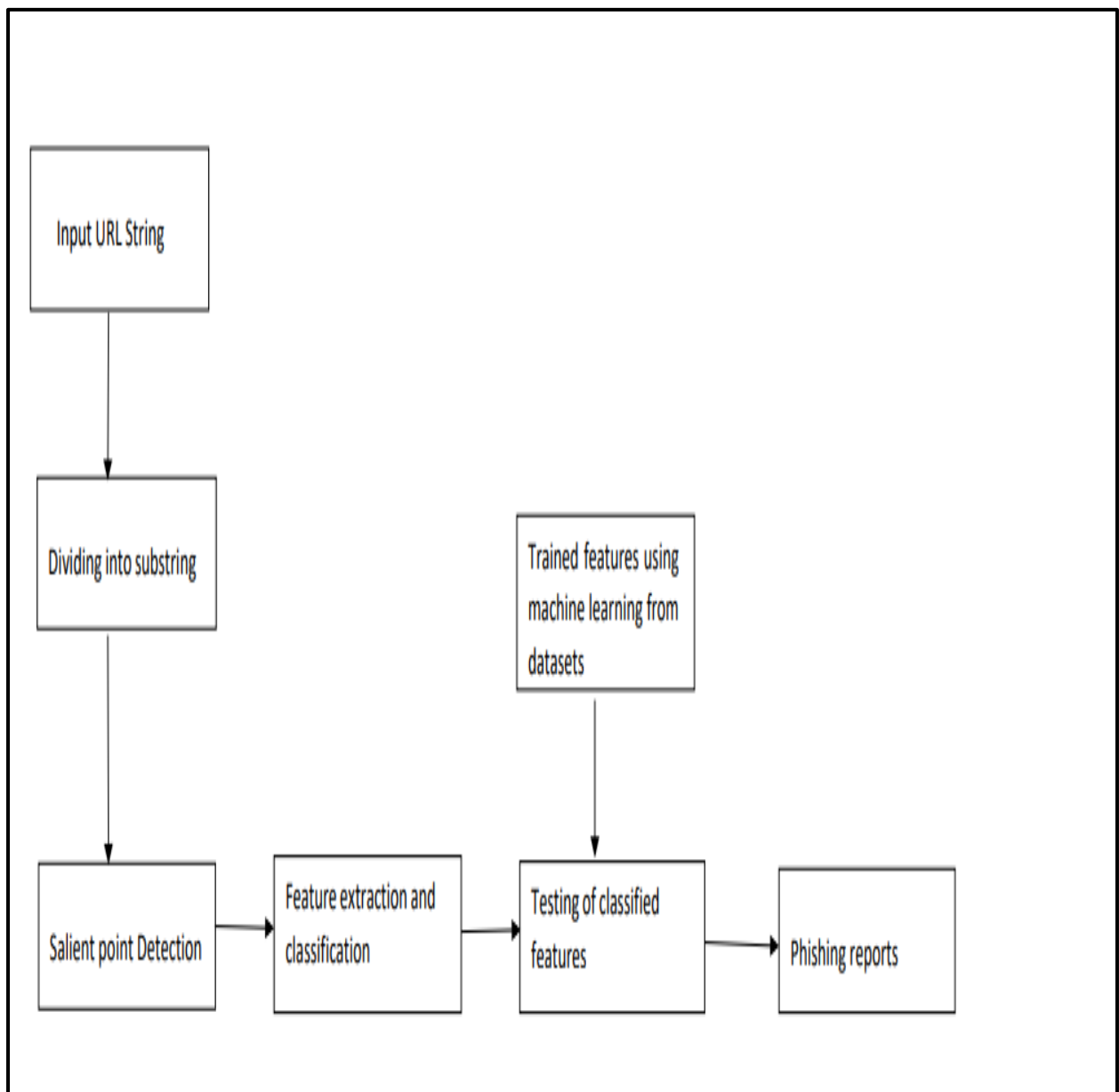
This solution is scalable enough to fit the Security issues by constructing the best website. The cost of establishing the website and maintaining all the programs may be high. It is acceptable to fit them over any place and any resources.

CHAPTER 5

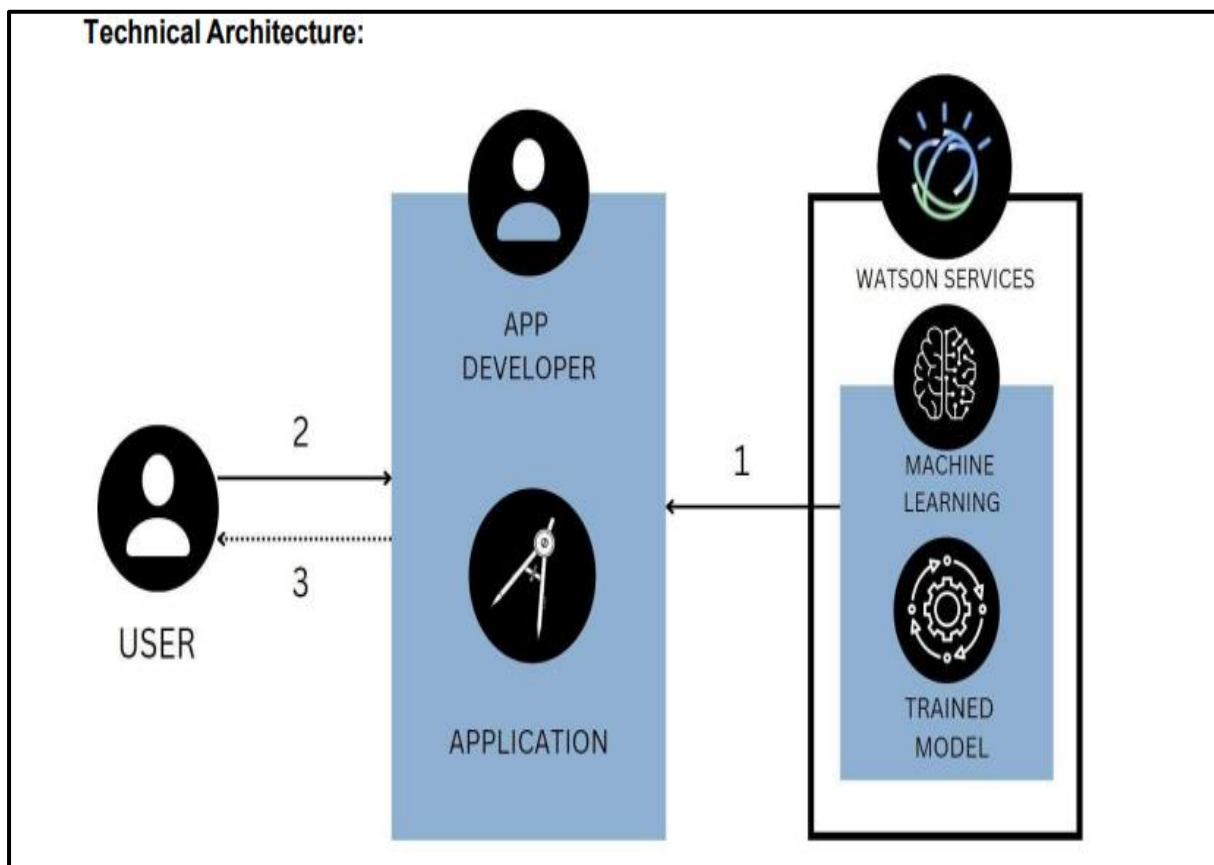
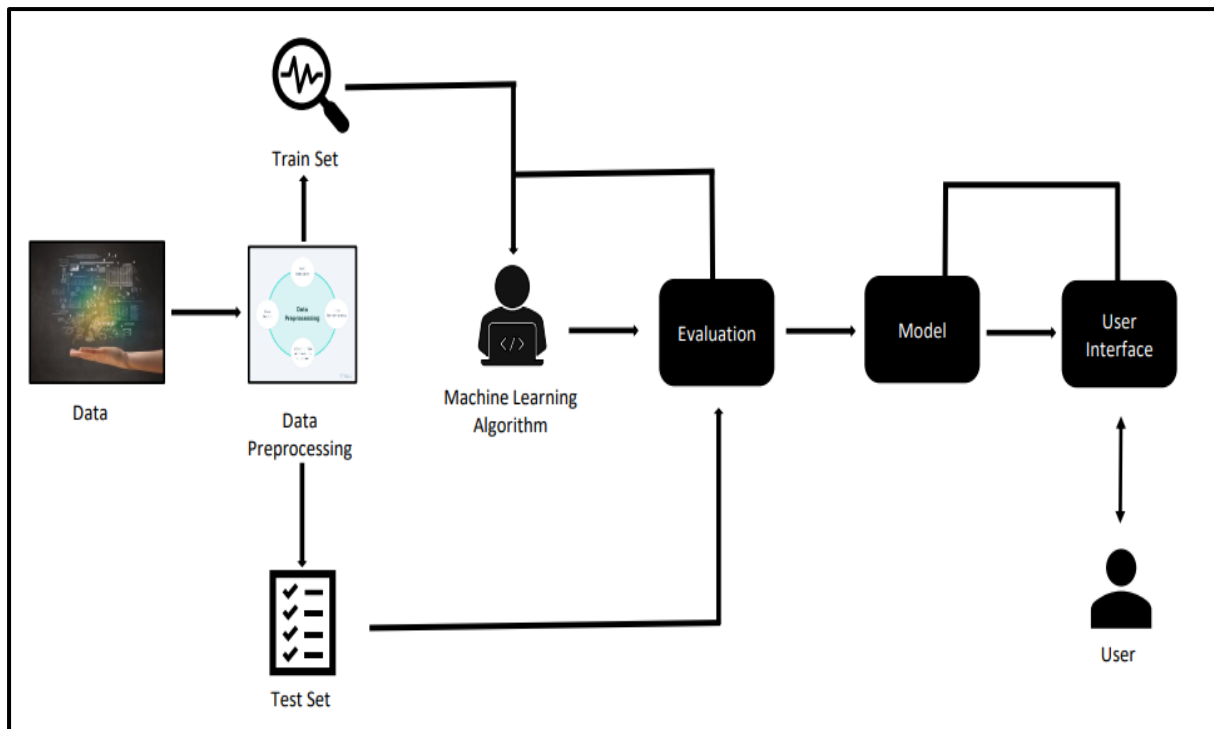
PROJECT DESIGN

5.1 Data Flow Diagram

A Data Flow Diagram (DFD) is a visual representation of the information flows within a system. It can be manual, automated, or a combination of both.



5.2 Solution & Technical Architecture



5.3 Components, Technologies, and Application characteristics:

Table-1: Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	A web application with information about phishing and a form field to get the URL from user to check genuineness.	HTML, CSS, JavaScript, Bootstrap
2.	Application Logic	Predict if the given URL is legitimate or not.	Flask API, Python
3.	Database	Store user input links in the database.	MySQL
4.	File Storage	Store training and testing datasets.	Local Filesystem
5.	Machine Learning Model	Classify legitimate and phishing URLs using XGBoost	Classification model
6.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud	Local, Cloud

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	High-level open-source frameworks	Flask, Bootstrap
2.	Security Implementations	It is the security discipline that makes it possible for the right entities (people or things) to use the right resources (applications or data) when they need to, without interference, using the devices they want to use.	IAM Controls
3.	Scalable Architecture	Compose is a tool for defining and running multi-container Docker applications. With a single command, can create and start all the services from the configuration.	Docker, Docker Compose
4.	Availability	It can balance the load traffic among the servers to help improve uptime. Can scale applications by adding or removing servers, with minimal disruption to traffic flows.	IBM Cloud load balancers
5.	Performance	It provides performance feedback such as page size and how long it takes to load a page, and can show the impact new features have on the performance of the site.	IBM's SpeedCurve and Delivery Pipeline

5.4 User Stories

A user story is a note that captures what a user does or needs to do as part of his/her work. Each user story consists of a short description written from the user's point of view, with natural language.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Rel
Customer (C-suite executive, CEO, mobile user, web user)	Detect and predict phishing websites	USN-1	As a user, I detect phishing websites	I can protect my data from getting stolen	High	Spr
Customer (C-suite executive, CEO, mobile user, web user)	Identify Fraudulent URL	USN-2	As a user, I need to identify the URL that looks suspicious	I can protect my data from hackers	High	Spr
Customer (C-suite executive, CEO, mobile user, web user)	Identification of valid or invalid URL	USN-3	As a user, I need to identify whether a URL is valid or not	I can prevent online money theft	High	Spr
Customer (C-suite executive, CEO, mobile user, web user)	Identification of the accuracy level of detected phished domains	USN-4	As a user, I need to know the accuracy level of detected phished domains	I can ensure the safety	Medium	Spr
Customer (C-suite executive, CEO, mobile user, web user)	Identify false positives and false negatives	USN-5	As a user, I need to identify false positives and false negatives	I can prevent unwanted malware	Low	Spr

CHAPTER 6

PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Home Page	USN-1	As a user, a simple user interface with a search bar and a detect button provides an immense user experience.	10	Medium	Anoj arul dhas, Dinesh Evaristus Zen
Sprint-1	Dashboard	USN-2	As a user, the information concerning the The functionalities of the webpage and its uses can be known.	10	Medium	Divya Lifna, Kaval Vizhi
Sprint-2	Dataset Collection	USN-3	Two datasets comprise legitimate websites and the other comprises phished and blacklisted websites which are featured to train and test the model.	10	Medium	Evaristus Zen, Divya Lifna, Kaval Vizhi
Sprint-2	Data Pre-processing	USN-4	Fetches the features of the data and pre-processes the data for training and testing.	10	High	Anoj arul dhas, Dinesh
Sprint-3	Training the data	USN-5	The featured data is trained using several Machine Learning Algorithms and among the best Algorithm is preferred for the model.	10	High	Anoj arul dhas, Dinesh, Evaristus Zen
Sprint-3	Testing the data	USN-6	The Best Algorithm is implemented to detect the phished sites.	10	High	Divya Lifna, Kaval Vizhi
Sprint-4	Prediction	USN-7	The user would be able to analyze whether the website is a real website or a phishing website.	5	High	Anoj arul dhas, Dinesh
Sprint-4	Result page	USN-8	Users can able to see the results web page of the analysis.	5	High	Evaristus Zen
Sprint-4	User query	USN-9	Users can able to reply any queries regarding the results of the analysis.	5	Low	Divya Lifna
Sprint-4	Contact	USN-10	Users can able to contact the developer directly for any other queries.	5	Low	Kaval Vizhi

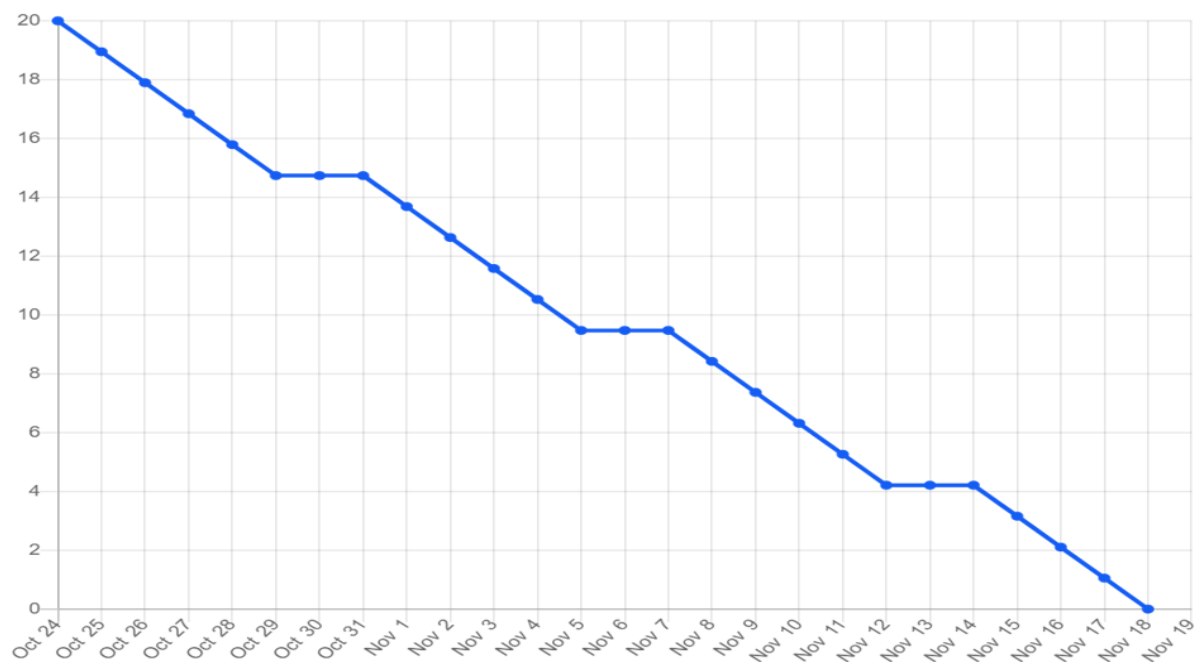
6.2 Sprint Delivery Schedule

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Nov 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Nov 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.3 Burndown Chart:

A Burndown chart is a project management chart that shows how quickly a team is working through a customer's user stories. This agile tool captures the description of a feature from an end-user perspective and shows the total effort against the amount of work for each iteration or agile sprint.



CHAPTER 7

CODING & SOLUTIONING

7.1 Code

```
app.py

import flask
from flask import request,render_template
from flask_cors import CORS
import joblib
from urllib.parse import urlparse,urlencode
import ipaddress
import re
import socket
import sklearn

app=flask.Flask(__name__,static_url_path='')
CORS(app)

@app.route('/',methods=['GET'])
def SendIndexPage():
    return render_template('index.html')

@app.route('/result',methods=['POST'])
def predictResult():
    url=request.form['URL']
    a=getDomain(url)
    b=havingIP(url)
    c=haveAtSign(url)
    d=getLength(url)
    e=getDepth(url)
    f=redirection(url)
    g=httpDomain(url)
    h=tinyURL(url)
    i=prefixSuffix(url)
    x=[[b,c,d,e,f,g,h,i]]
    model=joblib.load('Selected_Model.pkl')
    res=model.predict(x)[0]
    if(res==1):
        res1="Legitimate Website"
    elif(res==0):
        res1="Phished Website"
```

```

Appp_IBM.py

import flask
from flask import request,render_template
from flask_cors import CORS
from urllib.parse import urlparse,urlencode
import ipaddress
import re
import socket
import sklearn
import requests

# NOTE: you must manually set API_KEY below using information retrieved from your IBM
Cloud account.
API_KEY = "wDtCjNwnNmDwAyp3EkqOTdZFmealKljICmh_Xd4ZF0eF"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',
data={"apikey":API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

app=flask.Flask(__name__,static_url_path='')
CORS(app)

@app.route('/',methods=['GET'])
def SendIndexPage():
    return render_template('index.html')

@app.route('/result',methods=['POST'])
def predictResult():
    url=request.form['URL']
    a=getDomain(url)
    b=havingIP(url)
    c=haveAtSign(url)
    d=getLength(url)
    e=getDepth(url)
    f=redirection(url)
    g=httpDomain(url)
    h=tinyURL(url)
    i=prefixSuffix(url)
    x=[[b,c,d,e,f,g,h,i]]
    payload_scoring = {"input_data": [{"fields": [[b,c,d,e,f,g,h,i]], "values": x}]}
    response_scoring = requests.post('https://us-
south.ml.cloud.ibm.com/ml/v4/deployments/49108a1a-e0a8-48e3-862d-
9531ceb51ee3/predictions?version=2022-11-21',

```

Index.html

```
<html>
  <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">
    <title>Home Page</title>
    <style>
      * {
        margin: 0;
        padding: 0;
        box-sizing: border-box;
        font-family: 'Poppins', sans-serif;
      }

      .header {
        background: linear-gradient(rgba(0, 8, 51, 0.9), rgba(0, 8, 51, 0.9), rgb(45,
45, 150));
        text-align: center;
      }

      .header h1 {
        flex: 1;
        border: 0;
        outline: none;
        padding: 24px 20px;
        font-size: 40px;
        color: #cac7ff;
      }

      .container {
        width: 100%;
        min-height: 100vh;
        padding: 5%;
        background-image: linear-gradient(rgba(0,8,51,0.9),rgba(0,8,51,0.9));
        background-position: center;
        background-size: cover;
        display: flex;
        align-items: center;
        justify-content: center;
      }

      .search-bar {
        width: 100%;
```

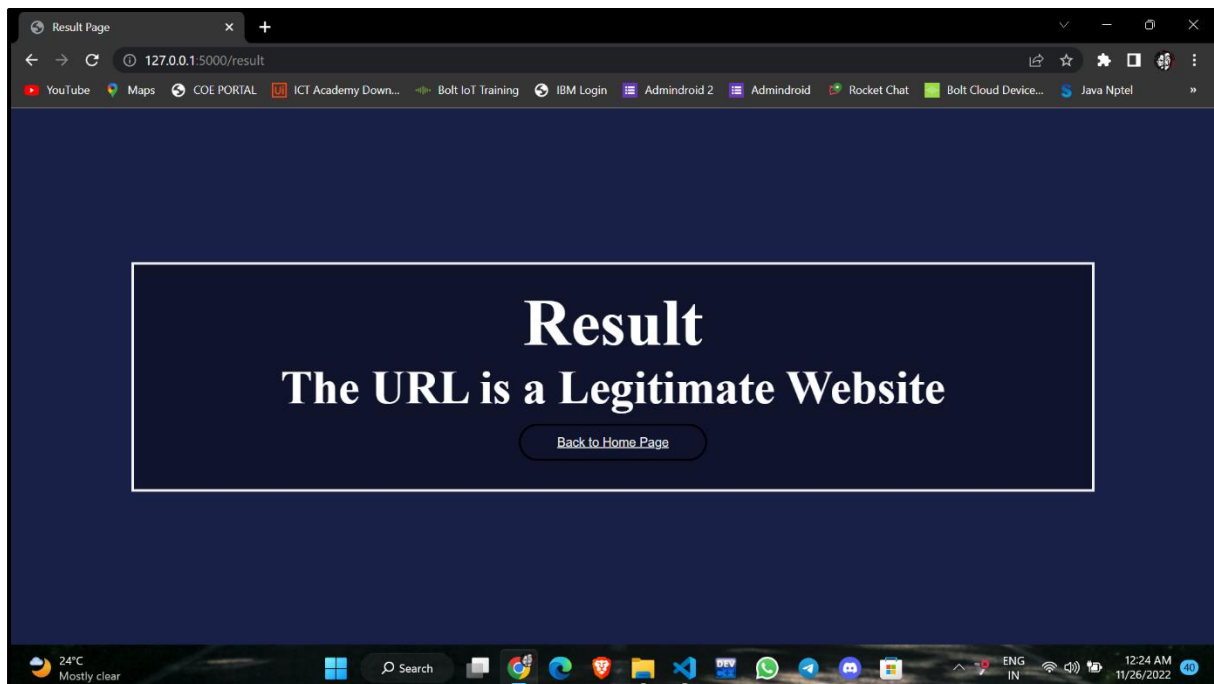
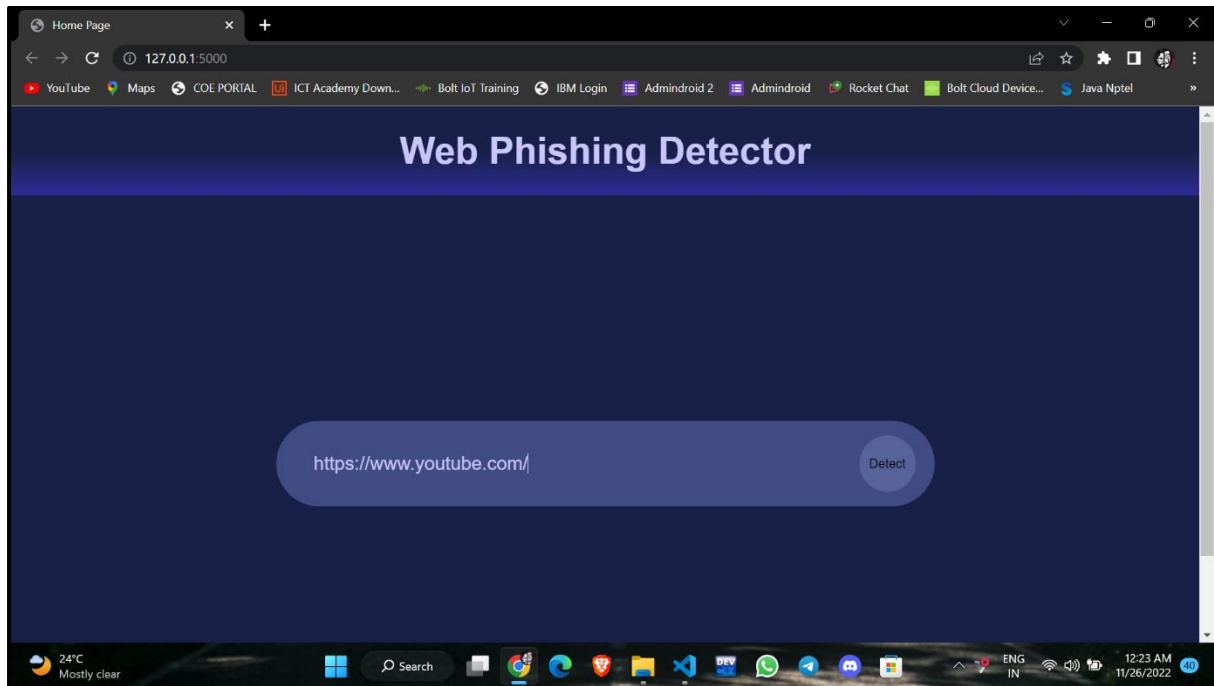
result.html

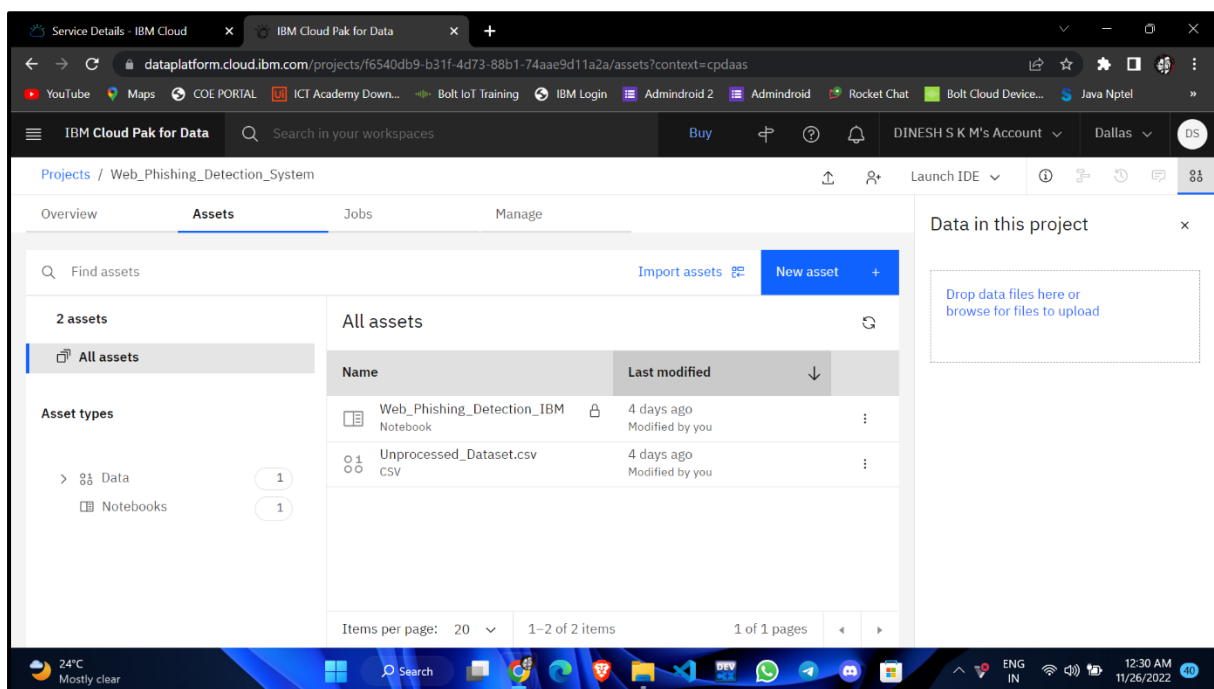
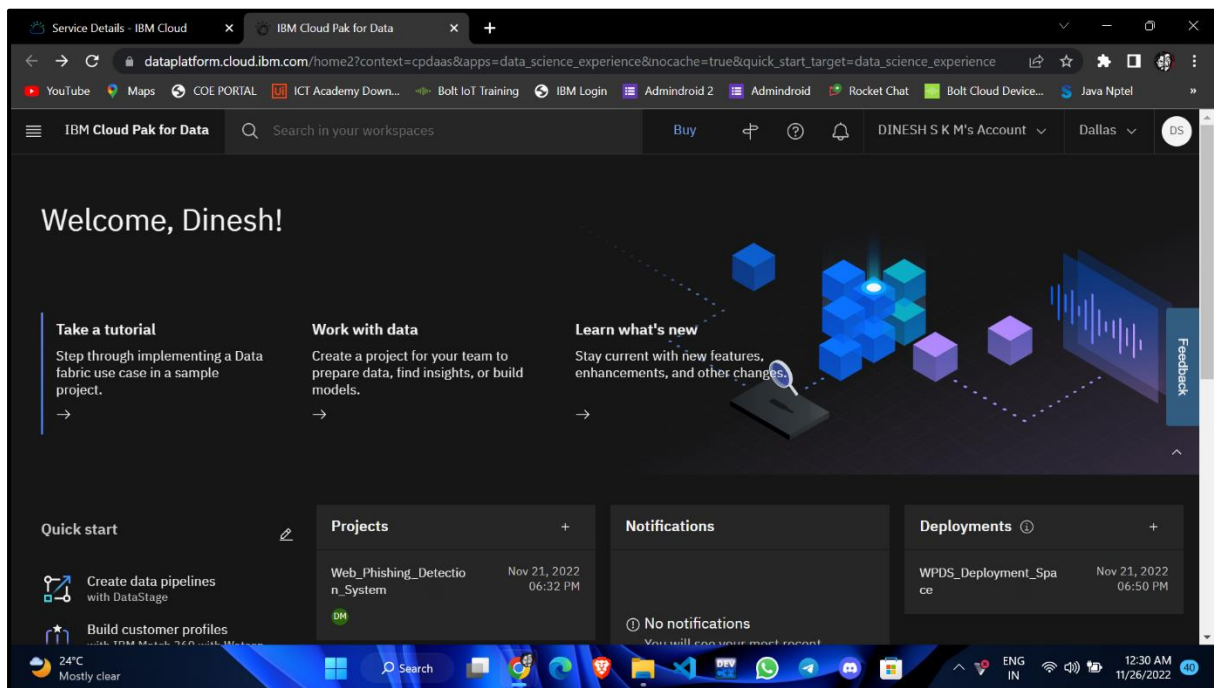
```
<!DOCTYPE html>
<html lang="en">
  <head>
    <title>Result Page</title>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <style>
      * {
        padding: 0;
        margin: 0;
        box-sizing: border-box;
      }

      .bg-img {
        width: 100%;
        min-height: 100vh;
        padding: 5%;
        background-image: linear-gradient(rgba(0,8,51,0.9),rgba(0,8,51,0.9));
        background-position: center;
        background-size: cover;
        display: flex;
        align-items: center;
        justify-content: center;
      }

      .bg-text {
        background-color: rgb(0, 0, 0);
        /* Fallback color */
        background-color: rgba(0, 0, 0, 0.4);
        /* Black w/opacity/see-through */
        color: white;
        font-weight: bold;
        border: 3px solid #f1f1f1;
        position: absolute;
        top: 50%;
        left: 50%;
        transform: translate(-50%, -50%);
        z-index: 2;
        width: 80%;
        padding: 20px;
        text-align: center;
      }
    </style>
  </head>
  <body>
    <div class="bg-img">
      <div class="bg-text">
        <h1>Result Page</h1>
      </div>
    </div>
  </body>
</html>
```

7.2 Output





CHAPTER 8

ADVANTAGES & DISADVANTAGES

8.1 Advantages:

- This system can be used by many E-commerce or other websites to have a good customer relationships.
- Users can make online payments securely.
- K Nearest Neighbor Algorithms used in this system provides better performance as compared to other traditional classification algorithms.
- With the help of this system, users can also purchase products online without any hesitation.

8.2 Disadvantages:

- If the internet connection fails, this system won't work.
- The model is an external site, which means for each time the web app should be opened to verify the website.
- No accurate result for the URL of extended features.

9. CONCLUSION:

Users commonly have many user accounts on various websites including social networks, email, and also accounts for banking. Therefore, innocent web users are the most vulnerable targets for this attack since the fact that most people are unaware of their valuable information, which helps to make this attack successful. The proposed system emphasized the phishing technique in the context of classification, where phishing websites are considered to involve the automatic categorization of websites into a predetermined set of class values based on several features and the class variable. The ML-based phishing techniques depend on website functionalities to gather the information that can help classify websites for detecting phishing sites. The problem of phishing cannot be eradicated, nonetheless can be reduced by combating it in two ways, improving targeted anti-phishing procedures and techniques and informing the public on how fraudulent phishing websites can be detected and identified. To combat the ever-evolving complexity of phishing attacks and tactics, ML anti-phishing techniques are essential. The outcome of this system reveals that the proposed method presents superior results rather than the existing deep learning methods. It has achieved better accuracy.

10. References:

IBM Project GitHub Repository: <https://github.com/IBM-EPBL/IBM-Project-30874-1660191732>

Video Link: <https://clipchamp.com/watch/qHDMo0lvtNx>