Detection of Malicious Websites using Machine Learning Based on URL

Durgesh Kadwe[1] Prasad Pawar[2] Shirode Nikhil[3] Daradi Kajol[4]
[1,2,3,4]Amrutvahini College of Engineering, Sangamner, India

Abstract— Phishing is one kind of cyber-attack and at the same time, it is most dangerous and common attack to acquire personal information, account details, organizational details, credit card details or password of a user to conduct transactions. Phishing websites look similar to the appropriate ones which is difficult to differentiate between them. The motive of this study is to perform Extreme Learning Machine (ELM) based on different 30 features classification using Machine Learning approach. Most of the phishing URL's use HTTPS to avoid getting detected. There are three approaches for detection of phishing websites. The first approach analyzing different features of URL, second approach checking legitimacy of website and knowing where the website is hosted or not and it also check who are managing it, third approach checking genuineness of website.

Keywords: Phishing, Extreme Learning Machine, Features Classification, URL, Information Security

I. INTRODUCTION

Phishing is a malicious attack in online theft to steal the user personal information. It is a type of fraud in which attacker tries to gain user private information and thus user falls into such traps. The aim of our study is to propose framework which is safe for detecting phishing websites in less time with high accuracy. Now a days, people carry out most online transactions, transferring money, paying the bills i.e. everything is done through websites or applications. Therefore, identifying phishing websites is the great importance in our day to day life. Detecting the phishing websites is a challenging task[1]. After a detail survey on this problem we found the list-based anti-phishing approaches (blacklist or whitelist) which stores URLs in the database. This approach is used to compare the URLs which are stored in the database with the URL entered by users in browsers. The URLs which are not being included in the database i.e. the newly created phishing URLs are fail to detect using this approaches[2]. A phishing attack is occur, when a criminal sends an email or the URL in order to get sensitive information of users for misuse. The victim in regard to sense of urgency and they enter the details like username, password or credit card number they are likely to accept.

The Fig. 1 looks exactly like a Gmail sign-in form, but the URL is somewhat changed. But it not filling the Gmail sign-in form would give the attacker gains full access. The kind of fraud and theft that could takes place by just gaining the detail of users. Gmail account controls all other accounts. So, this could be a huge fraud. Other targets are Facebook, Bank logins and Paytm, Microsoft Outlook etc[3].

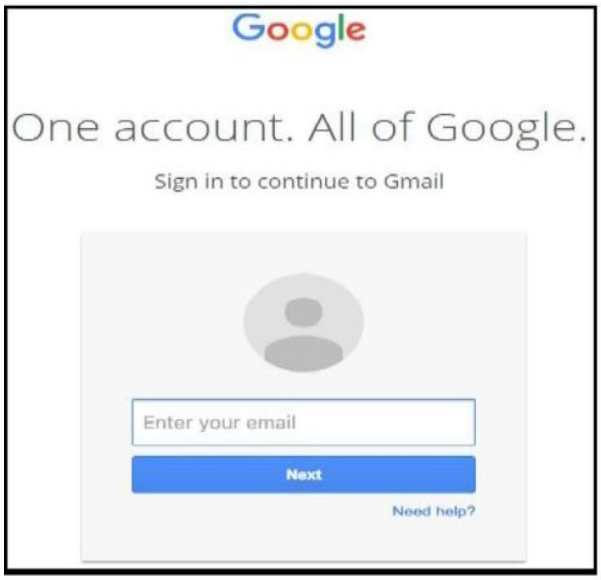


Fig. 1: Gmail Phishing Scam URL

II. LITERATURE SURVEY

Phishing is a security attack which is most common and dangerous attack to gain personal information, account details, credit card details or password of a user to conduct a transaction.

et.al Srushti Patil and Sudhir Dhage[1] uses different methods like Anti Phishing solutions. Anti Phishing solutions includes various approaches. Heuristic Approach is used to classifying the URLs. Features are extracted and they are classified using Machine Learning Techniques. Different techniques are combined to check whether the website is fake or real.

et.al Huaping Yuan, Xu Chen and Yukun Li[2] uses different algorithms for detecting the phishing websites. Different machine learning algorithms on phishing detection including k-Nearest Neighbor(KNN), Logistic Regression(LR), Random Forest(RF), Decision Tree(DT),

Gradient Boosting Decision Tree (GBDT), XGBoost(XGBST), and Deep Forest(DF).Authors introduce the statistical features and lexical features of URLs and links.

et.al Vaibhav Patil, Pritesh Thakkar and Chirag Shah[3] proposes hybrid solution which uses three approaches – blacklist and whitelist, heuristics and visual similarity. This approaches provides a three level security blocks and hence this system is more effective and accurate.

et.al Anu Vazhayil, Vinaya Kumar R and Soman KP[7] focuses on combination of CNN with the Long Short Term memory(LSTM) and Convolutional Neural Network(CNN) to derive the accuracy in classifying the phishing URLs. LSTM extracts sequential information and CNN helps to extract special information among the characters. CNN used to learn the special co-relationship among the characters.

et.al Martyn Weedon and Dimitris Tsaptsinos[8] focuses on the Random Forest(RF) algorithm to classify URLs are either malicious or gentle. The distribution of URL will be lexical base, which means features directly will be extracted from the URL itself.

et.al R. Dhamija and J. D. Tygar[9] proposed another solution which is called as Dynamic Security Skins. This method used shared secret image that allows a remote server to prove an identity of an user. In such a way that it is very easy for user to verification, but hard for attacker to spoof.

III. RELATED WORK PHISHING

Phishing is the process of an Internet fraud. Phishing is a type of technique that utilizes a combination of technology and social engineering to gather personal and sensitive information such as online shopping like selling or purchasing products, sending mail, chat with friends etc.

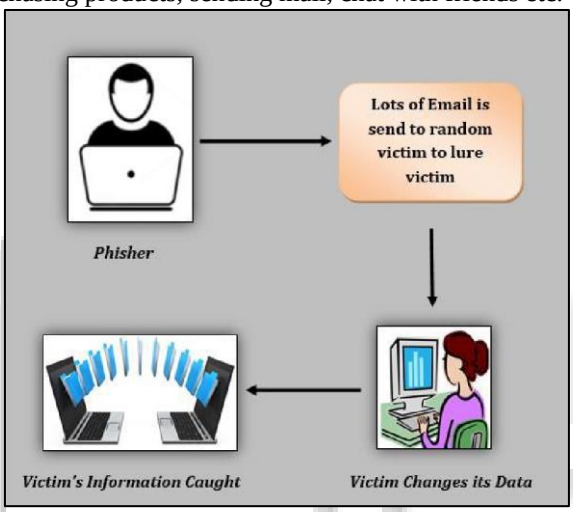


Fig. 2: Process of Phishing

In general, phishing attacks are done with the following four steps:

1) A fake website which looks exactly like the legitimate website which is set up by the phisher.
2) Phisher then send link to the fake website to target the users in the name of legitimate organizations and companies, trying to convince the potential victims to visit their web sites.
3) By clicking on the link, victim visit the fake website and gives personal information there.

Phishers then steal the personal information which is entered by the victim and perform their fraud such as transferring money from victim's account.

A. Blacklist and Whitelist Approach

Blacklist and Whitelist approach is used to identify the currently visited website is either phishing or legitimate. Blacklist involves list of websites that are declared as a spam. Such blacklist maintain by organizations like Google [1]. Whitelist approach is used to different phishing sites by comparing the current URL with predefined list of URL's. The main disadvantage of blacklist and whitelist approach is that it cannot differentiate the newly created phishing websites from legitimate websites.

B. Machine Learning Approach

In this technique, features are extracted and they are classified using the machine learning techniques. Machine learning focuses on developing the computational algorithms and motivate rules and patterns in order to produce general models. The machine learning is called as supervised learning if no labels are given with in the training phase. There are some popular machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), BackPropagation Neural Network (BPNN), Naïve Bayes Classifier (NB), and k-Nearest Neighbour (kNN).

C. Heuristic Based Approach

Heuristic is one of the problem solving technique that uses shortcuts to produce good-enough solutions given a limited time, deadline or frame. It uses heuristic to classify URLs. Heuristic is a type of feature that are consider to check the websites[1]. In this approach, some features of websites are gathered and evaluate them to select most influential features of website, they plays an important role in detecting the website phishing. Heuristic approach uses common features of legitimate and phishing sites based on URL, Search Engine, Lookup, HTML DOM and website traffic. Heuristic design of the website matches the predefined rules then websites are declared as phishing sites.

D. Hybrid Approach

In this approach, different techniques are combine to detect whether a website is real or fake. For e.g. blacklisting and heuristics of URL can be combine to form a better system[1]. To solve the phishing websites problems the Hybrid Model uses 30 features. To detect the websites a single model is not sufficient. Therefore, it enhance the efficiency, accuracy, and performance rate. To form a more robust classifier, two or more models are combined together. Firstly, performance of individual classifier is checked and then the high accuracy and less rate of best classifier is evaluated. After that the best classifier model is combined with other classifiers and finally, the better hybrid classification model is achieved.

E. Anti-Phishing Approach

It is a technological service that helps to prevent unauthorized access to secure and sensitive information. Antiphishing services protects different type of data in other ways beyond the variety of stages. Anti-phishing software comprise of computer programs that attempt to identify phishing contents.

IV. PROPOSED WORK

The proposed algorithm is based on machine learning process and automated real-time phishing detection. By using this features phishing URL's are extracted. For a machine learning classification the extracted features are used to detect phishing websites on real time.

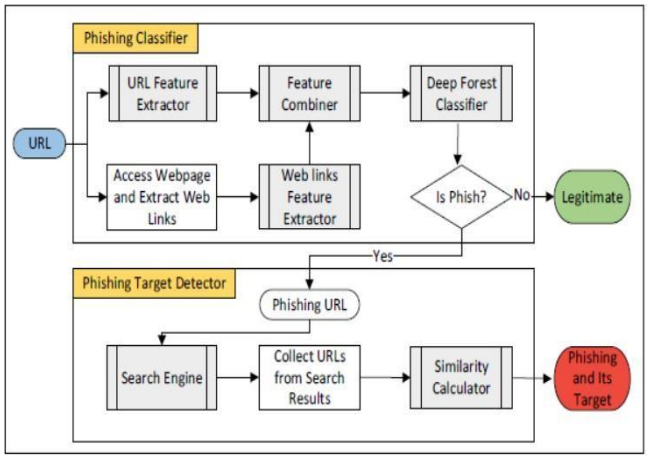


Fig. 3: Overview of the proposed work

A. Extreme Learning Machine

Extreme Learning Machine is a feed-forward Artificial Neural Network(ANN) and it also has a single hidden layer. The ANN is one of the main tool used in Machine Learning. Neural Network which consist of input and output layers as well as hidden layers. Extreme Learning Machine algorithm overcomes the slow training speed and over-fitting problems. ELM is based on its learning process and empirical risk minimization theory. The ELM avoids local minimization and multiple iterations. In ELM process is differently from Artificial Neural Network because it renew its parameters and input weight are randomly selected while output weight is calculated analytically. In order to generate the cells in the hidden layer of ELM.

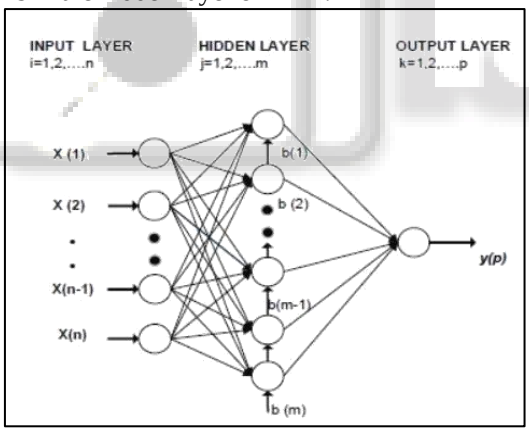


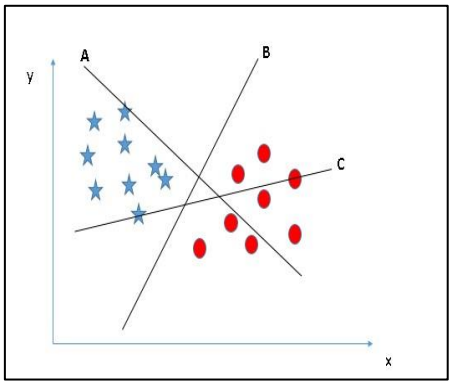
Fig. 4: An ANN model with a single hidden layer

B. Support Vector Machine

Support Vector Machine follows supervised learning. SVM is used to avoid the Internet user from a victim of phisher do not loss personal and financial information.

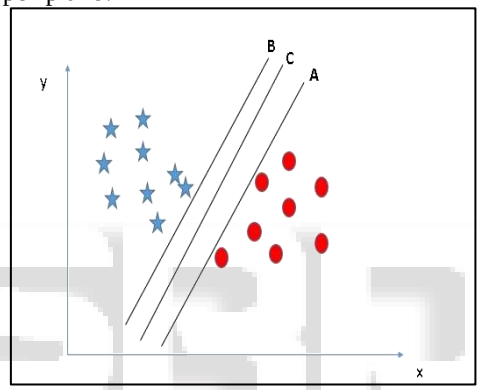
Identify the right hyper-plane (situation-1)

Here, we take hyper-planes(A,B, and C) now, we need to identify the right hyper-plane to classify star and circle.



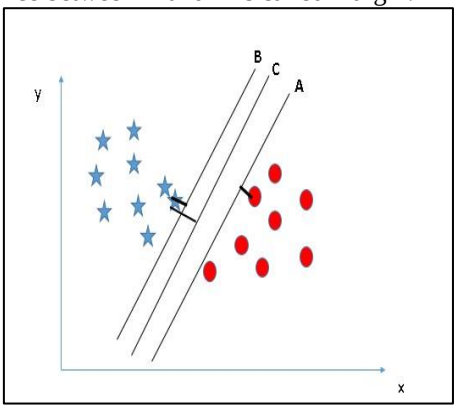
After that we need to remember a thumb rule to identify the right hyper-plane: "select the hyper-plane which divides two classes better". In this situation, hyper-plane B has excellently perform this task. Identify the right hyper-plane (situation-2)

Here, we take three hyper-planes (A,B and C) and all are divide classes well. Now, how we can identify the right hyper-plane?



Here, maximizing the distance during nearest data point and hyper-plane which helps to decide the right hyper-plane.

The distance between A and B is called margin.



Margin for hyper-plane C is compared with both A and B. Therefore, the hyper-plane with higher margin is robustness.

Features of website

Feature category	Attributes	Values
Address based Features	having_IP_Address	{ 1,0 }
	URL_Length	{ 1,0,-1 }
	Shortining_Service	{ 0,1 }
	having_At_Symbol	{ 0,1 }
	double_slash_redirecting	{ 1,0 }
	Prefix_Suffix	{ -1,0,1 }
	having_Sub_Domain	{ -1,0,1}
	SSLfinal_State	{ -1,1,0 }
	Domain_registeration_length	{ 0,1,-1}
	Favicon	{ 0,1 }
	Non-standard port	{ 0,1 }
	HTTPS_token	{ 1,0 }
Abnormal Features	Request_URL	{ 1,-1 }
	URL_of_Anchor	{ -1,0,1 }
	Links_in_tags	{ 1,-10 }
	SFH	{ -1,1 }
	Submitting_to_email	{ 1,0 }
	Abnormal_URL	{ 1,0 }
HTML, JavaScript Features	Redirect	{ 0,1 }
	on_mouseover	{ 0,1 }
	RightClick	{ 0,1 }
	popUpWidnow	{ 0,1 }
	Iframe	{ 0,1 }
Domain Features	age_of_domain	{ -1,0,1 }
	DNSRecord	{ 1,0 }
	web_traffic	{ -1,0,1 }
	Page_Rank	{ -1,0,1 }
	Google_Index	{ 0,1 }
	Links_pointing_to_page	{ 1,0,-1 }
	Statistical_report	{ 1,0 }

Table 1: Attributes and values for phihing feature

Table 1 represents the feature category, its attributes and values. Some attributes have 1 value or 2 values or 3 values which represent its strength ranging from low, medium and high. Te dataset is used to extract the phishing features for each URL under four categories: Addressed based features, Abnormal features, HTML, JavaScript features and Domain features. This features have 30 characteristics of phishing websites which is used to differentiate from legitimate website. Each category has its own characteristics of phishing i.e. attributes and values are defined. In this dataset, input attributes can take 3 different values which are 1, 0, and -1. Output attribute can take 2different values which are 1,and -1.

V. CONCLUSION

We have studied the different phishing attacks on URLs. The detection of the phishing websites can be performed using Extreme Learning Machine. Extracting the features of the website via URL when the user visits it which is done by using this technique. The obtained features will act as test data for the model. The main task of this technique is to detect the phishing website and alert the user beforehand to prevent the users from getting their credentials misused

REFERENCES

[1] Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an Anti-Phishing Framework," 2019 5th International Conference On Advanced Computing & Communication System(ICACCS), pp. 1-6,

[2] Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang and Wenyin Liu, "Detecting Phishing Websites and Targets Based On URLs and Webpage Links," 2018 24th International Conference on Pattern Recognition(ICPR) Beijing, China, August 20-24, 2018

[3] Vaibhav Patil, Pritesh Thakkar, Cjirag Shah, Tushar Bhat, Prof. S. P.Godse, "Detection and Prevention of Phishing Websites using Machine Learning Approach", 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] Mustafa AYDIN and Nazifa BAYKAL, "Feature Extraction and Classification Phishing Websites Based on URL,"2015

[5] Brad Wardman, Gaurang Shukla and Gary Warner, "Identifying Vulnerable Websites By Analysis of Common Strings in Phishing URLs," 978-1-4244-4626-1/09/2009 IEEE.

[6] Shraddha Parekh, Dhwanil Parikh, Srushti Kotak and Prof. Smita Sankhi, "A New Mthod For Detection of Phishing Websites: URL Detection,"Proceedings of the 2nd International Conference on Inventice Communicastion and Computational Technologies (ICICCT 2018) IEEE Explorer Complaint-Part Number: CFP18BAC-ART: ISBN: 978-1-5386-1974-2

[7] Anu Vazhayil, Vinaya Kumar R and Soman KP, "Comparative Srudy Of The Detcetion Of Malicious URLs Using Shallow and Deep Netoworks, "9th ICCCNT2018 July 10-12,2018, IISC, Bangluru,India.

[8] Martyn Weedon, Dimitris Tsaptsinos and James DenholmPrice, "Random Forest Explorations for URL Classification,"2017

[9] Ee Hung Chanj, Kang Leng Chiew, San Meh Sze and Wei King Tiong, "Phishing Detection via Identification of Website Identity," 978-1-4799-2845-3/13/2013 IEEE.

[10] Chuan Pham, Luong A.T. Nguyen, Nguyenh. Tran, Euinam Huh and Choong Seon Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks," DOI 10.1109/TNSM. 2018. 2831197, IEEE Transactions on Netowork and Service Management.

[11] Jayashri Haggude and Lata Ragha, "Phish Mail Guard: Phishing Mail Detection Technique By Using Textual and URL Analysis," 978-1-4673-4805-8/12/2012 IEEE.

[12] Varsharani Ramdas Hawanna, V. Y. Kulakarni and R.A. Rane, "A Novel Algorithm to Detect Phishing URL's," 9781-5090-2080-5/16/2016 IEEE.

[13] Xueni Li, Guanggang Geng, Zhiwei Yan, Yong Chen and Xiaodong Leee, "Phishing Detection Based on Newly Registered Domains," 2016 IEEE International Conference On Big Data(Big Data).

[14] Ebubekir Buber, Onder Demir and Ozgur Koray Sahingoz, "Feature Selections For The Machine Learning Based Detection of Phishing Websites," 978-1-5386-18806/17/2017 IEEE.

