EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

Email Phishing: An Enhanced Classification Model to Detect Malicious URLs

Shweta Sankhwar[1, *], Dhirendra Pandey[1] and R.A Khan[1]

[1]Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

Abstract

Phishing is the process of enticing people into visiting fraudulent websites and persuading them to enter their personal information. Number in phishing email are spread with the aim of making web users believe that they are communicating with a trusted entity or organization. Phishing is deployed by the use of advanced and harmful tactics like malicious or phishing URLs. So, it becomes necessary to detect malicious or phishing URLs in the present scenario. Numerous anti-phishing techniques are in vogue to discriminate fake and the authentic website but are not effective. This research, focuses on the relevant URLs features that discriminate between legitimate and malicious/phishing URLs. The impact of email phishing can be largely reduced by adopting an appropriate combination of all these features with classification techniques. Therefore, an Enhanced Malicious URLs Detection (EMUD) model is developed with machine learning techniques for better classification and accurate results.

Keywords: Email, Phishing, Machine Learning Techniques, Information security, Cybercrime.

Received on 28 February 2019, accepted on 26 April 2019, published on 06 May 2019

doi: 10.4108/eai.13-7-2018.158529

1. Introduction

Over the last decade phishing attacks have grown considerably in the internet. E-mail Phishing is presently amongst the latest, very tricky and problematic of trends in network security threats. Phishing is a process of gaining the sensitive information of user through generating a fake or counterfeit webpage, which appears to be a legitimate one that actually comes under cybercrime. Malicious URL are challenging threat in cyber space which steals the user's sensitive information. Phishing is a serious threat that intent to use the vulnerabilities or weakness found in system process as caused by online users. Phishing refers to sending of spurious emails which are usually forged by the phishers to lure a user in their snares leading the user to lose sensitive data or credential, identity theft, pecuniary loss etc. Phishing URLs are challenging threat in cyber space which steal the user's sensitive information. The phishers are using numerous phishing URLs crafting tactics pointing to the same phishing website to bypass the detection techniques [1]. Therefore, it becomes necessary to detect the suspicious or malicious URLs in the present scenario. A lot of anti-phishing techniques are in vogue to draw a dividing line or identify between the fake and the authentic websites, however due to the vast amount and new harmful tactics of phisher, the challenges are yet being faced. [2] [3]

For instance, a system can be technically safe and secure enough against password theft, however naive users may leak their sensitive information if an attacker lead them to update their sensitive information such as username, passwords via a given Hypertext Transfer Protocol (HTTP) link [4]. It could ultimately breach the security of the system, web vulnerabilities like obfuscated/phishing URLs can be used by phishers to craft far more influencing socially-engineered messages. Fraudsters or phishers use spoofed domain names which can be persuading instead using legitimate domain names. [5] [6]

Therefore, to reduce the phishing attack Enhanced Malicious URL Detection (EMUD) model is developed

[*]Corresponding author. Email: Shweta.sank@gmail.com

which includes EMUD algorithm for detection and classification of URL.

This EMUD algorithm selects 14 heuristics to detect malicious or phishing URL. Machine Learning (ML) techniques such as Naïve Bayes (NB), Support Vector Machine (SVM) is employed as classifiers to do phishing and legitimate URLs/ sites classification. EMUD model analyses the URL set to detect withier the website is genuine or malicious. The remaining of this paper is organized as follows: Section- 2 covers the background of the problem section-3 explains architecture of proposed model and throws light on URL feature set. Implementation of EMUD model is done in section-4 and the comparison of proposed and existing approach is done on the basis of accuracy and performance. At last, in section-5 the paper is concluded.

2. State-of-the-art of E-mail Phishing

Phishing is the process of enticing people into visiting fraudulent websites and persuading them to enter their personal information. Numbers of phishing email are spread with the aim of making web users believe that they are communicating with a trusted entity [2]. Phishing deployed by use advanced technical means. Phishing refers to sending of spurious emails which are usually forged by the phishers to lure a user in their snares leading the user to lose his/ her sensitive data or credential, identity theft, pecuniary loss etc. The cyber criminals have left no stone unturned in this regard and their advance, tenable way of cyber-attacks has given result to social engineering and phishing. In execution the cyber criminals have specifically used URLs and embedded links as their biggest weapon [3].

Phishing is a cyber-crime that occurs when a malicious webpage imitates as legitimate webpage so as to gain sensitive information from user. Phisher sends bulk emails containing links, to the naive users try to convince them to visit their fake site. The sender's mail server (Mail Transfer Agent or MTA) looks up the "@domain.com" portion of the recipient's email address in a Domain Name System (DNS) server to determine which destination mail server (referred to as a "Mail Exchanger," or MX) it should contact to deliver the message. The sending and receiving servers communicate using the SMTP protocol. The receiving server accepts the message so that it can be delivered to the recipient as shown in Figure.1. The recipient's email client retrieves the message using standards like the Post Office Protocol (POP) or Internet Message Access Protocol (IMAP) to download the message so it can be read. A Mail User Agent (MUA) is a program that allows you to receive and send e-mail messages; it's usually just called an e-mail program. MUA is sometimes called an e-mail agent or an e-mail client shown in Figure.1. The unaware victims of email phishing, unknowingly click the link/URL which takes them to the fake webpage which is replica of legitimate website. The phishers persuade victims to enters their sensitive data like credit card information, username or password etc. [7].

Obfuscated/malicious URLs are created to perform phishing attacks and generally, every legitimate URL has following common syntax: *<protocol>://<hostname><path>*

Phishers develops the tactics following the URL format as shown above to misguide the naive users to misuse their sensitive information for their own benefits in terms of money, forgery, identity theft etc. [8]. The state-of-the-art of URL based e-mail phishing is depicted in Figure 1.

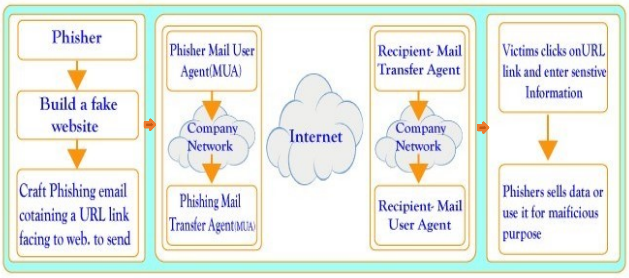
Phisher
Build a fake website
Craft Phishing email cotaining a URL link facing to web. to send
Phisher Mail User Agent(MUA)
Company Network
Phishing Mail Transfer Agent(MUA)
Internet
Recipient- Mail Transfer Agent
Company Network
Recipient- Mail User Agent
Victims clicks onURL link and enter senstive Information
Phishers sells data or use it for maificious purpose

Figure 1. State-of-the-art of URL based E-mail Phishing

2.1. URL Structure

A URL (Uniform Resource Locator) is a protocol used to specify the location of data on a web or network. The URL is composed of the protocol, primary domain, subdomain, Top- Level Domain (TLD), and path domain [6]. In this research paper, the subdomain, primary domain, and TLD are communally referred to as the domain and the individual components of a URL are shown in Figure 2.

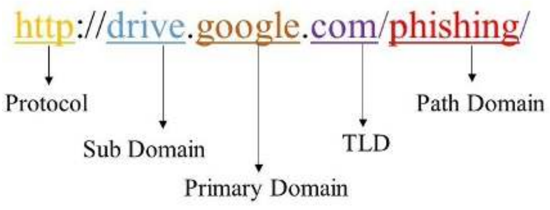

Figure 2. URL Structure

The protocol denotes to a communication protocol for exchanging information between information devices; e.g., Hyper Text Transfer Protocol (HTTP), File Transfer Protocol (FTP), Hyper Text Transfer Protocol Secure (HTTPS) etc. Protocols are of many types and used in accordingly with the desired way of communication. Domain includes the subdomain, is an ancillary domain and has many types depending on the services provided by the domain page. The domain is the name given to the real Internet Protocol (IP) address through the Domain Name System (DNS). The primary domain is the most important part of a domain. The TLD is the domain in the highest position in the domain name hierarchy architecture; e.g., .com, .net, .kr, .jp etc. We define features of each component of the URL, these features are used for phishing site detection. [9] [10]

3. E-mail Phishing URL Detection (EMUD) model

Enhanced Malicious URL Detection (EMUD) model is developed to detect the phishing email. It consists of EMUD algorithm for detection and Machine Learning technique for classification of phishing or malicious URLs. Two algorithms of the classification improve accuracy and performance of the EMUD model. EMUD model focuses on the relevant 14 heuristics that discriminate between legitimate and malicious/phishing URLs. These 14 heuristics are selected on the basis of rigorous literature review and keen observation of phishing attack patterns. Unnecessary URL features are not considered to reduce the execution time and false rate. The proposed model gives high performance and accuracy with less false rate.

3.1 Architecture of the EMUD Model

The core idea behind the proposed Enhanced Malicious URL Detection (EMUD) Model is to attempt to control or detect the phishing attacks as user faces difficulty to distinguish the legitimate and the malicious emails via URL. The EMUD model works basically in two-phase. In the first phase, the developed Enhanced Malicious URL Detection (EMUD) algorithm is applied to URLs set to detect malicious URLs with identified fourteen URL heuristics. EMUD algorithm detects almost all the obfuscated, phished/malicious URL which could result phishing attack. In second phase, the Machine learning techniques is used for classification to check the accuracy of EMUD algorithm. The Naïve Bayes (NB) classifier as well as Support Vector Machine (SVM) is employed, latter compared also to test which sense the nature of URL Set efficiently. This classifier improves the accuracy and performance of the classification. At last, an alert is generated to the user through pop-up if malicious URL is detected as shown in Figure 3.

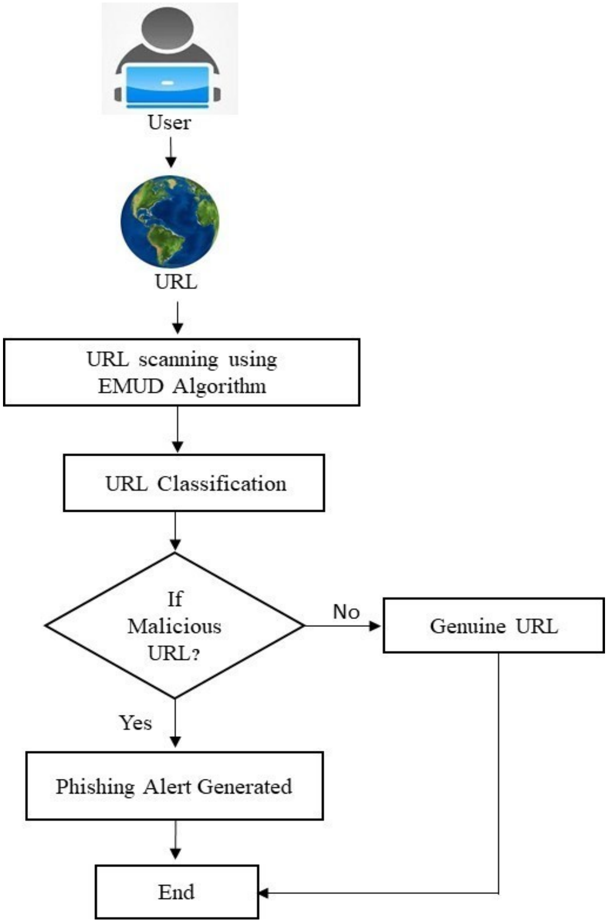

Figure 3. Architecture of Enhanced Malicious URL Detection Model

3.2 URL Feature Set

Heuristic approach is used to determine whether the URL is genuine or malicious. A rigorous literature review is done and observed the phishing URLs patterns. Here, we have selected the weighty heuristics which effectively identify the phishing URLs or phishing website. The list of Heuristic used by the EMUD algorithm is explained below:

Check Blacklisted domain: There are many resources or sites which maintain the backlisted domains. The proposed system will also maintain the backlisted domain to avoid repeated detection [11] [12].

Number of Dots in URL: Legitimate URLs do not contain more than five, but phishing URLs usually have many dots to gain user trust. Phishers insert some sub-domains after the domain name. This increases the number of dots in the URL. Another reason for the increase of dots is when phishers have a redirect script in the URL. User click on this URL trusting the legitimate site and URL but is deceived by visual similarities created by the phishers.

To check the URL, count the number of dots in the URL. If the count is more than 5, the URL is related to phishing. [13] [14].

Visual Similarity: The actual link domain name does not match with the visual link. Consider an example, this hyper- link: <a href="http://www.baroda.com/login.php"> http:// www.secure.onlinebaroda.in/login.php </a>which looks like it is going to navigate to secure.onlinebaroda.in, which is the portal of original site, instead it is pointing to the attacker website www.baroda.com. [13] [14]

Double slash in URL: The correct syntax of a URL is (protocol identifier) :// (resource name). As is obvious, the double slash occurs only once and that is after the colon. If a double slash exists again in URL, it is an indication towards a phishing attack [5]. The initial step is to check if the path segment of a URL contains double slashes (//). The phisher inserts legitimate domain names as a small part of a longer URL.
Thereby, phishers try to create an impression that the URL is authentic.
Example:http://blizzard.freel.coml/logindiablo3.vicp.net/log in/zh/?ref?https://us.badsite.net/ac-count/...
In the path segment, a double slash exists, and it includes a legitimate URL. This is a badsite.net member login page's phishing URL [13] [14].

IP-based URLs: All authentic websites have URLs that contain a domain name, which is associated with an IP address in the DNS. So, in an email, if a user comes across a domain name in a URL, the user can rely that the website is safe. However, if the user comes across a URL that does not have a domain name, but contains an IP address, this is an indication of phishing attack. Typically, phishers use compromised computers to develop a phishing website. The reason is that IP addresses of such computers' might not be found in the DNS. In this way, the phisher's identity is hidden. The sole method of referring to the website developed on compromised computers is by using the IP addresses of these computers [13] [14]

Length of Domain Name/ Long URL: Phishers mostly use long URL to hide the malicious part in the address bar. If the length of the URL is greater than or equal to 54 characters then the URL classified as phishing [14]. For Instance:
http://fabadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b73a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

Invalid Port Number: The port number part of the domain name is compared with stated port no. If these two numbers match the possibility is that the URL is of an authentic website. However, if the port number is mentioned in the domain part and differs from the stated port number of the protocol, this raises suspicion of a phishing URL. [15]
Example:
http://61.128.197.81:5800/signin.htm?_encoding¼UTF8&...
In this case, the port number is 5800. However, the stated protocol being 'http', and the default port number is either 80 or 8080.

Country-code Validation: The rule is that it is illegal to host a website in one country and has another country code in the domain name. Thus, if the country code in the domain name and the hosting country code are the same, then the URL is authentic. If not, it may be a phishing website [16]. If URL have Multiple top-level domains (TLDs) then also it sounds phishi. Generally, a legitimate URL has one domain name. If the domain name has a combination of two or more distinct domains or a blend of a Top-level domain and a second-level domain, this is indicative of a phishing URL [2]. As per RSA's fraud report, "In April 2012, these types of phishing attacks were seen in a large number [17].

@ Symbol: To outwit the users, phishers use the ampersand symbol (@) in a URL. The browser is designed to ignore the text prior to the ampersand symbol and consider the text after the ampersand symbol. The user wrongly concludes that the link will direct to the website that is preceding the ampersand symbol. However, the user is tricked because the link directs him/her to the website following the ampersand symbol, which is a phishing website. [13] [18]

In the following format <userinfo>@<host>, the browser will link to the <host> site and ignore the <userinfo>. That is, it is checked if the syntax of the URL is 'http(s)://username.password@domain_name or http://www.sbi.account.com@example.net/ Prior to the ampersand symbol is the user information, and following the ampersand symbol is the domain name to retrieve a webpage. The fact is that genuine websites never use this syntax. Only malicious ones use it to open an illegitimate webpage that appears an authentic website.

Special Character: Special character is basically used for adding prefix or suffix to the domain by phishers. The dash symbol (-) is used by phishers to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate URL. For example, http://www.Confirme-paypal.com/ or "www-Fsecure.com"," wwwf-secure.com" and "www.f-secue.com" are websites of phishers' whereas the actual legitimate website is www.fsecure.com. For Instance- http://paypal-update.com sounds legitimate but it is phished URL, redirecting spoofed website but the genuine is https://paypal.com. [13] [18]

Hexadecimal in URL: If an IP address is used as an alternative of the domain name in the URL, such as "http://125.94.5.140/phishi.html", users can be sure that someone is trying to steal their personal information. the IP address is also converted in hexadecimal code as shown

in the following link-
http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html or in ASCII code to confuse the user and redirect to phishing site [18].

HTTP in Domain Part: Phishers add https in domain part of URL to fool user and gain trust to input his sensitive information.
For example-
http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/. [17] [18].

Age of Domain Name: Phishers develop phishing websites, the names of which are similar to authentic websites. For example, for www.icicibank.com, phishers might use www.icici.com or www.icici_bank.com. To avoid being caught by legitimate organizations that keep a watch on such domain name registrations, the phishers try to leverage their phishing websites in a small-time duration. According to the Anti-Phishing working group's report (2012) the life of a phishing website as 46 hours and 3 minutes on an average [70]. A WHOIS search returns the age of a domain name in the number of months. If the age of a domain name is less than the threshold value, one must become suspicious that this domain name that might be of a phishing website. [13] [14]

Phishers try name-based attacks. They used fraudulently-procured credit cards to register a domain name that is similar to an existing domain name. To enhance the authenticity, they use trademarks of genuine organizations. Such domains have a short life. Most organizations monitor domain name registrations that include their trademarks. Here, there is possibility that phisher can be caught. So, phishers are in a hurry to leverage these domain names. Naive users are cheated by the similar appearance of these domains and fall prey to this well-designed phishing attack.

3.3 Enhanced Malicious URL Detection (EMUD) Algorithm

Enhanced Malicious URL Detection (EMUD) algorithm is developed to detect the phishing email. It is used for classification of phishing or malicious URLs. EMUD model focuses on the relevant 14 heuristics that discriminate between legitimate and malicious/phishing URLs. These 14 heuristics are selected on the basis of rigorous literature review and keen observation of phishing patterns. Unnecessary URL features are not considered to reduce the execution time and false rate. The proposed algorithm classifies phishing and legitimate URLs efficiently and results high performance, accuracy with less false rate. Enhanced Malicious URL Detection (EMUD) algorithm is inbuilt with relevant heuristics as discussed in above section to detect obfuscated, phished/malicious URLs. In this algorithm, step-by-step all fourteen heuristics are considered with input of URL and all the heuristics or steps are already explained in section 3.2. and EMUD Algorithm explained.

Enhanced Malicious URL Detection (EMUD) Algorithm:

```
Input = URL
Output = Phishing/ Legitimate

EMUD
{
Step 1: Check in Blacklist (H1)
if (Domain or IP address exists in blacklisted URL database)
Return 1 else Return 0
Step 2: Number of Dots (H2) If (Dots in Domain Name > 5)
Return 1 else Return 0
Step 3: Visual-similarity Redirection (H3)
If (Visual URL/link mismatched with redirected URL)
Return 1 else Return 0
Step 4: Double slash (H4)
If (double slash exists more than one in URL)
Return 1 else Return 0
Step 5: Port Number (H5)
If (valid Port Number in URL)
Return 1 else Return 0
Step 6: Length of Domain Name (H6)
If (Domain Name Length is more than 53 > Characters)
Return 1 else Return 0
Step 7: Check Country-code (H7) Get IP address of domain
name
Get Geographical location of the IP address through Geolite IP
query for country database Convert geolocation to hosting
country code and find CcTLD form domain name.
if (Hosting country-code does not match with CcTLD)
Return 1 else Return 0
Step 8: @ Symbol (H8)
 If (URL contains @symbol) Return 1 else Return 0
Step 9: Special Characters (H9)
If (Domain contains Special characters)
Return 1 else Return 0
Step 10: IP Address in URL (H10)
If (Domain name is in the form of IP Address)
Return 1 else Return 0
Step 11: ASCII code (H11)
If (Domain name is in the form of ASCII code)
Return 1 else Return 0
Step 12: Hexadecimal (H12)
If (Domain name is in the form of Hexadecimal)
Return 1 else Return 0
Step 13: Http in Domain (H13) If (http is present in URL)
Return 1 else Return
Step 14: Age of the Domain (H14)
Extract Registered Date of Domain Name from WHOIS record
and calculate its age (Current – registered date & time)
If (domain Age < 46 hours 3 minutes)
Return 1 else Return 0
}

Start
If (Return value is 1 in any Function)
Display URL is Phished/ Malicious else Display URL is
Genuine
End
```

4. Implementation of the EMUD Model

The implementation of EMUD model is done with two experimental setups. The first experimental setup comprises a dataset of small sample of same website where scanning of phishing URL is done with the help of EMUD algorithm with
14 heuristics of URLs and performance of is calculated. Thereby, NB classifier is employed and evaluation is done to check the performance of EMUD model. Thereafter, the comparison of proposed model i.e. EMUD with existing method i.e. EPCMU (Enhanced Probing Classification of Malicious URL) model is done [19]. Further, the second experimental setup comprise large real-world dataset. The experiment also works same as like first only; the difference is that SVM is used for classification. Thereby, confusion metrics and k-fold cross-validation (with 10-fold cross-validation) is done for performance evaluation.

4.1 Experimental Setup with Dataset-I

Dataset-I consist of 13 malicious URLs and two genuine URLs as listed below in Table 1. The EMUD algorithm test these data as input and employed machine learning to evaluate the accuracy and confusion matrix. NB classifier is used with confusion matrix used for accuracy or performance evaluation.

4.1.1 Performance Analysis

The proposed EMUD Algorithm scanned URL dataset and the experimental results shows that the 13 URLs are malicious and two are genuine as listed in Table 2. The existing approach i.e. Enhanced Probing Classification of Malicious URL (EPCMU) model detects 9 URLs are malicious and 6 URLs are genuine. But actually, in data set 13 URLs are malicious and two are genuine, as shown in Table 1. The reason may be EPCMU Algorithm is not modern as it has not considered all the heuristic from H1 to H14 in their algorithm the mentioned Table 3, whereas, EMUD algorithm have considered all 14 heuristic to predict exact classification shown in Table 2.

4.1.2 Performance Evaluation with Naïve Bayes Classifier

In the classification phase, the Naïve Bayes classifier is used to detect the phishing URLs. The classifier has high classification abilities as it is based on Bayes theorem [20] [21]. This classifier detects whether the URLs are phishing (malicious) or legitimate. Bayes formula is given below:

P(C|X) = (P(X|C) *P(C)) / P(X) ………………………. (1)
Where, P(C) is Probability of the occurrence of the class C.
P(X) is Probability of the occurrence of the attributes.

P(C|X) is the probability of the attribute X belongs to the class C.
P(X|C) is the probability of Class C having attribute X.

The class C have two parts phishing and legitimate in which URLs classification is done through the URL feature vector the analyses, the classifier decides the appropriate category of the URLs based on higher posterior probability by using the formulas given below. [21] [22] [19]

Probability of URL to be Phishing is given below:
P (C1|X) = (P (X|C1) *P (C1)) / P(X) …………………….. (2)

Probability of URL to be legitimate is given below
P (C2|X) = (P (X|C2) *P(C2)) / P(X) ………………... (3)

P (C1|X) > P (C2|X) then the URL is malicious else.… (4)
The URL is genuine.

Naïve Bayes classifier is used for classification for proposed Mechanism. The URL set is given to NB algorithm. Both the algorithm detects the following URL heuristics. Enhanced Malicious URL Detection (EMUD) algorithm with Naive Bay algorithm analyze the values of each independent Heuristic as given below in Table 4.

After applying the equation number (1), (2), (3) and (4) in the dataset, the calculations and results are given below:

P(Class$_{Malicious}$=Yes | X) = [P(H1='Y' | Class$_{Malicious}$ =Yes) *P(H2='Y' | Class$_{Malicious}$ =Yes) * (H3='Y' | Class$_{Malicious}$=Yes)* (H4='Y' | Class$_{Malicious}$=Yes) P(H5 | Class$_{Malicious}$ =Yes) * P(H6 | Class$_{Malicious}$=Yes) * P(H7='Y' | Class$_{Malicious}$=Yes) * P(H8='Y' | Class$_{Malicious}$ =Yes) * P(H9='Y' | Class$_{Malicious}$ =Yes) * (H10='Y' | Class$_{Malicious}$=Yes) * P (H11 | Class$_{Malicious}$=Yes) * P (H12 | Class$_{Malicious}$ =Yes) * P (H13='Y' | Class$_{Malicious}$ =Yes) * P (H14='Y' | Class$_{Malicious}$ =Yes)]
P(Class$_{Malicious}$=Yes|X) = [1/13*5/13*1/13*1/13*1/13*1/13*1/13*1/13*2/13*2/13*1/13*1/13*1/13*1/13*8/13] = 0.00000000000004

P(Class$_{Genuine}$=Yes | X) = [P(H1='Y' | Class$_{Genuine}$ =Yes) * P(H2='Y' | Class$_{Genuine}$ =Yes) * (H3='Y' | Class$_{Genuine}$ =Yes)*P(H4 | Class$_{Genuine}$ =Yes) * P(H5 | Class$_{Genuine}$ =Yes)*P(H6='Y' | Class$_{Genuine}$ =Yes) * P(H7='Y' | Class$_{Genuine}$=Yes)*P(H8='Y' | Class$_{Genuine}$ =Yes) * (H9='Y' | Chast$_{ening}$=Yes) * P(H10 | Class$_{Genuine}$ =Yes) * P(H11 | Class$_{Genuine}$=Yes) * P(F12='Y' | Class$_{Genuine}$ =Yes)] * P(H11 | Class$_{Genuine}$=Yes) * P(F13='Y' | Class$_{Genuine}$ =Yes)]
P (Class$_{Genuine}$ =Yes | X) = [0/2 * 0/2 * 0/2 * 0/2 * 0/2 * 0/2*0/2* 0/2 * 0/2 * 0/2 * 0/2 * 0/2* 0/2 * 0/2 * 0/2] = 0

0.00000000000004 > 0

P (Class$_{Malicious}$=Yes | X) > P (Class$_{Genuine}$=Yes | X) so the website is malicious and alert is generated (or website is blocked).

Table 1. List of URLs

SN	URL
1.	https://larapo.org@74.125.121.150
2.	http://www.larapo.org.ar/clientele/space-far.cs.3dsecureclient.as
3.	http://www.larapo.org.ar
4.	“http://www.larapo.org.ar//http://www.lavapo.com
5.	http//61.128.197.8:5800/
6.	http://larapo.org.ar/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&amp;dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@laropa.html
7.	http// www.larapo.org.ar.uk/cgi
8.	www.larapo.org.ar@larapo.html
9.	http://www.larapo.org/wpadmin/loadscripts.php?c=1&load[]=swfobject,jquery,utils&ver=3.5
10.	http:// 122.163.158.9/index.html
11.	http:// 119 119 119 46 108 97 114 97 112 111 46 99 111 109
12.	http://0x58.0xCC.0xCA.0x62/2/laraapo.ca/index.html
13.	http://https-www-larapo.org
14.	https://larapo.org/home.html
15.	https://larapo.org

Table 2. Experimental Result of EMUD Algorithm (proposed algorithm)

SN	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	Class
1	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes	Malicious
2	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Malicious
3	No	No	Yes	No	No	No	No	No	No	No	No	No	No	Yes	Malicious
4	No	Yes	No	Yes	No	No	No	No	No	No	No	No	No	Yes	Malicious
5	No	Yes	No	No	Yes	No	No	No	No	No	No	No	No	Yes	Malicious
6	No	Yes	No	No	No	Yes	No	No	No	No	No	No	No	No	Malicious
7	No	No	No	No	No	No	Yes	No	Yes	No	No	No	No	Yes	Malicious
8	No	No	No	No	No	No	No	Yes	No	No	No	No	No	Yes	Malicious
9	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No	Malicious
10	No	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes	Malicious
11	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	Malicious
12	No	Yes	No	No	No	No	No	No	No	No	No	Yes	No	Yes	Malicious
13	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	Malicious
14	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Genuine
15	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Genuine

Table 3. Experimental Result of EPCMU Algorithm (existing approach)

SN	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	Class
1	Yes	No	-	No	-	-	-	No	No	Yes	-	-	-	-	Malicious
2	No	Yes	-	No	-	-	-	No	No	No	-	-	-	-	Malicious
3	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
4	No	Yes	-	Yes	-	-	-	No	No	No	-	-	-	-	Malicious
5	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
6	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
7	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
8	No	No	-	No	-	-	-	Yes	No	No	-	-	-	-	Malicious
9	No	No	-	No	-	-	-	No	Yes	No	-	-	-	-	Malicious
10	No	No	-	No	-	-	-	No	No	Yes	-	-	-	-	Malicious
11	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
12	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
13	No	No	-	No	-	-	-	No	No	No	-	-	-	-	~~Genuine~~
14	No	No	-	No	-	-	-	No	No	No	-	-	-	-	Genuine
15	No	No	-	No	-	-	-	No	No	No	-	-	-	-	Genuine

Table 4. Naïve Bayes Classification of URLs

URL Heuristic	Frequency		Probability in class	
	Class=Malicious	Class=Genuine	Class=Malicious	Class=Genuine
H1. Check Blacklisted Domain				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H2. No. of Dots				
Yes	5	0	5/15	0/15
No	8	2	8/15	2/15
Total	13	2		
H3. Abnormal URL				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H4. Number of slashes				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H5. Length of domain name				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H6. Length of domain name				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H7. Invalid country code				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H8. @ Symbol				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H9. Special Character				
Yes	2	0	2/15	0/2
No	11	2	11/15	2/2
Total	13	2		
H10. IP Address				
Yes	2	0	2/15	0/15
No	11	2	11/15	2/15
Total	13	2		
H11. ASCII code				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H12. Hexadecimal in URL				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H13. HTTP in domain name				
Yes	1	0	1/15	0/15
No	12	2	12/15	2/15
Total	13	2		
H14. Combinational Check				
Yes	8	0	13/15	2/15
No	5	2	0/15	0/15
Total	13	2		

4.1.3 Comparation Study of Proposed and Existing Performance Metric

In this data analysis, confusion matrix is used to evaluate the performance of the proposed approach [24][25]. Here, the True Positive Rate (TPR) and False Positive Rate (FPR) is considered for evaluation [26]. In addition, we used standard measure i.e. Accuracy.

TP: Number of phishing URLs correctly classified as Phishing.
TN: Number of legitimate URLs correctly classified as legitimate.
FP: Number of legitimate URLs which are classified as phish
FN: Number of phishing URLs which are classified as legitimate
Here, the accuracy and performance are evaluated through confusion matrix as depicted in Table 5.

Table 5: Confusion matrix of EMUD and existing model

Model	Confusion Matrix		Result	
EPCMU	TPR	TP/(TP+FN)	6/6+7	6/13
	TNR	TN/(TN+FP)	2/2+0	1/1
	FPR	FP/(FP+TN)	0/0+2	0
	FNR	TP/(FN+TP)	7/7+6	7/13
	Accuracy	((TP+TN)/ (TP+TN+FP+FN)) *100	6+2/15	53%
EMUD	TPR	TP/(TP+FN)	13/13+0	1/1
	TNR	TN/(TN+FP)	2/2+0	1/1
	FPR	FP/(FP+TN)	0/0+2	0
	FNR	TP/(FN+TP)	0/0+13	0
	Accuracy %	((TP+TN)/ (TP+TN+F P+FN)) *100	13/13+0	100 %

Through this comparative analysis, EPCMU Classification Rate (%) is 53.3% and EMUD 100% Classification Rate which shows high accuracy in detection of malicious or phished URLs as shown in Table 5. In both the model i.e. EMUD and EPCMU employed NB classifier and it is also observed that the NB took long time for processing.
Therefore, in next section, some other ML techniques like SVM is employed for classification in the place NB in EMUD model.

4.2 Experimental Setup with Dataset-II

Dataset is collected from real world data of the 2000 phishing URLs and legitimate. Phishing URLs data source is Phishing tank (https://www.phishingtank.com) and legitimate URLs data source is DMOZ and Alexa (https://www.alexa.com/topsites). The EMUD algorithm test these data as input and employed machine learning to evaluate the accuracy and confusion matrix. The weka tool is used for the evaluation [27]. Specifically, the distribution ratio of phishing and legitimate data is in 60:40 ratio respectively. SVM classifier is used with confusion matrix and k-fold Cross validation (10-fold) is used for accuracy or performance evaluation.

4.2.1 Performance Evaluation with Support Vector Machine

In this experiment, Support Vector Machines (SVM) is adopted as it a popular classifier to achieve better classification. It is extensively used in text classification and specially in computer security field i.e. spam detection, hidden email construction, masquerade detection and phishing detection. The main benefit of this learning algorithm is that it is fully oblivious to the input feature numbers and focus to increase the separable margin [27] [28].

$$w' x_i + w_0 \_\ 0 \text{ if } t_i = +1;$$
$$w' x_i + w_0 < 0 \text{ if } t_i = -1;$$

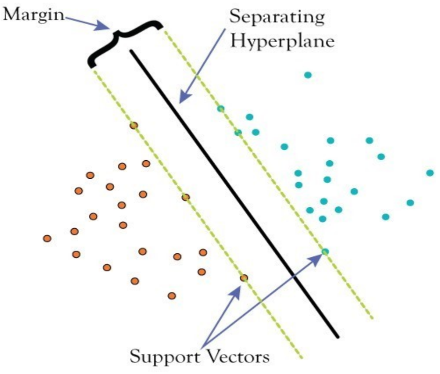

Figure 4. Support Vector Machine

Suppose that a linear discrimination function and two linear separable classes with target values +1 and -1. A discriminating hyperplane will satisfy, the distance of any point x to a hyperplane is |w' xi + w0| / ||w|| and distance to the origin is |w0 | / ||w ||. Support vectors are the points lying on the boundaries, and the middle of the margin is the optimal separating hyperplane that maximizes the margin of separation as shown in Figure 4. [28] [29][30]

4.2.2 Performance Metric

In this data analysis, we used confusion matrix to evaluate the performance of the proposed approach on which mainly the True Positive Rate (TPR) and False Positive Rate (FPR) is considered for evaluation as shown in Table 6. In addition, we used standard measure such as the Precision and Accuracy. [29] [31]

TP: Number of phishing URLs correctly classified as Phishing.
TN: Number of legitimate URLs correctly classified as legitimate.
FP: Number of legitimate URLs which are classified as phish
FN: Number of phishing URLs which are classified as legitimate [27]
Four metrics are calculated as follows:

True Positive Rate (TPR): It is the ratio of the phishing URLs that are correctly identified and the equation of the TPR is shown in eq. (5)

$$TPR = \frac{TP}{TP + FN}$$

True Negative Rate (TNR): It is the ratio of the legitimate URLs that are correctly identified and the equation of the TPR is shown in eq. (6)

$$TNR = \frac{TN}{TN + FP}$$

False Positive Rate (FPR): It is the ratio of the legitimate URLs that are classified as phishing and the equation of the FPR is shown in eq. (7)

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate (FNR): The number of phishing URLs classified as legitimate. The equation of the FNR is shown in eq. (8)

$$FNR = \frac{TP}{FN + TP}$$

Accuracy: The accuracy computation is shown in eq. (9)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: The precision is defined as the number of true positive (TP) over the sum of True positive and False positive number is shown in eq. (10)

$$Precision = \frac{TP}{TP + FP}$$

Recall: The recall is defined as the number of true positive (TP) over sum of TP and FN is shown in eq. (11)

$$Recall = \frac{TP}{TP + FN}$$

F-1 Measure: The F1- Measure defined as the harmonic mean of precision and recall is given in the eq. (12).

$$F1 - Measure = 2\,\frac{precision * recall}{precision + recall}$$

Table 6. Performance for proposed model

Metric	TPR (%)	FPR (%)	Precision	Accuracy
EMUD with SVM	90%	4.90%	91.26%	93.01%

The EMUD algorithm test phishing emails by putting all as input and employed machine learning to evaluate the accuracy and confusion matrix. Specifically, the distribution ratio of phishing and legitimate data is in 60:40 ratio respectively. SVM classifier is used with confusion matrix and k-fold Cross validation (10-fold) is used for accuracy or performance evaluation. After using equation 5, 7, 9 and 10; the proposed EMUD model achieves 93.01% accuracy with 90% of True Positive (TPR) and 4.90% False Positive Rate (FPR) as shown in Table 6.

4.2.3 Comparative Study of Proposed and Existing Model

After using equation 5, 7, 9 and 10 the proposed model-EMUD achieved 93.01% accuracy. The same experiment is also done with existing model i.e. EPCMU with the same testing set.

Table 7. Comparison of proposed model with existing model

Approach	EPCMU	EMUD (Proposed)
TPR (%)	83.63%	90.90%
FPR (%)	13.8%	4.90%
Precision	85.8%	91.26%
Accuracy	84.58%	93.01%

EPCMU achieved 84% in second experiment with SVM classifier. It has high false positive rate in comparison to EMUD. The experiment results are shown in Table 7. Though this experiment, it is obvious that the SVM has better performance result and accuracy than others supervised learning techniques as shown in Table 6 and Table 7.

5. Conclusion

Phishing URLs are challenging threat in cyber space which steal the user's sensitive information. The phishers are using numerous phishing URLs crafting tactics pointing to the same phishing website to bypass the detection techniques. Therefore, a reliable mechanism i.e. Enhanced Malicious URLs Detection (EMUD) model is proposed to combat against aforesaid challenge. In this research paper, supervised machine learning techniques (i.e. NB and SVM) to detect malicious URLs with the EMUD algorithm has been used with EMUD model. EMUD model has more effective detection capabilities as it includes more detection parameter (relevant URL heuristics) to catch and detect malicious URLs in Comparation with other existing URL detection algorithm. But the NB classifier used is not appropriate as the processing time was long. Therefore, the SVM is applied with EMUD for classification and it is evident from experiment, that it has less time in processing and gives better accuracy & results in comparison to others supervised learning techniques. Hence, it is concluded that EMUD model detects the phishing/ obfuscated URLs more accurately with SVM. EMUD model can be more effective by adding latest pertinent heuristics for zero-day phishing detection. Adoption of artificial neural network methods could be more acceptable. Deep neural network will be more promising for phishing detection in terms of enhanced accuracy and performance with large dataset.

References

[1] Hong, Jason. "The state of phishing attacks." Communications of the ACM 55.1 (2012): 74-81.

[2] Shah, Ripan, et al. "A proactive approach to preventing phishing attacks using Pshark." Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on. IEEE, 2009.

[3] Zhang, Yue, Jason I. Hong, and Lorrie F. Cranor. "Cantina: a content-based approach to detecting phishing web sites." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[4] Zhang, Jian, Phillip A. Porras, and Johannes Ullrich. "Highly Predictive Blacklisting." USENIX Security Symposium. 2008.

[5] Sankhwar, Shweta, Dhirendra Pandey, and R. A. Khan, "Phishing: A Critical Review", International pure Applied and Mathematics, ISSN: 1314-3395, Vol. 119 No. 15. 2018, pp. 2917-2923.

[6] Center, RSA Anti-Fraud Command. "RSA monthly online fraud report." (2012).

[7] Sankhwar, Shweta, and Dhirendra Pandey. "Defending Against Phishing: Case Studies." International Journal of Advanced Research in Computer Science 8.5 (2017).

[8] N. Chou, et al.," Client-side defense against web- based identity theft," in In Proc. 11th Annual Network and Distributed System Security Symposium (NDSS '04), San Diego, CA., 2004.

[9] McGrath, D. Kevin, and Minaxi Gupta. "Behind Phishing: An Examination of Phisher Modi Operandi." LEET 8 (2008): 4.

[10] Sankhwar, Shweta, Dhirendra Pandey, and R. A. Khan, "A Glance of Anti- Phish Techniques" International pure Applied and Mathematics, ISSN: 1314-3395, Vol. 119 No. 15. 2018, pp.2925-2936.

[11] Chandrasekaran, Madhusudhanan, Ramkumar Chinchani, and Shambhu Upadhyaya. "Phoney: Mimicking user response to detect phishing attacks." Proceedings of the 2006 International Symposium on on World of Wireless, Mobile and Multimedia Networks. IEEE Computer Society, 2006.

[12] Zhang, Yue, Jason I. Hong, and Lorrie F. Cranor. "Cantina: a content-based approach to detecting phishing web sites." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[13] Fette, Ian, Norman Sadeh, and Anthony Tomasic. "Learning to detect phishing emails." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[14] Sankhwar, Shweta, and Dhirendra Pandey. "A Comparative Analysis of Anti-Phishing Mechanisms: Email Phishing." International Journal of Advanced Research in Computer Science 8.3 Volume 8, No. 3, March – April 2017 (2017).

[15] Suriya, R., K. Saravanan, and Arunkumar Thangavelu. "An integrated approach to detect phishing mail attacks: a case study." Proceedings of the 2nd International Conference on Security of Information and Networks. ACM, 2009.

[16] Sankhwar, Shweta, Dhirendra Pandey, and R. A. Khan. "A Step Towards Internet Anonymity Minimization: Cybercrime Investigation Process Perspective." Information and Decision Sciences. Springer, Singapore, 2018. 257-265.

[17] Center, RSA Anti-Fraud Command. "RSA monthly online fraud report." (2012).

[18] J. Yearwood, et al.," Profiling Phishing E-mails Based extracted from emails," in Soc. Netw. Anal. Min. (2012) 2:5–16

[19] Jayakanthan, N., A. V. Ramani, and M. Ravichandran. "Two phase Classification Model to Detect Malicious URLs." International Journal of Applied Engineering Research 12.9 (2017): 1893-1898.

[20] Harrington, Peter. "Machine learning in action." Shelter Island, NY: Manning Publications Co (2012).

[21] Lotte, Fabien, et al. "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update."Journal of neural engineering 15.3 (2018): 031005.

[22] Manik Sharma, Samriti Sharma, Gurvinder Singh. "Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining". Data 2018, 3, 54.

[23] Kaur, Loveleen, and Ashutosh Mishra. "An Empirical Analysis for Predicting Source Code File Reusability Using Meta-Classification Algorithms." Advanced

Computational and Communication Paradigms. Springer, Singapore, 2018. 493-504.

[24] Gomez, Juan Carlos, Erik Boiy, and Marie-Francine Moens. "Highly discriminative statistical features for email classification." Knowledge and information systems 31.1 (2012): 23-53.

[25] A. G. K. Janecek and W. N. Gansterer, ``E-mail classification based on NMF," in Proc. 9th SIAM Int. Conf. Data Mining (SDM), Sparks, NV, USA, 2009, pp. 1345_1354.

[26] Gansterer, Wilfried N., and David Pölz. "E-mail classification for phishing defense." European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2009.

[27] Abu-Nimeh, Saeed, et al. "A comparison of machine learning techniques for phishing detection." Proceedings.

[28] Manik Sharma, Gurvinder Singh, Rajinder Singh. "Accurate Prediction of Life Style Based Disorders by Smart Healthcare Using Machine Learning and Prescriptive Big Data Analytics." Data Intensive Computing Applications for Big Data 29 (2018): 428.

[29] Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." Information Processing & Management 45.4 (2009): 427-437.

[30] Sankhwar, Shweta, Dhirendra Pandey, and R. A. Khan. "A Novel Anti-Phishing Effectiveness Evaluator Model." International Conference on Information and Communication Technology for Intelligent Systems. Springer, Cham, 2017.

[31] Sharma, M., G. Singh, and R. Singh. "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques." IRBM (2017).