PROJECT REPORT

WEB PHISHING DETECTION

Submitted by-

ILAKIYA B(TEAM LEADER) - 412419205031
MOHITAVARSHINI A S - 412419205059
NITHILASRI T - 412419205062
YUVAHARINI A - 412419205103

TEAM ID: PNT2022TMID20856
BATCH ID: B10-4A6E

CHAPTER 1

INTRODUCTION

1.1 PROJECT OVERVIEW:

Nowadays 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. Main aim of the attacker is to steal banks account credentials. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol(IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear as original URL. To overcome the drawbacks of blacklist and many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero-hour phishing websites.

1.2 PURPOSE:

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

CHAPTER 2

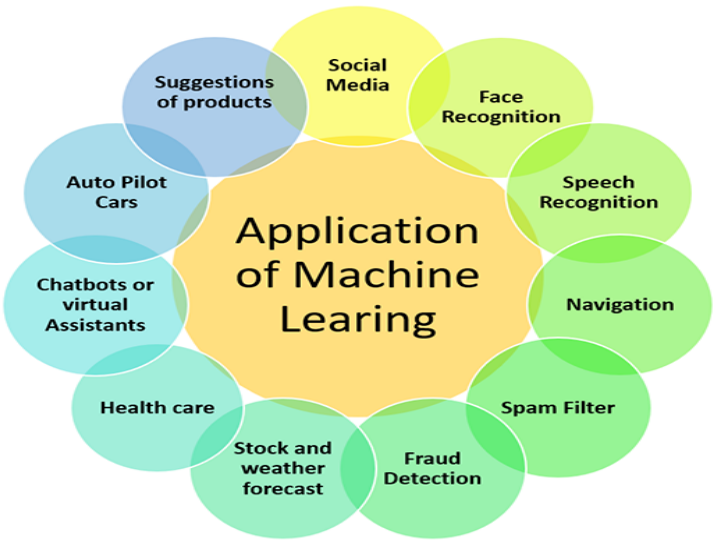
LITERATURE SURVEY

The purpose or goal behind phishing is data, money or personal information stealing through the fake website. The best strategy for avoiding the contact with the phishing web site is to detect real time malicious URL. Phishing websites can be determined on the basis of their domains. Also this model is depends on the quality and quantity of the training set and Broken links.

2.1 MACHINE LEARNING:

AI (ML) is a class of calculation that enables programming applications to turn out to be progressively precise in anticipating results without being expressly customized. The fundamental reason of AI is to assemble calculations that can get input information and utilize factual examination to foresee a yield while refreshing yields as new information winds up accessible.

Simplifies Product Marketing and Assists in Accurate Sales Forecasts.
Utilization and efficiency improvement.
Very high Scalability
High Computing power



2.2 2.2

TECHNOLOGIES USED:

- NUMPY
- PANDAS
- SCIKIT
- FLASK
- MATPLOTLIB
- ANACONDA

2.3 PROBLEM STATEMENT:

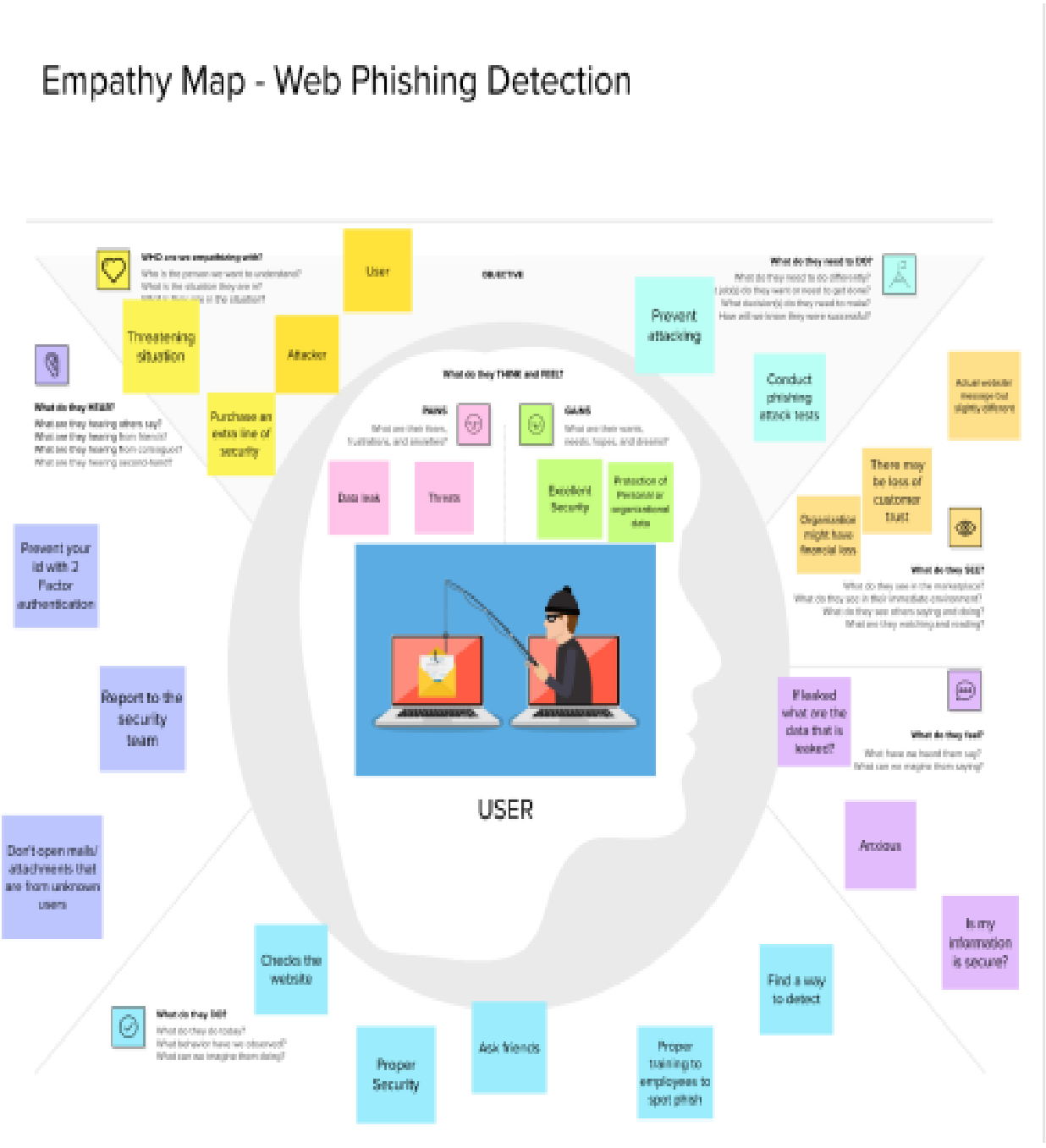
The interaction between the user/customer and the phish attack is described in the problem statement.

I AM (CUSTOMER) - user
I AM TRYING - to protect my information/details.
BUT - I think my information is leaked.
BECAUSE - I accidentally gave my details to wrong webstie.
WHICH MAKES ME FEEL - frightened/confused.

CHAPTER 3

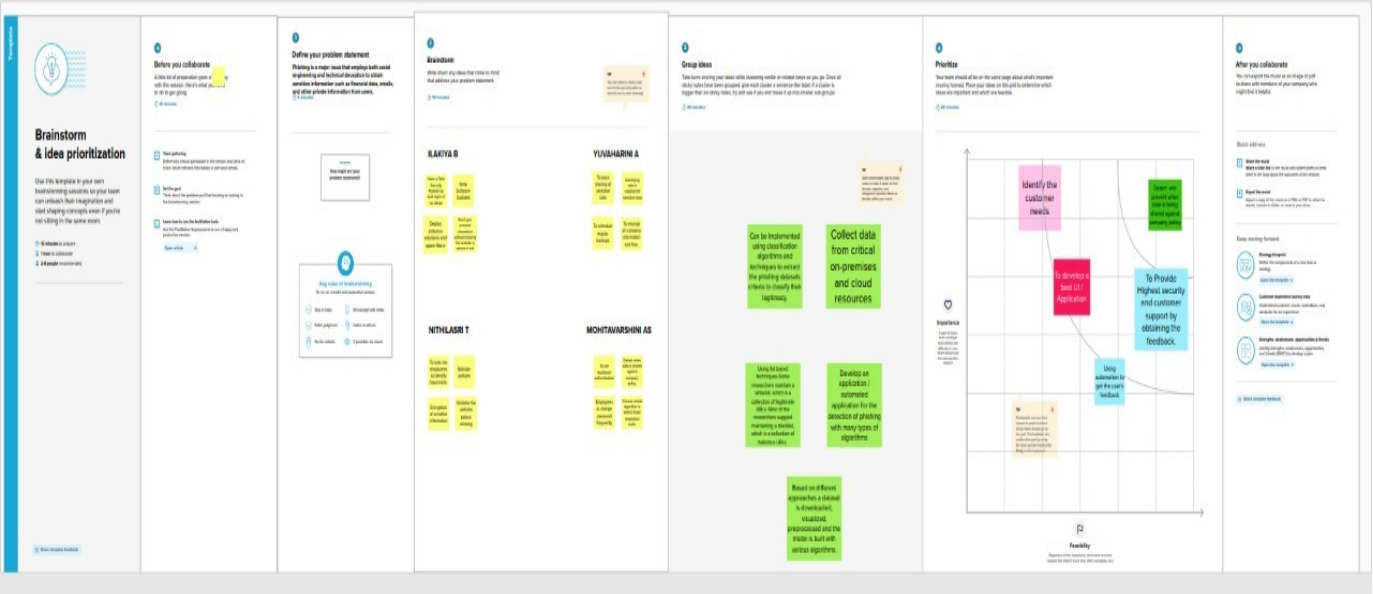
IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP:

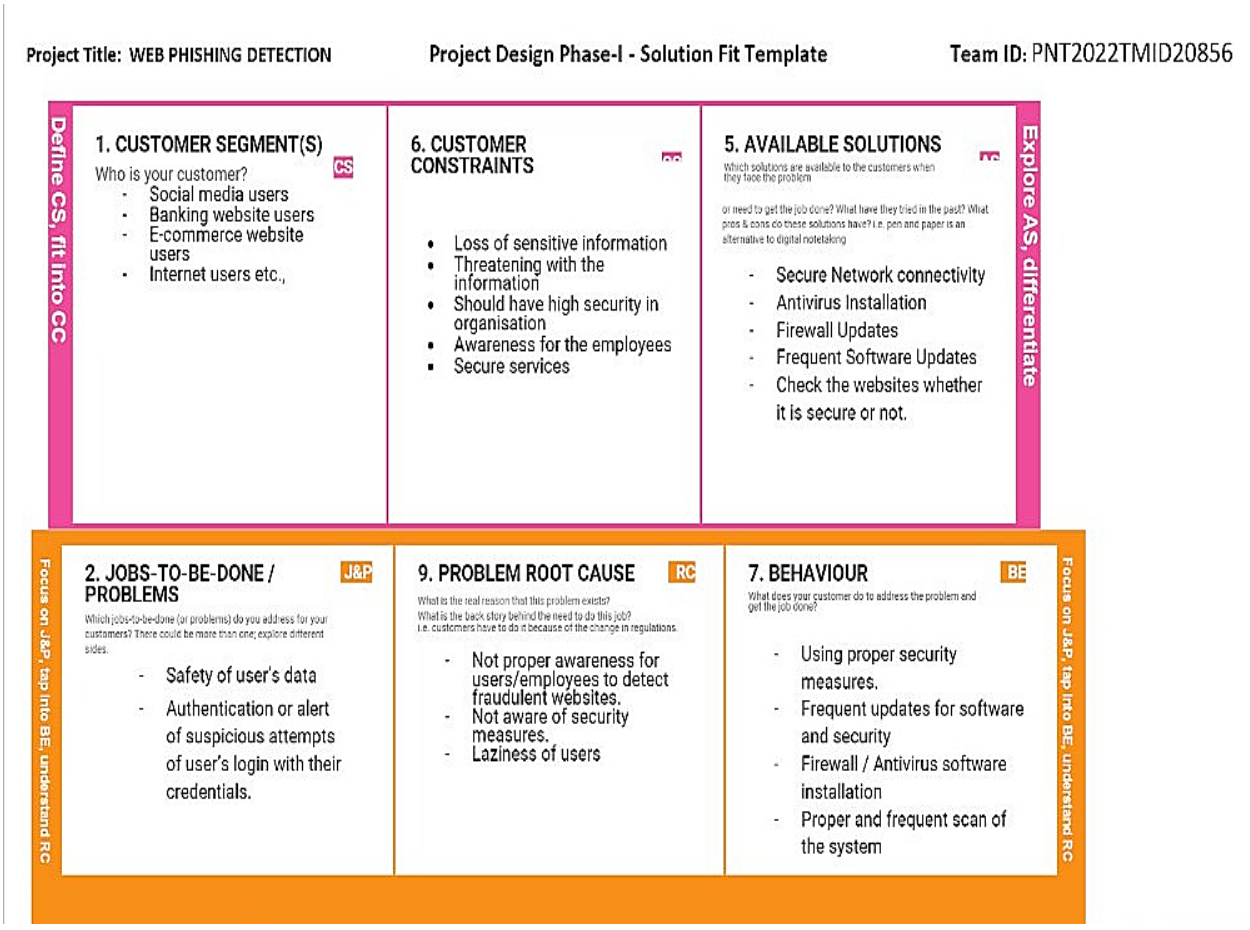


3.2 BRAINSTORMING IDEAS:

Brainstorming is a method design teams use to generate ideas to solve clearly defined design problems. In controlled conditions and a free-thinking environment, teams approach a problem by such means as “How Might We” questions. Brainstorming is part of design thinking. It is generally used in Ideation phase of the project. Everyone in a design team should have a *clear* definition of the target problem.



3.3 PROBLEM SOLUTION FIT:



CHAPTER 4

REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS:

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration/log in	Check validation of the URL
FR-2	Validating and Comparison of websites	With the help of list-based approaches.
FR-3	The built model will able to detect with the help of machine learning algorithms	Algorithms like K-Nearest Neighbour (KNN), Random Forest, Support Vector Machine etc.,
FR-4	Generates output	Based on the above it will classify and produce the output.
FR-5	Result	If the website is illegal it will classify as phishing website and sends to blacklist.
FR-6	For Email phishing detection/ If the user clicks the link that as been sent to his/her mail.	Validate the URL
FR-7	Compare, Predict and generates output	- If it is an illegal website/link then it will generate the output.

4.1 NON-FUNCTIONAL REQUIREMENTS:

Non-functional Requirements:

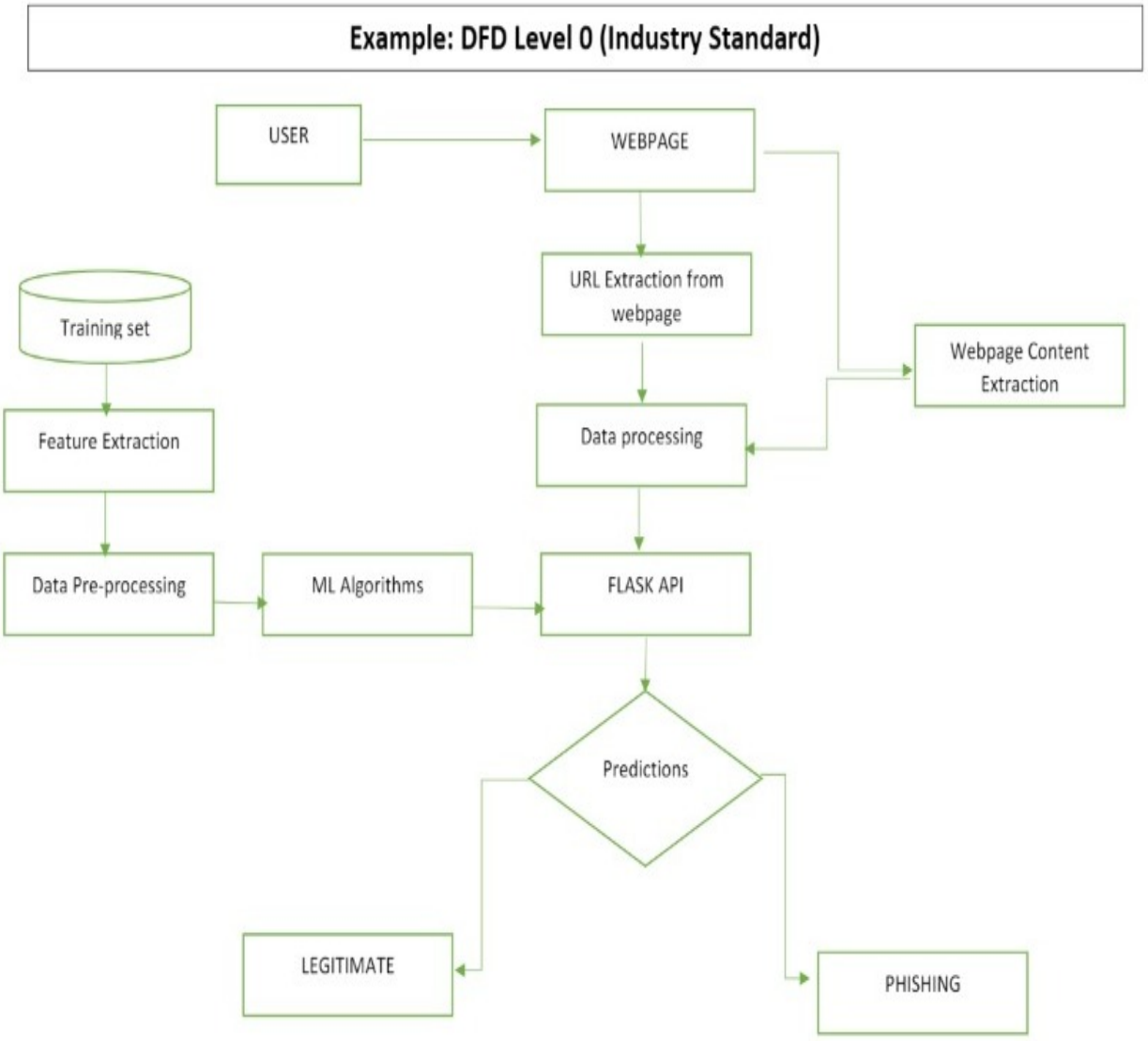
Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	It should satisfy the users requirements.
NFR-2	**Security**	The system should have strengthened security with up-to-the-minute insights
NFR-4	**Performance**	It should perform based on requirements made by the user.
NFR-5	**Availability**	It should be available all time (like 24/7)

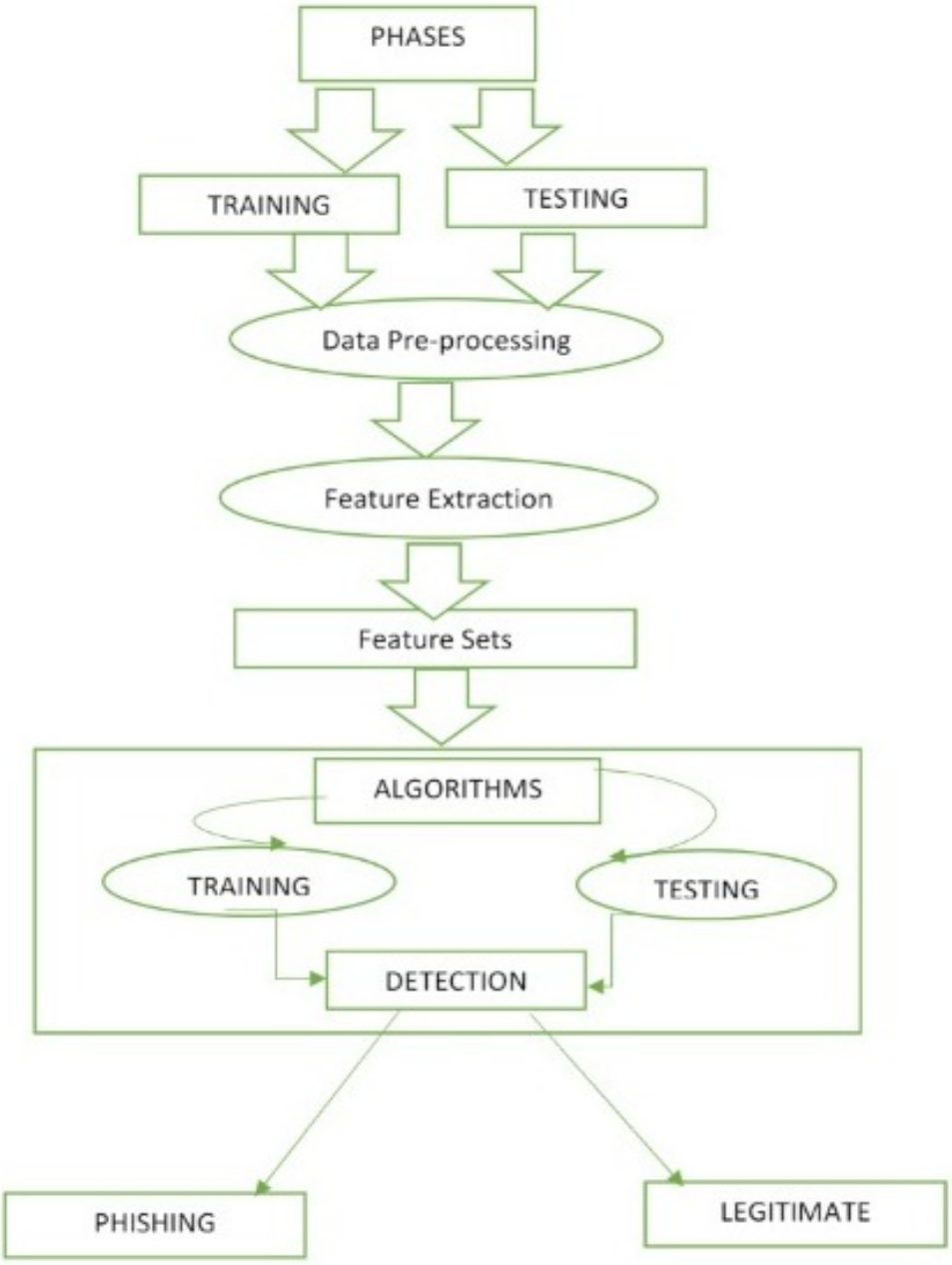
CHAPTER 5

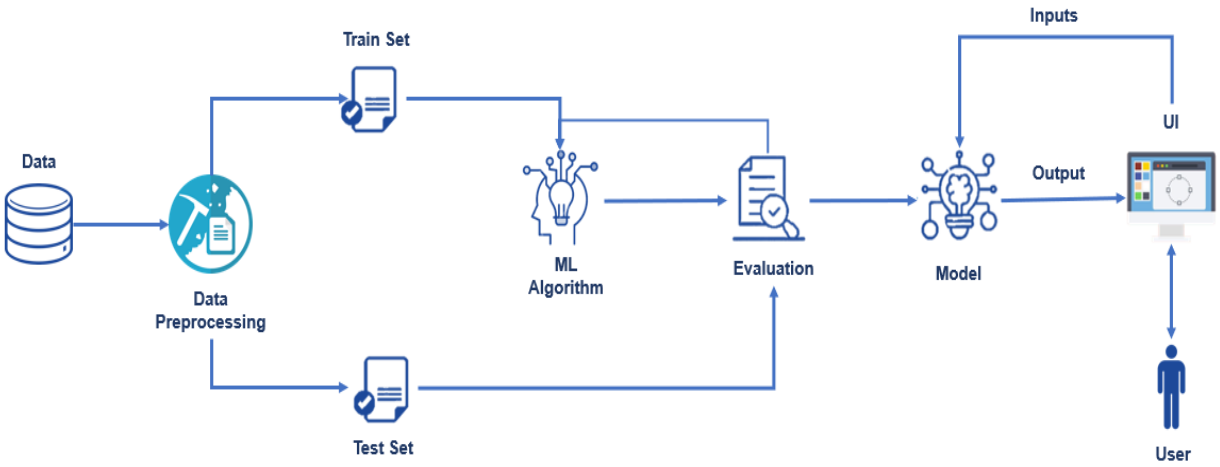
PROJECT DESIGN

5.1 DATA FLOW DIAGRAM:



5.2 SOLUTION AND TECHNICAL ARCHITECTURE:





5.3 USER STORIES:

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, for validating the URL or link I will enter in the validation text box in the application.	I can now easily access the website since there is no problem in checking	High	Sprint-1
Customer Care Executive	Feature	USN-1	I can extract feature if no problem was found while comparing.	I can compare with the different websites	High	Sprint-1
Administrator	Prediction	USN-1	Machine algorithms are used to build the model like Logistic Regression etc.,	I can see the algorithm gives the correct prediction	High	Sprint-1
	Classifier	USN-2	Final end result is produced by the classifier.	To predict the correct output, I will use the fitting classifier	Medium	Sprint-2

CHAPTER 6

PROJECT PLANNING & SCHEDULING

6.1 SPRINT PLANNING & ESTIMATION:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task - As a User,	Story Points	Priority	Team Members
Sprint-1	Registration/ User Input	USN-1	Register for the application or login using the credentials	2	High	Nithilasri T
Sprint-1	Website Comparison	USN-2	If new user confirmation mail is sent	1	High	Mohitavarshini A S
Sprint-1	Storage	USN-3	Inputs the URL in the box where it is checked	2	Low	Yuvaharini A
Sprint-2	Feature Extraction	USN-4	Extraction process-Checking whether URL is suspicious or not	2	Medium	Ilakiya B
Sprint-2	Prediction	USN-5	Model predicts the URL using Machine Learning algorithm	1	High	Nithilasri T
Sprint-3	Classifier	USN-6	Classification models	1	Medium	Mohitavarshini AS
Sprint-4	Announcement	USN-7	Whether the website is phishing website or legitimate website.	1	High	Yuvaharini A
Sprint-4	Events	USN-7	Alert is given if it is a phishing website	1	High	Ilakiya B

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	9 November
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	10 November
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 November
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	16 November

CHAPTER 7

CODING AND SOLUTIONING

7.1 Feature 1(Explain the features added in the project along with code)
7.2 Feature 2
7.3 Database Schema

CHAPTER 8

TESTING

8.1 Test Cases
8.2 User Acceptance Testing

UAT REPORT:

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	11	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	3	4	21	39
Not Reproduced	0	0	2	0	2
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	25	15	14	27	81

CHAPTER 9

RESULT

Accuracy of various model used for URL detection:

ML Model	Accuracy	F1_SCORE	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Cat Boost Classifier	0.972	0.975	0.994	0.989
XG Boost Classifier	0.969	0.973	0.993	0.984
Random Forest	0.967	0.971	0.993	0.990
Multi - layer Perceptron	0.969	0.973	0.995	0.981
Support Vector Machine	0.964	0.968	0.980	0.965
Logistic Regression	0.934	0.941	0.943	0.927
K-Nearest Neighbours	0.956	0.961	0.991	0.989
Naive Bayes Classifier	0.605	0.454	0.292	0.997
Decision Tree	0.960	0.964	0.991	0.993

CHAPTER 10

ADVANTAGES AND DISADVANTAGES

ADVANTAGES:

- Measure the degrees of corporate and employee vulnerability.
- Eliminate the cyber threat risk level.
- Increase user alertness to phishing risks.
- Instill a cyber security culture and create cyber security heroes.
- Fast in classification process.
- Higher level of accuracy.

DISADVANATAGES:

- Time consuming.
- Huge number of features.
- Consuming features.
- Time consuimg because this feature has many layer to make a final result.
- Higher in cost.

CHAPTER 11

CONCLUSION

The importance to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker's tool, phishing detection tools play a vital role in ensuring a secure online experience for users.

These critical issues have drawn many researchers to work on various approaches to improve detection accuracy of phishing attacks and to minimize false alarm rate. The inconsistent nature of attacks behaviors and continuously changing URL phish patterns require timely updating of the reference model.

Therefore, it requires an effective technique to regulate retraining as to enable machine learning algorithm to actively adapt to the changes in phish patterns.

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 and complexity of phishing attacks and tactics, ML anti-phishing techniques are essential.

CHAPTER 12

FUTURE SCOPE

The future direction of this study is to develop an unsupervised deep learning method to generate insight from a URL.

In addition, the study can be extended in order to generate an outcome for a larger network and protect the privacy of an individual.

The means by which hackers access user information have quickly evolved beyond traditional phishing emails.

Phishing has always had the aim of baiting users to take an action or share a piece of sensitive information by appearing as a non-threat – but awareness has since grown.

In future if we get structured dataset of phishing we can perform phishing detection much more faster than any other technique.

In future we can use a combination of any other two or more classifier to get maximum accuracy.