FINAL REPORT

Web Phishing Detection

TEAM MEMBERS:

E.GAJALAKSHMI

KIRUTHIKA.P

MAGHNAA.S

HARIHARAN.P.K

TABLE OF CONTENTS

Web Phishing Detection

1. INTRODUCTION

1.1 Project Overview

Phishing is a type of cybersecurity attack during which malicious actors send messages pretending to be a trusted person or entity. Phishing messages manipulate a user, causing them to perform actions like installing a malicious file, clicking a malicious link, or divulging sensitive information such as access credentials. Phishing is the most common type of social engineering, which is a general term describing attempts to manipulate or trick computer users. Social engineering is an increasingly common threat vector used in almost all security incidents. Social engineering attacks, like phishing, are often combined with other threats, such as malware, code injection, and network attacks.

1.2 Project Description

We have developed our project using a website as a platform for all the users. This is an interactive and responsive website that will be used to detect whether a website is legitimate or phishing. This website is made using different web designing languages which include HTML, CSS, Javascript and Django. The basic structure of the website is made with the help of HTML. CSS is used to add effects to the website and make it more attractive and user-friendly. It must be noted that the website is created for all users, hence it must be easy to operate with and no user should face any difficulty while making its use. Every naïve person must be able to use this website and avail maximum benefits from it. The dataset consists of different features that are to be taken into consideration while determining a website URL as legitimate or phishing.

The components for detection and classification of phishing websites are as follows:

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript Based Features
4. Domain Based Features

2. LITERATURE SURVEY

2.1 Problem Statement Definition

Phishing attacks are becoming more and more sophisticated, and our algorithms are suffering to keep up with this level of sophistication. They have low detection rate and high false alarm especially when novel phishing approaches are use. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier. Moreover, comprehensive blacklist can ensure a perfect up-to-date database. Various other techniques such as page content inspection algorithms have been used to combat the false negatives but as each algorithm uses a different approach, their accuracy varies. Therefore, a combination of the two can increase the accuracy while implementing different error detection methods.

2.2 References

[1] Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon Lee, 'Heuristic based approach for phishing site detection using URL features' in Proceedings of 3rd International Conference in Computing, Electronics and Electrical Technology - CEET 2015.

[2] Sadia Afroz, Rachel Greenstadt Department of Computer Science Drexel University Philadelphia, PA 19104 Email: sa499@drexel.edu, 2011 'PhishZoo: Detecting Phishing Websites By Looking at Them' in 2011 Fifth IEEE International Conference on Semantic Computing.

[3] Michael Blasi Iowa State University, 'Techniques for detecting zero day phishing websites', 2009.

[4] Chen Jin, Luo De-lin* School of Information Science and Technology Xiamen University Xiamen, 361005, China, 'An Improved ID3 Decision Tree Algorithm' in Proceedings of 2009 4th International Conference on Computer Science & Education.

[5] Zou Futai, Gang Yuxiang, Pei Bei Key Lab of Information Network Security Ministry of Public Security, Pan Li, Li Linsen, 'Web Phishing Detection Based on Graph Mining' in 2016 2nd IEEE International Conference on Computer and Communications.

[6] Phoebe Barraclough, Graham Sexton, 'Phishing Website Detection Fuzzy System Modelling', in proceedings of Science and Information Conference 2015.

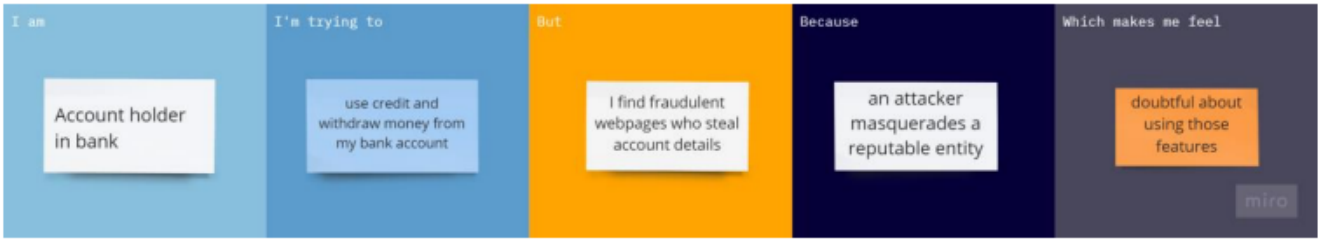
[7] https://www.phishtank.com/

[8] http://dmoztools.net/

2.3 Problem Statement

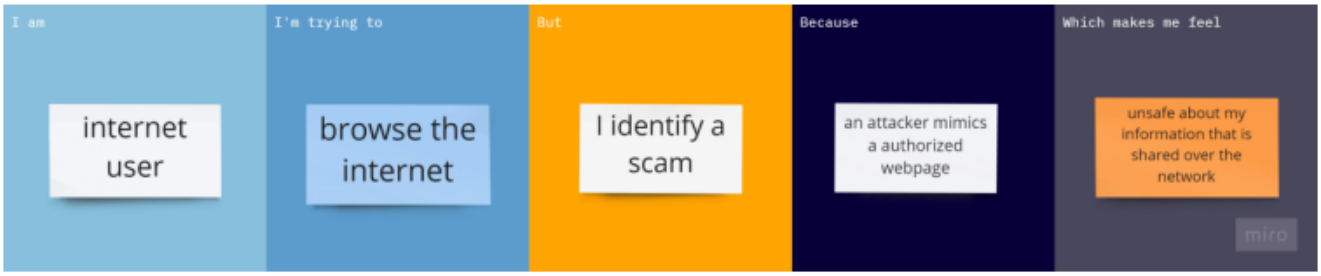
Problem Statement 1:

Problem Statement 2:



Problem Statement 3:

Problem Statement 4:



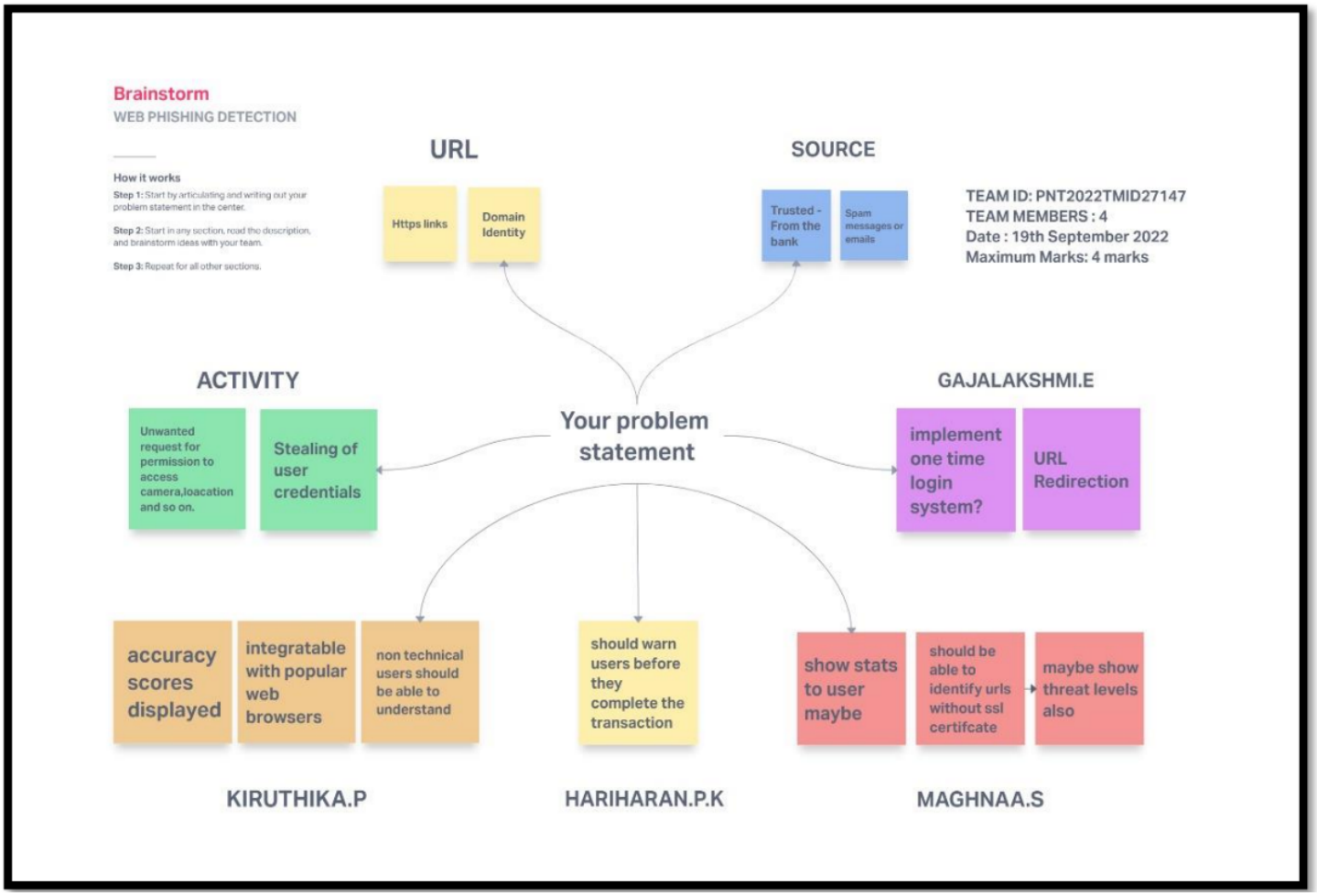
Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	Vendor	Use online transactions	I find illegal pages who indulge in bankruptcy	Of counterfeit websites who steal credentials	Unsafe about online transactions
PS-2	Account holder in Bank	Use credit and withdraw money from bank account	I find fraudulent webpages who steal account details	an attacker masquerades a reputable entity	doubtful about using those features
PS-3	enterprise user	open mails in the cloud server	I detect malicious mails	they are not cryptographically signed	Emails are not verified and third party intrusions
PS-4	internet user	browse the internet	I identify a scam	an attacker mimics a authorized webpage	unsafe about my information that is shared over the network

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

The phishing attack is used to steal confidential information of a user. Fraud websites appear like genuine websites with the logo and graphics of genuine websites. This project aims to detect fraud or phishing website using machine learning techniques.

3.2 Ideation & Brainstorming



3.3 Proposed Solution

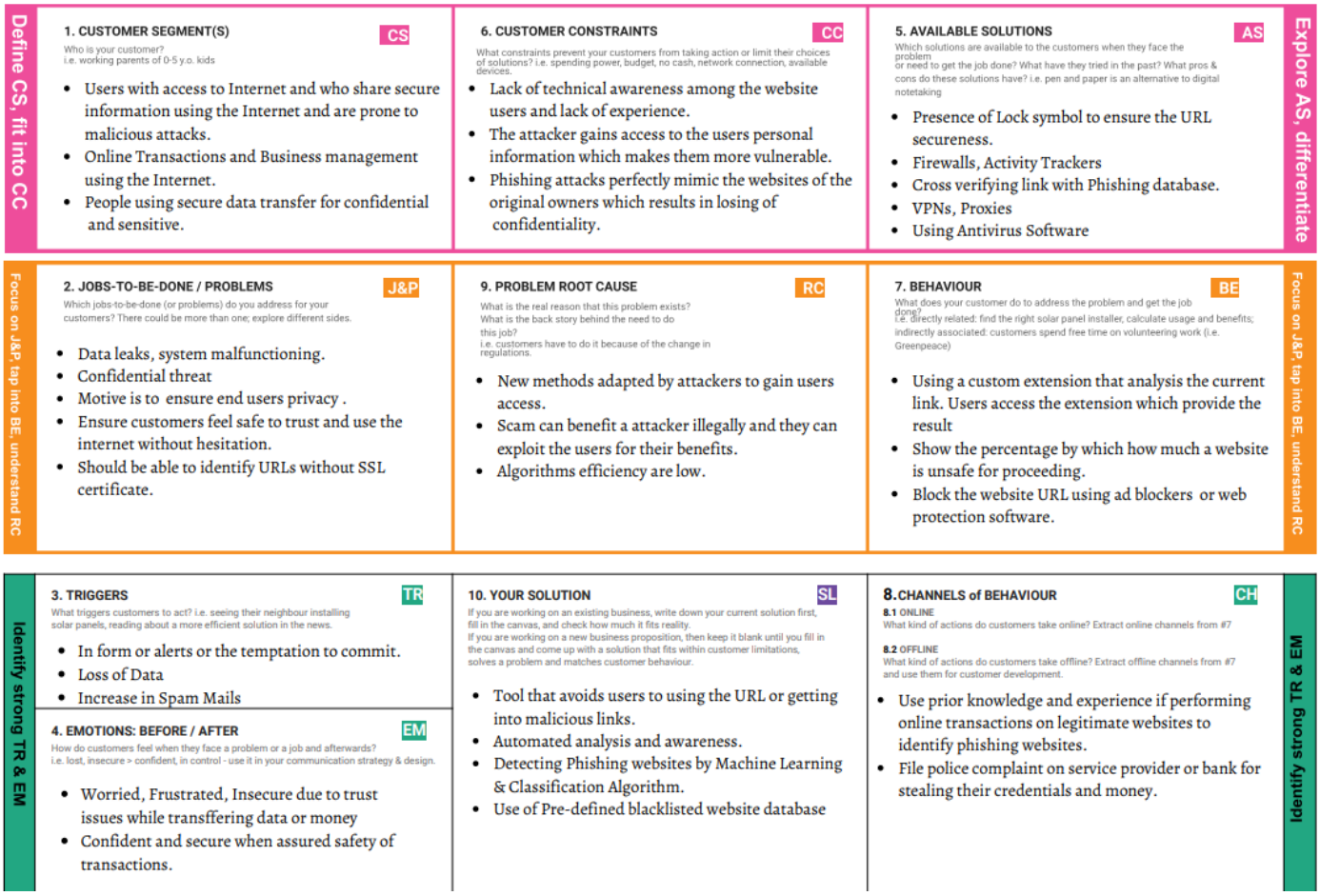
S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	This is an interactive and responsive website that will be used to detect whether a website is legitimate or phishing. Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack, as an example, passwords associate degree open-end credit unpretentious elements by presumptuous the highlights of a reliable individual or business in electronic correspondence. Phishing makes use of parody messages that square measure created to seem substantial and instructed to start out from true blue sources like money connected institutions, online business goals, etc, to draw in customers to go to phoney destinations through joins gave within the phishing websites.
2.	Idea / Solution description	Determine whether the provided URL is real or a phishing URL, and then output the answer with the proportion of risk factors.

3.	Novelty / Uniqueness	• Proposed web technology features improve phishing detection accuracy. • The usage of 10 machine learning algorithms produces the results with an accuracy of 96% approximately. • Simple, Easy-to-Understand UI. • A successful detection mechanism is developed by using an ideal dataset
4.	Social Impact / Customer Satisfaction	• It is based on URL feature extraction that helps in detecting phishing attacks that are relatively new and which is not possible for most of the other phishing detectors. • The system involves just ten algorithms that act as filters to determine the legitimacy of the URL • Users just need to provide the URL of the website whose legitimacy needs to be determined. Nothing else needs to be done by the user.
5.	Business Model (Revenue Model)	• B2C Model (end product sold to individuals such as children's gadgets and senior citizens at risk of assaults) and B2B Model (Machine Learning model/API can be sold to multiple enterprises for their employees) • The Application Programming Interface can be purchased in bulk by businesses at a subsidised rate (API) • Premium subscribers will get access to the URL's data and

		the justifications for a site's "unsafe" rating.
6.	Scalability of the Solution	• When there are more users and activity, the solution may require more hardware resources. • The API can make sure that several requests are processed in parallel at once.

3.4 Problem Solution fit

Project Title: Project Design Phase-I - **Solu on Fit Template** **Team ID:** PNT2022TMIDxxxxxx



4. REQUIREMENT ANALYSIS

4.1 Functional requirement

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Users must sign up using their email and Google accounts.
FR-2	User Confirmation	User has to Confirm their Email.
FR-3	User Input	The necessary field is filled up by the user with a suspect URL.
FR-4	URL Processing	The model will handle the new input by utilising the dataset and the proper machine learning methods.
FR-5	Classification	The URL will be marked as a legitimate or phishing URL.
FR-5	Result	The user will see the model's projected result, and if the URL turns out to be harmful, they will be warned about the website and given further instructions.

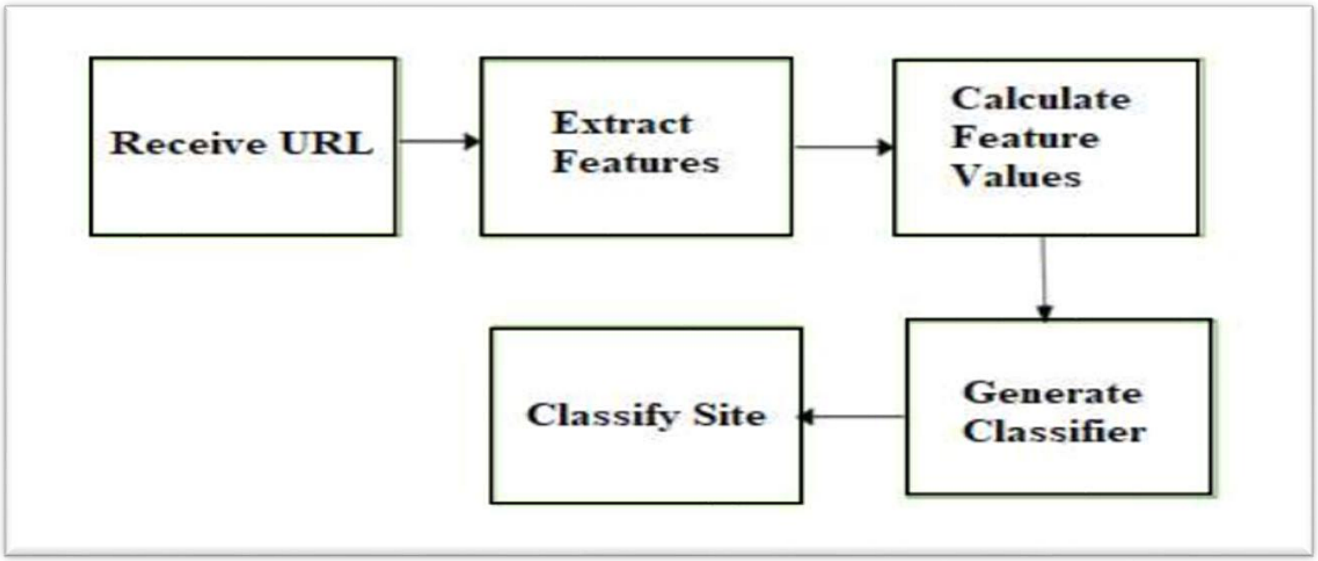
4.2 Non-Functional requirements

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	Users won't have any trouble using the solution and navigating the system with an effective, simple-to use UI
NFR-2	Security	Using Google authentication, which automatically offers email-based and multi-factor authentication.
NFR-3	Reliability	Probability of error-free operations in the designated usage environment.
NFR-4	Performance	For effectiveness and efficiency, the performance should be quicker and more user-friendly.
NFR-5	Availability	The model should always be usable, exportable to users, and executable on local computers.
NFR-6	Scalability	This might be turned into an API that other people could use and incorporate.

5. PROJECT DESIGN

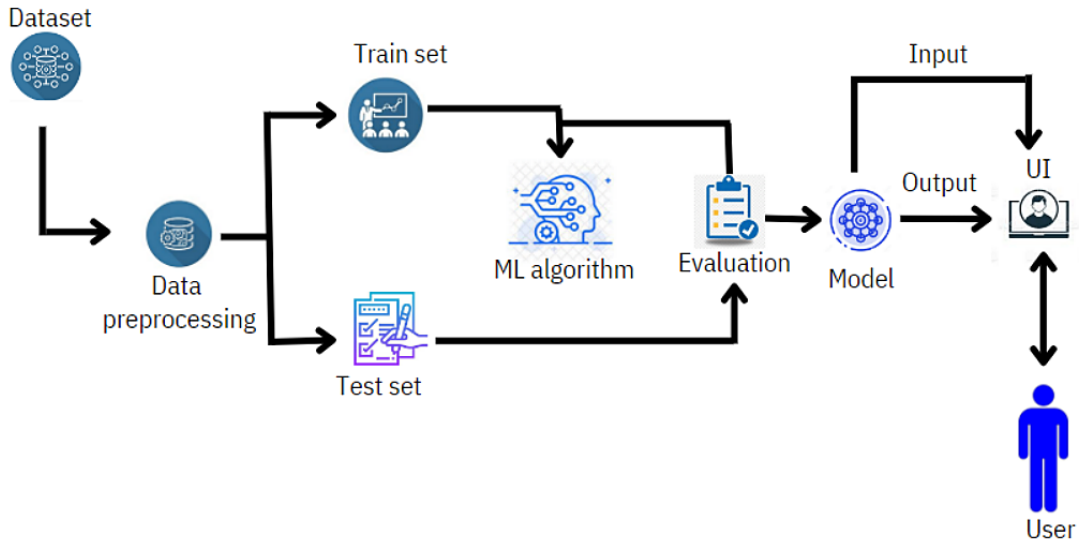
5.1 Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



5.2 Solution & Technical Architecture

Solution Architecture Diagram for Phishing Web Detection:



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Registration	USN-1	I can sign up for the application as a user by providing my email address, a password, and a password confirmation.	I can access my dashboard or account.	High	Sprint-1
		USN-2	When I register for the application as a user, I will get a confirmation email.	I can get a confirmation email and confirm it.	High	Sprint-1
		USN-3	I can sign up for the application as a user using Google authentication.	I may use Google Login to sign up and access the dashboard.	Medium	Sprint-2
	Login	USN-4	I can access the application as a user by providing my email		High	Sprint-1

			address and password.			
	Dashboard	USN-5				
	Website	USN-6	I can enter the suspect URL and view the forecast as a user.		High	Sprint-1
Customer Care Executive	Feature Extraction	USN-1	I am able to apply feature selection algorithms to extract URL-based features and Interaction Features (Indegree of URL, Outdegree of URL, etc.)	I can obtain a feature matrix for training that solely contains numerical data.	High	Sprint-1
Administrator	Machine Learning Prediction	USN-1	Use of 10 machine learning models to forecast the real-world data using the chosen feature matrix.	The classifier can provide me with an accurate ground truth class label.	High	Sprint-1

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	User input	USN 1	User inputs an URL in the required field to check its validation.	1	Medium	Hariharan P K
Sprint-1	Registration	USN 1	The user registers their details in the website	1	Medium	Gajalakshmi E
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach.	1	High	Kiruthika P
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extracts feature using heuristic and visual similarity.	2	High	Maghnaa S

Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	1	Medium	Hariharan P K
Sprint-3	Classifier	USN-5	Model sends all the output to the classifier and produces the final result.		Medium	Maghnaa S
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site.	1	High	Kiruthika P

6.2 Sprint Delivery Schedule

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 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 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

7. CODING & SOLUTIONING

7.1 Libraries to be installed

1. beautifulsoup4==4.9.3
2. Flask==2.0.2
3. googlesearch_python==1.0.1
4. numpy==1.21.4
5. pandas==1.3.4
6. python_dateutil==2.8.2
7. requests==2.25.1
8. scikit_learn==1.0.1
9. whois==0.9.13
10. gunicorn==20.1.0

8. TESTING

8.1 Test Cases

Testing is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is defect free. It involves the execution of a software component or system component to evaluate one or more properties of interest. Software testing also helps to identify errors, gaps, or missing requirements
in contrary to the actual requirement

1. **Unit Testing:**

When the testing happens for some individual group or some related units then that type of testing is called Unit Testing. It is often done by a programmer to test the part of the program he or she has implemented. Unit Testing is successful means all the modules have been successfully tested and it can proceed further.

2. **Functional Testing:**

This type of testing is tested to check the functional components, or the functionality required from the system is gained or not. It falls under the testing of Black Box testing of Software Engineering. This part includes the feeding of the inputs in the system or the project and checking if that system or the project is getting the same value or not as expected if not then calculate the error as wanted and check for more. Functional Testing of this project mainly involves below things. All of these are tested successfully, and errors are also calculated.

a. (i) Verifying the input image
   (ii) Verifying the workflow

b. Correct recognition and calculate the error

a. **Integration Testing:**

In a total project or system, many groups of components are getting added or summed up for the purpose of the project query. Integration testing is about checking the interaction between various modules of the project or the system. This module also includes the hardware and software requirements of the project. All the individual modules are integrated and tested together. All the best and extreme cases that the modules are interacting or not are successfully checked and passed, and errors are calculated for the deep learning platforms.

b. **System Testing:**

This type of testing is actually meant for the system or the project and also the platform and the integrated software and tools, technologies are also tested. The idea or purpose behind the system testing is to check all the requirements that will be provided by the system. This application of the project along with the tools and technologies has been tested in both windows and Linux. It passed successfully.

8.2 User Acceptance Testing

This is a type of system or software testing where a system has been tested for availability. The purpose of this test is to check the business requirements and assess whether it will be accepted for delivery.

9. RESULTS

9.1 Performance Metrics

The performance metrics of the classifiers are calculated and the result are as follows:

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	XGBoost Classifier	0.969	0.973	0.993	0.984
3	Multi-layer Perceptron	0.969	0.973	0.995	0.981
4	Random Forest	0.967	0.971	0.993	0.990
5	Support Vector Machine	0.964	0.968	0.980	0.965
6	Decision Tree	0.960	0.964	0.991	0.993
7	K-Nearest Neighbors	0.956	0.961	0.991	0.989
8	Logistic Regression	0.934	0.941	0.943	0.927
9	Naive Bayes Classifier	0.605	0.454	0.292	0.997

The percision of the model is **98.01**

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

1. The system not only produces a classification of the phishing URLs but also a rich description of the features contributing to the co-relation mapping
2. The method involves a relatively small number of parameters and hence training is relatively easy and fast.

The use of the Random Forest Model helps with identifying the necessary corelation between features that help classify the URL.

DISADVANTAGES:

1. It can only classify malicious links which are direct links to malicious websites
2. Cannot extract latent features from the URL that characterizes certain URLs as malicious
3. Cannot check for batch inputs

11. CONCLUSION

Web Phishing Detection using Deep Learning methods has been implemented. Random Forest Model have been trained and tested on the same data in order to acquire the comparison between the classifiers. Utilizing these deep learning techniques, a high amount of accuracy can be obtained.

12. FUTURE SCOPE

The proposed system cannot extract the latent features from certain malicious URLs hence classifying them as safe, future works can include identifying these latent features and extracting them for the model to learn to provide more accurate predictions.

The following are the features that can be added in our application:

- A communication app can be built with the same set of features. The user can choose the appropriate mode (speech to sign or sign to speech) and accordingly the real time detection would take place on both the end users' application.
- The accuracy of the model shall be increased.
- Customization of languages shall be added.
- Users shall be allowed to write notes while on call.
- Customization of signs can also be added as a feature.

13. APPENDIX

Python:

Python is an interpreted, high-level, general-purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant White space. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed, and garbage is collected. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

Keras:

Keras is a powerful and easy-to-use free open-source Python library for developing and evaluating deep learning models. It wraps the efficient numerical computation libraries Theano and TensorFlow and allows you to define and train neural network models in just a few lines of code. It uses libraries such as Python, C#, C++, or standalone machine learning toolkits. Theano and TensorFlow are very powerful libraries but difficult to understand for creating neural networks

Numpy:

NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. Numpy which stands for Numerical Python is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. It is an open-source project, and you can use it freely. NumPy stands for Numerical Python. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray

Machine Learning:

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

Deep Learning:

Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision-making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabelled. Also known as deep neural learning or deep neural network.

13.1 GitHub & Project Demo Link

GitHub : https://github.com/IBM-EPBL/IBM-Project-489-1658303709

Demo :