e-ISSN:2582 - 7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 4, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 5.928

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

Smarteye - Fishy URL Detection Using URL Features and CNN

Nayan Suresh Ambolikar[1], AtulMadan Thorat[2], Ashutosh Mahendra Sakhare[3], Prof. Vijay Sonawane[4]

U.G. Students, Dept. of Computer Engineering, JSPM's BSIOTR, Pune University - Pune, Maharashtra, India[1,2,3]

Professor, Dept. of Computer Engineering, JSPM's BSIOTR, Pune University - Pune, Maharashtra, India[4]

ABSTRACT: Phishing is a criminal scheme to steal the user's personal data and other credential information. It is afraud that acquires victim's confidential information such as password, bank account detail, credit card number,financial username and password etc. and later it can be misused by attacker.

First of all, the phisher has to create a phishing website to lure the victim which seems as legitimate one. Then, hostthe site on the internet for use of victim secrete information. If victim visit phishing website, it convinces the victim toenter some confidential information. Phisher then acquire some entered data and later it can be misuse by phisher.

We aim to use WhoIs features of URL as the basis of detecting phishing websites. We propose a novel solution,Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URLand WhoIs features. The convolution Neural Network is used to train the network and finally detect the site is Phishingor not.

KEYWORDS: Fishy URL Detection, WHOIS features, URL Features, Convolutional Neural Network, URL Length,IP address, Avoiding Phishing Attacks.

I.INTRODUCTION

Phishing is defined as the fraudulent acquisition of confidential data by the intended recipients and the misuse ofsuch data. The phishing attack is often done by email. An example of Phishing; as if e-mail appear to be from knownweb sites, from a user's bank, credit card company, e-mail, or Internet service provider. Generally, personalinformation such as credit card number or password is asked to update accounts.

These emails contain a URL link that directs users to another website. This site is actually a fake or modifiedwebsite. When users go to this site, they are asked to enter personal information to be forwarded to the phishingattacker.

PHISHING ATTACKS

The aim is to steal sensitive data such as credit card and login information or to install malicious software on thevictim's machine.Phishing is a common type of cyber-attack that everyone must learn to protect them. Phishing is start with a fake e-mail or other type of transmission designed to attract a victim. In this type of attack, the message appears to come froma trusted source.

In a phishing attack, attackers can use social engineering and other public information resources, including socialnetworks like LinkedIn, Facebook and Twitter, to gather background information about the victim's personal and

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

workhistory, interests and activities. With this pre-discovery, attackers can identify potential victims' names, job titles and email addresses, information about the names of key employees in their colleagues and organizations.

Phishing is also used to learn someone's password or credit card information. With the help of e-mail prepared as ifcoming from a bank or official institution, computer users are directed to fake sites.

The common information that is stolen by a phishing attack is listed as follows:

- User account number
- User passwords and user name
- Credit card information

II.LITERATURE REVIEW

A Jian Mao1* , JingdongBian et al.[1] proposed a system where the key aim is to enable automated page-layout based phishing detection techniques using machine learning techniques.

The given aggregation analysis mechanism decides page layout similarity, which is used to detect phishing pages. Itevaluates four popular machine learning classifiers on their accuracy and the factors affecting their results.

SHAFI'I MUHAMMAD ABDULHAMID et al.[2] uses soft computing approach called Artificial Neural Network(ANN) algorithm with confusion matrix analysis for the detection of e-banking phishing websites. The ANN algorithmproduces a significant accuracy and reduced false positive rate during detection. This signifies that ANN algorithmwith confusion matrix analysis can generate a competitive results that is fit for detecting phishing in e-bankingwebsites.

NedaAbdelhamid et al.[3] experimentally compare large numbers of ML techniques on real phishing datasets andwith respect to different metrics. The main purpose of the comparison is to disclose the advantages and disadvantagesof ML predictive models and to show their actual performance when it comes to phishing attacks. The experimentalresults show that Covering approach models are more suitable as anti-phishing solutions, particularly for novice users,since of their simple yet effective knowledge bases in addition to their good phishing detection rate.

Longfei Wu, et al.[4] designed for web phishing attacks on PCs cannot efficiently point out the several phishingattacks on mobile devices. Henceforth, the author presented MobiFish, a novel automated lightweight anti-phishingscheme for mobile platforms. MobiFish verifies the validity of web pages, applications, and persistent accounts bycomparing the actual identity to the claimed identity.

Mohammed NazimFeroz et al.[5] describes an approach that categorizes URLs repeatedly based on their lexicaland host based parameters. Clustering is used on the whole dataset and a cluster ID (or label) is calculated for eachURL, which in turn is used as a predictive feature by the classification system.

LuongAnh Tuan Nguyen et al.[6] proposing a new approach to detect phishing site by using the features of URL.Mostly,we develop different components from URL and compute a metric for every component. So, the page rankingwill be shared with the achieved metrics to decide whether the websites are phishing websites.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

III.METHODOLOGY OF PROPOSED SURVEY

The phishing attack is often done by email. An example of Phishing; as if e-mail appears to be from known websites, from a user's bank, credit card company, e-mail, or Internet service provider. Generally, personal informationsuch as credit card number or password is asked to update accounts.The system is solution for avoiding such phishing attacks.

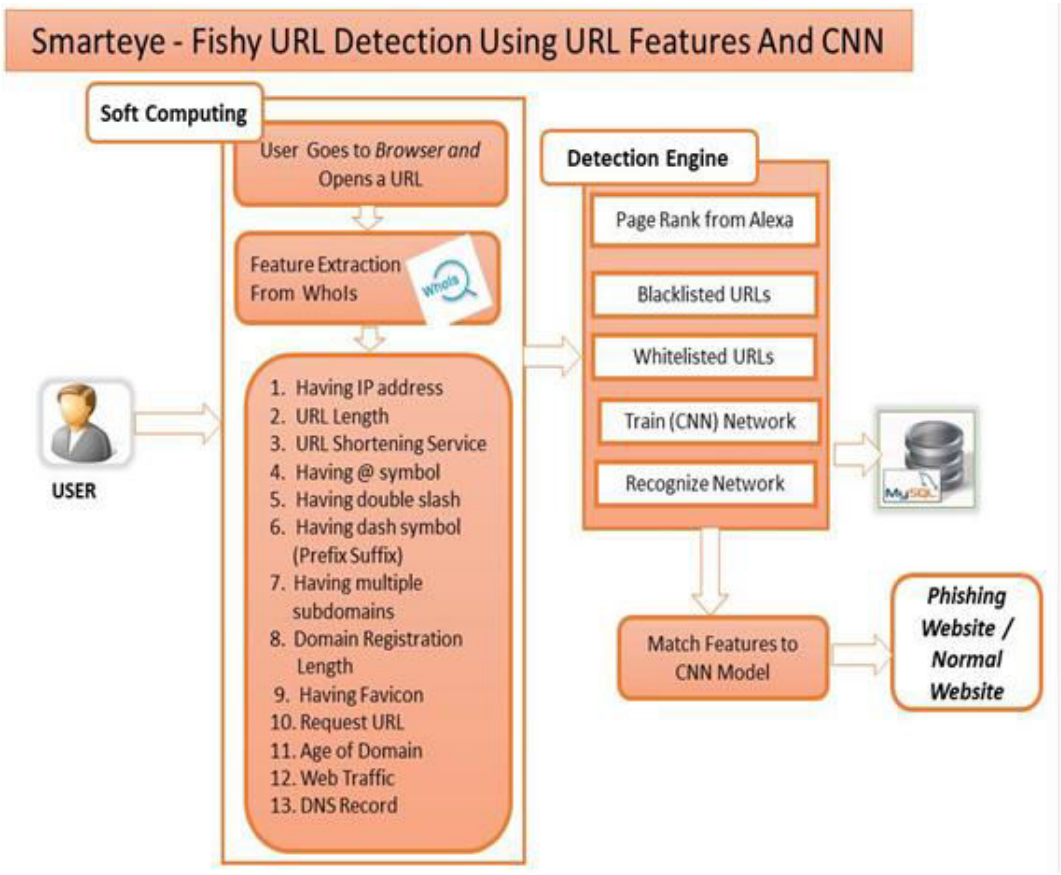

Figure 3.1: System Architecture

Whois Domain: WHOIS is a question and reaction convention that is broadly utilized for questioning databases thatstore the enlisted clients or trustees of an Internet asset, for example, an area name, an IP address square or a self-governing framework, but on the other hand is utilized for a more extensive scope of other data. The conventionstores and conveys database content in a comprehensible format. A WHOIS is a way for you to search the publicdatabase for information about a specific domain, such as the expiration date, current registrar, registrant information,etc.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

The following features are extracted from Whois and URLs:

Sr. No.	Features	Significance
01.	Having IP Address	If IP address is used in domain name
02.	URL Length	Legitimate URLs have length of nearly 75 characters, URL length morethan 75 is Phishing sites.
03.	Shortening Service	Shortening Service
04.	Having @ Symbol	Websites having @ symbol are Phishy in general
05.	Double slash redirecting	If there is '/1' then it can be categorized
06.	Having Sub Domain	Legitimate Websites use only domain generally upto two level.
07.	URL of Anchor	In legitimate websites the anchor tag is connected to the same domain asthe source code, Phishy
08.	Links in tags	Links in tags lead to some fraudulent websites
09.	Abnormal URL	This feature is extracted from Who is Database, Legitimate websites' mainidentity is in the URL
10.	Age of domain	Legitimate websites have an age of six months; websites with more thanthis age can be classified as Phishing.
11.	Page Rank	Phishing websites will have low page rank due to lack of links pointing tothem.
12.	Links Pointing to page	Phishing websites have links pointing to zip files that automatically getdownloaded containing malware.
13.	Favicon	Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity inthe address bar websites.
14.	DNS (Domain Name System) Record	If the DNS record is empty or not found then the website is classified as"Phishing", otherwise it is classified as "Legitimate".
15.	Web Traffic	Web traffic is the amount of data sent and received by visitors to a website.Phishing websites will create huge web traffic.
16.	Website Traffic	This feature measures the popularity of the website by determining thenumber of visitors and the number of pages they visit.

Table 3.1: Websites Features

Built Detection Model using Convolution Neural Network:

The system can detect the phishing site usingConvolution Neural Network (CNN) technique. A CNN consists of an input and an output layer, as well as multiplehidden layers. The hidden layers of a CNN typically consist of Convolutional layers, pooling layers, fully connectedlayers and normalization layers. CNN will be used to train the data analytics engine for recognizing the phishing siteURL.

Avoiding phishing attacks: A whitelist in the context of phishing detection is simply a list of trusted websites.

– The URL of the trusted site:

The URL of the trusted site is used to periodically update the information in the database. This is the URL ofthe site such as "https://signin.ebay.com".

– The domain of the site:

The domain of the trusted site is the domain of the URL such as "signin.ebay.com" and is used to determinewhether the current page displayed in the browser is on the whitelist or not.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

– The title of the site:

The title of the trusted site is the page title of the site such as "Welcome to eBay" and can be used to speed upthe matching potential of phishing site titles with titles in the whitelist Database.

– Alexa Ranking

In case your site is ranked relative to other sites, changes in traffic to other sites affect your site's rank. Everyday, Alexa estimates the average daily visitors and page views to every site over the past 3 months. The sitewith the highest combination of visitors and page views over the past 3 months is ranked 1. As phishingwebsites live for a short period of time, they may not be recognized by the Alexa database (Alexa the WebInformation Company., 1996). if the domain has no traffic or is not recognized by the Alexa database, it isclassified as "Phishing". Otherwise, it is classified as "Suspicious".

IV. PROPOSED ALGORITHM AND MATHEMATICALMODEL

4.1 ALGORITHM DETAILS

Convolutional Neural Networks (CNNs)

Convolution Neural Network Traditional feature learning methods rely on semantic labels of images as supervision.They usually assume that the tags are evenly exclusive and thus do not pointing out towards the complication of labels.The learned features endow explicit semantic relations with words. We also develop a novel cross-modal feature thatcan both represent visual and textual contents. CNN is a method of categorizing the images as a part of deep learning.In which we apply a single neural network to the full image. The steps in CNN are as follows: convolution,subsampling, activation and full connectedness.

Step 1: Convolution it is the primary layers that accept an input signal are called convolution filters. Convolution is aprocedure where the network tries to tag the input signal by referring to what it has learned in the past.

Step 2: Subsampling Inputs from the convolution layer can be smoothened to decrease the sensitivity of the filters tonoise and variations. This smoothing procedure is labeled as sub- sampling, and can be attained by taking averages orconsidering the maximum over a sample of the signal.

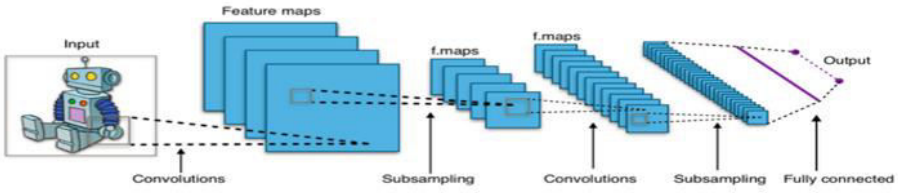

Figure 4.1: Working of Convolutional Neural Network Algorithm

Step 3: Activation the activation layer manages the signal flows from one layer to the subsequent Output signals whichare strongly connected with past references would activate more neurons, enabling signals to be propagated moreefficiently for identification.

Step 4: Fully connected the final layers in the network are fully connected, such that the neurons of preceding layers areconnected to every neuron in subsequent layers. This imitates high Level reasoning where all feasible path ways fromthe input to output is measured.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

4.2 MATHEMATICAL MODEL

Let us consider S be a Systems such that

S= U, ES, SS, K, DE, DS, where

- U= {U1, U2, U3. . . . . .Un | 'U' is a Set of all USERS }

  U is the users of the system. Users of the system may grow as the system is used by more and more people. User is infinite set.

- ES =ES1,ES2 | 'ES' is a Set of user visit to the browser and opens the a URL}

  These are the data to be entered in URL of the system, so this is also Finite Set.

- SS= {SS1,SS2,SS3,....SSn | SS is a Set of features checked for detection}

  SS are the main features like DNS Test, IP address, URL encode, Shorten URL, WhiteList and Black List URL so this is also Finite Set.

- K= {K1, K2, K3. . . . . .Kn | K is a Set of train network}

  This set is used for training the network. This is also an infinite Set.

- B= {B1, B2, B3. . . . . .Bn | Bn is a set for Recognize Network }

V. DATA FLOW AND ENTITY RELATIONSHIP

A data flow diagram (DFD) is a graphical representation of the flow of data through aninformation system, modeling its process aspects. It shows data is processed by a system interms of inputs and outputs.

DFD Level-0

It only contains one process node (Process 0) that generalizes the function of the entiresystem in relationship to external entities.

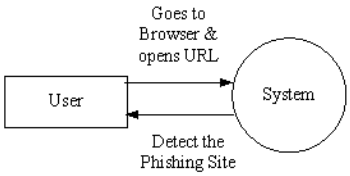

Figure 5.1: DFD Level-0

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

DFD Level-1

DFD level 1 diagram expands the DFD 0 and shows the detailed flow of the proposed system.

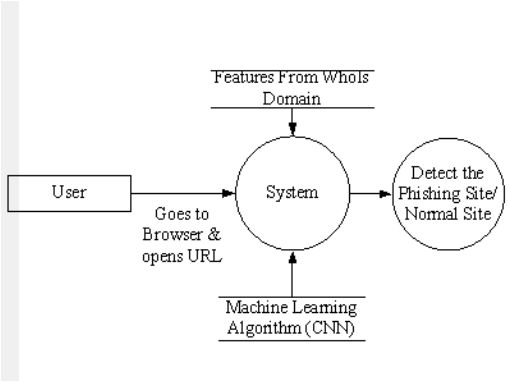

Figure 5.2: DFD Level-1

DFD Level-2

DFD level 2 diagram expands the DFD 1 and shows the detailed flow in the proposed system.It shows the different processes that take place to perform the authentication.

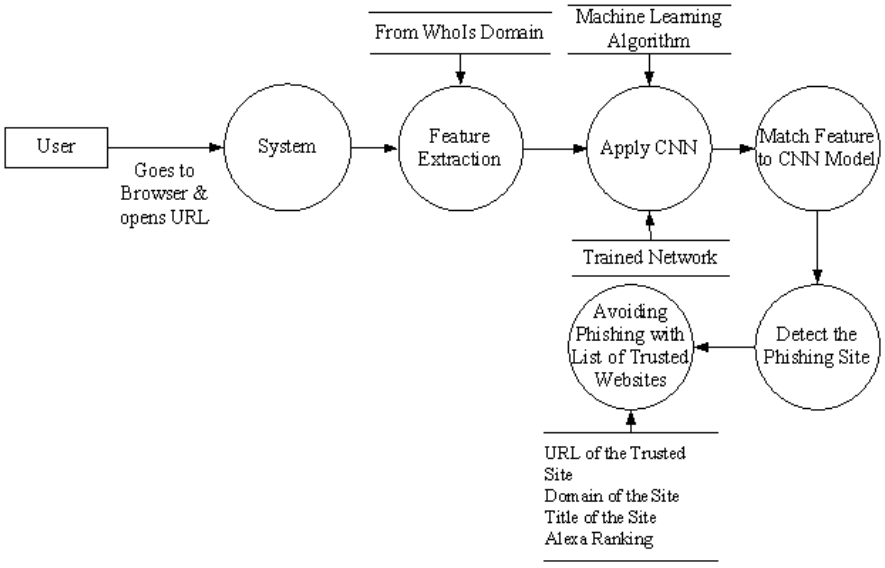

Figure 5.3: DFD Level-2

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

Entity Relationship And Sequence Of The Project

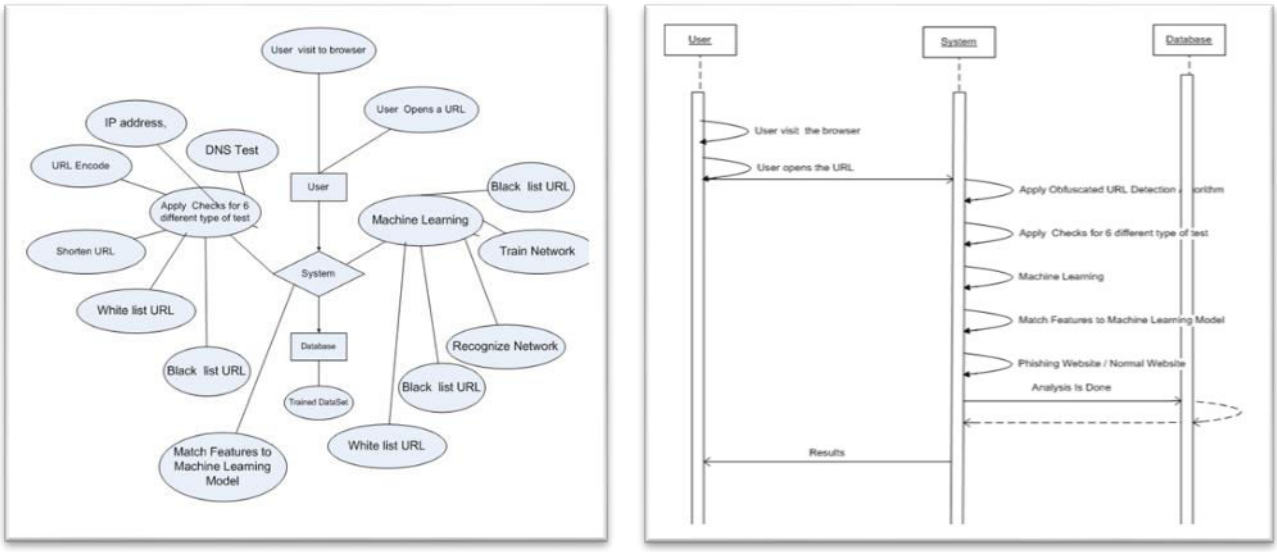

Figure 5.4: Entity Relationship and Sequence Diagram

VI. PROJECT IMPLEMENTATION AND RESULTS

The system's GUI was designed using java JSP. Core Technologies used were Java, JSP. Theoverall development was done in the Eclipse 3.3 Indigo and for DB we used MY SQL GUIbrowser.The database basically used for user storing user details like Username, user identity, the tool used for Database functionalities was MYSQL GUI Browser. Convolutional Neural Network Algorithm was implemented using Python3.

OVERVIEW OF PROJECT MODULES

We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and WhoIs features. The convolutionNeural Network is used to train the network and finally detect the site is Phishing or not.

The main modules involved in the systemare :

1. Whois feature extraction of the input website.
2. Training
3. Avoiding phishing attacks

OUTCOMES

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bankaccount detail, credit card number, financial username and password etc. and later it can bemisuse by attacker.We aim to use WhoIs features of URL as the basis of detecting phishing websites. We propose a novel solution, Phishing Detection using Soft Computing andMachine Learning, to efficiently detect phishing web pages using URL and WhoIs features.The convolution Neural Network is used to train the network and finally detect the site isPhishing or not.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

RESULT

Non Phishing Site:

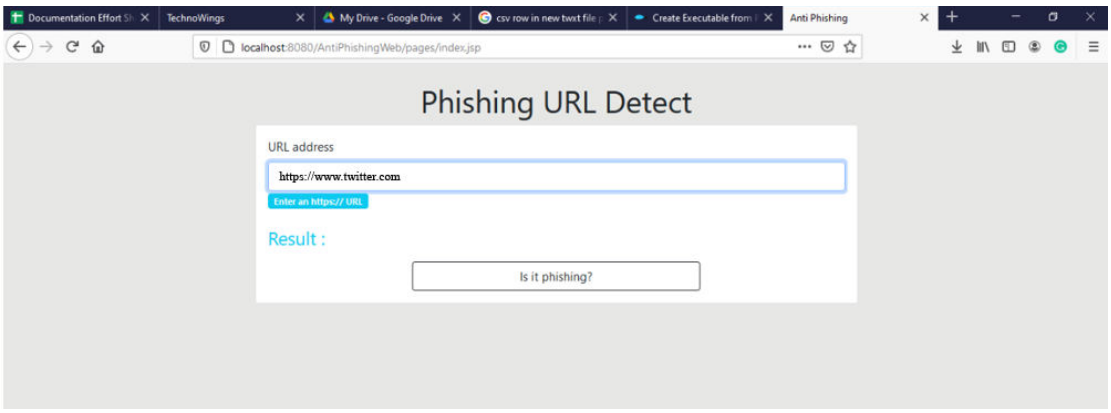

Figure 6.1: Input to check the URL address

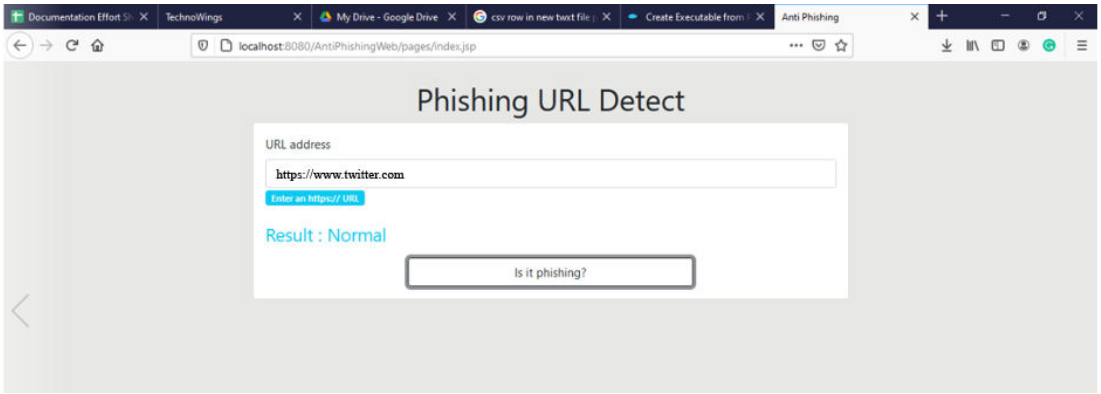

Figure 6.2: The result- Non phishing (Normal Site)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

Phishing Site:

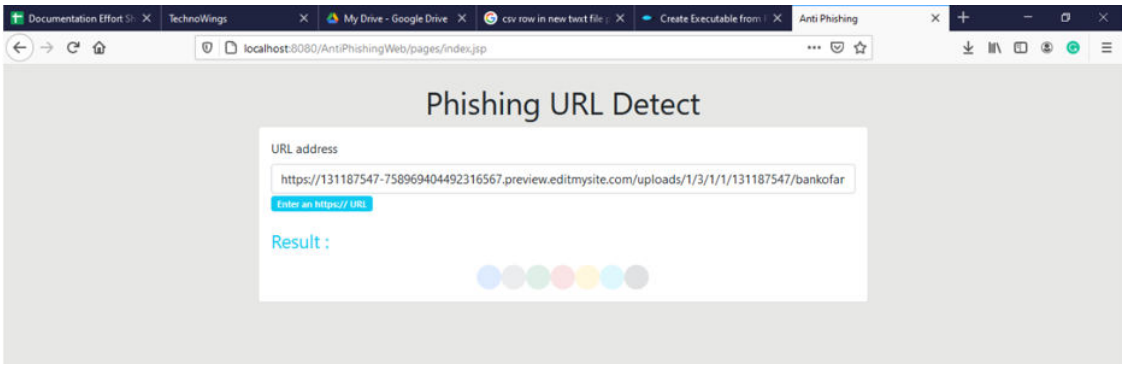

Figure 6.3: Input to check the URL address

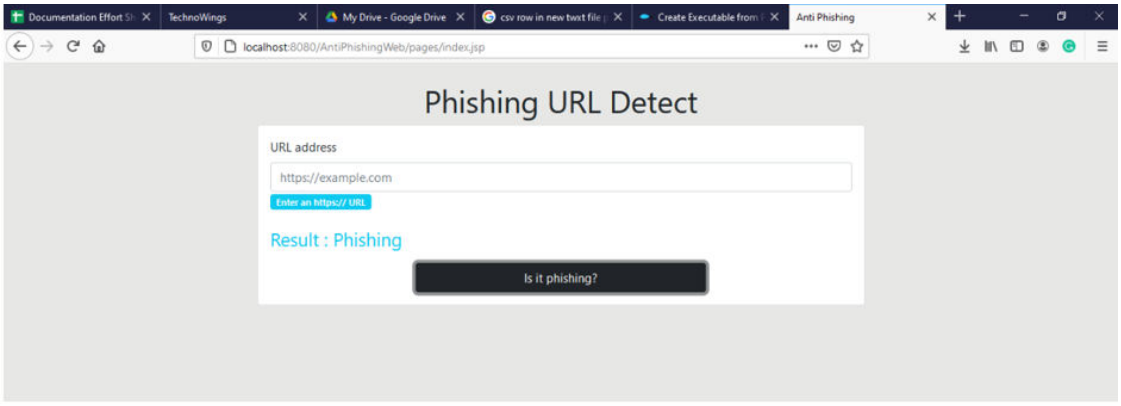

Figure 6.4: The result- Phishing (Abnormal Site)

VII. CONCLUSION AND FUTURE WORK

Conclusion:-

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud thatacquires victim's confidential information such as password, bank account detail, credit card number, financialusername and password etc. and later it can be misused by attacker.We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detectphishing web pages using URL and CSS features. Features are extracted for Blacklisted and white listed URL featuresare used as dataset for machine learning algorithms.

Future Work:-

The software designed can also be integrated within Software Application, Web application or Plugins of the WebBrowsers.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 5.928|

| Volume 4, Issue 7, July 2021 |

|DOI:10.15680/IJMRSET.2021.0407005|

REFERENCES

1. S. Jahan, "Human Resources Information System (HRIS): A Theoretical Perspective", Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129, 2014.\
2. Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber "Phishing Detection: A Recent Intelligent MachineLearning Comparison based on Models Content and Features" IEEE 2017.
3. Longfei Wu, Xiaojiang Du, and Jie Wu "MobiFish: A Lightweight Anti-Phishing Scheme for Mobile Phones"IEEE 2014.
4. LongfeiWu, Xiaojiang Du, and Jie Wu, "Effective Defense Schemes for Phishing Attacks on MobileComputing Platforms" IEEE 2015.
5. Guang-Gang Geng, Zhi-Wei Yan, Yu Zeng and Xiao-Bo Jin "RRPhish- Anti-Phishing via Mining BrandResources Request" 2018 IEEE International Conference on Consumer Electronics (ICCE).
6. SadiaAfroz and Rachel Greenstadt "PhishZoo: Detecting Phishing Websites By Looking at Them" IEEE2011.
7. Muhammet Baykara and Zahit Ziya Gurel "Detection of phishing attacks" IEEE 2018.
8. Mohammed NazimFeroz, SusanMengel "Phishing URL detection using URL Ranking" InternationalCongress on Big Data 2015 IEEE.
9. LuongAnh Tuan Nguyen†, Ba Lam To†, HuuKhuong Nguyen† and Minh Hoang Nguyen* † Faculty ofInformation Technology "Detecting Phishing Web sites: A Heuristic URL-Based Approach" InternationalConference on Advanced Technologies for Communications 2013.
10. JiHua 1,2, Zhang Huaxiang 1,2 "Analysis on the Content Features and Their Correlation of Web Pages forSpam Detection" IEEE 2015.
11. Samuel Marchal, Jer´ ome Franc ¸ois, Radu State, and Thomas Engel "PhishStorm: Detecting Phishing WithStreaming Analytics" IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, 2014.
12. LuongAnh Tuan Nguyen1, Ba Lam To2, HuuKhuong Nguyen1 and Minh Hoang Nguyen31 Faculty ofInformation Technology "A Novel Approach for Phishing Detection Using URL-Based Heuristic" IEEE 2014.
13. Jian Mao1,2, Pei Li 1, Kun Li1, Tao Wei3, and Zhenkai Liang4 "BaitAlarm: Detecting Phishing Sites UsingSimilarity in Fundamental Visual Features" 5th International Conference on Intelligent Networking andCollaborative Systems 2013.
14. T. G. Gregory Paul and T. Gireesh Kumar, A Framework for Dynamic Malware Analysis Based onBehaviorArtifacts. Singapore: Springer Singapore, 2017.
15. Mohammad, R., M., Thabtah, F., and McCluskey, L., 2014 Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications.
16. Kaveh, A,"Cuckoo search optimization," Advances in Metaheuristic Algorithms for Optimal Design ofStructures, 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 5.928

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

9710 583 466 9710 583 466 ijmrset@gmail.com

www.ijmrset.com