A hybrid approach for phishing detection in web application through machine learning

[1]Pradip Maliwad, [2]Prof. Deepak Upadhyay, [3]Mr. Kaushal Bhavsar

[1]Student of M.E Cyber Security, [2]Assistant Professor, [3]Founder of BugSkan Cyber Security solutions

[1]M.E cyber security,

[1]GTU –School of Engineering and Technology, Gandhinagar, India

Abstract : Over the last few years, phishing has become an important threat for companies and other kinds of organizations, making them lose millions of pounds every year. There are many researches who study different methods to detect and stop these attacks. Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It is usually via a email, SMS or social medial application to attract computer users to reveal sensitive personal information. We proposed lexical based, HTML based, JavaScript based and page reputation based features for detection of phishing websites through machine learning. We then apply various machine learning algorithms to build models from training data, which is comprised of pairs of feature values and class labels using WEKA. After evaluating the classifiers, a Random forest get higher accuracy so we use Random Forest algorithm for classify website is phishing or legitimate. Our Proposed method is highly effective in detecting phishing URLs with 95.043 accuracy and 0.052 false positive rate.

Keywords: Phishing, Hybrid features, Application security, Machine learning, Phishing detection, Cyber security.

I. INTRODUCTION

Recently, there has been a dramatic increase in phishing, a kind of attack in which victims are tricked by spoofed emails and fraudulent web sites into giving up personal information[1]. The online payment services, e-commerce, and social networks are the most affected sectors by this attack[1][2].

A phishing attack is performed by taking advantage of the visual resemblance between the fake and the authentic web pages. The attacker creates a web-page that looks exactly similar to the legitimate web page[3]. The link of phishing web page is then send to thousands of Internet users through emails and other means of communication. Usually, the fake email content shows some sense of fear, urgency or offer some price money and asks the user to take urgent action. E.g., the fake email will impel user to update their PIN to avoid debit/credit card suspension. When the user unknowingly updates the confidential credentials, the cyber criminals acquire user's details[4][5]. Phishing attack performed not only for gaining information; now it has become the number 1 delivery method for spreading other types of malicious software like ransomware[4][6]. According to APWG report, 291,096 unique phishing web-sites were detected between January to June 2017. The per month attack growth has also increased by 5753% over 12 years from 2004 to 2016[7]. Figure 1 presents the growth of phishing attack from 2005 to 2016.

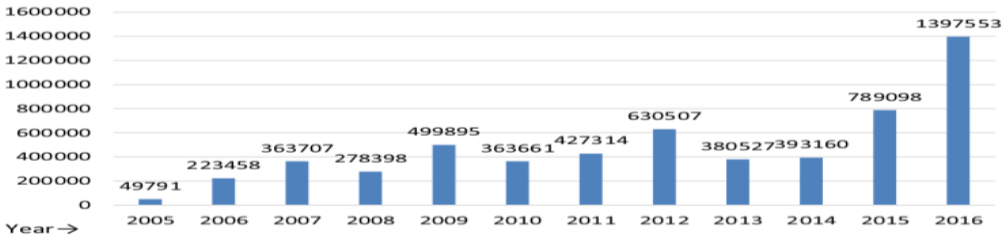

Figure 1 : Growth of phishing attack[7]

II. LITERATURE SURVEY

In this section, we present an overview of various anti-phishing solutions proposed in the literature. Phishing detection approaches are divided into two categories. First, based on user education, and another relies on the software. In the user education-based approaches, Internet users are educa-

ted to understand the characteristics of phishing attacks, which eventually leads them to appropriately identifying phishing and legitimate websites and emails[8][9][10][11]. Software-based approaches are further classified into machine learning, blacklist, and visual similarity based approaches. Machine learning based approach trains a classification algorithm with some features and a web-site is declared as phishing, if the design of the websites matches with the predefined feature set[12]. Visual similarity based approaches compare the visual appearance of the suspicious website and its corresponding legitimate website [1]. Blacklist matches the suspiciousm domain with some predefined phishing domains which are blacklisted. The negative aspect of the blacklist and visual similarity based schemes is that they usually do not cover newly launched (i.e. zero hour attack) phishing websites. Most of the phishing URLs in the blacklist are updated only after 12 h of phishing attack[7]. Therefore, machine learning based approaches are more effective in dealing with phishing attacks. Some of the machine learning based approaches given in the literature are explained below.

In [13] proposed an anti-phishing method, which inspects the anomalies in the website. The approach extracts the anomalies from the various sources like URL, page title, cookies, login form, DNS records, SSL certificates, etc. The approach used SVM and achieved 88% true positive rate and 29% false positive rate. However, the proposed scheme used a dataset of only

379 websites. in [1] proposed a content specific approach CANTINA that can detect the phishing webpage by analysing text content and using TF-IDF algorithm. Top five keywords with highest TF-IDF are submitted into the search engine to extract the relevant domains. CANTINA also uses some heuristic like the special symbol in URL "@" (at sign), "–"(dash) symbol, dot count, domain age, etc. However, the accuracy of the scheme depends on TF-IDF algorithm and language used on the website. CANTINA achieved 6% of false positive rate, which is considered very high compared six [10]machine learning algorithm for phishing e-mail detection namely Logical regression, Bayesian additive regression trees, SVM, RF, Neural network, and Regression trees. The result shows that there are no standard machine learning algorithms which can efficiently detect phishing attack. In [14] proposed a technique based on phishing URLs. The given approach discussed four different kinds of obfuscation techniques of phishing URLs. The approach uses logistic regression as a classifier. However, this technique cannot identify tiny URL based phishing websites. In [15] proposed an intelligent phishing detection system using the self-structuring neural network. Authors have collected 17 features from URL, source code and the third party to train the system using the neural network. Back propagation algorithmis used to adjust the weights of the network. Nevertheless, the design of network was a little bit complex. However, the training and testing set accuracy were 94.07 and 92.18, respectively on 1000 epochs. In [17] have used 27 features to construct a model based on fuzzy-logic for detection of phishing attack in banking websites. The authors used the features from the URL, page content, SSL certificates, etc., to identify the phishing attack. This approach focused only on e-banking websites and did not discuss the detection results on another type of websites. Whittaker et published research on a large-scale classification of phishing websites, which uses the features from URL, page hosting, and page content. The TPR and FPR of the approach is 90 and 0.1%, respectively. In [1] proposed CANTINA+, which takes 15 features from URL, HTML DOM (Document object model), third party services, search engine, and trained these features using support vector machine (SVM). Although, the performance of the scheme is affected by third party services like WHOIS lookup and search results. In [4] have used 12 features from the legitimate and phishing websites and achieved 97% true positive rate and 4% false positive rate. Most of approach for banking while our approach can filter all kinda of phishing and legitimate website.

III. PROPOSED METHOD

First we collecting phishing from PhishTank and benign URLs From Alexa dataset. The HTML based, Javascript based, Page reputation based and lexical based feature extract and make training dataset. We then apply various machine learning algorithms to build models using training dataset, which is comprised of pairs of feature values and class labels using WEKA. After evaluating the classifiers, a particular classifier is selected and is implemented in Python. At last using selected classifier we can classify URLs is Phishing or not.

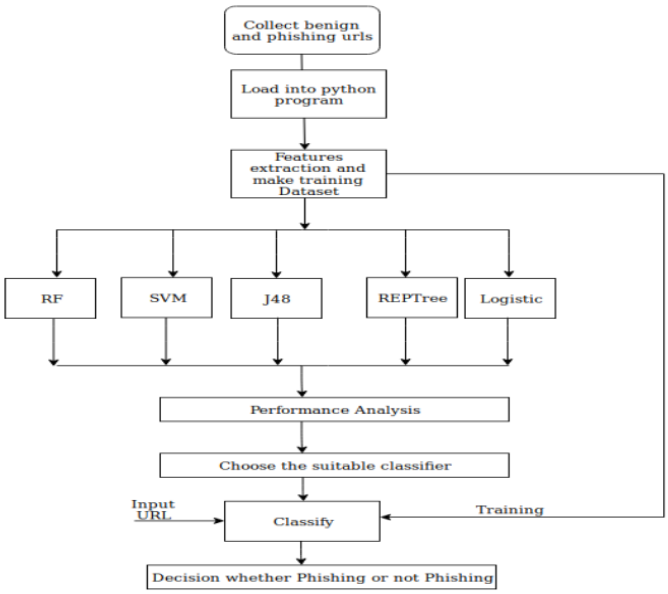

Figure 2 : proposed method

IV. IMPLEMENTATION

First, We collected the non-phishing URLs from alexa and collect phishing URLs taken from PhishTank. PhishTank is a collaborative clearing house for data and information about phishing on the Internet.

We used Python script for Feature extraction. We developed our set of 24 features based on related works, drawing primarily from [14], [15], [18], [12], and [9]. Some of these features are modified to fit our needs, while others are newly proposed. The use of relatively small number of fixed set of features makes the decision boundaries less complex, and therefore less prone to over fitting as well as faster to evaluate for most batch algorithms.

We group features that we gather into 4 broad categories. Table 1 summarizes each category and the number of features from that category that we use in our data sets for classifying phishing URLs.

The proposed approach build a binary classifier based features, which classify phishing and legitimate websites correctly. Our training dataset consists of 6000 phishing and 5000 legitimate websites.

Table 1 : Feature Categories and Number of Features in Each Category

Feature Category	Feature Count
Lexical based	10
HTML based	4
Page Reputation based	5
Javascript based	5

Classification Models :

Since no single classifier is perfect, we evaluate several supervised batch-learning classifiers. As researchers, we have no vested interest in any particular classifier. These classifiers are chosen mostly because they have been applied to problems similar to ours, such as in detecting: spam and phishing emails, phishing and malicious URLs, phishing webpages, etc. We simply want to empirically compare a number of classifiers based on their availability in implementation and determine the one that yields the best performance in terms of both training and testing time and accuracy to the problem of detecting phishing URLs.

We evaluate the following five classifiers implemented in WEKA (Waikato Environment for Knowledge Analysis):

1. Random Forest
2. SVM
3. REP Tree
4. Logistic
5. J48

EMPIRICAL EVALUATIONS:

In order to evaluate our methodology, we perform 2 major experiments. We use 10 times 10-fold cross-validation (unless otherwise stated) to evaluate the classifiers.

Experiment 1- Classifier Evaluation

In this experiment, we evaluate classification performance of five classifiers on dataset using the 24 feature set. Table (2) compare the overall accuracy of five classifiers. machine learning algorithms considered for processing the feature set are: Random forest, SVM, Logistic, J48 and REPTree.

Table 2: Compare the overall accuracy of classifiers

ML algorithm	Accuracy
Random Forest	95.043%
REPTree	92.3021 %
Logistic	88.4758 %
SVM	88.8105 %
J48	93.5866 %

From Table 2 it clear that Random Forest classifier achieves has one of the best overall accuracy . So we choose RF classifier for the second experiment.

Experiment 2 – Feature Evaluation

In this experiment, we compare various combinations of feature sets to evaluate how effective each feature category is in detecting phishing URLs. Specifically, we compare individual feature category and combine it with the lexical based feature category – the most commonly used feature category in phishing detection. We use RF classifier on data set as it has sufficiently good number of phishing and non-phishing URLs with varieties in URL structures covering most feature categories. Results on these experiments are displayed in Table 3.

Table 3: FP and TP of combinations of feature sets

Feature Category	Random Forest	
	FP(%)	TP(%)
Lexical based	0.241	0.747
HTML based	0.142	0.867
Javascript based	0.537	0.560
Page reputation based	0.274	0.733
lexical+host based	0.105	0.903
Lexical+hosted+javascript	0.099	0.909
All Features	0.052	0.952

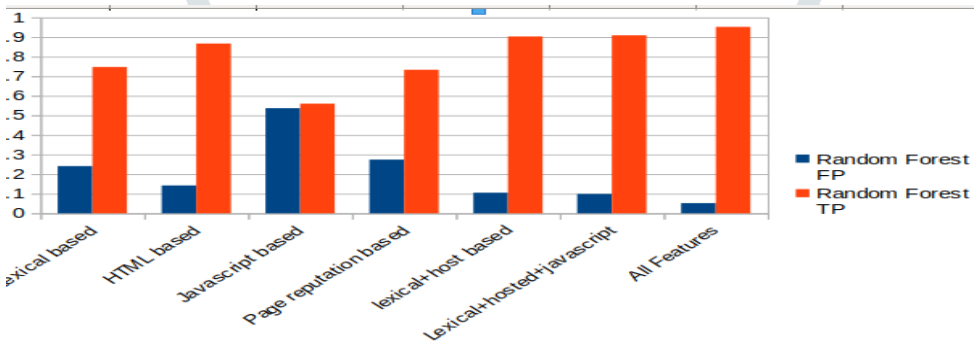

Figure 3: FP rate and TP rate Comparison Graph of different dataset using Random forest

When using lexical based feature type alone, RF classifier achieves an 0.01% FP. Similarly, Javascript based feature type, achives an 0.537% FP. When combined lexical with HTML based features, the FP 0.105%. From Table 3 it clear that Random Forest classifier achieves has one of the Lowest 0.052 FP on all features.

Result:

Figure 6.5 Compare the accuracy, overall error rates and false positive rates result between proposed system and Base paper.

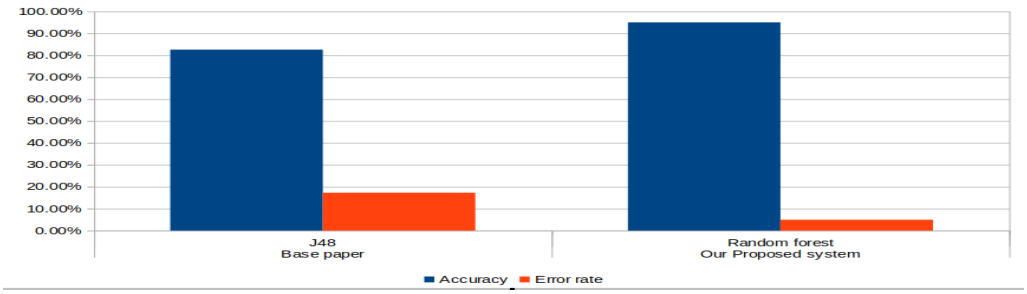

Figure 4: Compare the accuracy, overall error rates and false positive rates result between proposed syetem and Base paper

When we compare with base paper[3] Our method achieves 95.043 % accuracy and error rate 4.957% using Random Forest algorithm. however their approach achieves 82.6% accuracy and 17.3% error rate using J48 algorithm.

Classify :

We can observe that Random Forests (RF) has one of the best overall accuracy and worst error rate so we use Random Forests algorithm for classifying URL is Phishing or benign.

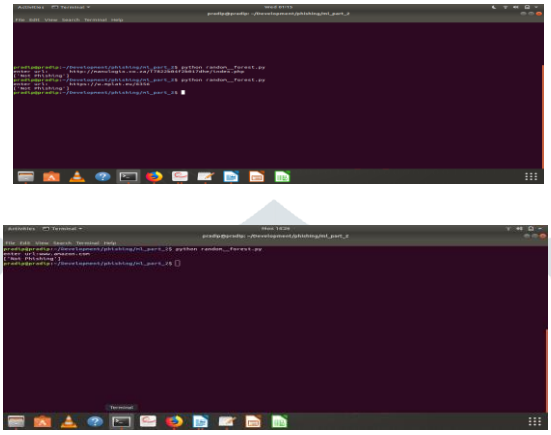

Figure 4: New phishing URLs are tested

V. CONCLUSION

The dissertation addresses issue of identifying the Phishing URLs. We can achieve by using machine learning supervised classification algorithm. We proposed Lexical based, HTML based, JavaScript based and Page Reputation based features for classifying phishing URLs. We empirically demonstrated that the proposed features are highly relevant to classification of phishing URLs. We evaluated our approach on dataset by comparing performance results of several popular supervised learning methods. Experimental results showed that the proposed solution was able to detect phishing URLs with an accuracy of 95.043 % and false positive rate 0.052.

We have used 6231 non-phishing URLs from alexa database and 4824 phishing URLs taken from PhishTank. We made our training dataset consists of 24 features and 11055 instances. We have trained Random forest, SVM, J48, Logistic and REP Tree classifiers using training dataset in weka tool. Random Forest achieved 95.043% accuracy and 0.052 false positive rate.

VI. BIBLIOGRAPHY

[1] J. Hong and L. Cranor, “CANTINA : A Content-Based Approach to Detecting Phishing Web Sites,” in *Www 2007*,, pp. 639–648.

[2] S. Smadi, N. Aslam, and L. Zhang, “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning,” *Decis. Support Syst.*, vol. 107, pp. 88–102, 2018.

[3] E. Medvet, E. Kirda, and C. Kruegel, “Visual-similarity-based phishing detection,” p. 1, 2009.

[4] A. A. Ahmed and N. A. Abdullah, “Real time detection of phishingwebsites,” *7th IEEE Annu. Inf. Technol. Electron. Mob. Commun.Conf. IEEE IEMCON 2016*, 2016.

[5] M. Baykara and Z. Z. Gürel, “Detection of phishing attacks,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018 Janua, pp. 1–5, 2018.

[6] N. Chandru, “A Review on Phishing Attacks and Anti- Phishing Browser Plugins,” Int. J. Comput. Appl., vol. 9, no. 05, pp. 51–58, 2018.

[7] T. Chin, K. Xiong, and C. Hu, “Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking,” IEEE Access, vol. 6, no. c, pp. 42513–42531, 2018.

[8] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing url detection using association rule mining," Human-centric Comput. Inf. Sci., vol. 6, no. 1, 2016.

[9] V. Preethi and G. Velmayil, "Automated Phishing Website Detection Using URL Features and Machine Learning Technique," Int. J. Eng. Tech., vol. 2, no. 5, pp. 107–115, 2016.

[10] N. Moradpoor and B. Clavie, "Machine Learning Techniques for the Detection and Classification of Phishing Emails," no. July, pp. 149–156, 2017.

[11] P. Ying and D. Xuhua, "Anomaly based web phishing page detection," *Proc. - Annu. Comput. Secur. Appl. Conf. ACSAC*, pp. 381–390, 2006.

[12] C. Juan and G. Chuanxiong, "Online detection and prevention of phishing attacks (invited paper)," *First Int. Conf. Commun. Netw. China, ChinaCom '06*, no. October 2006, 2007.

[13] A. K. Jain and B. B. Gupta, "Towards detection of phishing websites on client-side using machine learning based approach," *Telecommun. Syst.*, vol. 68, no. 4, pp. 687–700, 2018.

[14] A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature based phishing URL detection using online learning," p. 54, 2010.

[15] J. James, L. Sandhya, and C. Thomas, "Detection of phishing URLs using machine learning techniques," *2013 Int. Conf. Control Commun. Comput. ICCC 2013*, no. February, pp. 304–309, 2013.

[16] A. Priya and E. Meenakshi, "Detection of phishing websites using C4.5 data mining algorithm," *RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.*, vol. 2018–Janua, pp. 1468–1472, 2018.

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

[18] E. Kidmose, E. Lansing, S. Brandbyge, and J. M. Pedersen, "Detection of malicious and abusive domain names," *Proc. - 2018 1st Int. Conf. Data Intell. Secur. ICDIS 2018*, pp. 49–56, 2018.