An Efficient Model for Detecting Uniform Resource Locator (URL) Phishing using Machine Learning Techniques

Nkue Dumka[#1]
Department of Computer Science, Rivers State University, Nigeria.
dumkaforgod21@gmail.com

Dr. Daniel Matthias[#2]
Department of Computer Science, Rivers State University, Nigeria.
daniel.matthias@ust.edu.ng

Dr. E.O Bennett[#3]
Department of Computer Science, Rivers State University, Nigeria.
bennett.okoni@ust.edu.ng

ABSTRACT

One of the major challenges that researchers face in the field of e-fraud detection is a rapid shift in fraud trends. There are three key forms of e-fraud, namely: spam email, phishing and infiltration of the network. As many organizations and individuals worldwide have become victims, phishing remains a major threat to global security and the economy. As technology progresses, it is a lucrative trade that will keep blooming. The rapid pace of development of new websites for phishing and distributed phishing attacks has made it difficult to keep blacklists up to date. This paper concentrated on developing an efficient model for detecting phishing URL using machine learning technique. Link Guard algorithm was used to extract real and visual links from the Domain Name System (DNS) and compared the actual link and the visual link to check if the links are the same. Support Vector Machine has been used to train and classify URL into genuine and phishing URL. The system was implemented in Python programming language. Experiments were conducted using two publicly accessible website databases to test the efficiency of the identification of phishing websites. 3200 websites samples were used, 2127 genuine websites and 1036 websites were phishing. The LinkGuard and SVM yielded an accuracy of 98.8%.

Keywords: Efficient Model, Phishing URL, Machine Learning Techniques.

I. INTRODUCTION

Phishing is a type of social engineering in which a phisher is referred to as an intruder. By imitating electronic messages from a trustworthy or public entity in an automatic way, Phisher tries to fraudulently obtain registered users' private or sensitive credentials. The term "phishing" was coined when Internet scammers used email lures to "fish" for Internet users' passwords and financial information from the sea; "ph" is a common hacker replacement for "f," and derives from the primary method of hacking, "phreaking" on telephone switches. Early phishers copied the code from the AOL website and created websites that appeared to be part of America Online (AOL), sending a Standardized Resource Locator (URL) link to this fake web page to take off emails or instant messages, requesting potential victims to reveal their login credentials, especially passwords [1].

For several things such as online shopping, online bill paying, online smart phone recharge, and banking transfers, the internet is now a global site. Customers face numerous security threats, such as cybercrime, because of this use. There are various cybercrimes that are carried out, including spam, theft, cyber bullying, and phishing. Among these crimes, phishing has now become an excellent and very successful cybercrime. A phishing attacker aims to access confidential user data. Phishers design a website that looks the same as any real website and fake users for different purposes to access private user information such as username, password, banking data.

The overall number of phish reported in the first quarter of 2018 was 263,538 according to the Anti-Phishing Working Group (APWG) first quarter survey. This was up 46 percent from the 180,577 found in the 4th quarter of 2017.

It was also higher than the 190,942 seen in the third quarter of 2017. APWG received 262,704 specific phishing accounts in the first quarter of 2018, compared to 233,613 in the fourth quarter of 2017 and 296,208 in the third quarter of 2017.

This study is focused on identifying malicious Uniform resource locator (URL) using machine learning algorithm. The research focused on building an Anti-phishing system on URL features. The system is made up of three phases:

1. Feature selection
2. Feature extraction
3. Classification using machine learning algorithm.

II. LITERATURE REVIEW

According to the Anti-Phishing Working Group, the term phishing was coined in 1996 as a result of social engineering attempts by web scammers against America Online (AOL) accounts [2]. The term phishing comes from fishing, where fishers (i.e. attackers) use a bait (i.e. socially coded messages) to capture prey (e.g. steal personal information of victims). Since one of the first methods of hacking was against mobile networks, and was known as Phone Phreaking, the ph substitution of character f in fishing was born. As a result, ph has become a common hacking character replacement for f. According to the Anti-Phishing Working Group, compromised accounts from phishing attacks were also used as a currency by hackers in 1997, when they traded hacking software for stolen accounts (APWG). Phishing attacks started with the hacking of America On-line (AOL) accounts and progressed to more lucrative targets over time, such as online banking and e-commerce services. Phishing attacks now target not only scheme participants, but also service providers' technical staff, and may use sophisticated tactics such as Man-in-the-Browser (MITB) attacks.

Pratiwi *et al*. [3] proposed a system in Phishing Site Detection Analysis Using Artificial Neural Network, analyze phishing sites using the perceptron algorithm of the neural network. The Dataset was collected from the archive of the University of California, Irvine (UCI), using 2,455 phishing sites and split into two parts; training data as much as 80% or 1964 data and evaluation data as much as 20% or 491. There are 18 features in the dataset, namely: IP address, Uniform Resource Locator (URL) length, service shortening, @ mark, double slash redirection, prefix suffix, subdomains, Stable Socket Layer (SSL) final state, domain registration length, favicon, port, Secure Hyper Text Transmission Protocol (HTTPS) token, request a Uniform Resource Locator (URL), the Uniform Resource Locator (URL) of anchor, links in tags, SFH, submitting to email and abnormal Uniform Resource Locator(URL). They obtained results 83.38% accuracy, precision 83.36% and recall 83.36%... not clear.

Desai *et al*. [4] proposed a system in Malicious Web Content Detection Using Machine Leaning, With the support of machine learning algorithms, a Google Chrome plugin was developed to identify content from phishing websites. University of California, Irvine (UCI) Dataset-Used Machine Learning Library and derived 22 characteristics for this dataset. For consistency, recall, f1-score and precision comparison, KNN, SVM and Random Forest algorithms were selected. The best score was achieved by Random Forest and Hypertext Markup Language (HTML), JavaScript, Cascading Style Sheet (CSS) used along with python for applying chrome extension. This extension has a downside with a declared list of malicious sites that grows regularly.

Machado *et al*. [5] proposed a system in Phishing Sites Detection Based On C4.5 Decision Tree Algorithm, It recommends an appropriate way to use the c4.5 decision tree approach to identify phishing Uniform Resource Locator (URL) websites. This approach extracts functionality and measures heuristic values from the pages. To decide whether the site is phishing or not, these values were given to the C4.5 decision tree algorithm. Phish Tank and Google are collecting the dataset. This approach entails two stages: the pre-processing stage and the detection stage. In which characteristics are extracted based on pre-processing step rules and the characteristics and their valued values were entered into the algorithm c4.5 and achieved 89.40 percent accuracy.

Parekh *et al*. [6] proposed a system using New Method for Detection Of Phishing Websites: Uniform Resource Locator (URL) Detection, Using the Uniform Resource Locator (URL) recognition technique using the Random Forest algorithm, a model with response was proposed to identify phishing sites. It has three steps, namely Parsing, Output Measurement, Heuristic and Data Classification. For evaluating the feature collection, parsing is used. The phish tank dataset was collected. Just 8 characteristics out of 31 features are considered for parsing. The random forest approach obtained a 95 percent precision level.

Shrivas *et al*. [7] proposed a system using Decision Tree Classifier for Classification Of Phishing Website With Information Gain Feature, collected phishing data set from University of

California, Irvine (UCI) The phishing data set was stored in a cache, and the quick miner function was used to compare decision trees, random trees, and random forest algorithms for phishing and non-phishing classification. When compared to other algorithms, decision tree has the highest accuracy (91.8%), followed by random tree (66.7%), random forest (78.8%), and decision stump (84%).

Sönmez *et al*. [8] proposed a phishing Web Sites Features Classification, Based on Extreme Learning Machine,proposes a classification model to classify phishing attacks. This model utilizes website extraction and grouping of features. In the extraction of features, 30 features were taken from the machine learning repository data collection of the University of California, Irvine (UCI) and the extraction rules for phishing features were clearly specified. Help Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM) algorithms were used to characterize these characteristics. Six activation functions were used in the Extreme Learning Machine (ELM) and attained 95.34 percent precision than SVM and Naïve Bayes NB. With the aid of MATLAB, the results were collected.

Sudhanshu *et al*. [9] proposed a system in Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison, The association's data mining method was used. They also suggested a rule-based classification system for the identification of phishing websites. Thanks to their basic rule transformation, they have concluded that the association classification algorithm is better than all other algorithms. By extracting 16 features, they achieved 92.67 percent accuracy, but this is not up to mark so that the proposed algorithm can be improved for successful detection rate.

Features of a Phishing URL

Since phishing involves tweaking legitimate Uniform Resource Locator (URL) over time, some patterns have been adopted by attackers in tweaking Uniform Resource Locator (URL). This section examines some of the common Uniform Resource Locator (URL) tweaking patterns as follows:

i. **Uniform Resource Locator (URL) IP address presence**: Most innocuous domains do not use the Uniform Resource Locator (URL) IP address to download a webpage. Using the IP address in the Uniform Resource Locator (URL) means the confidential information is being stolen by the intruder.

ii. **Presence of the @ symbol in the Uniform Resource Locator (URL):** Phishers apply a special @ symbol to the Uniform Resource Locator (URL) that allows the browser to disregard anything that precedes the "@" symbol and the "@" symbol is sometimes accompanied by the actual address.

iii. **Amount of Host Name Dots**: Phishing Uniform Resource Locator (URL) have many dots in URL. For example, http://shop.fun.amazon.phishing.com, in this URL The real domain name is phishing.com, while the use of the term "amazon" is to confuse people into clicking on it. The Benevolent Standardized Resource Locator average number of dots is 3.

iv. **Domain-separated prefix or suffix:** The dash symbol is occasionally found in a legal Standardized Resource Locator: Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example, Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

v. **Uniform Resource Locator (URL) redirection**: Within the URL path, the presence of '//' indicates that the consumer would be routed to another website.

vi. **Hyper Text Transfer Protocol Secure (HTTPS) token in Uniform Resource Locator (URL):**Phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-mpp-home.soft-hair.com

vii. **Email Submission Information**: To redirect user details to his personal account, Phisher will use the functions "mail()" or "mailto:."

viii. **Uniform Resource Locator (URL) Shortening Services “Tiny URL”**: Tiny URL Phisher's service makes long phishing URL to be obscured by keeping it short. The aim is to divert users to phishing channels.

ix. **Host name range:** The total length of the URLs of the benign Uniform Resource Locator is calculated to be 2510.

x. **Presence of Sensitive Words in URL**: Phishing sites use sensitive terms in URL to trick users into thinking they're dealing with a legitimate website. The terms 'confirm ',' account ',' banking ',' secure ',' ebyisapi ',' webscr ',' signin ',' mail ',' install ',' toolbar ',' backup ',' paypal ',' login ',' username ', etc. may be contained up many phishing URLs.

xi. **Uniform Resource Locator Slash Number:** The number of slashes used in benevolent URLs is 5.

xii. **Presence of Unicode in Uniform Resource Locator:** Unicode characters in URL can be used by phishers to trick users into clicking on them. The "xn—80ak6aa92e.com" domain, for instance, is similar to "apple.com." The user's visible URL is "apple.com," but the user can access "xn—80ak6aa92e.com," which is a phishing site, after clicking on this URL.

xiii. **Age of Stable Socket Layer (SSL) Certificate**: In providing the impression of website credibility, the presence of Hyper Text Transmission Protocol Secure (HTTPS) is very critical. But the minimum age of the benevolent website Stable Socket Layer (SSL) certificate is between 1 year and 2 years.

III. PROPOSED ALGORITHM AND METHODOLOGY

The methodology adopted for this study was constructive research. It is a discipline that explores how analysis is done. The constructive study contribute to a new technique, Algorithm, framework, domain for solving specific problem and a study work plan. An efficient model was developed using the combination of Support Vector Machine(SVM) and Link guard Machine Technique to detect phishing URL.

Link Guard

Link Guard works by analyzing the variations between the visual link and the real link. The algorithm is built on real-world visual and interaction data. The actual relation is the one to which the user is guided as they select a link, while the visual link is one that the user can see. The link Guard algorithm derives true and visual connections from the Domain Name System (DNS) and compares them to see if they are the same. Furthermore, the LinkGuard algorithm checks if the Blacklist or Whitelist databases have a valid domain name. As a result, though White list has a real URL, Blacklist has a phishing URL.

Link Guard Algorithm

```
v_dns = GetDNSName(v_link);
a_dns = GetDNSName(a_link);
if v_dns and a_dns are not empty and v_dns!=a_dns
return PHISHING and goto end
if a_dns is dotted decimal
return possible_PHISHING and goto next phase image based webpage matching
if a_linkorv_linkis encoded
v_link2= decode (v_link);
a_link2= decode (a_link);
returnLinkGuard(v_link2, a_link2);
if v_dnsis NULL
returnAnalyzeDNS(a_link);
end
```

However, v_link is visual link (the link that is seen by the user).a_link is actual link (the link to which user is redirected when clicked). v_dns is visual Domain Name System (DNS) name. a_dns is actual DNS name.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning system that has three stages: URL feature extraction, training phase, and classification.

SVM Algorithm for Websites Classification

```
i.     Import and Preprocess Dataset.
ii.    Extract the features of URL
iii.   Compute attribute values.
iv.    if Attribute X present value = 1
v.     then legitimate is detected
vi.    end if
vii.   else if Attribute Y absent value = -1
viii.  then phishing is detected
ix.    End if
x.     Select attribute X and Y
xi.    Compute threshold value for attribute X and Y
xii.   Find Range value.
xiii.  Select Attribute to get threshold value.
xiv.   Classify phishing and legitimate site using attribute value.
xv.    Compute Sensitivity and Specificity.
```

IV. RESULTS AND DISCUSSION

The security check in websites is achieved through URL and webpage matching. websites were gotten online, the legitimate and phishing websites. 3200 websites samples were used, 2127 were genuine websites and 1036 websites were phishing as shown in Table 1. The performance accuracy of LinkGuard and support vector machine systems were estimated using precision, recall and accuracy.

Table1: URL Performance Rate with LinkGuard+SVM

Total URL	Phishing	Legitimate	FN	FP	Precision	Recall	Accuracy
	TP	TN					
3200	1036	2127	23	14	98.6%	97.8%	98.8%
3200	1062	2091	28	19	98.2%	97.4%	98.5%

The LinkGuard and SVM were tested in URL having a total number of 3200, precision of 98.6%, recall of 97.8% and accuracy of 98.8%. The evaluations produced good results. The LinkGuard and SVM yielded a classification accuracy of 98.8% than the existing system (support vector machine) with an accuracy of 96.30%.

We have tested 1005 web pages and analyzed the performance on the system for various thresholds as depicted in Table 2. To evaluate the effectiveness of our solution, we have used three metrics such as precision, recall and accuracy.

Precision is the proportion of the number of web pages that are correctly detected as phishing to the number of total detected URL. Formular: Precision = TP/(TP+FP) *100%

The proportion of URLs accurately identified as phishing to the total number of phishing samples is known as recall. Formular : Recall = TP/(TP+FN) *100%

The proportion of URLs correctly identified as phishing or genuine to the total number of sample URLs is known as accuracy. Formular: Accuracy = (TP+TN)/(TP+TN+FP+FN) *100%.

However, when tested in webpages and images containing hyperlinks in actual URL and visual URL form, the recall and accuracy decreased, having a recall of 86% and accuracy of 88.8%.

Table 2: Web pages Performance Rate

Total URL	Phishing TP	Legitimate TN	FN	FP	Precision	Recall	Accuracy
1005	593	300	96	16	97.3%	86%	88.8%
928	358	511	52	7	98%	87.3%	93.6%

The comparison analysis of proposed and existing systems; LinkGuard+SVM and SVM were evaluated using precision, recall and accuracy as depicted in Table 3.

Table3: Comparison Performance of Proposed System (LinkGuard+SVM) and Existing System (SVM)

Models	Precision	Recall	Accuracy
LinkGuard+SVM	98.6%	97.8%	98.8%
SVM	95.9%	94.6%	96.30%

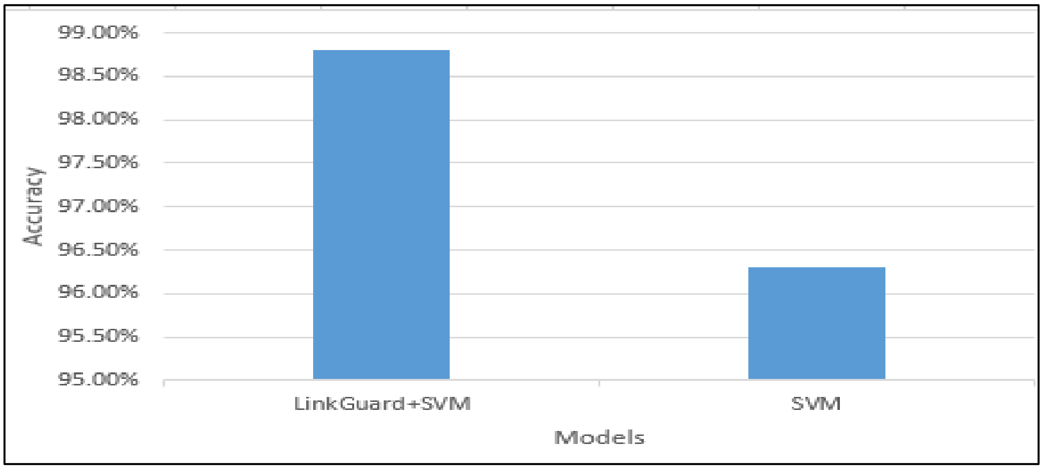

Figure 1: Comparison of Performance of LinkGuard+SVM and SVM using Accuracy

V. CONCLUSION

Due to the huge financial losses suffered by many organizations and individuals due to online fraud, the urgent need for a robust, secure, fast and reliable e-fraud detection system can not be overemphasized. One of the major challenges that researchers face in the field of e-fraud detection is a rapid shift in fraud trends. There are three key forms of e-fraud, namely: spam email, phishing and infiltration of the network. This study concentrates on detecting phishing. As many organisations and individuals worldwide have become victims, phishing remains a major threat to global security and the economy. As technology progresses, it is a lucrative trade that will keep blooming. The rapid pace of development of new websites for phishing and distributed phishing attacks has made it difficult to keep blacklists up to date. A detection technique that can adapt to new trends of fraud is therefore highly needed.In this report, we used support vector machine and LinkGuard algorithms to detect phishing websites, with the primary goal of creating a complex and stable phishing detection system with increased classification accuracy and decreased erroneous phishing detection.

The LinkGuard algorithm based on visual and real interactions. The real connection is the connection to which the user is routed when a link is clicked, and the visual link is a noticeable link to the user. The Link Guard algorithm extracts real and visual links from the Domain Name System (DNS) and compares the actual link and the visual link to check if the links are the same. In addition, the LinkGuard algorithm verifies whether Blacklist or Whitelist databases contain an actual domain name. Therefore, though Whitelist contains a valid URL, Blacklist contains a phishing URL.

The machine learning method known as Support Vector Machine (SVM) consists of three major stages: feature extraction of URL, training phase and URL classification phase. In the feature extraction phase, the feature vectors of all the features were extracted and saved in a file for easy access. However, SVM was used to carry out the classification.

REFERENCES

[1]. Nivedha, S.,Gokulan,K. C.,Gopinath,R. and Gowshik,R.(2017). *Improving phishing URL Detection using Fuzzy Association mining*,2(3), 12319-1805.

[2]. APWG “Phishing activity trends report 3rd quarter.”*US.* Vol 1Issue 11, 2018.

[3]. Pratiwi, M.E., Lorosae, T. A. and Wibowo, F. W. (2018). “Phishing Site Detection Analysis Using Artificial Neural Network IC-ELINVO IOP Conference Series: *Journal of Physics: Conference Series 1140 (2018) 012048 IOP Publishing doi:10.1088/1742-6596/1140/1/012048.*

[4]. Desai, J. A., Jatakia, N. and Raul, A. (2018). Malicious web content detection using machine leaning. *2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol.*, 2(3), 1432–1436.

[5]. Machado, L. and Gadge J. (2017). Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” *International Conference on Computing, Communication, Control and Automation*, 1, 1–5.

[6]. Parekh S., Parikh, S. D. Kotak, and Sankhe, P. S. (2018). A New Method for Detection of Phishing Websites: URL Detection,” in *2018 Second International Conference on Inventive Communication and Computational Technologies* (*ICICCT*), 949–952.

[7]. Sönmez, Y. T., Tuncer, H. Gökal, and Avci, E. (2018). Phishing web sites features classification based on extreme learning machine,” *6th Int.Symp. Digit. Forensic Security*,1, 1–5.

[8]. Shrivas , A. K. andSuryawanshi, R. (2017). Decision tree classifier forclassification of phishing website with info gain feature selection. *Int. J. Res. Appl. Sci. Eng. Technol.*, 5(5), 780–783.

[9]. Sudhanshu, G., Kritika, R. and Bansidhar. J. (2018). Detecting Phishing Websites Using Rule-Based Classification Algorithm.