Detection and Prevention of Phishing Websites using Machine Learning Approach

Vaibhav Patil
Dept. of Computer Engineering
Sinhgad Academy of Engineering
Pune, India
vaibhav95.patil@gmail.com

Pritesh Thakkar
Dept. of Computer Engineering
Sinhgad Academy of Engineering
Pune, India
priteshthakkar53@gmail.com

Chirag Shah
Dept. of Computer Engineering
Sinhgad Academy of Engineering
Pune, India
chirag041@gmail.com

Tushar Bhat
Dept. of Computer Engineering
Sinhgad Academy of Engineering
Pune, India
tusharbhat002@gmail.com

Prof. S. P. Godse
Dept. of Computer Engineering
Sinhgad Academy of Engineering
Pune, India
sachin.gds@gmail.com

***Abstract*—Phishing costs Internet user's lots of dollars per year. It refers to exploiting weakness on the user side, which is vulnerable to such attacks. The phishing problem is huge and there does not exist only one solution to minimize all vulnerabilities effectively, thus multiple techniques are implemented. In this paper, we discuss three approaches for detecting phishing websites. First is by analyzing various features of URL , second is by checking legitimacy of website by knowing where the website is being hosted and who are managing it, the third approach uses visual appearance based analysis for checking genuineness of website. We make use of Machine Learning techniques and algorithms for evaluation of these different features of URL and websites. In this paper, an overview about these approaches is presented.**

Keywords— phishing, security, blacklist, whitelist, URL, anti-phishing, web-page

I. INTRODUCTION

Phishing is a social engineering attack that aims at exploiting the weakness found in the system at the user's end. For example, a system may be technically secure enough for password theft but the unaware user may leak his/her password when the attacker sends a false update password request through forged (phished) website. For addressing this issue, a layer of protection must be added on the user side to address this problem.

A phishing attack is when a criminal sends an email or the url pretending to be someone or something he's not, in order to get sensitive information out of the victim. The victim in regard to his/her curiosity or a sense of urgency, they enter the details, like a username, password, or credit card number, they are likely to acquiesce. The recent example of a Gmail phishing scam that targeted around1 billion Gmail users worldwide.

The Fig. 1 looks exactly like a Gmail sign-in form, the URL is slightly changed, but it's not the . Filling in this form would give the attacker full access to the victim's Gmail account. The kind of theft and fraud that could take place by just acquiring the details of someone's or some organizations' account couldn't really be imagined. All other account are controlled by the Gmail account. That could be a huge threat. Microsoft Outlook fraud is the second-most targeted and Google drive being the third. Other targets are facebook, bank logins and paytm, paypal etc.

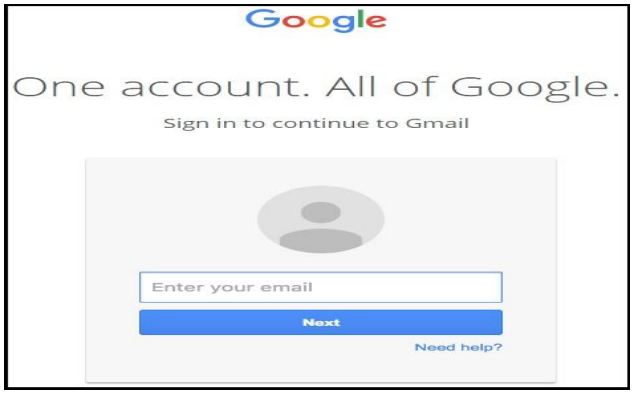

Fig. 1. Gmail Phishing Scam Url

II. RELATED WORK

Many researchers have previously been carried out in this field of phishing detection. We have gathered the information from various such works and have profoundly reviewed them which has helped us in motivating our own methodologies in the process of making a more secure and accurate system.

A. Blacklist Approach and Whitelist Approach

In [13], Pawan Prakash, Manish Kumar, Ramana Rao Kompella, Minaxi Gupta (2010) proposed a predictive blacklist approach to detect phishing websites. It identified new phishing URL using heuristics and by using an appropriate matching algorithm. Heuristics created new URL's by combining parts of the known phished websites from the available blacklist. The matching algorithm then calculates the score of URL .If this score is more than a given threshold value it flags this website as phishing website. The score was evaluated by matching various parts of the URL against the URL available in the blacklist.

In [14], Jung Min Kang and DoHoon Lee described approach which detected phishing based on users online activities. This method maintained a white list as a part of users' profile. This profile was dynamically updated whenever a user visited any website. An engine used here identified a website by evaluating a score and then comparing it with a threshold score. The score was calculated from the entries available in the user profile and details of the current website.

B. Heuristic Approach

In [7], Aaron Blum, Brad Wardman, Thamar Solorio proposed a work which focused on the exploration of surface level features from URLs to train a confidence-

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

weighted learning algorithm. The idea is to restrict the source of possible features to the character string of the URL and avoid having the vulnerability of extracting host-based information. Every URL is displayed as a vector of binary feature. These vectors are fed to the online algorithm where at time of testing, previously unseen URLs in the binary feature vector is then mapped to it. The learner continues this new vector and output into the final result, either phish or non phish.

In [15], Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor proposed CANTINA+, a comprehensive feature-based approach inthe literature including eight novel features, which exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phish. Also two other filters are designed in it to help reduce FP and achieve good runtime speedup. The first is a near-duplicate phish detector that uses hashing to catch highly similar phish. The second is a login form filter, which directly classifies webpages with no identified login form as legitimate.

In [8] , Joby James , Sandhya L, Ciza Thomas proposed a work which with the combined help of blacklisting approach and the Host based Analysis applied certain classifiers which can be used to help detect and take down various phishing sites. The host based, popularity based and lexical based feature extractions are applied to form a database of feature values. The database is knowledge mined using different machine learning methods. After evaluating the classifiers, a particular classifier was selected and was implemented in MATLAB.

In [9], APWGM published a case study citing the importance of the WHOis tool and how invaluable it has been for the rapid phishing site shutdown over the past few years all around the globe.

C. Visual Similarity Approach

In [2], A. Mishra and B. B. Gupta presented a hybrid solution based on URL and CSS matching. In this approach it can detect embedded noise contents like an image in a web page which is used to sustain the visual similarity in the webpage. They used the technique used in [3] by Jian Mao, Pei Li, Kun Li,Tao Wei, and Zhenkai Liang to compare the CSS similarity and used it in their technique. The different types of visual features are - text content and text features. Text features are like font colour, font size, background colour, font family and so forth. This approach matches the visual features of different websites because the attacker copies the page content from the actual website.

In [5] Matthew Dunlop, Stephen Groat, and David Shelly proposed a browser based plug in called goldphish to identify phishing websites. It uses the website logos to identify the fake website. The attacker can use the real logo of the target website to trap the internet users. Three stages to it is:

- *Logo Extraction :* Goldphish is used to extracts the website logo from the suspicious website. Then it converts it into text using optical character recognition (OCR) software.
- *Legitimate website extraction* **:** The text obtained is used as a query for the search engine. Generally, search engine "google" is used because it always return genuine websites in their top results.
- *Comparisons :* Suspicious website is compared with the top result obtained from the search engine based on different features. If any domain is matched with the current website then it is declared legitimate or else make it phishing site.

III. PROPOSED WORK

A. Overview of our approach

Out of all the previous work, only the blacklist and whitelist are implemented which has a drawback of not being updated in long time. The basic idea of our proposed solution is the hybrid solution which uses all the three approaches – blacklist and whitelist, heuristics and visual similarity. Our proposed system has the following algorithm.

1. Monitor all "http" traffic of end-user system by creating a browser extension. The benefit of an extension over an application or software is that the system will be based purely in real time and at the same time will also be quite agile in delivering the outputs.
2. Compare domain of each URL with the white-list of trusted domains and also the black-list of illegitimate domains. The data required for both the lists would be extracted dynamically by web scraping and stored on the server. If domain of the URL is found under the white-list, mark the URL as innocent (Exact Matching), else go further and use the other approaches.
3. Furthermore, the whole website analysis would now be done by considering various details (features). The set of features we took are : website protocol (secure or unsecure), length of the URL, number of hyphen (-) in URL, number of @ symbol in URL, number of dots in the URL, using direct IP address or not, alexa rank, bounce rate, daily page view, whois availability, registration and expiration date of website, alexa.com availability, rank2traffic.com availability, siterankdata.com availability, daily unique visitor, favicon icon similarity and google indexing.
   Example :
   If hyphen in URL > 1 – Phished website
   If hyphen in URL = 1 – Suspicious website
   If hyphen in URL < 1 – Legitimate website
   All the feature take into consideration at the same time increases the accuracy of the system.
4. Intuitively, the higher similarity between the phishing page and the target page indicates a greater chance of the users being deceived. This is the reason, attackers always try their best to clone the target pages.
5. To counter such antics, our next approach would be to extract and compare CSS of suspicious URL and compare it with the CSS of each of the legitimate domains in queue. This method will look into visual based features of the phished websites.
6. The machine learning classifiers such as decision tree, logistic regression and random forest will be applied to the collected data and a score is generated.
7. The match score and similarity score is calculated. If the score is greater than threshold then we mark the URL as phishing and block it.

8. This approach basically provides a three level security block and hence can prove to be more effective and accurate than any of the other existing systems

B. Requirement Analysis

The System which deals with providing security concern using new and effective technology like Machine Learning with the help of user's personal computer and the browser extension.

(i) Software Requirements

- Python 3.6
- Beautiful Soup (Package in Python)
- Scikit-learn (Package in Python)
- JavaScript
- Browser (Chrome)

(ii) Hardware Requirements

- Windows 7 above
- Hard disk of at least 64 GB

C. Design Phase

The flow of the proposed system is shown in Fig. 2.

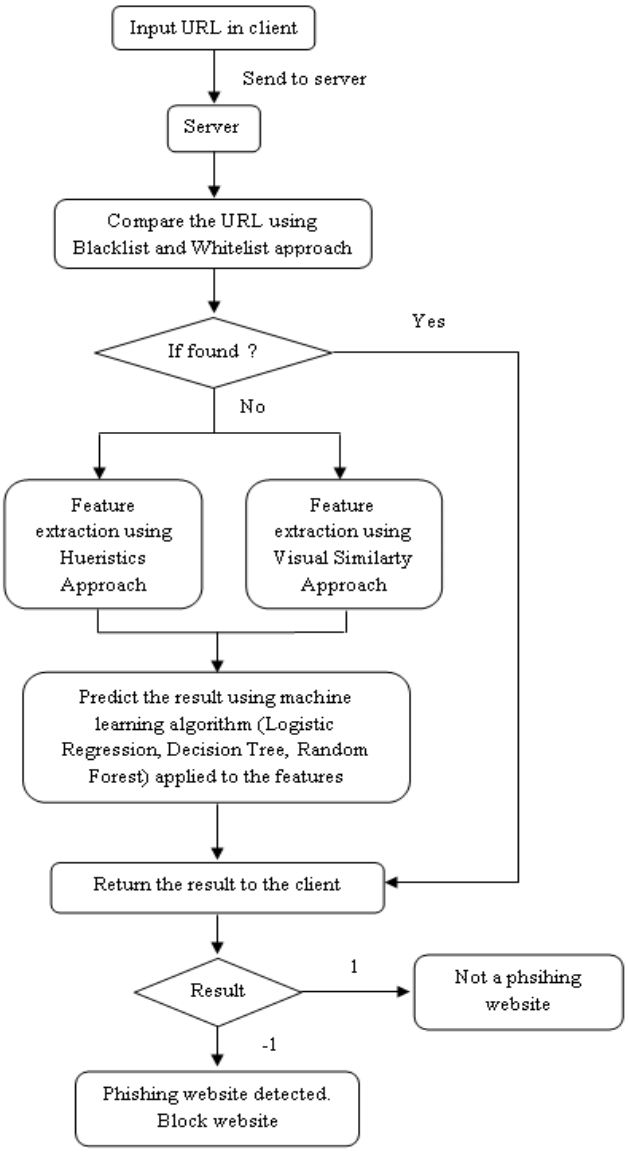

Fig. 2. Flow chart of the proposed system

D. Analysis Phase

For different features we put different rules based on the analysis of phished and non-phished website scraped over from internet.

For example Fig. 3 and Fig. 4 shows the hyphen count of the phished and legitimate websites respectively. Y axis denotes count of websites and X axis denotes count of hyphens in the website. Based on this analysis, we concluded that phished websites do consist of hyphen in the domain part of the URL and legitimate websites don't.

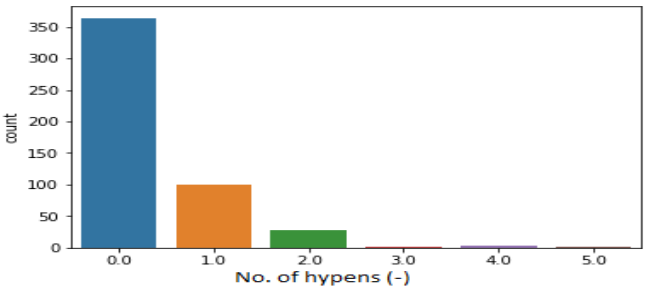

Fig. 3. Hyphen count of phished websites

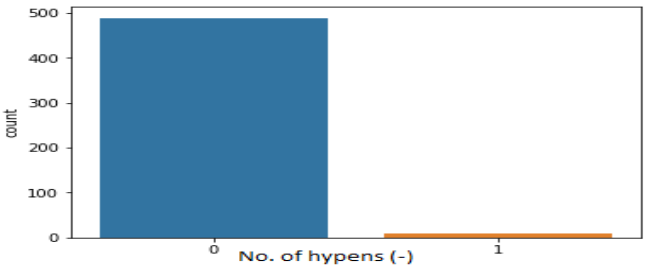

Fig. 4. Hyphen count of legitimate websites

Following are the detailed rules that we created based on analysis :

1. Has protocol ?
   If yes then legitimate, else suspicious
2. Length of domain in url
   If length in between 3 to 20, then legitimate
   Else if length in between 20 to 24, then suspicious
   Else if length greater than 24, then phished.
3. Number of hyphen in domain
   If number of hyphen is 0, then legitimate
   Else phished
4. @ symbol in domain
   If number of @ symbol is 0, then legitimate
   Else phished
5. In between domain the keyword 'http'
   If 'http' found in domain, then phished
   Else legitimate
6. Direct IP Address
   If url is a numeric IP address, then suspicious
   Else legitimate
7. Alexa.com, rank2traffic.com, and siterankdata.com availability

If the website is available in the database of any of these website, then legitimate
Else suspicious

8. Time difference of date of expiration and data of registration of the website
   If time difference is greater than 90 days then legitimate
   Else suspicious
9. Daily unique visitors
   If daily unique visitor details are available on internet, then legitimate
   Else suspicious
10. Google indexing using title
    If the title of the website queried on google search engine shows the exact same url of website in the top results, then legitimate
    Else suspicious
11. Google indexing using url
    If the url of the website queried of google search engine shows the exact same url of website in top result, then legitimate
    Else phished
12. Favicon similarity using google indexing
    If the favicon of the two websites are similar and domain of url is different, then phished
    Else legitimate

All these features combined together will lead to accurate results.

IV. Result

The linear regression plot of expected output versus predicted output is show in Fig. 5. This was predicted by the random forest algorithm. It has a slight deviation from the expected output for the phished websites.

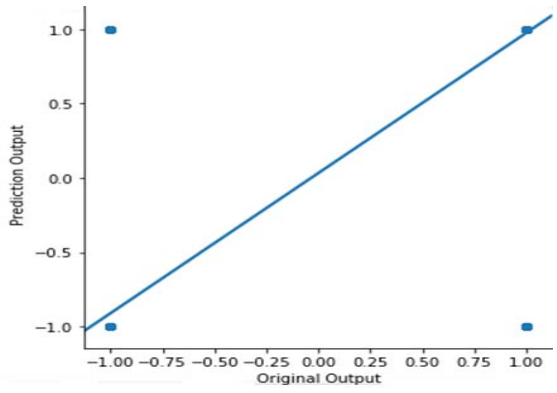

Fig. 5. Linear regression plot of original output versus predicted output

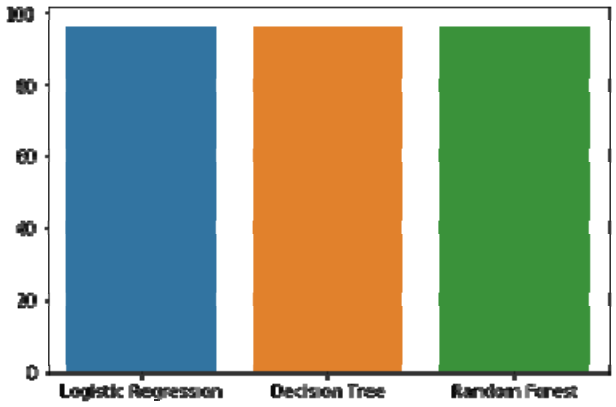

Fig. 6. Machine learning accuracy bar plot

The true positive, false positive, true negative, false negative count and accuracy results of 9076 test websites is as shown in Table I.

TABLE I. Confusion Matrix Results

Algorithm	TN	TP	FP	FN	Accuracy
Logistic Regression	6447	2287	325	17	96.23 %
Decision Tree	6393	2341	326	16	96.23 %
Random Forest	6392	2374	297	13	96.58 %

V. Conclusion

The proposed system enables the internet users to have a safe browsing and safe transactions. Its helps users to save their important priivate details that should not be leaked. Providing our proposed system to users in the form of extension makes the process of delevering our system much easier. The results points to the efficiency that can be achieved using the hybrid solution of hueristic features, visual features and blacklist and whitelist approach and feeding these features to machine learning algorithms. A particular challege in this domains is that criminals are constantly making new strategies to countrer our defense meausres. To succeed in this context, we need algorithms that continually adapt to new examples and features of phishing URL's. And thus we use online learning alogrithms. This new system can be designed to avail maximum accuracy. Using different approaches altogether will enhance the accuracy of the system, providing an efficient protection system. The drawback of this system is detecting of some minimal false positive and false negative results. These drawbacks can be eliminated by introducing much richer feature to feed to the machine learning

algorithm that would result in much higher accuracy.

REFERENCES

[1] Ankit Kumar Jain and B. B. Gupta, “Phishing Detection Analysis of Visual Similarity Based Approaches”, Hindawi 2017.

[2] A. Mishra and B. B. Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks”, ERCICA 2014.

[3] Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang, “Bait Alarm Detecting Phishing Sites Using Similarity in Fundamental Visual Features”, INCS 2013.

[4] Eric Medvet, Engin Kirda and Christopher Kruegel, “Visual-Similarity-Based Phishing Detection”, ACM 2015.

[5] Matthew Dunlop, Stephen Groat, and David Shelly, “GoldPhish Using Images for Content-Based Phishing analysis”, IEEE 2010.

[6] Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, “Textual and Visual Content-Based Anti-Phishing A Bayesian Approach”, IEEE 2011

[7] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner; “Lexical Feature Based Phishing URL Detection Using Online Learning”, Department of Computer and Information Sciences The University of Alabama at Birmingham, Alabama, 2016

[8] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, MinaxiGupta,Purdue University, Indiana University "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".

[9] The Anti-Phishing Working Group, DNS Policy Committee;” Issues in Using DNS Whois Data for Phishing Site Take Down”,The Anti-Phishing Working Group Memorandum, 2011.

[10] Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor ,”CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites”, School of Computer Science Carnegie Mellon University, ACM Society of computing Journal, 2015.

[11] Joby James,SandhyaL,Ciza Thomas "Detection of phishing websites using Machine learning techniques", 2013 International Conference on Control Communication and Computing (ICCC).

[12] Mohsen Sharifi and Seyed Hossein Siadati "A Phishing Sites Blacklist Generator".

[13] JungMin Kang and DoHoon Lee "Advanced White List Approach for Preventing Access to Phishing Sites".

[14] Y. Zhang, J. I. Hong, and L. F. Cranor. Cantina: a content-based approach to detecting phishing web sites. In WWW ’07: Proceedings of the 16th international conference on World Wide Web, pages 639–648, New York, NY, USA, 2007. ACM.