A

Major Project

On

DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

Submitted to

Jawaharlal Nehru Technological University, Hyderabad

In partial fulfillment of the requirements for the award of Degree

Of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

A. VARUN KUMAR	**(187R1A0504)**
E.UMA MAHESHWARI	**(187R1A0520)**
K. VISHANTH	**(187R1A0531)**

Under the Guidance of

DR. K SRUJAN RAJU

(Head of the Department)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi)
Recognized Under Section 2(f) & 12(B) of the UGCAct.1956, Kandlakoya (V), Medchal Road,
Hyderabad-501401.
2018-2022

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING**" is being submitted by**A. VARUN KUMAR (187R1A0504), E.UMA MAHESWARI (187R1A0520), K. VISHANTH (187R1A0531)** inpartial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2021-22.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dr. K. Srujan Raju
Head of the Department
INTERNAL GUIDE

Dr. A. RajiReddy
DIRECTOR

Dr. K.Srujan Raju
HOD

EXTERNALEXAMINER

Submitted for viva voice Examination heldon________________

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

Dr. K. Srujan Raju, head of the department for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing,help and guidance given by him shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) **Mr. A. Uday Kiran,Mr. J. Narasimha Rao, Dr. T. S. Mastan Rao,Mrs.G. Latha,Mr. A. Kiran Kumar,** for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. K. Srujan Raju,** Head, Department of Computer Science and Engineering for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We also express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

A. VARUN KUMAR (187R1A0502)
E.UMA MAHESHWARI(187R1A0520)
K. VISHANTH (187R1A0531)

ABSTRACT

The risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays.

There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a bigdata technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection

LIST OF FIGURES

FIGURE NO	FIGURE NAME	PAGE NO
Figure 3.1	Project Architecture	7
Figure 3.2	Use case diagram	8
Figure 3.3	Class diagram	9
Figure 3.4	Sequence diagram	10
Figure 3.5	Activity diagram	11

LIST OF SCREENSHOTS

SCREENSHOT NO	SCREENSHOT NAME	PAGE NO
Screenshot 5.1	Landing Page	19
Screenshot 5.2	Malicious URL Page	20
Screenshot 5.3	Legitimate URL Page	21
Screenshot 5.4	Domain Information	22
Screenshot 5.5	User History	23

TABLE OF CONTENTS

ABSTRACT i

LIST OF FIGURES ii

1. INTRODUCTION

1.1 PROJECT SCOPE 1

1.2 PROJECT PURPOSE 1

1.3 PROJECT FEATURES 1

2. SYSTEM ANALYSIS

2.1 PROBLEM DEFINITION 2

2.2 EXISTING SYSTEM 2

2.2.1 LIMITATIONS OF THEEXISTINGSYSTEM 3

2.3 PROPOSED SYSTEM 3

2.3.1 ADVANTAGES OFPROPOSED SYSTEM 3

2.4 FEASIBILITY STUDY 4

2.4.1 ECONOMIC FEASIBILITY 4

2.4.2 TECHNICAL FEASIBILITY 5

2.4.3 BEHAVIORAL FEASIBILITY 5

2.5 HARDWARE &SOFTWAREREQUIREMENTS 5

2.5.1 HARDWARE REQUIREMENTS 5

2.5.2 SOFTWARE REQUIREMENTS 6

3. ARCHITECTURE

3.1 PROJECT ARCHITECTURE 7

3.2 DESCRIPTION 7

3.3 USE CASE DIAGRAM 8

3.4 CLASS DIAGRAM 9

3.5 SEQUENCE DIAGRAM 10

3.6 ACTIVITY DIAGRAM 11

4. IMPLEMENTATION 12

4.1 SAMPLE CODE 12

5. SCREENSHOTS 19

6.TESTING 23

6.1 INTRODUCTION TO TESTING 23

6.2 TYPES OF TESTING 23

6.2.1 UNIT TESTING 23

6.2.2 INTEGRATION TESTING 24

6.2.3 FUNCTIONAL TESTING 24

6.3 TEST CASES 25

6.3.1 CLASSIFICATION 25

7. CONCLUSION & FUTURE SCOPE 26

7.1 PROJECT CONCLUSION 26

7.2 FUTURE SCOPE 26

8. REFERENCES 27

8.1 REFERENCES 27

8.2 WEBSITES 27

8.3 GITHUB LINK 27

1. INTRODUCTION

1. INTRODUCTION

1.1 PROJECTSCOPE

The risk of network information in security is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. In short, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes

1.2 PROJECTPURPOSE

Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or another phishing site, or malware download

1.3 PROJECTFEATURES

The main feature of this project is that the system prevents users from visiting malicious websites by displaying a pop-up, it also displays the information regarding the website such as domain details and also gives suggestions for user on how to be safe from malicious URLS

2. SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

SYSTEM ANALYSIS

System Analysis is the important phase in the system development process. The System is studied to the minute details and analyzed. In analysis, a detailed study of these operations performed by the system and their relationships within and outside the system is done. A key question considered here is, “what must be done to solve the problem?” The system is viewed as a whole and the inputs to the system are identified. Once analysis is completed the analyst has a firm understanding of what is to be done.

2.1 PROBLEMDEFINITION

Compromised URLs that are used for cyber-attacks are termed as malicious URLs. In fact, it was noted that close to one-third of all websites are potentially malicious in nature, demonstrating rampant use of malicious URLs to perpetrate cyber-crimes. A Malicious URL or a malicious web site hosts a variety of unsolicited content in the form of spam, phishing, or drive-by download in order to launch attacks.

2.2 EXISTINGSYSTEM

The Traditional classification techniques like blacklisting, regular expression, and signature matching approach are lacking the ability to detect newly generated malicious URLs

2.2.1 LIMITATIONS OF EXISTINGSYSTEM

Following are the limitations of the existing system:

- It is hard and expensive
- New types of attacks and vulnerabilities emerge continuously, so they maybe misclassified.
- Lack of Security

2.3 PROPOSEDSYSTEM

Given the URL we extract the following features:

1. Lexical features

2. Host-based features

3. popularity features

using these features, we predict if the given URL is a malicious URL or a legitimate URL

2.3.1 ADVANTAGES OF THE PROPOSEDSYSTEM

The following are the advantages of the proposed system:

- Detects if the given URL is malicious
- Displays suggestions if the URL is malicious
- Ensures safe surfing
- Gives the details of the particular Domain, IP rating

2.4 FEASIBILITYSTUDY

The feasibility of the project is analyzed in this phase and the business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis, the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. Three key considerations involved in the feasibility analysis:

- Economic Feasibility
- TechnicalFeasibility
- Social Feasibility

2.4.1 ECONOMIC FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on a project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require. The following are some of the important financial questions asked during the preliminary investigation:

- The costs conduct a full system investigation.
- The cost of the hardware and software.
- The benefits are in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also, all the resources are already available, which gives an indication that the system is economically possible for development.

2.4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

2.4.3 BEHAVIORAL FEASIBILITY

This includes the following questions:

- Is there sufficient support for theusers?
- Will the proposed system causeharm?

The project would be beneficial because it satisfies the objectives when developed and installed. All behavioral aspects are considered carefully and conclude that the project is behaviorally feasible.

2.5 HARDWARE &SOFTWAREREQUIREMENTS

2.5.1 HARDWARE REQUIREMENTS:

Hardware interfaces specify the logical characteristics of each interface between the software product and the hardware components of the system.

The following are some hardware requirements:

- CPU: Intel core i3 and above
- RAM: 4 GB and above
- Hard disk: 8 GB and above
- Input devices: Keyboard, Mouse

2.5.2 SOFTWAREREQUIREMENTS:

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements:

- Operating System: Windows – 8 andabove
- Programming Language: Python 3.7, Html, CSS, JS
- IDE: Anaconda - Jupyter notebook and Spyder

3. ARCHITECTURE

3. ARCHITECTURE

3.1 PROJECTARCHITECTURE

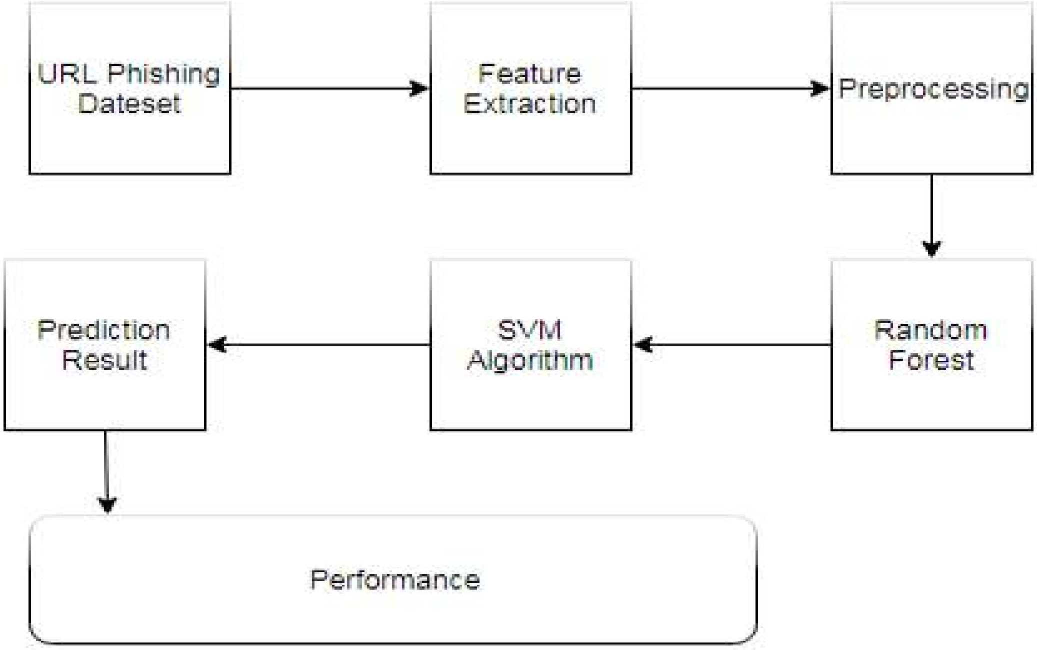


Figure 3.1: Architecture of detection of phishing websites using random forest classifier

3.2 DESCRIPTION

The most common method to detect malicious URLs deployed by many antivirus groups is the blacklist method. Blacklist are essentially a database of URLs that have been confirmed to be malicious in the past.

Machine learning approaches try to analyze the information of URL and its corresponding websites or webpages, by extracting good feature representations of URLs, and training a prediction model on training data of both malicious and benign URLs. In this we can use static and dynamic features can be used - static features can perform the analysis of a webpage based on information available without executing the URLs that execute the JavaScript and includes the lexical and host based featurestatic analysis techniques have been extensively explored by applying machine learning techniques.

3.3 USE CASE DIAGRAM

In the use case diagram, we have two actors who are the user, the admin, . The user uploads the URLs in the tool.

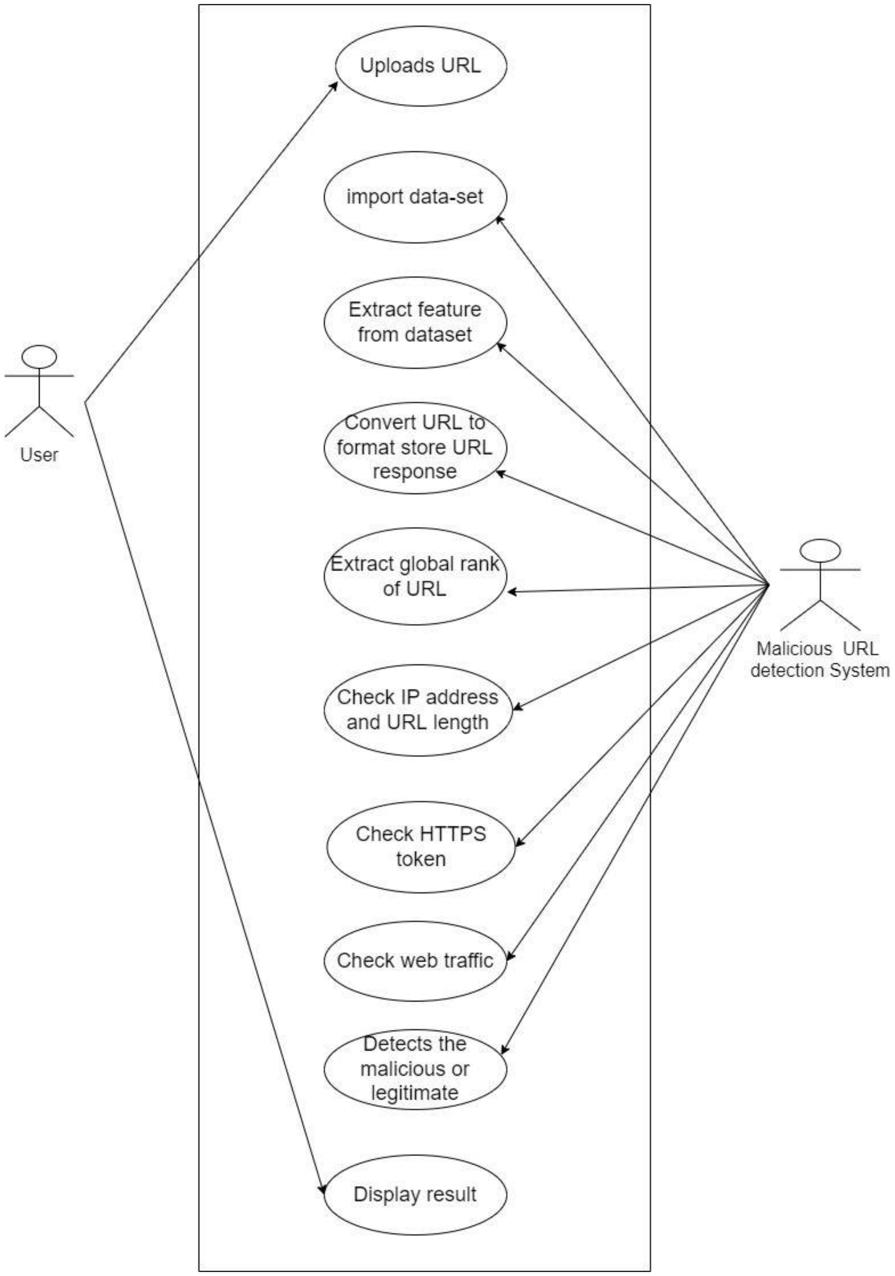

Uploads URL
import data-set
Extract feature from dataset
Convert URL to format store URL response
Extract global rank of URL
Check IP address and URL length
Check HTTPS token
Check web traffic
Detects the malicious or legitimate
Display result
User
Malicious URL detection System

Figure 3.2: Use Case Diagram for detection of phishing websites using random forest classifier

3.4 CLASSDIAGRAM

Class Diagram is a collection of classes and objects.

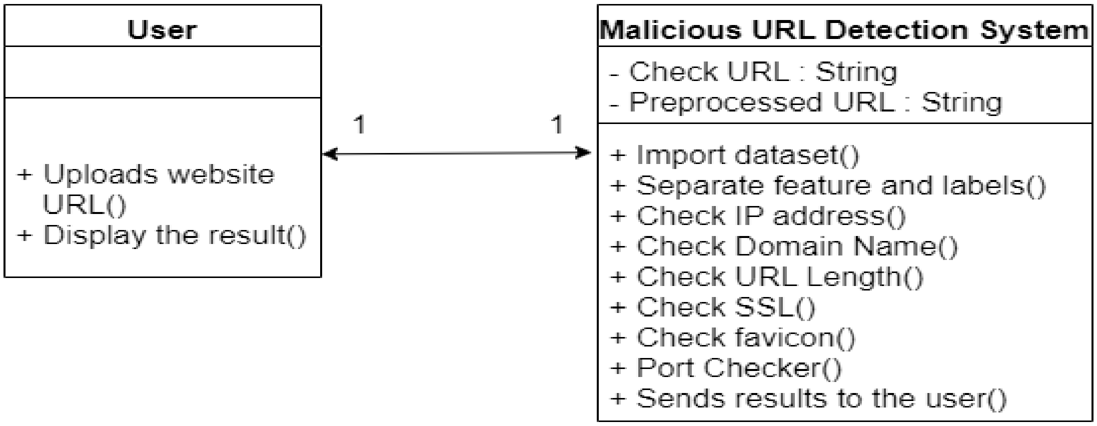


Figure 3.3: Class Diagram for detection of phishing websites using random forest classifier

3.5 SEQUENCEDIAGRAM

The below Figure 3.4 depicts the Sequence diagram of vulnerability detection using random forest classifier.

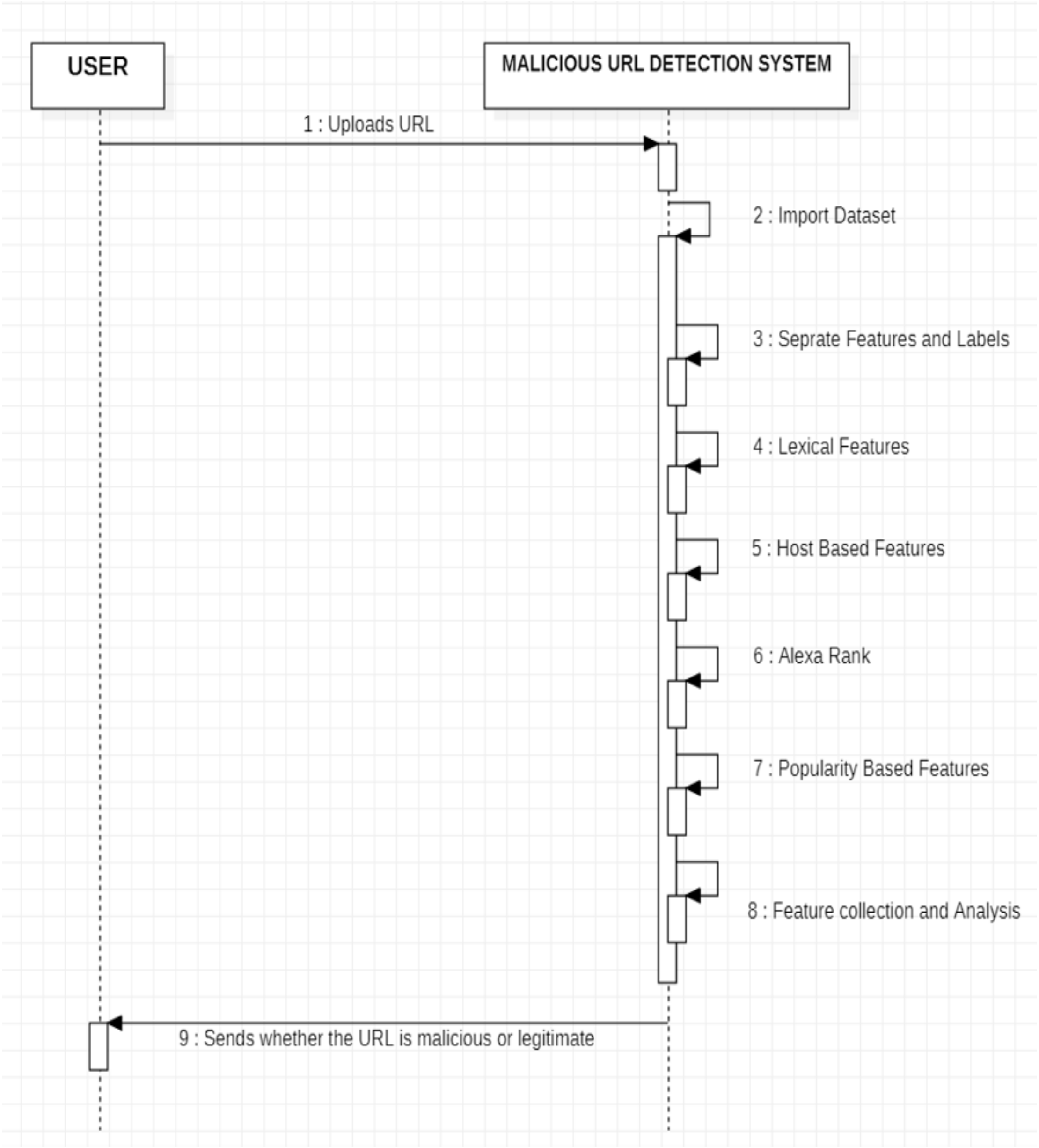

USER
MALICIOUS URL DETECTION SYSTEM
1 : Uploads URL
2 : Import Dataset
3 : Seprate Features and Labels
4 : Lexical Features
5 : Host Based Features
6 : Alexa Rank
7 : Popularity Based Features
8 : Feature collection and Analysis
9 : Sends whether the URL is malicious or legitimate

Figure 3.4: Sequence Diagram for detection of phishing websites using random forest classifier

3.6 ACTIVITY DIAGRAM

The activity diagram describes the flow of activity states.

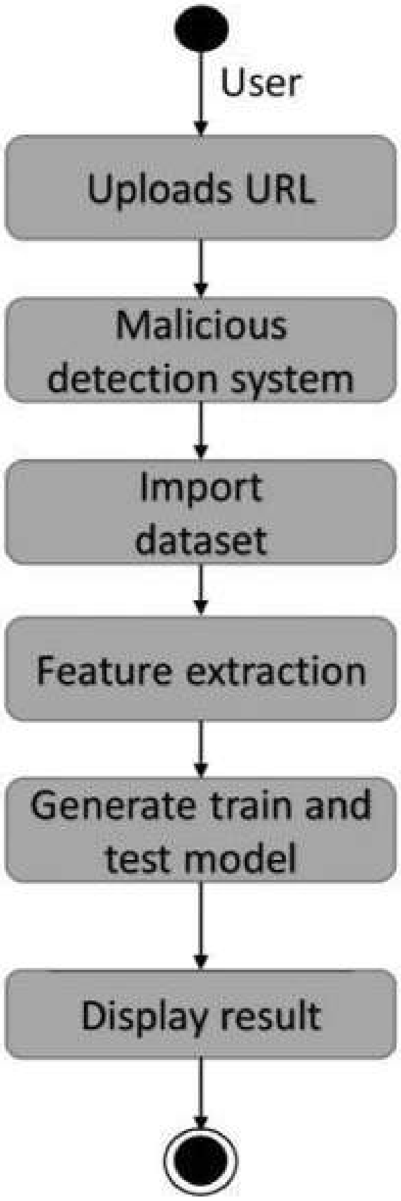


Figure 3.5: Activity Diagram for detection of phishing websites using random forest classifier

4. IMPLEMENTATION

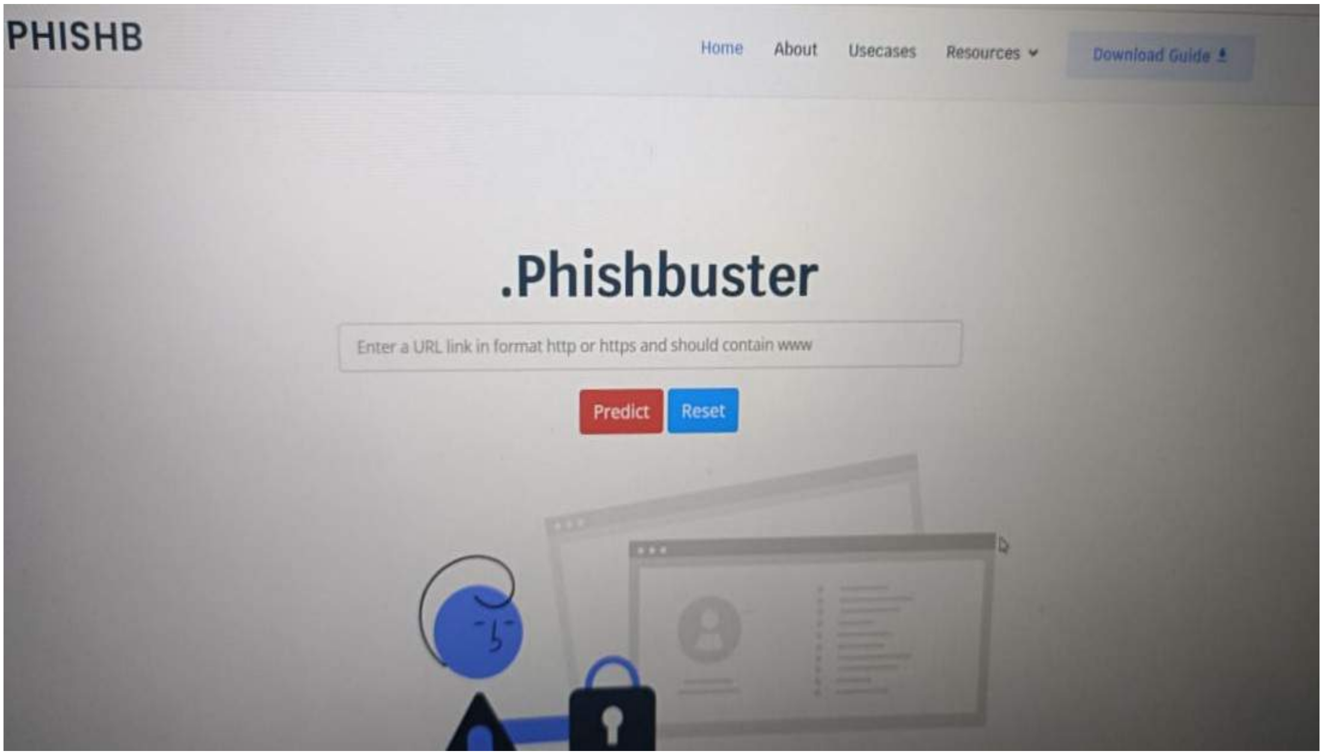
4.1 SAMPLE CODE

```
#importing basic packages
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

Index(['Domain', 'Have_IP', 'Have_At', 'URL_Length', 'URL_Depth',
'Redirection', 'https_Domain', 'TinyURL', 'Prefix/Suffix', 'DNS_Record',
'Web_Traffic', 'Domain_Age', 'Domain_End', 'iFrame', 'Mouse_Over',
'Right_Click', 'Web_Forwards', 'Label'],
dtype='object')

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Domain         10000 non-null  object
 1   Have_IP        10000 non-null  int64
 2   Have_At        10000 non-null  int64
 3   URL_Length     10000 non-null  int64
 4   URL_Depth      10000 non-null  int64
 5   Redirection    10000 non-null  int64
 6   https_Domain   10000 non-null  int64
 7   TinyURL        10000 non-null  int64
 8   Prefix/Suffix  10000 non-null  int64
 9   DNS_Record     10000 non-null  int64
 10  Web_Traffic    10000 non-null  int64
 11  Domain_Age     10000 non-null  int64
 12  Domain_End     10000 non-null  int64
 13  iFrame         10000 non-null  int64
 14  Mouse_Over     10000 non-null  int64
 15  Right_Click    10000 non-null  int64
 16  Web_Forwards   10000 non-null  int64
 17  Label          10000 non-null  int64
dtypes: int64(17), object(1)
memory usage: 1.4+ MB

# Sepratating& assigning features and target columns to X & y
y = dfsa['Label']  #target variable
X = dfsa.drop('Label',axis=1)   #independent variable
# Decision Tree model
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
tree = DecisionTreeClassifier(max_depth = 5)
# fit the model
tree.fit(X_train, y_train)
```

5. SCREENSHOTS

5.1 HOMEPAGE

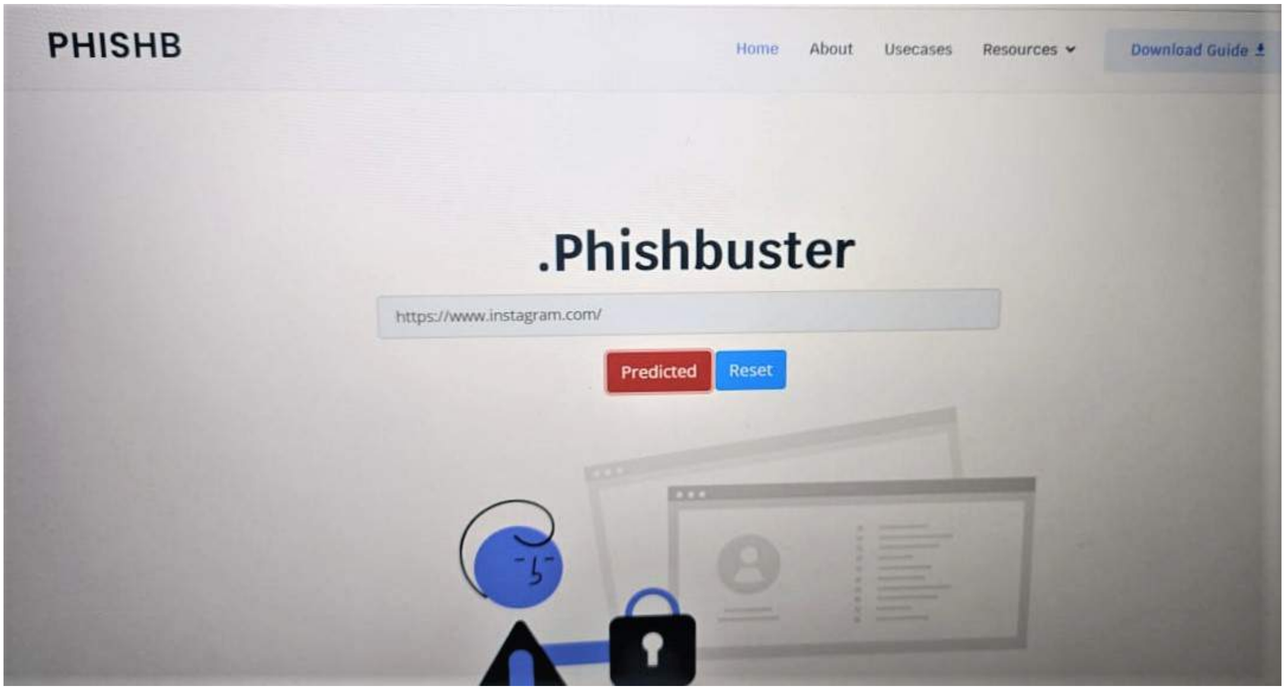
This home page indicates the user uploads the URL to check whether the URL is malicious or legitimate.



Screenshot 5.1: Landing page

5.2 MALICIOUSURL

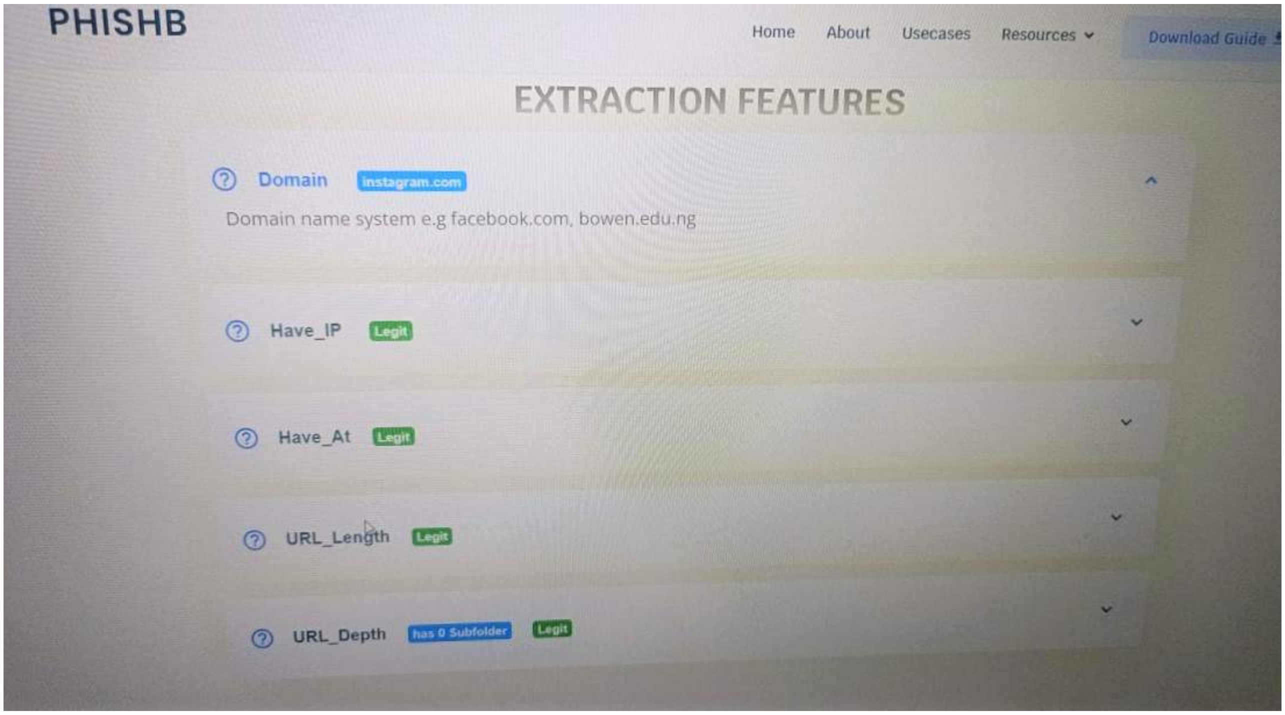
When a user uploads the URL if it is malicious the detection system displays the result.



Screenshot 5.2: Malicious URL Page

5.3 DOMAININFORMATION

When the detection system detects the malicious URL and also displays the domaindetails of the URL.



EXTRACTION FEATURES

Domain instagram.com

Domain name system e.g facebook.com, bowen.edu.ng

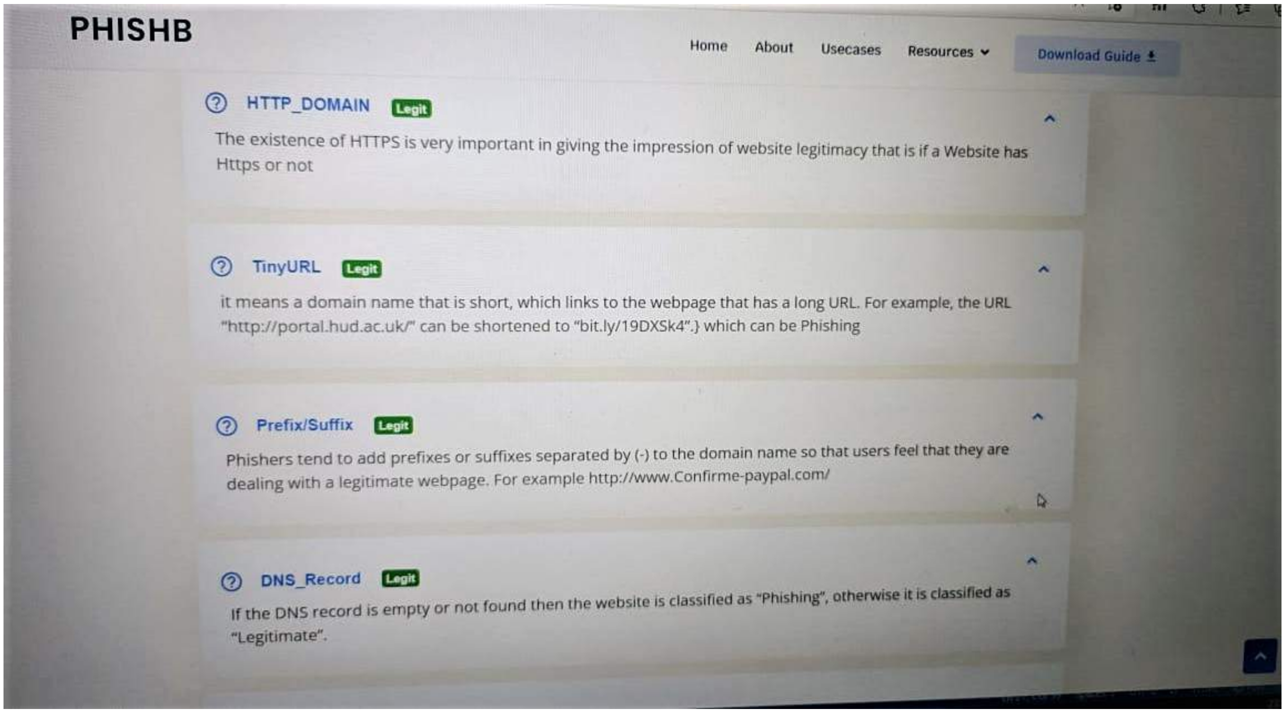
Have_IP Legit

Have_At Legit

URL_Length Legit

URL_Depth has 0 Subfolder Legit

Screenshot 5.4: Domain Information



HTTP_DOMAIN Legit

The existence of HTTPS is very important in giving the impression of website legitimacy that is if a Website has Https or not

TinyURL Legit

it means a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".} which can be Phishing

Prefix/Suffix Legit

Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example http://www.Confirme-paypal.com/

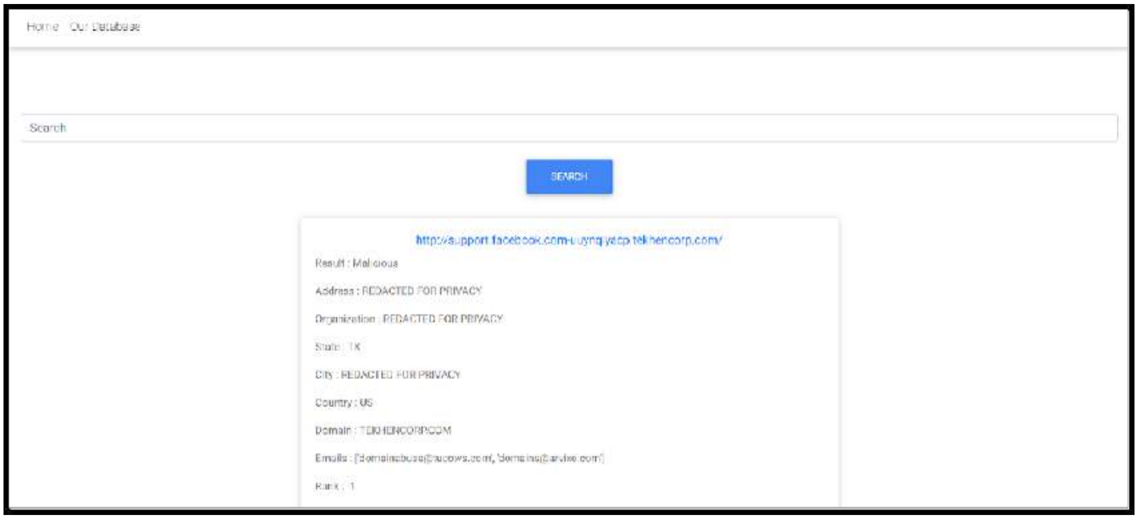
DNS_Record Legit

If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

Screenshot 5.4.1: Domain Information

5.4 HISTORY

Our database contains all the searches that the user has made.



http://support.facebook.com-uuyngiyacp.tekhencorp.com/

Result : Malicious

Address : REDACTED FOR PRIVACY

Organization : REDACTED FOR PRIVACY

State : TX

City : REDACTED FOR PRIVACY

Country : US

Domain : TEKHENCORP.COM

Emails : ['domainabuse@tucows.com', 'domains@arvixe.com']

Rank : 1

Screenshot 5.5: User History

6. TESTING

6. TESTING

6.1 INTRODUCTION TOTESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6.2 TYPES OFTESTING

6.2.1 UNITTESTING

Unit testing involves the design of test cases that validate that the internal programlogic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

6.2.2 INTEGRATIONTESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components are correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.2.3 FUNCTIONALTESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, anduser manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes

6.3 TEST CASES

Test case ID	Test case name	Purpose	Test Case	Output
1	Uploading URL	For processing the URL in the detection system	The user uploads the URL and checks whether malicious or not	Uploaded Successfully
2	Malicious URL	Checking whether the URL is malicious or not	From the feature extraction it checks the URL	Malicious URL detected
3	Legitimate URL	Checking the URL malicious or not	From the feature extraction it checks the URL	The URL looks safe
4	Domain details	It displays the URL host details	It displays the host, email, IP address of URL	Deatils of the domain

7. CONCLUSION

7. CONCLUSION & FUTURESCOPE

7.1 PROJECTCONCLUSION

The detection model achieves the expected effect in experiments. However, considering that the network traffic in the test environment and the real network are different, and with the development of the Internet, types of malicious URL are more diverse. It is necessary to timely update the model in the actual scenario. Therefore, to better adapt to the requirements of various complex application scenarios, we plan to study how to simplify the detection model's architecture and shorten the training timewhile keeping the detection performance unchanged in the future.

7.2 FUTURESCOPE

Creating Google-chrome extension so that users can directly interact with theapplication without any installation process and users can get the results instantly on the same web page on which they are working. Vulnerabilities are rapidly increasingas new technologies are getting evolved so there can be numerous number of loop holes so new algorithms must be implemented for best results.

8. BIBLIOGRAPHY

8. BIBLIOGRAPHY

8.1 REFERENCES

[1] Sadia Afroz and Rachel Greenstadt. 2011. Phishzoo: Detecting phishing websites by looking at them. In Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE.

[2] A Astorino, A Chiarello, M Gaudioso, and A Piccolo. 2016. Malicious URL Detection via spherical classification. Neural Computing and Applications (2016)

8.2 WEBSITES

[1] https://www.irjet.net/archives/V8/i4/IRJET-V8I4274.pdf

[2] https://blog.keras.io/building-autoencoders-in-keras.html

8.3 GitHubLink

https://github.com/goodydeves/Detection-of-Phishing-Website-Using-Machine-Learning