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. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual world impact of phishing could be as high as $5 billion Phishing attacks are becoming successful because lack of user awareness.

Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

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 legitimate via obfuscation and many other simple techniques including:fast-flux, in which proxies are automatically generated to host the webpage;algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high . To overcome the drawbacks of blacklist and heuristics based method, 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

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are

gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.The purpose of a phishing attack is to steal data, money or both.

- **Data**–The type of data that cybercriminals are most often interested in are usernames and passwords, identity information (e.g., social security numbers), and financial data (e.g., credit card numbers or bank account information). Login credentials can be used to breach the victim's systems to steal intellectual property or inject malware for other malicious purposes. Data of any kind can also be monetized by selling it on the dark web to other criminals.
- **Money**—If the intent of the phishing attack is to steal money, the cybercriminals may send a fake invoice, try to convince the victim to wire money, or ask the victim to input financial account information into a fake website.

2.LITERATURE SURVEY

Author 1:

Andy Jones Abstract—

This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyberattacks are spread via mechanisms that exploit weaknesses found in end users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

AUTHOR 2:

Mahmoud Khonji: This article surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyberattacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently proposed phishing mitigation techniques. A high-

level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

2.1 EXISTING PROBLEM:

A poorly structured NN model may cause the model to under fit the training dataset. On the other hand, exaggeration in restructuring the system to suit every single item in the training dataset may cause the system to be over fitted. One possible solution to avoid the Over fitting problem is by restructuring the NN model in terms of tuning some parameters, adding new neurons to the hidden layer or sometimes adding a new layer to the network.ANN with a small number of hidden neurons may not have a satisfactory representational power to model the complexity and diversity inherent in the data. On the other hand, networks with too many hidden neurons could over fit the data. However, a certain stage the model can no longer be improved, therefore, the structuring process should be terminated.

- data loss
- data manipulation
- security issues
- authentication problems
- privacy problems
- data abduction

2.2 REFERENCES

[1] Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.

[2] https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attack-statistics/#gref

[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[4] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

[5] http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[6] http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

[7] https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

[8] www.alexa.com

[9] www.phishtank.com

2.3 Problem statement definition

There are e-banking websites that requests the users to provide more sensitive information such as credit card details, password etc., for malicious reasons. These websites that mimics trustful URLs and webpages are known as phishing websites. Common causes for web phishing attacks involve: • Users lack of security awareness • Not performing sufficient due diligence • Low-cost phishing and ransomware tools are easy to get hold of • Malware is becoming more sophisticated and so on Web phishing is considered to be a threat in various aspects of security on the internet, which might involve scams and private information disclosure. Some of the common threats of web phishing are: • Attempt to fraudulently solicit personal information from an individual or organization. • Attempt to deliver malicious software by posing as a trustworthy organization or entity. • Installing those malwares infects the data that cause a data breach or even nature's forces that takes down your company's data headquarters, disrupting access. For this purpose, the objective of our project involves building an efficient and intelligent system to detect such websites by applying a machine-learning algorithm which implements classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy and as a result of which whenever a user makes a transaction

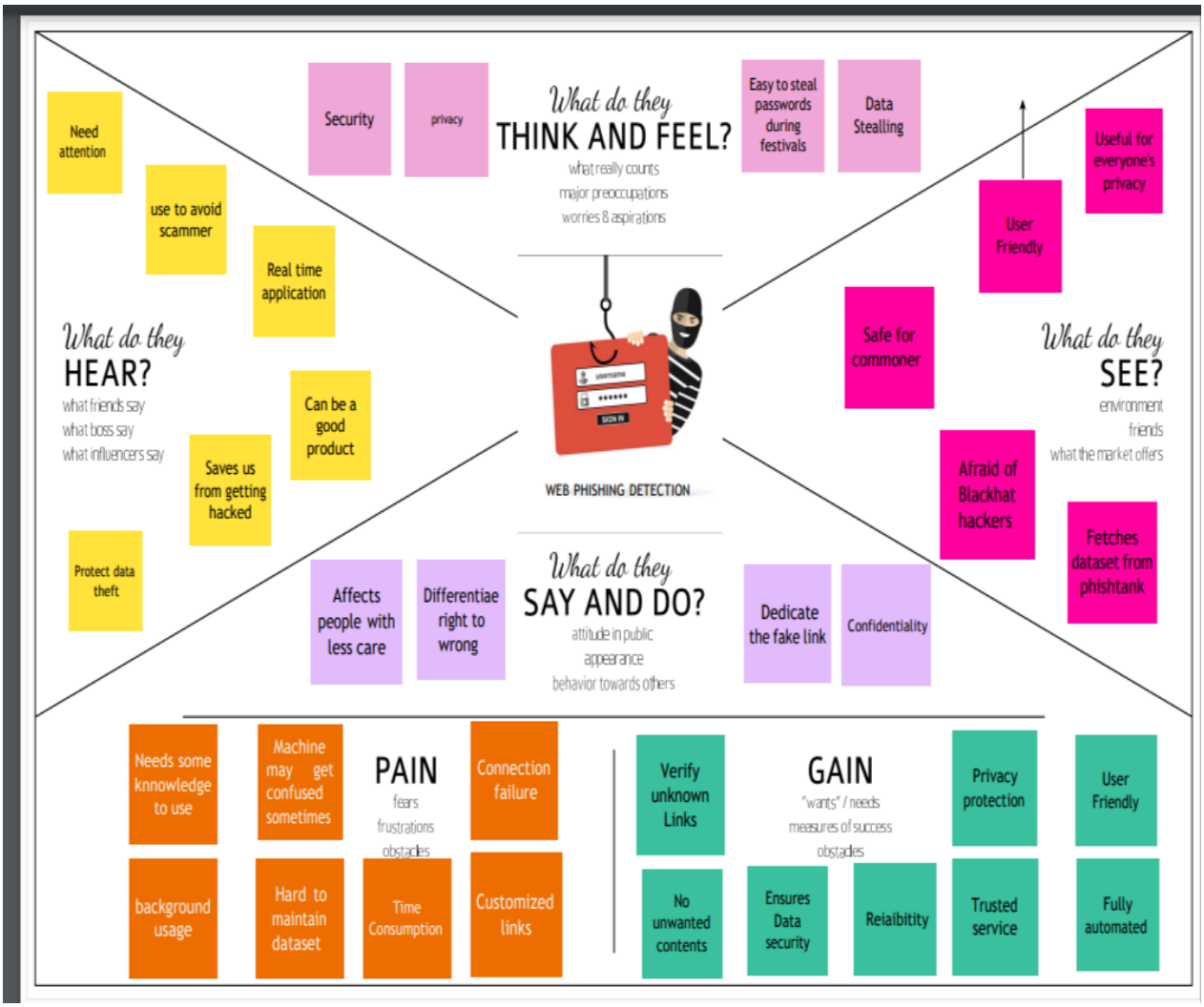
online and makes payment through an e banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not. This project can be further extended by creating a browser extension or develop a GUI which takes the URL and analyze its nature to determine if it is a legitimate or a phishing website.

3.IDEATION AND PROPOSED SOLUTION

S.NO.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Novel phishing approaches suffer low detection accuracy. The most common technique used is the blacklist-based method. It has become inefficient since registering a new domain has become easier. No comprehensive blacklist can ensure a perfect up-to-date database.
2.	Idea / Solution description	Our solution is to build an efficient and intelligent system to detect phishing sites by applying a machine learning algorithm which implements classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy.

3.	Novelty / Uniqueness	We have carefully analyzed and identified various factors that could be used to detect a phishing site. These factors fall under the categories of address bar based features, domain based features, HTML & Javascript based features. Using these features we can identify a phishing site with high accuracy.
4.	Social Impact / Customer Satisfaction	By using this application the customer has the sense of safety whenever he attempts to provide sensitive information to a site.
5.	Business Model (Revenue Model)	By generating leads we can improve our business model. By detecting the phishing sites, people won't access them which will reduce the revenue of malicious site owners.
6.	Scalability of the Solution	This application can be accessed online without paying. It can be accessed via any browser of your choice. It can detect any site with high accuracy.

3.1 Empathy map canvas



3.2 IDEATION AND BAINSTORMING

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD) and “hud” is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as “Suspicious” since it has one sub domain. However, if the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.

Malicious Web sites largely promote the growth of Internet criminal activities and constrain the development of Web services. As a result, there has been strong motivation to develop systemic solution to stopping the user from visiting such Web sites. We propose a learning based approach to classifying Web sites into 3 classes: Benign, Spam and Malicious. Our mechanism only analyzes the Uniform J

Benign: Safe websites with normal services Spam: Website performs the

act of attempting to flood the user with advertising or sites such as fake

surveys and online dating etc. Malware: Website created by attackers to disrupt computer operation, gather sensitive information, or gain access to private computer systems.Phishing is one of the most common and most dangerous attacks among cybercrimes. The aim of these attacks is

to steal the information used by individuals and organizations to conduct transactions. Phishing websites are fake websites that contain various hints among their contents and web browser-based information. When a user opens a fake webpage and enters the username and protected password, the credentials of the user are acquired by the attacker which can be used for malicious purposes. Phishing websites look very similar in appearance to their corresponding legitimate websites to attract large number of Internet users.The study uses a dataset which contains approximately 11,000 data containing the 30 features extracted based on the features of websites in UC Irvine Machine Learning Repository database. For classification, a neural network named Extreme Learning Machine (ELM) will be used. Extreme Learning Machine (ELM) is a feed-forward artificial neural network (ANN) model with a single hidden layer. In ELM Learning Processes, differently from ANN that renews its parameters as gradient-based, input weights are randomly selected while output weights are analytically calculated. The given data set will be divided into three parts as training, validation and test data by three-phase division in K-Fold method, and model selection and performance status will be simultaneously performed. This way the performance of the model will be measured in a reliable manner.

The procedures engaged with AI are like that of information mining and prescient displaying. Both require scanning through information to search for examples and modifying program activities as needs be. Numerous individuals know about AI from shopping on the web and being served advertisements identified with their buy. This happens on the grounds that suggestion motors use AI to customize online promotion conveyance in practically continuous. Past customized

advertising, other regular AI use cases incorporate misrepresentation location, spam separating, arrange security risk identification, prescient support and building news sources.

3.3 proposed solution

S.NO.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Novel phishing approaches suffer low detection accuracy. The most common technique used is the blacklist-based method. It has become inefficient since registering a new domain has become easier. No comprehensive blacklist can ensure a perfect up-to-date database.
2.	Idea / Solution description	Our solution is to build an efficient and intelligent system to detect phishing sites by applying a machine learning algorithm which implements classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy.
3.	Novelty / Uniqueness	We have carefully analyzed and identified various factors that could be used to detect a phishing site. These factors fall under the categories of address

		bar based features, domain based features, HTML & Javascript based features. Using these features we can identify a phishing site with high accuracy.
4.	Social Impact / Customer Satisfaction	By using this application the customer has the sense of safety whenever he attempts to provide sensitive information to a site.
5.	Business Model (Revenue Model)	By generating leads we can improve our business model. By detecting the phishing sites, people won't access them which will reduce the revenue of malicious site owners.
6	Scalability of the Solution	This application can be accessed online without paying. It can be accessed via any browser of your choice. It can detect any site with high accuracy

3.4 problem solution fit

Problem – Solution Fit Template: The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you 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 Purpose: Solve complex problems in a way that fits the state of your customers. Succeed faster and increase your solution adoption by tapping into existing mediums and channels of behavior. Sharpen your communication and marketing strategy with the right triggers and messaging. Increase touch-points with your company by finding the right problem-behavior fit and building trust by solving frequent annoyances, or urgent or costly problems. Understand the existing situation in order to improve it for

your target group. Template:

4.Requirement Analysis

4.1 Functional Requirement

S.NO.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
1.	User Input	User inputs an URL in required field to check its validation
2.	Website Comparison	Model compares the websites using Blacklist and White list approach.
3.	Feature extraction	After comparing, if none found on comparison then it extracts feature using heuristic and visual similarity approach.
4.	Prediction	Model predicts the URL using Machine Learning algorithms such as Logistic Regression, KNN
5.	Classifier	Model sends all output to classifier and produces final result.
6.	Announcement	Model then displays whether website is a legal site or a phishing site.
7.	Events	This model needs the capability of retrieving and displaying accurate result for a website.

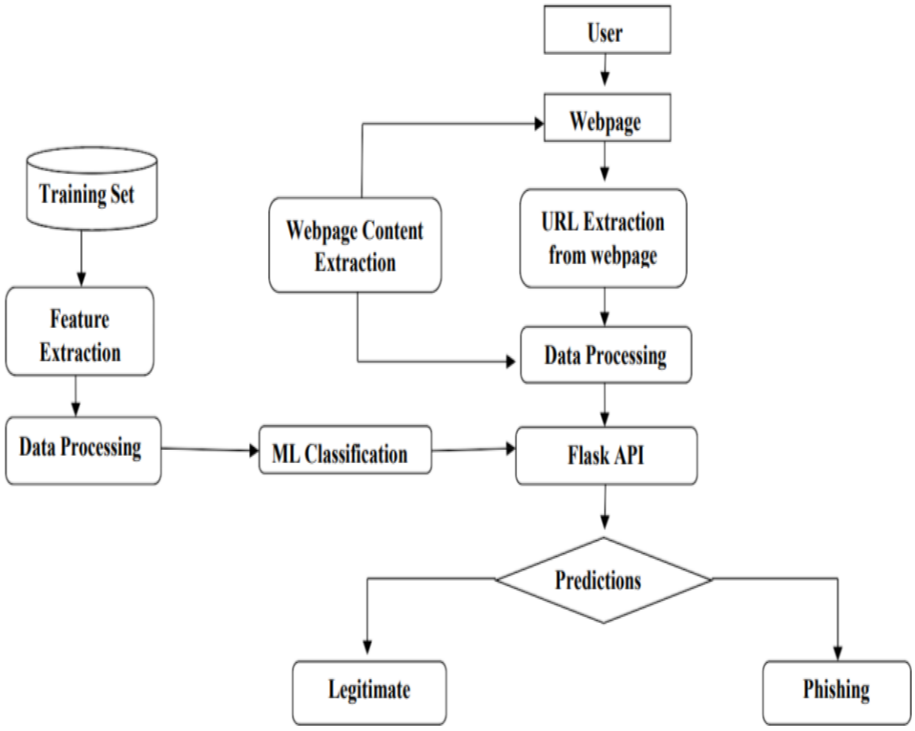
4.2 Non Functional-Requirements

- Usability
- Integrity
- Efficiency
- Test ability
- Reusability
- Portability

5.Project Design

Data Flow Diagrams:

Predictions 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



User Stories

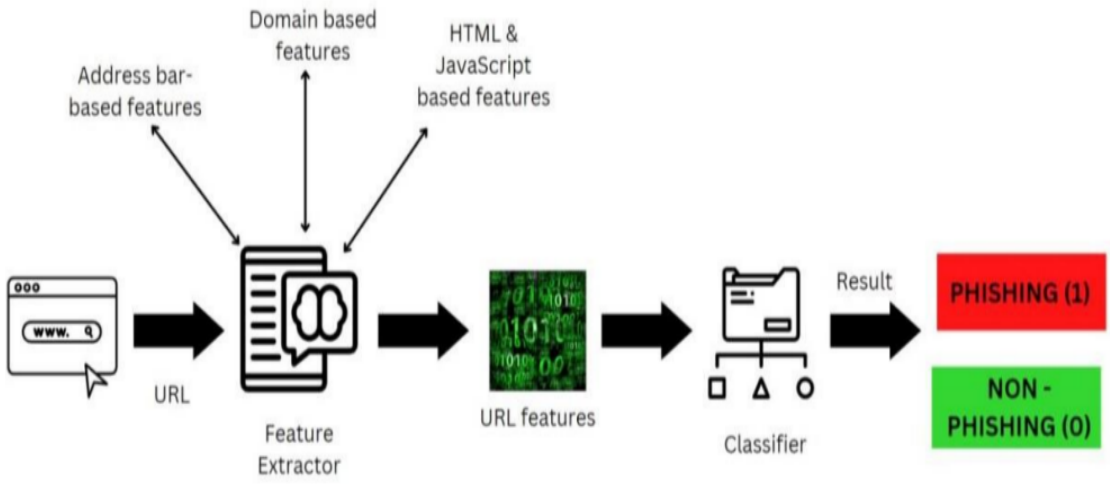
Use the below template to list all the user stories for the product.

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 input the particular URL in the required field and waiting for validation.	I can go access the website without any problem	High	Sprint-1
Customer Care Executive	Feature extraction	USN-1	After i compare in case if none found on comparison then we can extract feature using heuristic and visual similarity approach.	As a User i can have comparison between websites for security.	High	Sprint-1
Administrator	Prediction	USN-1	Here the Model will predict the URL websites using Machine Learning algorithms such as Logistic Regression, KNN	In this i can have correct prediction on the particular algorithms	High	Sprint-1
	Classifier	USN-2	Here i will send all the model output to classifier in order to produce final result.	I this i will find the correct classifier for producing the result	Medium	Sprint-2

5.2 Solution and Technical Architecture

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to: • Find the best tech solution to solve existing business problems. • Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders. • Define features, development phases, and solution requirements. • Provide specifications according to which the solution is defined, managed, and delivered.

Solution Architecture Diagram:



5.3 User Stories

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	Vaka Vasantha
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach.	1	High	Vaka Reshma
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extract feature using heuristic and visual similarity.	2	High	Vaka Vasantha
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	1	Medium	Narala Praveena
Sprint-3	Classifier	USN-5	Model sends all the output to the classifier and produces the final result.	1	Medium	Vaka Reshma
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site.	1	High	Naralla praveena
Sprint-4	Events	USN-7	This model needs the capability of retrieving and displaying accurate result for a website.	1	High	R. Lakshmi susmitha

6 .Project Planning and Scheduling

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. They usually are related to URL which needs to be

registered (low-level domain and upper-level domain, path, query). Recently acquired status of intra-URL relationship is used to evaluate it using distinctive properties extracted from words that compose a URL based on query data from various search engines such as Google and Yahoo. These properties are further led to the machine-learning based classification for the identification of phishing URLs from a real dataset. This paper focus on real time URL phishing against phishing content by using phish-STORM.

6.1 SPRINT PLANNING AND ESTIMATION

The Product Owner proposes how the product could increase its value and utility in the current Sprint. The whole Scrum Team then collaborates to define a Sprint Goal that communicates why the Sprint is valuable to stakeholders. The Sprint Goal must be finalized prior to the end of Sprint Planning.

Topic Two: What can be Done this Sprint?

Through discussion with the Product Owner, the Developers select items from the Product Backlog to include in the current Sprint. The Scrum Team may refine these items during this process, which increases understanding and confidence.

Selecting how much can be completed within a Sprint may be challenging. However, the more the Developers know about their past performance, their upcoming capacity, and their Definition of Done, the more confident they will be in their Sprint forecasts.

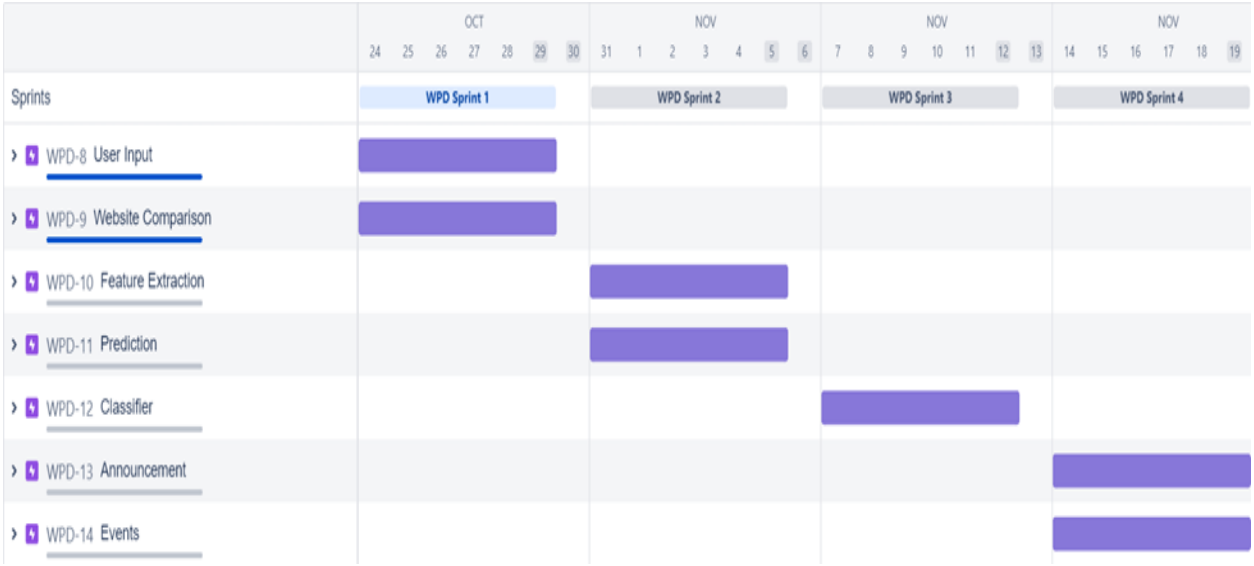
Topic Three: How will the chosen work get done?

For each selected Product Backlog item, the Developers plan the work necessary to create an Increment that meets the Definition of Done. This is often done by decomposing Product Backlog items into smaller work items of one day or less. How this is done is at the sole discretion of the Developers. No one else tells them how to turn Product Backlog items into Increments of value.The Sprint Goal, the Product Backlog items selected for the Sprint, plus the plan for delivering them are together referred to as the Sprint Backlog.Sprint Planning is time boxed to a maximum of eight hours for a one-month Sprint. For shorter Sprints, the event is usually shorter.

6.2 Sprint Delivery Schedule

sprint	Duration	Tell story points	Sprint start date	Sprint End date	Story points Completed date	Sprint released date
1.	20	24October 2022	24 October 2022	29 October 2022	20	29 October 2022
2.	20	31October 2022	31 October 2022	5 October 2022	20	15 Nov 2022
3.	20	07 Nov 2022	07 Nov 2022	12 Nov 2022	20	9 Nov 2022
4.	20	14 Nov 2022	14 Nov 2022	19 Nov 2022	20	12 Nov 2022

6.3 Reports from jira



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	Vaka Vasantha
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach.	1	High	Vaka Reshma
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extract feature using heuristic and visual similarity.	2	High	Vaka Vasantha
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	1	Medium	Narala Praveena
Sprint-3	Classifier	USN-5	Model sends all the output to the classifier and produces the final result.	1	Medium	Vaka Reshma
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site.	1	High	Naralla praveena
Sprint-4	Events	USN-7	This model needs the capability of retrieving and displaying accurate result for a website.	1	High	R. Lakshmi susmitha

7. CODING AND SOLUTIONING

7.1 Feature 1

URL-Based Features

URL is the first thing to analyze a website to decide whether it is a phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

- Digit count in the URL

- Total length of URL
- Checking whether the URL is Typosquatted or not. (google.com → goggle.com)
- Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)
- Number of subdomains in URL
- Is Top Level Domain (TLD) one of the commonly used one?

Domain-Based Features

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

- Its domain name or its IP address in blacklists of well-known reputation services?
- How many days passed since the domain was registered?

- Is the registrant name hidden?

7.2 Feature 2

Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how much reliable a web site is. Some of Page-Based Features are given below.

- Global Pagerank
- Country Pagerank
- Position at the Alex Top 1 Million Site

Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

- Estimated Number of Visits for the domain on a daily, weekly, or monthly basis

- Average Page views per visit
- Average Visit Duration
- Web traffic share per country
- Count of reference from Social Networks to the given domain
- Category of the domain
- Similar websites etc.

Content-Based Features

Obtaining these types of features requires active scan to target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

- Page Titles
- Meta Tags
- Hidden Text
- Text in the Body

- Images etc.

By analyzing these information, we can gather information such as;

- Is it required to login to website
- Website category
- Information about audience profile etc.

All of features explained above are useful for phishing domain detection. In some cases, it may not be useful to use some of these, so there are some limitations for using these features. For example, it may not be logical to use some of the features such as Content-Based Features for the developing fast detection mechanism which is able to analyze the number of domains between 100.000 and 200.000. Another example would be, if we want to analyze new registered domains Page-Based Features is not very useful. Therefore, the features that will be used by the detection mechanism depends on the purpose of the detection mechanism. Which features to use in the detection mechanism should be selected carefully.

8.Testing

Model Performance Testing:

Project team shall fill the following information in model performance testing template.

S.No.	Parameter	Values
1.	Metrics	**Random Forest Classifier Accuracy score-96.653**
2.	Tune the Model	**Hyperparameter Tuning - Validation Method -**

1. METRICS

Classification Report:

```
[ ]  #classification report of Randomm Forest model
     print(metrics.classification_report(y_test,y_test_rf))
```

```
                  precision    recall  f1-score   support

              -1       0.98      0.95      0.96      1014
               1       0.96      0.98      0.97      1197

        accuracy                           0.97      2211
       macro avg       0.97      0.97      0.97      2211
    weighted avg       0.97      0.97      0.97      2211
```

TESTCASES REPORT

Date	17-Nov-22
Team ID	PNT2022TMID23787
Project Name	Project - Web Phishing Detection
Maximum Marks	4 marks

Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Expected Result	Actual Result	Status	Comments	TC for Automation(Y/N)	BUG ID
LoginPage_TC_OO1	Functional	Home Page	Verify user is able to see the Landing Page when user can type the URL in thebox		1.Enter URL and click go 2.Type the URL 3.Verify whether it is processing or not.	Should Display the Webpage	Working as expected	Pass		N	
LoginPage_TC_OO2	UI	Home Page	Verify the UI elements is Responsive		1. Enter URL and click go 2. Type or copy paste the URL 3. Check whether the button is responsive or not 4. Reload and Test Simultaneously	Should Wait for Response and then gets Acknowledge	Working as expected	Pass		N	
LoginPage_TC_OO3	Functional	Home page	Verify whether the link is legitimate or not		1. Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Observe the results	User should observe whether the website is legitimate or not.	Working as expected	Pass		N	
LoginPage_TC_OO4	Functional	Home Page	Verify user is able to access the legitimate website or not		1. Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Continue if the website is legitimate or be cautious if it is not legitimate.	Application should show that Safe Webpage or Unsafe.	Working as expected	Pass		N	
LoginPage_TC_OO5	Functional	Home Page	Testing the website with multiple URLs		1.Enter URL (https://phishingshield.herokuapp.com/) and click go 2. Type or copy paste the URL to test 3. Check the website is legitimate or not 4. Continue if the website is secure or be cautious if it is not secure	User can able to identify the websites whether it is secure or not	Working as expected	Pass		N	

2. Tune the model

[ ]

	ML Model	Accuracy	f1_score	Recall	Precision
0	Logistic Regression	91.814	92.567	94.496	94.496
1	Random Forest	96.653	96.942	100.000	100.000
2	XgbClassifier	94.754	95.207	96.714	96.714
3	Decision tree	95.206	95.605	100.000	100.000

[ ]
```
sorted_result=result.sort_values(by=['Accuracy', 'f1_score'],ascending=False).reset_index(drop=True)
sorted_result
```

	ML Model	Accuracy	f1_score	Recall	Precision
0	Random Forest	96.653	96.942	100.000	100.000
1	Decision tree	95.206	95.605	100.000	100.000
2	XgbClassifier	94.754	95.207	96.714	96.714
3	Logistic Regression	91.814	92.567	94.496	94.496

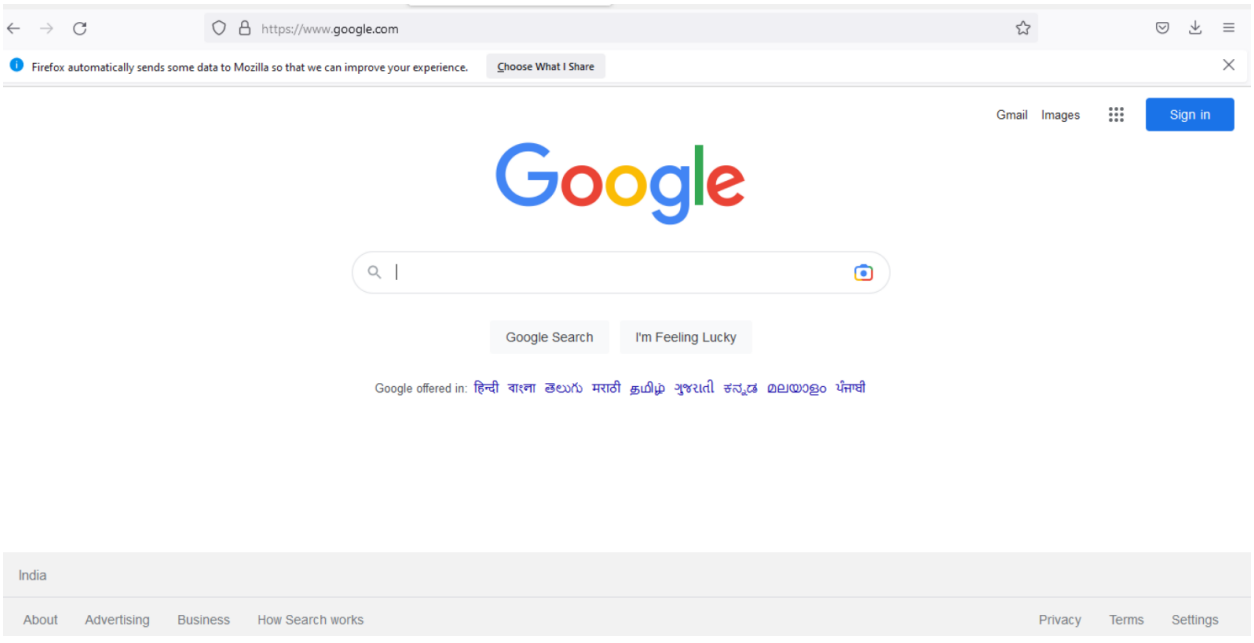
9.RESULTS

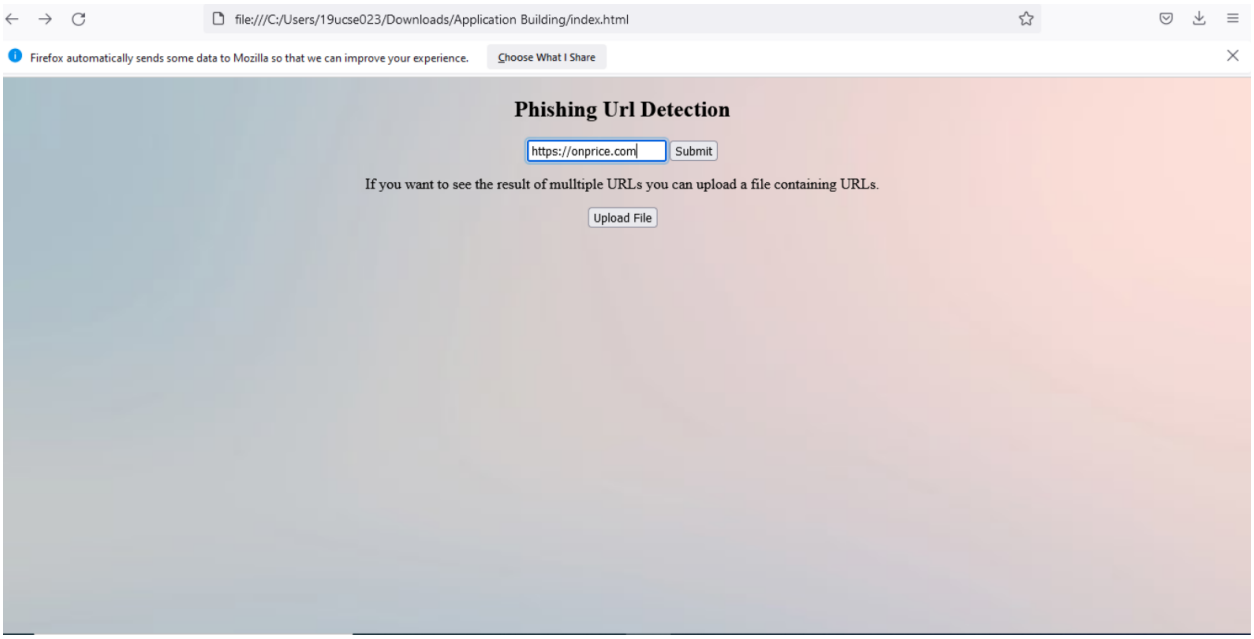
9.1 PERFORMANCE METRICSEvaluating the performance of a Machine learning model is one of the important steps while building an effective ML model. To evaluate the performance or quality of the model, different metrics are used, and these metrics are known as performance metrics or evaluation metrics. These performance metrics help us understand how well our model has performed for the given data. In this way, we can improve the model's performance by tuning the hyper-parameters. Each ML model aims to generalize well on unseen/new data, and performance metrics help determine how well the model generalizes on the new dataset.

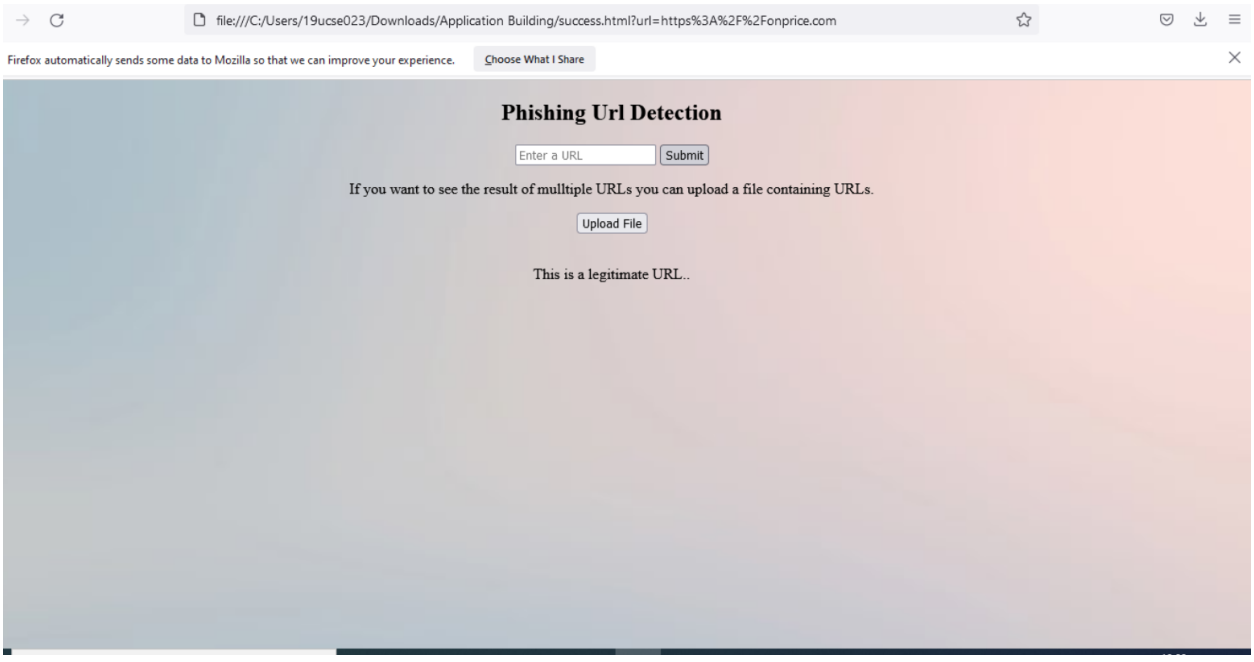
Phishing Detection

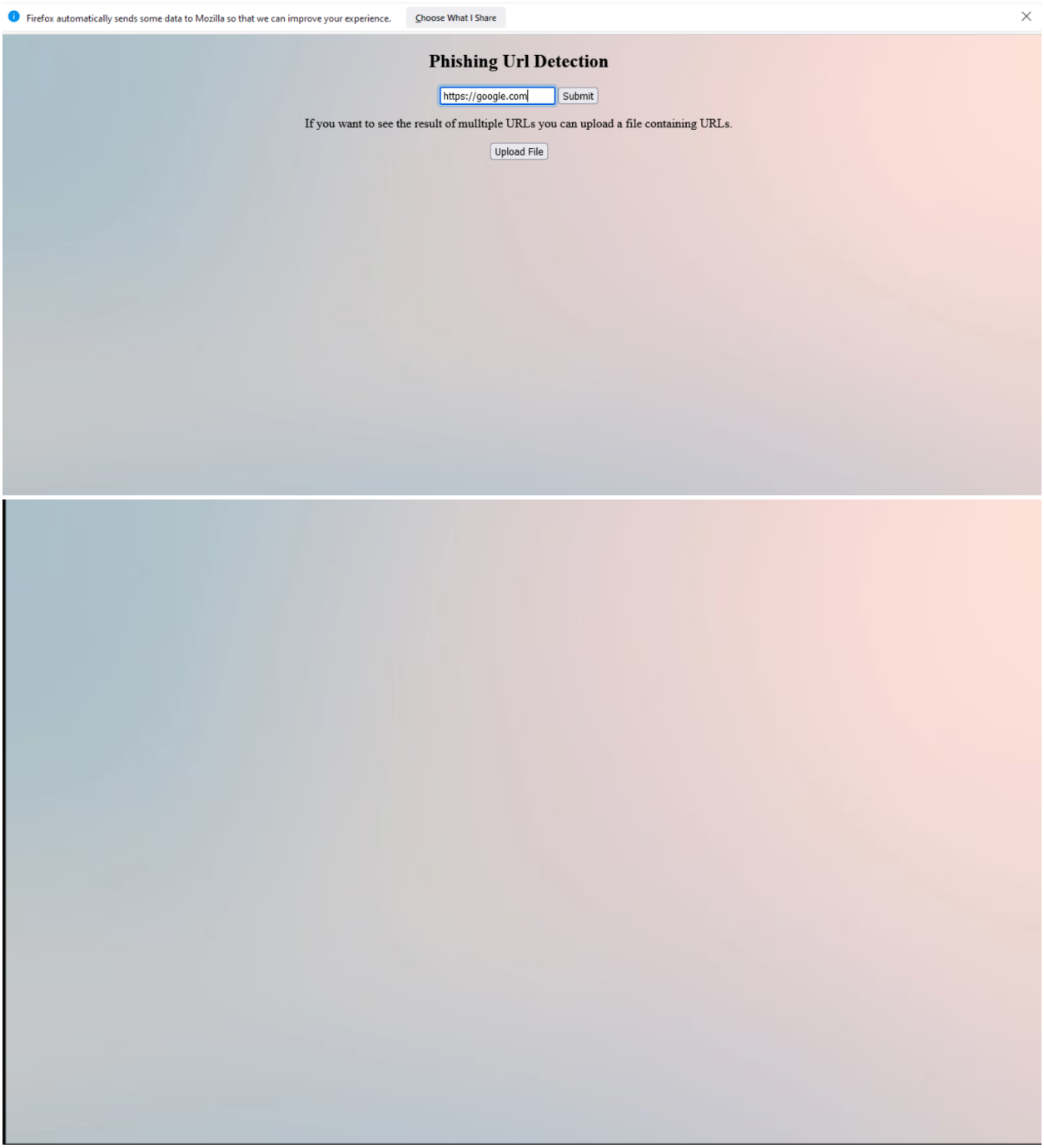
```
Import numpy as np
from sklearn.ensemble
import RandomForestClassifier as rfc
from sklearn.model_selection
import train_test_split
import feature extraction
def getResult(url):
 #Importing dataset data= np.loadtxt("dataset.csv", delimiter = "")
 #Seperating features and labels X = data[:, -1] y = data[:, -1]
#Seperating training features, testing features, training labels & testing
labels x_train, x_test, y_train, y_test train_test_split(x, y, test_size = 0.2)
clf- rfc() clf.fit(x_train, y_train) score clf.score(x_test, y_test) print(score
100)
 X_new = []
```

```
 X input url
 X_new-feature extraction.generate_data_set(X_input)
 X new np.array(X_new).reshape(1,-1)
 try:
 prediction clf.predict(X_new)
if prediction -1:
 return "Phishing Url"
 else:
return "Legitimate Url"
 except:
 return "Phishing Url"
```




file:///C:/Users/19ucse023/Downloads/Application Building/index.html
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Phishing Url Detection
https://onprice.com
Submit
If you want to see the result of mulltiple URLs you can upload a file containing URLs.
Upload File


file:///C:/Users/19ucse023/Downloads/Application Building/success.html?url=https%3A%2F%2Fonprice.com
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Phishing Url Detection
Enter a URL
Submit
If you want to see the result of mulltiple URLs you can upload a file containing URLs.
Upload File
This is a legitimate URL..


Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Phishing Url Detection
https://google.com
Submit
If you want to see the result of mulltiple URLs you can upload a file containing URLs.
Upload File

10.ADVANTAGES AND DISADVANTAGES

No	Techniques Used	Advantages	Disadvantages
1	*Methods based on Bag-of-Words model*	-Build secure connection between user's mail transfer Agent (MTA) and mail user agent (MUA)	-Time consuming - huge number of features -consuming memory
2	*Compared multi Classifiers algorithms*	-Provide clear idea about the effective level of each classifier on phishing email detection	Non standard classifier
3	*hybrid system*	-High level of accuracy by take the advantages of many classifiers	-Time consuming because this technique has many layers to make the final result
4	*Classifiers Model-Based Features*	- High level of accuracy - create new type of features like Markov features	-huge number of features -many algorithm for classification which mean time consuming -higher cost -need large mail server and high memory requirement
5	*Clustering of Phishing Email*	-Fast in classification process	-Less accuracy because it depend on unsupervised learning , need feed continuously
6	Evolving Connectionist System (ECOS) for phishing email detection	fast ,less consuming memory, high accuracy, Evolving with time, online working	Need feed continuously

11. CONCLUSION

Finally, phishing attacks are a major problem. It is important that they are countered. The work reported in this thesis indicates how understanding of the nature of phishing may be increased and provides a method to identify phishing problems in systems. It also contains a prototype of a system that catches those phishing attacks that evaded other defences, i.e. those attacks that have slipped through the net=. An original contribution has been made in this important field, and the work reported here has the potential to make the internet world a safer place for a significant number of people.

In the future we provide some technical solution by improve the efficiency of spam filters. By which too many mails are classified correctly and properly. By this legitimate user can surf internet with less fear. The user-phishing interaction model was derived from application of cognitive walkthroughs. A large-scale controlled user study and follow on interviews could be carried out to provide a more rigorous conclusion. The current model does not describe irrational decision making nor address influence by other external factors such as emotion, pressure, and other human factors. It would be very useful to expand the model to accommodate these factors. we have theoretically and experimentally evaluated of Phish Limiter. We have evaluated the trustworthiness of each SDN üow to identify any potential hazards based on each deep packet inspection. Likewise, we have observed how the proposed inspection approach of two SF and FI modes within Phish Limiter detects and mitigates phishing attacks before reaching end users if the üow has been determined untrustworthy. Using our real-world experimental evaluation on GENI and phishing dataset, we have demonstrated that Phish Limiter is an effective and efûcient solution to detect and mitigate phishing attacks with its accuracy of98.39%.

12.FEATURE SCOPE

Phishing attacks are growing in the similar manner as e-commerce industries are growing. Prediction and prevention of phishing attacks is a very critical step towards safeguarding online transactions. Data mining tools can be applied in this regard as the technique is very easy and can mine millions of information within seconds and deliver accurate results. With the help of machine learning algorithms like, Random Forest, Decision Tree, Neural network and Linear model we can classify data into phishing, suspicious and legitimate. This can be done based on unique features of phishing websites and user does not need to check individual websites. Rather we can identify and predict phishing, suspicious and legitimate websites by extracting some unique features. The aim of this work was to develop model to safeguard users from phishing attack.

13.APPENDIX

13.1 Source code

```
<!DOCTYPE html>
<html>
<head>
	<title>Phishing</title>
	<style>
	 body {
		background-image: url('img.jpg');
		background-repeat: no-repeat;
		background-attachment: fixed;
```

```
            background-size: 100% 100%;
        }
        </style>
</head>
<body>
        <center>
        <h2>Phishing Url Detection</h2>
        <form action="success.html">
            <input type="text" id="url" name="url" placeholder="Enter a
URL">
            <button>Submit</button>
            <p>If you want to see the result of multiple URLs you can
upload a file containing URLs.</p>
            <button>Upload File</button>
        </form>
        </center>
</body>
</html>
```

13.2 Github link and Demo Link

https://github.com/IBM-EPBL/IBM-Project-34490-1660236434

https://"Demo link IBM.approj-wal"