**PHISHING AWARE: SVM CLASSIFICATION FOR ANTI-PHISHING**

**A PROJECT REPORT**

*Submitted by*

**AKILA. R (810015104006)**

**NITHYA.R (810015104055)**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**



**UNIVERSITY COLLEGE OF ENGINEERING – BIT CAMPUS,**

**TIRUCHIRAPPALLI**

**ANNA UNIVERSITY::CHENNAI 600 025**

**APRIL 2019**

**UNIVERSITY COLLEGE OF ENGINEERING,**

**BIT CAMPUS,**

**TIRUCHIRAPPALLI-620 024**

**BONAFIDE CERTIFICATE**

Certified that this project report “**Phishing aware: SVM classification for anti-phishing on fog networks”** is the bonafide work of **“Ms. R. AKILA(810015104006)** and **Ms. R. NITHYA (810015104055) “** who carried out the project work under my supervision.

**SIGNATURE SIGNATURE**

**Dr. D. Venkatesan Mrs. M. Revathi**

HEAD OF THE DEPARTMENT SUPERVISOR

Assistant Professor Teaching Fellow

Computer Science & Engineering Computer Science &Engineering

University College of Engineering, University College of Engineering,

Anna University-BIT Campus, Anna University-BIT Campus,

Tiruchirappalli-620 024 Tiruchirappalli-620 024

Submitted for the project Viva voce examination held on ………………

**Internal Examiner External Examiner**

**DECLARATION**

We hereby declare the work entitled “**PHISHING AWARE:SVM CLASSIFICATION FOR ANTI-PHISHING ON FOG NETWORKS** ” is submitted in partial fulfillment of the requirement for the award of the degree in B.E., Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli, is a record of our own work carried out by us during the academic year 2018-2019 under the supervision and guidance of Mrs. M.Revathi, Teaching Fellow, Department of Computer Science and Engineering, University College of Engineering(BIT Campus), Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any degree, either in this or any other University.

SIGNATURE OF THE CANDIDATES

R. AKILA (810015104006)

R. NITHYA (810015104055)

I certify that the declaration made above by the candidate is true.

SIGNATURE OF THE GUIDE

Mrs. M. REVATHI

Teaching Fellow,

Department of CSE,

University College of Engineering,

BIT Campus, Anna University,

Tiruchirappalli-620 024.

**ACKNOWLEDGEMENT**

I would like to convey my heartfelt thanks to our honorable Dean **Dr. T. SENTHILKUMAR,** Associate Professor for having provided me with all required facilities to complete my project without hurdles.

I would like to express my sincere thanks and deep sense of gratitude to guide **Dr. D. VENKATESAN,** Assistant Professor and Head, Department of Information Technology, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of this project work.

I would like to thank my project guide **Mrs. M. REVATHI,** Teaching Fellow, Department of Computer Science and Engineering, for his valuable guidance throughout the phase of the project. It is our responsibility to thank our project coordinator **Mr. P. KARTHIKEYAN,Mr. C. SANKAR RAM,** Assistant Professor, Department of Computer science and Engineering for his constant inspiration that he has all through the project period.

I would like to thank **Mr. C. SURESH KUMAR,** Teaching Fellow, Department of Computer Science and Engineering, for his encouragement for this work.

I extend my thanks to all other teaching and non-teaching staffs for their encouragement and support.

I thank my beloved parents and friends, for their full support in my career development of this project.

**ABSTRACT**

Phishing detection is recognized as a criminal issue of Internet security. By deploying a gateway anti-phishing in the networks, these current hardware-based approaches provide an additional layer of defense against phishing attacks. However, such hardware devices are expensive and inefficient in operation due to the diversity of phishing attacks. With promising technologies of virtualization in fog networks, an anti-phishing gateway can be implemented as software at the edge of the network and embedded robust machine learning techniques for phishing detection. In this paper, we use uniform resource locator features and Web traffic features to detect phishing websites based on a designed SVM classification. Based on the new approach, fog computing as encouraged by Cisco, we design an anti-phishing model to transparently monitor and protect fog users from phishing attacks. The experiment results of our proposed approach, based on a large-scale dataset collected from real phishing cases, have shown that our system can effectively prevent phishing attacks and improve the security of the network

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER**  **NO.** | **TITLE**  **ABSTRACT**  **LIST OF FIGURES**  **LIST OF ABBREVIATIONS** | **PAGE NO**.  v  viii  ix |
| 1 | **INTRODUCTION** | 1 |
|  | * 1. DATA MINING | 1 |
|  | * 1. SIGNIFICANCE OF E-C0MMERCE | 1 |
|  | * 1. EMERGE OF ONLINE SHOPPING | 3 |
|  | * 1. REVIEWR AND RATINGS | 6 |
|  | 1.5 COLLABORATIVE FILLTERING | 8 |
|  | 1.6 ISSUE IN E-COMMERCE | 11 |
|  |  |  |
| 2 | **LITERATURE SURVEY** | 13 |
|  |  |  |
| 3 | **SYSTEM ANALYSIS** | 21 |
|  | 3.1 EXISTING SYSTEM | 21 |
|  | 3.1.1 Limitations | 21 |
|  | 3.3 PROPOSED SYSTEM | 22 |
|  |  |  |
| 4 | **SYSTEM SPECIFICATION** | 23 |
|  | 4.1 SOFTWARE REQUIREMENTS | 23 |
|  | 4.2 HARDWARE REQUIREMENTS | 23 |
|  | 4.3 ABOUT THE SOFTWARE | 23 |
|  |  |  |
| 5 | **SYSTEM DESIGN** | 27 |
|  | 5.1ARCHITECTURAL OVERVIEW | 27 |
|  | 5.2 MODULE DESCRIPTION | 28 |
|  | 5.2.1 Data Preparation | 29 |
|  | 5.2.2 Data cleaning | 29 |
|  | 5.2.3 Data Collection | 30 |
|  | 5.2.4 User Sentimental Measurement | 35 |
|  | 5.2.5Manipulated Review | 36 |
|  |  |  |
| 6 | **CONCLUSION AND FUTURE WORKS** | 40 |
|  | 6.1 CONCLUSION | 40 |
|  |  |  |
| **APPENDICES** |  |  |
|  | A1 SAMPLE SOURCE CODE | 41 |
|  | A2 SAMPLE SCREENSHOTS | 45 |
|  |  |  |
| **REFERENCES** |  | 49 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **FIGURE NAME** | **PAGE NO.** |
| 5.1 | Architecture of the proposed system | 27 |
|  | |  | | --- | |  | |  | |  | |  | |  |  |  |  | | --- | --- | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |

**LIST OF ABBREVIATIONS**

**API** Application programming interface

**URL** Uniform resource locator

**AP** Access point

**BS** Base station

**NFV** Network function virtualization

**SVM** Support vector machine

**CHAPTER 1**

**INTRODUCTION**

**1.1 DATA MINING**

Data mining is an interdisciplinary subfield of computer science.It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence,machine learning,statistics, and database systems.The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.Aside from the raw analysis step,it involves database and data management aspects,data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures,visualization, and updating.Data mining is the analysis step of the “knowledge discovery in databases”process, or KDD.

1. **Phishing Websites**

PHISHING is a criminal activity that steals victims’ personal information using misleading emails or fake websites . The word “phishing” is originated from the word “fishing” . Online users can be easily deceived into entering their personal information because phishing websites are highly similar to real ones. Maliciously, by creating phishing sites, “phishers” use a number of techniques to fool their Manuscript received December 29, 2016; revised May 18, 2017, September 19, 2017, and February 7, 2018; accepted April 17, 2018. Date of publication April 30, 2018; date of current version September 7, 2018. victims, including email messages, instant messages, forum posts, phone calls, and social networking information .Phishing results in severe economic loss all over the world, and phishing sites are also growing rapidly in quantity and complexity. According to reports from the Anti-Phishing Working Group , the number of phishing attacks is increasing by 5% monthly. The urgency and importance of phishing identification in modern society, which is based on a phishing website report received in the first quarter of 2016 . However, at the edge of networks, the anti-phishing problem has not been well-addressed due to the following reasons. First, mobile users check their emails and use Web browsers more frequently than desktop users . Thus, they are much more likely to access on phishing sites that have not yet been detected or taken down by anti-phishing applications and firewalls at their local networks or on their devices. Second, mobile devices are always “hungry” for energy and computing resources (e.g., limitations of CPU, memory, and user interfaces), so anti-phishing tools are usually ignored or removed on these devices. Hence, it is hard for users to discern if an incoming link is legitimate or not. Third, existing antiphishing tools (e.g., default plug-ins on Web browsers or local anti-phishing applications) are inefficient in terms of detection (this will banalyzed concretely later in Section III), and mobile users may be exposed to phishing attacks when engaging in usual behaviors. According to the report [5], mobile users are three times more likely to submit their login information than desktop users do. Therefore, preventing phishing attacks against terminal users is a critical issue in the edge of networks.

1. **Fog Computing Paradigm and Anti-Phishing**

Services on Fog Focusing on the edge of networks, Cisco recently introduces the concept of fog computing, which extends the cloud so that it is closer to users . Fog computing also provides data, computing, storage and application services to end-users as well as cloud computing. In addition, at the edge of networks, fog supports high mobility and a dense geographic distribution . Strong characteristics of fog networks now pull services provided at places near the end-users, such as access gateways, or even the set-top-box. Can we use the advantages of fog networks to develop antiphishing tools at the edge of networks, where a base station (BS) or an access point (AP) can support security services for mobile devices? It is possible with the help of network function virtualization (NFV) in the fog network, where a fog node now has sufficient resource capacity to virtualize any network function, such as a firewall, an anti-virus, and an anti-phishing function. Further, it is also more tractable and significant to embed machine learning techniques on fog nodes than on hardware-based devices. A fog-based service can enhance the performance compared to existing cloud based methods since fog nodes are located at the edge of networks. Last but not least, deployment of an antiphishing service on fog nodes does not degrade the computing resources of mobile devices as much compared to installing directly anti-phishing applications on these devices.

**CHAPTER 2**

**LITERATURE SURVEY**

**Phishing attacks. Accessed sept 2015. [Online]. Available: https://securityintelligence.com/**

Internet technology is so pervasive today, for example, from online social networking to online banking, it has made people’s lives more comfortable. Due the growth of Internet technology, security threats to systems and networks are relentlessly inventive. One such a serious threat is “phishing”, in which, attackers attempt to steal the user’s credentials using fake emails or websites or both. It is true that both industry and academia are working hard to develop solutions to combat against phishing threats. It is therefore very important that organisations to pay attention to end-user awareness in phishing threat prevention. Therefore, aim of our paper is twofold. First, we will discuss the history of phishing attacks and the attackers’ motivation in details. Then, we will provide taxonomy of various types of phishing attacks. Second, we will provide taxonomy of various solutions proposed in literature to protect users from phishing based on the attacks identified in our taxonomy. Moreover, we have also discussed impact of phishing attacks in Internet of Things (IoTs). We conclude our paper discussing various issues and challenges that still exist in the literature, which are important to fight against with phishing threats.

**L. Wenyin,G. Liu,B. Qiu, and X. Quan, “Antiphishing through phishing target discovery”, IEEE Internet Computing, vol. 16,no. 2,pp. 52-61, March 2012.**

Phishing attacks are growing in both volume and sophistication. The antiphishing method described here collects webpages with either a direct or indirect association with a given suspicious webpage. This enables the discovery of a webpage's so-called "parasitic" community and then ultimately its phishing target — that is, the page with the strongest parasitic relationship to the suspicious webpage. Finding this target lets users determine whether the given webpage is a phishing page.

**Y. Zhang, J. I. Hong, and L. F. Cranor, “CANTINA: A content-based approach to detecting phishing Web sites,” in Proc. ACM 16th Int. Conf.World Wide Web, Banff, AB, Canada, 2007, pp. 639–648.** Phishing is a significant problem involving fraudulent email and web sites that trick unsuspecting users into revealing private information. In this paper, we present the design, implementation, and evaluation of CANTINA, a novel, content-based approach to detecting phishing web sites, based on the TF-IDF information retrieval algorithm. We also discuss the design and evaluation of several heuristics we developed to reduce false positives. Our experiments show that CANTINA is good at detecting phishing sites, correctly labeling approximately 95% of phishing sites.

**B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, “Fighting againstphishing attacks: State of the art and future challenges,” Neural Comput.Appl., vol. 28, no. 12, pp. 3629–3654, Dec. 2017.**

In the last few years, phishing scams have rapidly grown posing huge threat to global Internet security. Today, phishing attack is one of the most common and serious threats over Internet where cyber attackers try to steal user’s personal or financial credentials by using either malwares or social engineering. Detection of phishing attacks with high accuracy has always been an issue of great interest. Recent developments in phishing detection techniques have led to various new techniques, specially designed for phishing detection where accuracy is extremely important. Phishing problem is widely present as there are several ways to carry out such an attack, which implies that one solution is not adequate to address it. Two main issues are addressed in our paper. First, we discuss in detail phishing attacks, history of phishing attacks and motivation of attacker behind performing this attack. In addition, we also provide taxonomy of various types of phishing attacks. Second, we provide taxonomy of various solutions proposed in the literature to detect and defend from phishing attacks. In addition, we also discuss various issues and challenges faced in dealing with phishing attacks and spear phishing and how phishing is now targeting the emerging domain of IoT. We discuss various tools and datasets that are used by the researchers for the evaluation of their approaches. This provides better understanding of the problem, current solution space and future research scope to efficiently deal with such attacks.

**Real Time Anti-Phishing. Accessed: Sep. 2016. [Online]. Available:**

**http://www.brightcloud.com/services/real-time-anti-phishing.php** In a recent Webroot survey of IT Directors conducted in the USA and the UK, 55% of respondents reported the number one Internet security breach was phishing1. Phishing and spear phishing are now targeted at businesses of all sizes and are rapidly becoming the preferred way cybercriminals breach networks. Phishing attacks are now so sophisticated, they often fool even IT Security professionals. In a recent RSA study, the estimated cost to the global economy in fraud damages from phishing attacks increased to $1.5 billion (22%) between 2011 and 2012.In the face of this growing threat, you can fight back. The new Webroot Real-Time Anti-Phishing Service protects users by ensuring they visit only the websites they intend to interact with, not a phishing site.

**Anti Spam Hardware. Accessed: Sep. 2016. [Online]. Available:http://www.windowsnetworking.com/hardware/Anti-Spam-Hardware/**  Internet technology is so pervasive today, for example, from online social networking to online banking, it has made people’s lives more comfortable. Due the growth of Internet technology, security threats to systems and networks are relentlessly inventive. One such a serious threat is “phishing”, in which, attackers attempt to steal the user’s credentials using fake emails or websites or both. It is true that both industry and academia are working hard to develop solutions to combat against phishing threats. It is therefore very important that organisations to pay attention to end-user awareness in phishing threat prevention. Therefore, aim of our paper is twofold. First, we will discuss the history of phishing attacks and the attackers’ motivation in details. Then, we will provide taxonomy of various types of phishing attacks. Second, we will provide taxonomy of various solutions proposed in literature to protect users from phishing based on the attacks identified in our taxonomy. Moreover, we have also discussed impact of phishing attacks in Internet of Things (IoTs). We conclude our paper discussing various issues and challenges that still exist in the literature, which are important to fight against with phishing threats.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

The security issues regarding the fog network to enhance network safety. In particular, we study the phishing website problem and propose an identification architecture on the fog network. Based on the advantages of the fog architecture and the neuro-fuzzy approach, we propose a phishing identification model, called Fi-NFN, to protect local devices easily and quickly. Without consuming many resources from local devices, our Fi-NFN model not only transparently protects users in real time, but also improves the quality of services at the edge of the network. Without using an inefficient blacklist method, we design a five-layer neuro-fuzzy network with six heuristic input values (PrimaryDomain, SubDomain, PathDomain, PageRank, GoogleIndex and Alexareputation). Our simulation results indicate that the efficiency of phishing identification after training with the training dataset by improving the average accuracy to 98.36% and reducing the missed detection and false alarm rates to 0.9% and 0.74%, respectively. We also compare our approach with current methods [17], [20], and [24] to evaluate our model. Simulation results show that our method is more efficient, stable and accurate. Especially, various testing results indicate that our model in a fog computing environment is not only possible, but also can be applied practically.

**3.2 PROPOSED SYSTEM**

* The concept is an end host anti-phishing algorithm called the SVM ,by utilizing the generic characteristics of the web links.
* This algorithm is finding the phishing link provide by the attackers to grasp the information of the end user.
* SVM is used to detect and prevent not only known phishing attacks but also unknown ones.
* Classifier determines whether a requested site is a phishing site.
* When a page request occurs, the URL of the requested site is transmitted to the feature extractor, which extracts the feature values through the predefined URL-based features.
* The classifier determines whether a new site is a phishing site based on learned information then alerts the page-requesting user about the classification result.
* The collected URLs are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features.
* The extracted features are stored as input and passed to the classifier generator, which generates a classifier by using the input features and the machine learning algorithm.
* The phishing site based on learned information then alerts the page-requesting user about the classification result.
* The result is predicting whether Phishing or Legitimate sites.

**CHAPTER 4**

**SYSTEM SPECIFICATION**

**4.1 SOFTWARE REQUIREMENTS**

Operating System : Windows OS

Front End : PHP

IDE : Macromedia Dreamweaver 8

Back End : MYSQL SERVER

**4.2 HARDWARE REQUIREMENTS**

Processor : Dual core processor 2.6.0 GHZ

RAM : 1GB

Hard disk : 160 GB

Compact Disk : 650 Mb

Keyboard : Standard keyboard

Monitor : 15 inch color monitor

#### 4.3 ABOUT THE SOFTWARE

**WAMPSERVER**

WAMP server stands for “Windows, Apache, My SQL, and PHP”. WAMP is a variation of LAMP for windows systems and is often installed as a software bundle (Apache, My SQL and PHP).It is often used for web development and internal testing, but may also be used to serve live websites. The most important part of the WAMP package is Apache which is used run the web server within windows. By running a local Apache web server on the windows machine, a web developer can test web pages in a web browser without publishing them live on the internet.

WAMP also includes My SQL and PHP which are two of the most common technologies used for creating dynamic websites.

**Macromedia Dreamweaver**

Macromedia Dreamweaver is a professional HTML editor for designing and managing websites and pages. It is also called WYSIWYG=What You See Is What You Get editor. This means that it enables direct visual edition of web pages. That features permits edition of web pages without extensive knowledge of the HTML web code language. The code is automatically written by the editor. However it remains complicated to use for anyone who isn’t familiar with HTML.

**PHP**

Hypertext Preprocessor is a general purpose programming language originally designed for web development. It was originally created by Rasmus Lerdorf in 1994. PHP originally stood for personal home page, but it now stands for the recursive initialism.

**Functions**

PHP defines a large array of functions in the core language and many are also available in various extensions; this functions are well documented in the online PHP documentation. However, the build in libraury has a wide variety of naming conventions and associated in consistencie.

**PHP Objects**

Basic Object Oriented Programming functionality was added in PHP 3 and improved in PHP 4. This allowed for PHP to gain further abstraction, making creative task easier for programmers using the language. Object handling was completely rewritten for PHP 5, expanding the feature set and enhancing performance. In previous versions of PHP, Objects were handled like value types. The drawback of this method was that code had to make heavy use of PHP’s ”Reference “ variables if it wanted to modify an object it was passed rather than creating a copy of it. In the new approaches, Objects are referenced by handle, and not by value.

**Security**

In 2017, 3% of all vulnerabilities listed by the National Vulnerabilities Database were linked to PHP, historically, about 30% of all vulnerabilities listed since 1996 in this database are linked to PHP. Technical security flaws of the language itself or of its core libraries are not frequent.

Recognizing that programmers make mistakes, some languages include taint checking to automatically detect the lack of input validation which intuces many issues. Such a feature is being developed for PHP, but its inclusion into a release has been rejected several times in the past. There are advanced production batches such as Suhosin and Hardening-patch, especially designed for web hosting environments.

**Use**

PHP is a general purpose scripting language that is especially suited to server side web development, in which case PHP generally runs on a web server.It can also be used for command-line scripting and client side graphical user interface applications.PHP can be deployed on most web servers, many operating systems and platforms and can be used with many relational database management systems.

**CHAPTER 5**

**SYSTEM DESIGN**

**5.1 ARCHITECTURAL OVERVIEW**

The proposed work of this approach is predicting whether Phishing or Legitimate sites. The following figure 5.1 explains the overall architecture of the proposed work.

**Figure: 5.1 Architectural Diagram**

The above Figure 5.1 shows the implemental procedure of the proposed work. We collect the URL’s from different online websites.

**The Procedure is as follows**

**Step1:** Get set of URL’s for banking from different websites.

**Step2:** Acquisition of data sets.

**Step3:** Preprocessing the websites.

**Step4:** URLs are transmitted to feature extractor .

**Step5:** Classifier determines a new site is a phishing site.

**Step6:** Finally predicting whether Phishing or legitimates websites.

**5.2 MODULE DESCRIPTION**

* Data set Acquisition
* Preprocessing
* Feature Selection
* SVM Classification
* Phishing Website Prediction

**5.2.1. IDENTIFYING PHISHING SITES AND THE SYSTEM MODEL ON FOG NETWORKS**

To easily understand how to identify phishing sites, in this section, we briefly discuss some background knowledge related to phishing identification and illustrate our proposed anti-phishing model in fog.A. Preliminaries

1) Phishing Identification Tools: Following the approach using URL and Web traffic features, we show some famous techniques used to detect phishing websites and related to our work.• WHOIS: The WHOIS function provides details about the date of registration, update and expiration, the registrar . Phishing sites are often unstable, and their registration dates are often newer than those of the legitimate sites. Moreover, many phishing sites contain IP addresses in their URLs . Therefore, WHOIS is a helpful tool for detecting phishing sites.• DNS Blacklist: There are many blacklist providers that contain a list of phishing sites. These providers frequently update their database and support query methods for users, for example SORBS , URIBL , and SURBL .

• Browser toolbars: Browser toolbars provide a client-side defense for user browsers. Whenever a user visits a website, the browser toolbar will filter URLs from the address bar, then refer to a blacklist database. If the URL exists on that database, a special warning will be responsed to the user. Google Toolbar [20] is a popular tool integrated as a Firefox browser extension. In addition, there are several safe browsing tool bars that work with Chrome, Safari, and Internet Explorer .• Network Appliance: To combat phishing and other Internet attacks, a network appliance, such as a firewall or a gateway in a network, can be implemented.Trend Micro and Symantec have developed joint Internet access and security solutions as a safeguard . Such hardware tools often need to refer local or online URL blacklist databases. They can carry out well a small network; however they lack of flexibility in update that needs to adapt to diversity of phishing websites. Furthermore, some middle-layer defense models and third-party models are proposed in practice, even though they still raise much discomfort level in mobile users as they require many communication steps to protect mobile users.In the next part, we illustrate the identification features and discuss how to use them in phishing detection.2) Identification Features: Phishers usually try to make the Internet addresses (URLs) of phishing sites similar to legitimate sites to fool online users. However, they cannot reuse URLs of legitimate sites that are already registered. Based on various characteristics of URLs, we indicate the differences between a legitimate URL and a phishing URL.• Features of URL: The structure of URL is as follows:<protocol> :// <SubDomain>.<PrimaryDomain. <TLD>/<PathDomain>. For example, the URL: http://paypal.abc.net/ index.htm includes the following six elements: the protocol is http, the SubDomain is paypal, the PrimaryDomain is abc, the top-level domain (TLD) is net, the Domain is abc.net, and the PathDomain is index.htm.There exist many differences between phishing URLs and legitimate URLs that can be used to recognize easily based on URL features.In particular, we describe in detail three features: the PrimaryDomain, SubDomain and PathDomain of the URL. **IDENTIFYING PHISHING SITES AND THE SYSTEM**

**MODEL ON FOG NETWORKS**

To easily understand how to identify phishing sites, in this section, we briefly discuss some background knowledge related to phishing identification and illustrate our proposed anti-phishing model in fog.A. Preliminaries

1) Phishing Identification Tools: Following the approach using URL and Web traffic features, we show some famous techniques used to detect phishing websites and related to our work.

• WHOIS: The WHOIS function provides details about the date of registration, update and expiration, the registrar [34]. Phishing sites are often unstable, and their registration dates are often newer than those of the legitimate sites. Moreover, many phishing sites contain IP addresses in their URLs [35]. Therefore, WHOIS is a helpful tool for detecting phishing sites.

• DNS Blacklist: There are many blacklist providers that contain a list of phishing sites. These providers frequently update their database and support query methods for users, for example SORBS [36], URIBL [37], and SURBL [38].

• Browser toolbars: Browser toolbars provide a client-sidedefense for user browsers [23]. Whenever a user visits a website, the browser toolbar will filter URLs from the address bar, then refer to a blacklist database. If the URL exists on that database, a special warning will be responsed to the user. Google Toolbar [20] is a popular tool integrated as a Firefox browser extension. In addition, there are several safe browsing tool bars that work with Chrome, Safari, and Internet Explorer [20].

• Network Appliance: To combat phishing and other Internet attacks, a network appliance, such as a firewall or a gateway in a network, can be implemented.Trend Micro [19] and Symantec [39] have developed joint Internet access and security solutions as a safeguard [32]. Such hardware tools often need to refer local or online URL blacklist databases. They can carry out well a small network; however they lack of flexibilityin update that needs to adapt to diversity of phishing websites. Furthermore, some middle-layer defense models and third-party models are proposed in practice [40], even though they still raise much discomfort level in mobile users as they require many communication steps to protect mobile users.In the next part, we illustrate the identification features anddiscuss how to use them in phishing detection.

2) Identification Features: Phishers usually try to make theInternet addresses (URLs) of phishing sites similar to legitimate sites to fool online users. However, they cannot reuse URLs of legitimate sites that are already registered. Based on various characteristics of URLs, we indicate the differences between a legitimate URL and a phishing URL.

• Features of URL: The structure of URL is as follows:<protocol> :// <SubDomain>.<PrimaryDomain>. <TLD>/<PathDomain>.

For example, the URL:http://paypal.abc.net/ index.htm includes the following six elements: the protocol is http, the SubDomain is paypal, the PrimaryDomain is abc, the top-level domain (TLD) is net, the Domain is abc.net, and the PathDomain is index.htm. There exist many differences between phishing URLs and legitimate URLs that can be used to recognize easily based on URL features. In particular, we describe in detail three features: the PrimaryDomain, SubDomain and PathDomain of the URL.

**SIMULATION AND NUMERICAL RESULTS**In this section, we discuss our settings and datasets for the evaluation of our model. We implemented and performed sideby-side comparisons with prior works. Furthermore, we have built a test-bed as a fog network to validate the performanceof our proposed algorithm.

**A. Settings**We collected 11,660 URLs for phishing sites from PhishTank [7] and 10,000 URLs for legitimate sites from DMOZ [54] to make the training and testing datasets in both traditional and fog networks. We then randomly divide this dataset into the training dataset and the testing dataset following the ratio 75:25.

In order to set values for the parameters bi, we derive the proper settings of parameters (shown in Table III) based on the observation in the whole dataset. We also present the observation results TableII)to illustrate how we set thevalues for bi.Other parameters are shown in Table IV. We divide the testing dataset into three URL sets, including the long URL set (i.e., URLs with full features), the short URL set (i.e., URLs that lack features), and the random URL set. Finally, the **weights (w1, w2) are initialized randomly from −0.5 to 0.5.**

**B**.**Results**We conduct the simulation using the above datasets and settings. The convergence, accuracy of phishing identification, and response time are recorded as outputs of our simulation.We compare the performance of Fi-NFN to the current benchmark approaches, such as Fuzzy [24], Neural network [17], Google PageRank [42], eMCAC [22] and FACA [21]. First, we provide a brief outline of those methods that we compareFi-NFN with:

• Fuzzy: We compare Fi-NFN with an online algorithm that classifies URLs using the fuzzy approach. Essentially, Fuzzy is built by a rule set based on the URL characteristics [24].

• Neural network: Neural network is an approach proposed by [17] using the neural network model to identify phishing URLs. We implement a three-layer model and use our dataset to train this neural network.

**TRAINING DURATION**• Google: This is the popular tool of Google [20], which can be easily installed on Web browsers. Google Toolbar can detect phishing terms based on input keywords. In this work, to evaluate the performance of our approach, we develop an application that calls the Google API to detect phishing URLs at fog nodes instead of installing them on user devices.

• eMCAC [22] and FACA [21]: These are new approaches based on the rule set method to detect phishing URLs.

Similar to the Google API, we implement eMCAC and FACA on fog nodes for detection.

**Evaluation in a Traditional Environment (Convergence):**

To evaluate the convergence of Fi-NFN, we run with different learning rates from 0.1 to 0.9. Corresponding to our setting, we observe that our model obtained the most rapid convergence and the lowest error with the learning rate R = 0.7. Furthermore, we evaluate the duration of the training phase for different sizes of the dataset.

Accuracy: We first evaluate the impact of various learning rates on Fi-NFN. where we use three types of URLs. Our model achieves the best average accuracy when the learning rate R is set by 0.7.

**CHAPTER 6**

**CONCLUSION**

In this paper, we consider the security issues regarding the fog network to enhance network safety. In particular, we study the phishing website problem and propose an identification architecture on the fog network. Based on the advantages of the fog architecture and the SVM classification approach, we propose a phishing identification model, to protect local devices easily and quickly. Without consuming many resources from local devices, our model not only transparently protects users in real time, but also improves the quality of services at the edge of the network. Our simulation results indicate that the efficiency of phishing identification after training with the training dataset by improving the average accuracy to 98.36% and reducing the missed detection and false alarm rates to 0.9% and 0.74%, respectively. We also compare our approach with current methods to evaluate our model. Simulation results show that our method is more efficient, stable and accurate. Especially, various testing results indicate that our model in a fog computing environment is not only possible, but also can be applied practically.

**APPENDIX-I**

**SAMPLE SOURCE CODE**

<?php

include("include/protect.php");

include("include/dbconnect.php");

extract($\_POST);

$rdate=date("d-m-Y");

$fname="dataset.xlsx";

/\*function isDomainAvailible($domain)

{

//check, if a valid url is provided

if(!filter\_var($domain, FILTER\_VALIDATE\_URL))

{

return false;

}

//initialize curl

$curlInit = curl\_init($domain);

curl\_setopt($curlInit,CURLOPT\_CONNECTTIMEOUT,10);

curl\_setopt($curlInit,CURLOPT\_HEADER,true);

curl\_setopt($curlInit,CURLOPT\_NOBODY,true);

curl\_setopt($curlInit,CURLOPT\_RETURNTRANSFER,true);

//get answer

$response = curl\_exec($curlInit);

curl\_close($curlInit);

if ($response) return true;

return false;

}\*/

if(isset($btn))

{

//$q3=mysql\_query("select \* from payment where block\_st=0");

//while($r3=mysql\_fetch\_array($q3))

//{

//echo $url=$r3['web\_link'];

// if (isDomainAvailible($url))

// {

// mysql\_query("update payment set block\_st=1 where id=".$r3['id']."");

// }

// else

// {

// mysql\_query("update payment set block\_st=2 where id=".$r3['id']."");

// }

//}

?>

<script language="javascript">

alert("Blocked Successfully");

window.location.href="view\_block.php";

</script>

<?php

}

?>

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">

<head>

<title><?php include("include/title.php"); ?></title>

<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />

<link href="style.css" rel="stylesheet" type="text/css" />

<!-- CuFon: Enables smooth pretty custom font rendering. 100% SEO friendly. To disable, remove this section -->

<script type="text/javascript" src="js/cufon-yui.js"></script>

<script type="text/javascript" src="js/arial.js"></script>

<script type="text/javascript" src="js/cuf\_run.js"></script>

<!-- CuFon ends -->

</head>

<body>

<div class="main">

<div class="header">

<div class="header\_resize">

<div class="logo">

<?php include("include/header.php"); ?>

</div>

<div class="menu\_nav">

<ul>

<li class="active"><a href="upload.php">Home</a></li>

<li><a href="logout.php">Logout</a></li>

</ul>

</div>

<div class="clr"></div>

</div>

</div>

<div class="content">

<div class="content\_resize">

<div class="mainbar">

<div class="clr"></div>

<form action="" method="post" name="form1" id="form1">

<h2> Data </h2>

<table width="578" border="1">

<tr>

<th scope="col">Sno</th>

<th scope="col">Protocol</th>

<th scope="col">Domain</th>

<th scope="col">Path</th>

<th scope="col">Subdomain Item1 </th>

<th scope="col">Subdomain Item2 </th>

<th scope="col">TLD</th>

<th scope="col">Words</th>

</tr>

<?php

$i=0;

$q1=mysql\_query("select \* from payment group by payment");

while($r1=mysql\_fetch\_array($q1))

{

$q11=mysql\_query("select \* from payment where payment='".$r1['payment']."'");

while($r11=mysql\_fetch\_array($q11))

{

$i++;

$aa=explode("/",$r11['web\_link']);

$bb=explode(" ",$r11['payment']);

$cc=explode(".",$aa[2]);

if($bb[1]=="of" || $bb[1]=="off" || $bb[1]=="&")

{

}

else

{

$x=$bb[1];

}

if($bb[2]=="of" || $bb[2]=="off" || $bb[2]=="&")

{

}

else

{

$y=$bb[2];

}

if($bb[3]=="of" || $bb[3]=="off" || $bb[3]=="&")

{

}

else

{

$z=$bb[3];

}

/////////////////////////////

$e1=explode(".",$aa[2]);

$e1\_cn=count($e1);

$tld=$e1[$e1\_cn-1];

if(count($e1)>=5)

{

$ph1="1";

}

else

{

$ph1="0";

}

/////////////////////////////

if($aa[0]=="https")

{

$ph2="1";

}

else

{

$ph2="0";

}

////////////////////////////

$jj=0;

$arr\_tld=array("ae","ai","al","an","ao","bb","bd","bf","bv","cd","cf","cg","ci","ck","cm","cu","cv","cw","cy","dj","dm","do","dz","ec","eg","er","et","fj","fk","fm","ga","gb","ge","mu","ed");

for($j=0;$j<count($arr\_tld);$j++)

{

if($tld==$arr\_tld[$j])

{

$jj++;

}

}

if($jj>0)

{

$ph3="1";

}

else

{

$ph3="0";

}

////////////////////////////

$hi=explode("-",$e1[1]);

if(count($hi)>1)

{

$ph4="1";

}

else

{

$ph4="0";

}

/////////////////////////

$at=explode("@",$e1[1]);

if(count($at)>1)

{

$ph5="1";

}

else

{

$ph5="0";

}

//////////////////////////////////

$vi=$r11['visitor'];

if($vi<1000)

{

$ph6="1";

}

else

{

$ph6="0";

}

/////////////////////////////////

$po=$r11['portno'];

if($po>65535)

{

$ph7="1";

}

else

{

$ph7="0";

}

/////////////////////////////////

$ipp=$r11['ipaddress'];

$ipp1=explode(".",$ipp);

if($ipp1[0]<256 && $ipp1[1]<256)

{

$ph8="1";

}

else

{

$ph8="0";

}

/////////////////////////////////

$ti=$r11['ttl'];

if($ti>=3600)

{

$ph9="1";

}

else

{

$ph9="0";

}

/////////////////////////////////

$wi=$r11['web\_list'];

if($wi=="yes")

{

$ph10="1";

}

else

{

$ph10="0";

}

/////////////////////////////////

$tot=$ph1+$ph2+$ph3+$ph4+$ph5+$ph6+$ph7+$ph8+$ph9+$ph10;

if($tot>4)

{

$block\_st="1";

}

else

{

if($ph2=="1" || $ph4=="1" || $ph5=="1")

{

$block\_st="1";

}

else

{

$block\_st="0";

}

}

mysql\_query("update payment set block\_st=$block\_st where id=".$r11['id']."");

?>

<tr>

<th scope="row"><?php echo $i; ?></th>

<td><?php echo $aa[0]; ?></td>

<td><?php echo $e1[1]; ?></td>

<td><?php echo $aa[3]."/".$aa[4]; ?></td>

<td><?php echo $bb[0]; ?></td>

<td><?php echo $aa[5]; ?></td>

<td><?php echo $tld; ?></td>

<td><?php echo $cc[1];

if($x!="") { echo ",".$x; }

if($y!="") { echo ",".$y; }

if($z!="") { echo ",".$z; }

?></td>

</tr>

<?php

}

}

?>

</table>

<p align="center" class="msg">

<input type="submit" name="btn" value="Phishing Website" />

</p>

<p align="center">&nbsp;</p>

</form>

<p>&nbsp; </p>

<div class="clr"></div>

</div>

<div class="clr"></div>

</div>

</div>

<div class="fbg">

<div class="fbg\_resize">

<div class="clr"></div>

</div>

</div>

<div class="footer">

<div class="footer\_resize">

<!--<ul class="fmenu">

<li class="active"><a href="index.html">Home</a></li>

<li><a href="contact.html">Contacts</a></li>

</ul>-->

<?php include("include/footer.php"); ?>

<div class="clr"></div>

</div>

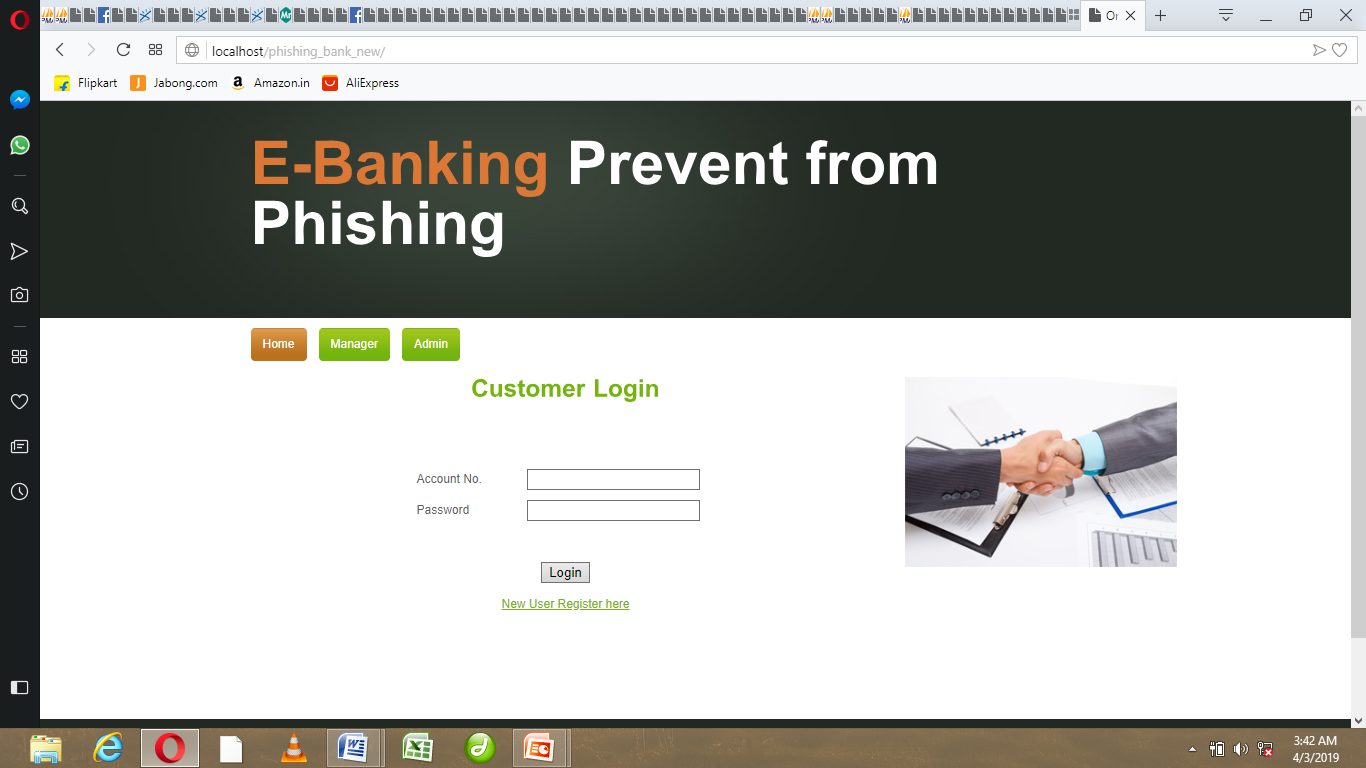
</div>

</body>

</html>

**APPENDIX-II**

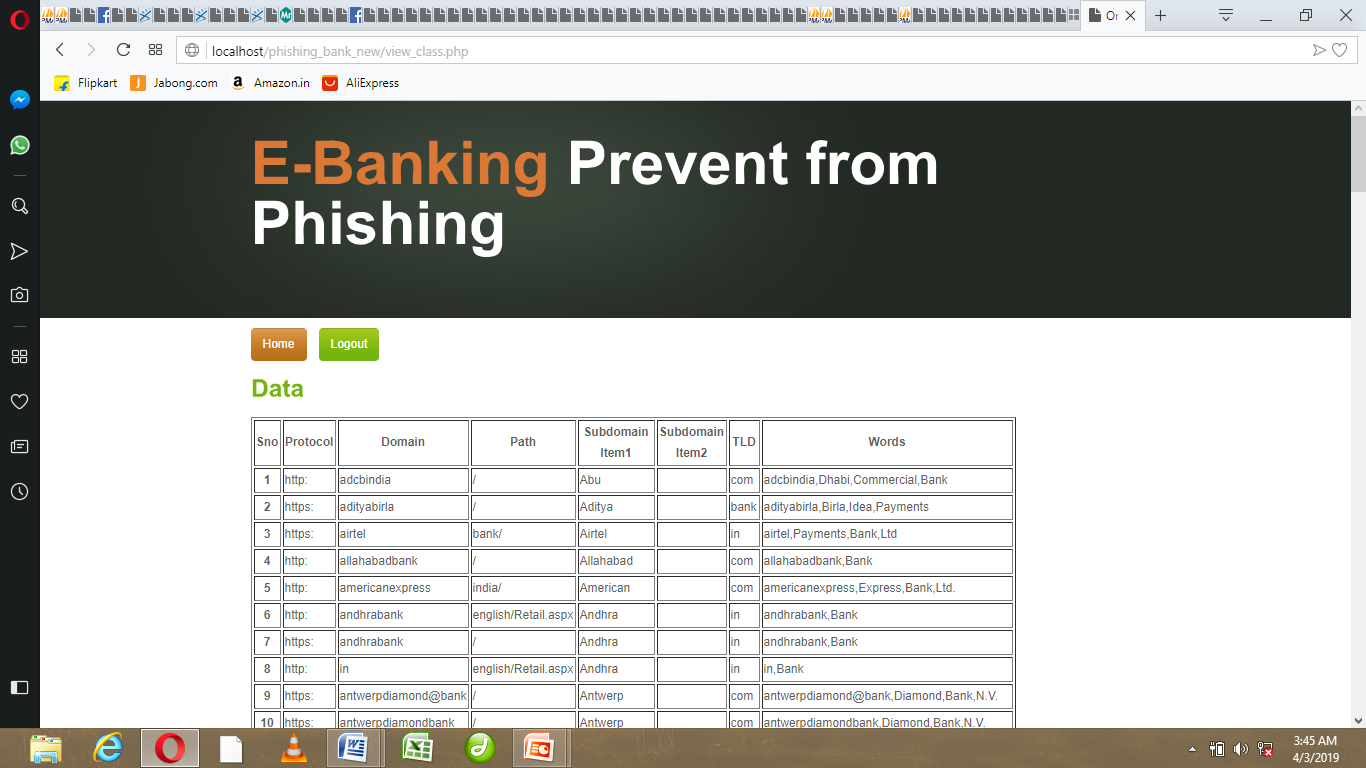
**SAMPLE SCREENSHOTS**



**S1:Homepage**

****

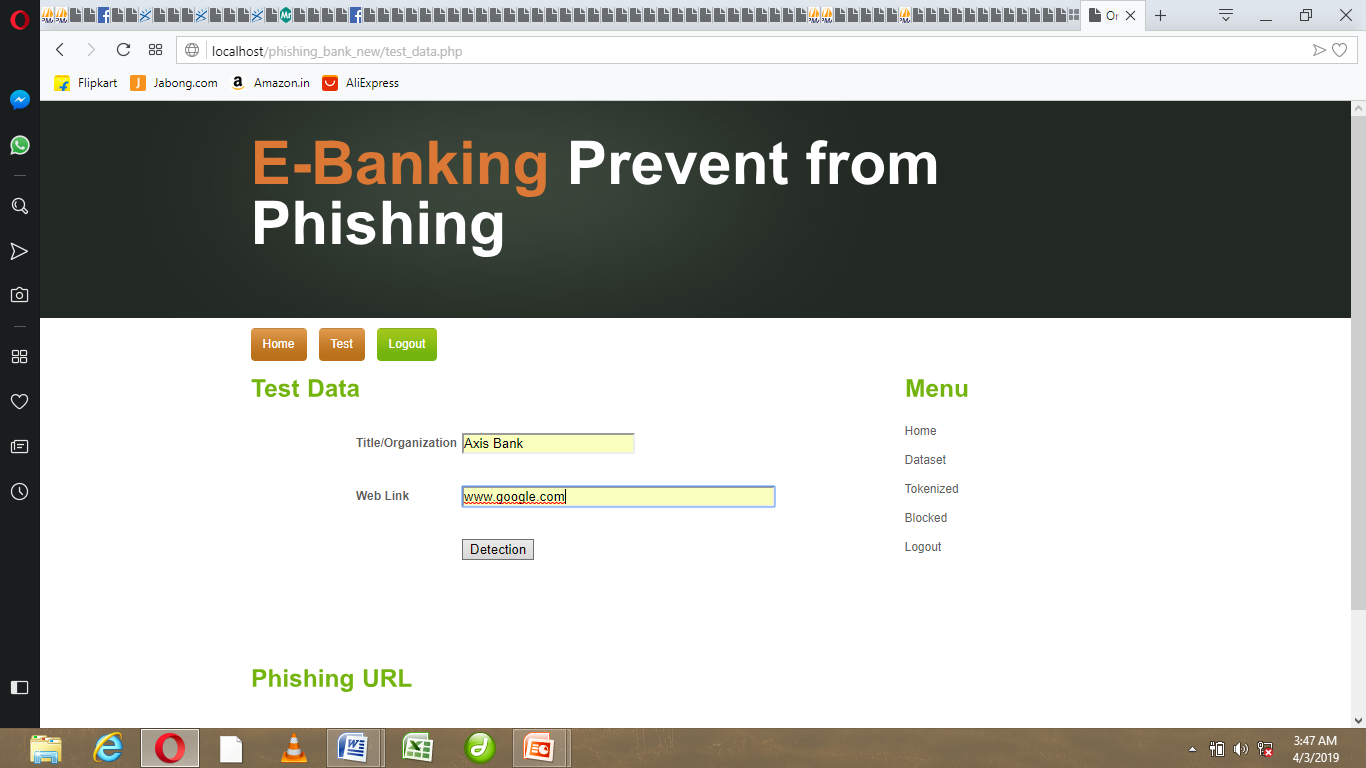
**S2: Administrative Login**

****

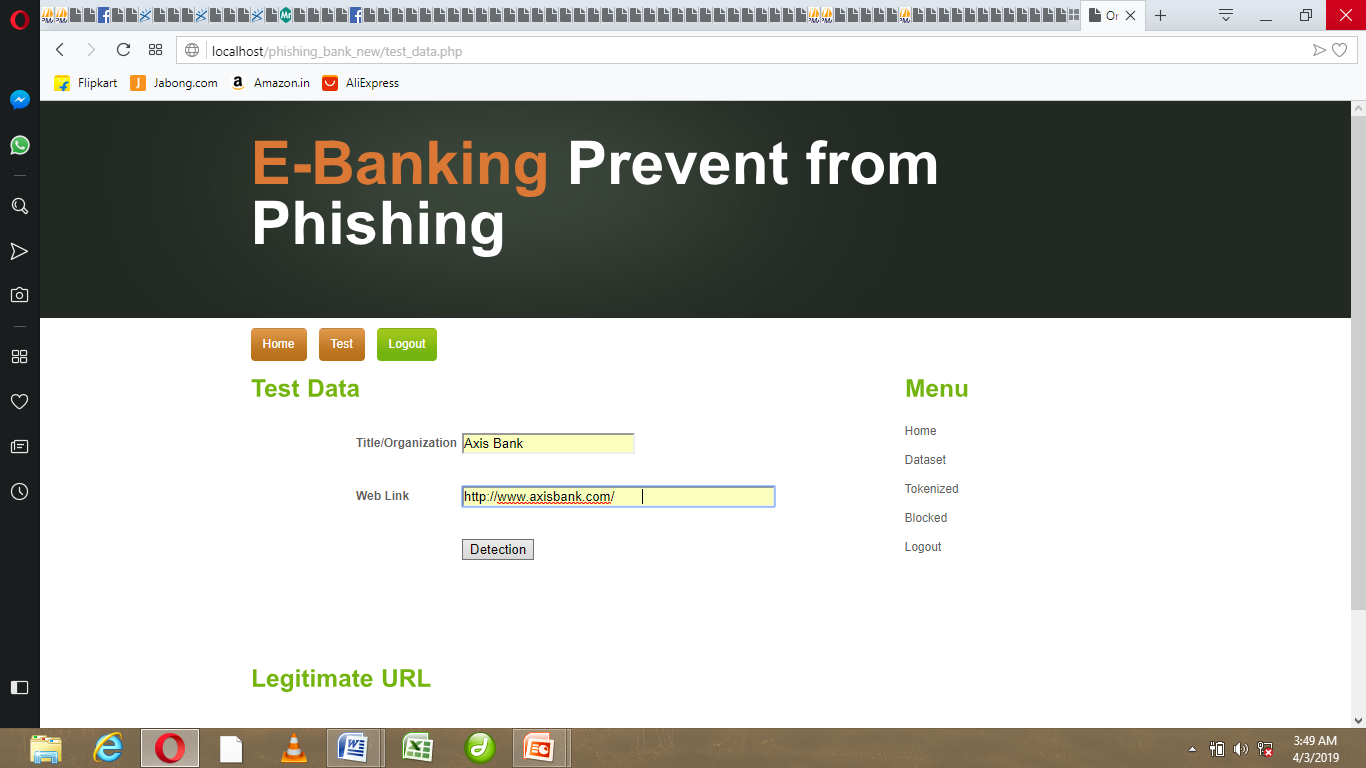
**S3: Feature Selection**

****

**S4: Blocked Websites**

****

**S5: Testing Data**

****

**S6: Legitimate Websites**

**REFERENCES**

[1] L. Wenyin, G. Huang, L. Xiaoyue, X. Deng, and Z. Min, “Phishing Web page detection,” in Proc. IEEE 8th Int. Conf. Document Anal. Recognit.,Seoul, South Korea, 2005, pp. 560–564.

[2] P. Stavroulakis and M. Stamp, Handbook of Information and Communication Security, 1st ed. Heidelberg, Germany: Springer, 2010.

[3] Anti-Phishing Working Group. Accessed: Sep. 2016. [Online]. Available http://www.antiphishing.org

[4]MobileMarketingStatistic.Accessed:Mar.2017.[Online].Availablehttp://www.smartinsights.com/mobile-marketing/mobilemarketing-analytics/mobile-marketing-statistics/

[5] Phishing Attacks. Accessed: Sep. 2015. [Online]. Available:https://securityintelligence.com/

[6] Y. Zhang, J. I. Hong, and L. F. Cranor, “CANTINA: A content-based approach to detecting phishing Web sites,” in Proc. ACM 16th Int. Conf.World Wide Web, Banff, AB, Canada, 2007, pp. 639–648.

[7] PhishTank. Accessed: Nov. 2015. [Online]. Available:http://www.phishtank.com/stats/2014/01/

[8] S. Sheng et al., “An empirical analysis of phishing blacklists,” inProc. 6th Conf. Email Anti-Spam (CEAS), 2009, pp. 1–6.

[9] K. Gandhimathi and M. S. Vijaya, “Identifying similar Web pages using scoring methods for Web community mining,” Int. J. Data Min. Knowl.Manag. Process, vol. 3, no. 6, p. 41, 2013.

[10] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, “Fighting againstphishing attacks: State of the art and future challenges,” Neural Comput.Appl., vol. 28, no. 12, pp. 3629–3654, Dec. 2017.

[11] APWG Reports. Accessed: Sep. 2016. [Online]. Available:http://docs.apwg.org/reports/

[12] Fog Computing and the Internet of Things: Extend the Cloudto Where the ThingsAccessed:Sep.2016.[Online].Available:http://www.cisco.com/c/dam/en\_us/solutions/trends/iot/docs/computing-overview.pdf

[13] Y. Li and M. Chen, “Software-defined network function virtualization:A survey,” IEEE Access, vol. 3, pp. 2542–2553, 2015.

[14]RealTimeAntiPhishing.Accessed:Sep.2016.[Online].Available:http://www.brightcloud.com/services/real-time-anti-phishing.php

[15]AntiSpamHardware.Accessed:Sep.2016.[Online].Available:http://www.windowsnetworking.com/hardware/Anti-Spam-Hardware/modusGate-Anti-Spam-Appliance.html

[16] Security Intelligence Blacklisting. Accessed: Sep. 2016.[Online].Available:https://www.cisco.com/c/en/us/td/docs/security/firepower/601/configuration/guide/fpmc-config-guide-v601/Security\_Intelligence\_Blacklisting.pdf

[17] N. Zhang and Y. Yuan. (2012). Phishing Detection UsingNeural Network, CS229 Lecture Notes. [Online]. Available:http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf

[18] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, “CANTINA+: A featurerich machine learning framework for detecting phishing Web sites,”ACM Trans. Inf. Syst. Security, vol. 14, no. 2, pp. 1–28, 2011.

[19] Vulnerability in Spotify Android-App. Accessed: Sep. 2015.[Online]. Available:http://blog.trendmicro.com/trendlabs-securityintelligence/vulnerability-in-spotify-android-app-may-lead-to-phishing/

[20] Google Safe Browsing. Accessed: Sep. 2016. [Online]. Available:https://safebrowsing.google.com/

[21] N. Abdelhamid, “Multi-label rules for phishing classification,” Appl.Comput. Informat., vol. 11, no. 1, pp. 29–46, 2015. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S2210832714000210

[22] W. Hadi, F. Aburub, and S. Alhawari, “A new fast associativeclassification algorithm for detecting phishing websites,” Appl. SoftComput., vol. 48, pp. 729–734, Nov. 2016. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S1568494616303970

[23] L. T. C. James, L. Sandhya, and C. Thomas, “Detection of phishing URLs using machine learning techniques,” in Proc. Int. Conf. Control Commun. Comput. (ICCC), 2013, pp. 304–309.

[24] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, “Intelligent phishing detection system for e-banking using fuzzy data mining,” Expert Syst. Appl., vol. 37, no. 12, pp. 7913–7921, 2010.

[25] A. N. V. Sunil and A. Sardana, “A PageRank based detection technique for phishing Web sites,” in Proc. IEEE Symp. Comput. Informat. (ISCI),Penang, Malaysia, 2012, pp. 58–63