Phishing Websites Detection Using Machine Learning

R. Kiruthiga, D. Akila

Abstract--- Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. This paper surveys the features used for detection and detection techniques using machine learning.

Keywords--- Phishing, Phishing Websites, Detection, Machine Learning.

I. INTRODUCTION

Phishing is the most unsafe criminal exercises in cyber space. Since most of the users go online to access the services provided by government and financial institutions, there has been a significant increase in phishing attacks for the past few years. Phishers started to earn money and they are doing this as a successful business. Various methods are used by phishers to attack the vulnerable users such as messaging, VOIP, spoofed link and counterfeit websites. It is very easy to create counterfeit websites, which looks like a genuine website in terms of layout and content. Even, the content of these websites would be identical to their legitimate websites. The reasonfor creating these websites is to get private data from users like account numbers, login id, passwords of debit and credit card, etc. Moreover, attackers ask security questions to answer to posing as a high level security measure providing to users. When users respond to those questions, they get easily trapped into phishing attacks. Many researches have been going on to prevent phishing attacks by different communities around the world. Phishing attacks can be prevented by detecting the websites and creating awareness to users to identify the phishing websites. Machine learning algorithms have been one of the powerful techniques in detecting phishing websites. In this study, various methods of detecting phishing websites have been discussed.

II. LITERARY REVIEW

Authors in this paper[1] explained a novel approach to detect phishing websites using machine learning algorithms. They also compared the accuracy of five machine learning algorithms Decision Tree (DT), Random Forest (RF)[1], Gradient Boosting (GBM), Generalized Linear Model (GLM) and Generalized Additive Model (GAM)[1]. Accuracy, Precision and Recall evaluation methods were calculated for each algorithm and compared. Website attributes (30) are extracted with the help of Python and performance evaluation done with open source programming language R. Top three algorithms namely Decision Tree, Random Forest and GBM performance were compared in table. From the tables of accuracy, recall and performance, it is shown that Random Forest algorithm has given highest 98.4% accuracy, 98.59% recall and 97.70% precision.

In this paper authors [2] proposes a classification mode[2]l in order to classify the phishing attacks. This model comprises of feature extraction from sites and classification of website. In feature extraction, 30 features has been taken from UCI Irvine machine learning repository data set and phishing feature extraction rules has been clearly defined. In order to classification of these features, Support Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM)[2] were used. In Extreme Learning Machine (ELM), six activation functions were used and achieved 95.34% accuracy than SVM and NB. The results were obtained with the help of MATLAB.

Authors [3] presents an approach to detect phishing email attacks using natural language processing and machine learning. This is used to perform the semantic analysis of the text to detect malicious intent. A natural Language Processing (NLP) technique is usedto parse each sentence and finds the semantic jobs of words in the sentence in connection to the predicate. In light of the job of each word in the sentence, this strategy recognizes whether the sentence is an inquiry or an order. Supervised machine learning[3] is used to generate the blacklist of malicious pairs. Authors defined algorithm SEAHound[3] for detecting phishing emails and Netcraft Anti-Phishing Toolbar is used to verify the validity of a URL. This algorithm is implemented with Python scripts and dataset Nazario phishing email set is used. Results of Netcraft and SEAHound[3] are compared and obtained precision 98% and 95% respectively.

Manuscript received September 16, 2019.

R. Kiruthiga, Ph.D. Research Scholar, Department of Computer Science, VELS Institute of Science, Technology & Advanced Studies, Chennai, Tamilnadu. (e-mail: kirthimrk238@gmail.com)

Dr.D. Akila, Associate Professor, Department of Information Technology, School of Computing Sciences, VELS Institute of Science, Technology & Advanced Studies, Chennai, Tamilnadu, India. (e-mail: akiindia@yahoo.com)

This result demonstrates that semantic data is a solid pointer of social designing.

Another approach by authors [4] proposes feature selection algorithms to decrease the components of dataset to get higher order execution [4]. It also compared with other data mining classification algorithms and results obtained. Dataset for phishing websites was taken from UCI machine learning repository[4]. From the outcomes, it is seen that some classification strategies increment the execution; some of them decline the execution with decreased component. Bayesian Network, Stochastic Gradient Descent (SGD), lazy.K.Star, Randomizable Filtered Classifier, Logistic model tree (LMT) and ID3 (Iterative Dichotomiser)[4] are useful for reduce phishing dataset and Multilayer Perception, JRip, PART, J48[4], Random Forest and Random Tree algorithms are not valuable for the diminished phishing dataset. Lazy.K.Star obtained 97.58% accuracy with 27 reduced features. This study is obtained with the help of WEKA software.

Authors [5]proposed a model with answer for recognize phishing sites by utilizing URL identification strategy utilizing Random Forest algorithm. Show has three stages, namely Parsing, Heuristic Classification of data, Performance Analysis [5]. Parsing is used to analyze feature set. Dataset gathered from Phishtank. Out of 31 features only 8 features are considered for parsing. Random forest method obtained accuracy level of 95%.

Authors [6] proposed a flexible filtering decision module to extract features automatically without any specific expert knowledge of the URL domain using neural network model. In this approach authors used all the characters included in the URL strings and count byte values. They not only count byte values and also overlap parts of neighbouring characters by shifting 4-bits. They embed combination information of two characters appearing sequentially and counts how many times each value appears in the original URL string and achieves a 512 dimension vector. Neural network model tested with three optimizers Adam, AdaDelta and SGD. Adam was the best optimizer with accuracy 94.18% than others. Authors also conclude that this model accuracy is higher than the previously proposed complex neural network topology.

In this paper authors [7] made a comparative study to detect malicious URL with classical machine learning technique – logistic regression using bigram, deep learning techniques like convolution neural network (CNN) and CNN long short-term memory (CNN-LSTM)[7] as architecture. The dataset collected from Phishtank, OpenPhish for phishing URLs and dataset MalwareDomainlist, MalwareDomains were collected for malicious URLs. As a result of comparison, CNN-LSTM obtained 98% accuracy. In this paper authors used TensorFlow[7] in conjuction with Keras[7] for deep learning architecture.

Authors in this paper [8] also proposed reduced feature selection model to detect phishing websites. They used Logistic Regression and Support Vector Machine (SVM)[8] as classification methods to validate the feature selection method. 19 features reduced from 30 site features have been selected and used for phishing detection. The LR and SVM calculations performance was surveyed dependent on precision, recall, f-measure and accuracy. Study shows that SVM algorithm achieved best performance over LR algorithm.

In this paper authors [9] proposed a phishing detection model to detect the phishing performance effectively by using mining the semantic features of word embedding, semantic feature and multi-scale statistical features[9] in Chinese web pages. Eleven features were extracted and categorized into five classes to acquire statistical features of web pages. AdaBoost, Bagging, Random Forest and SMO[9] are used to implement learning and testing the model. Legitimate URLs dataset obtained from DirectIndustry web guides and phishing data was obtained from Anti-Phishing Alliance of China. According to study, only semantic features well identified the phishing sites with high detection[9] efficiency and fusion model achieved the best performance detection. This model is unique to Chinese web pages and it has dependency in certain language.

This paper [10] proposes a efficient way to detect phishing URL websites by using c4.5 decision tree approach. This technique extracts features from the sites and calculates heuristic values. These values were given to the c4.5 decision tree algorithm[10] to determine whether the site is phishing or not. Dataset is collected from PhishTank and Google. This process includes two phases namely pre-processing phase and detection phase[10]. In which features are extracted based on rules in pre-processing phase and the features and their respected values were inputted to the c4.5 algorithm and obtained 89.40% accuracy.

Authors [11] in this paper created an extension to Google Chrome to detect phishing websites content with the help of machine learning algorithms. Dataset UCI-Machine Learning Repository used and 22 features were extracted for this dataset. Algorithms kNN, SVM and Random Forest were chosen for precision, recall,f1-score and accuracy comparison. Random Forest obtained a best score and HTML,JavaScript, CSS[11] used for implementing chrome extension along with python. This extension is having a drawback of declared malicious site list which is increasing every day.

This paper [12] approaches a framework to extract features flexible and simple with new strategies. Data is collected from PhishTank[12] and legitimate URLs from Google[12]. To obtain the text properties C# programming and R programming were used. 133 features were obtained from the dataset and third party service providers. CFS subset based and Consistency subset based feature selection[12] methods used for feature selection and analyzed with WEKA tool. Naïve Bayes and Sequential Minimal Optimization (SMO)[12] algorithms were compared for performance evaluation and SMO is preferred by the author for phishing detection than NB.

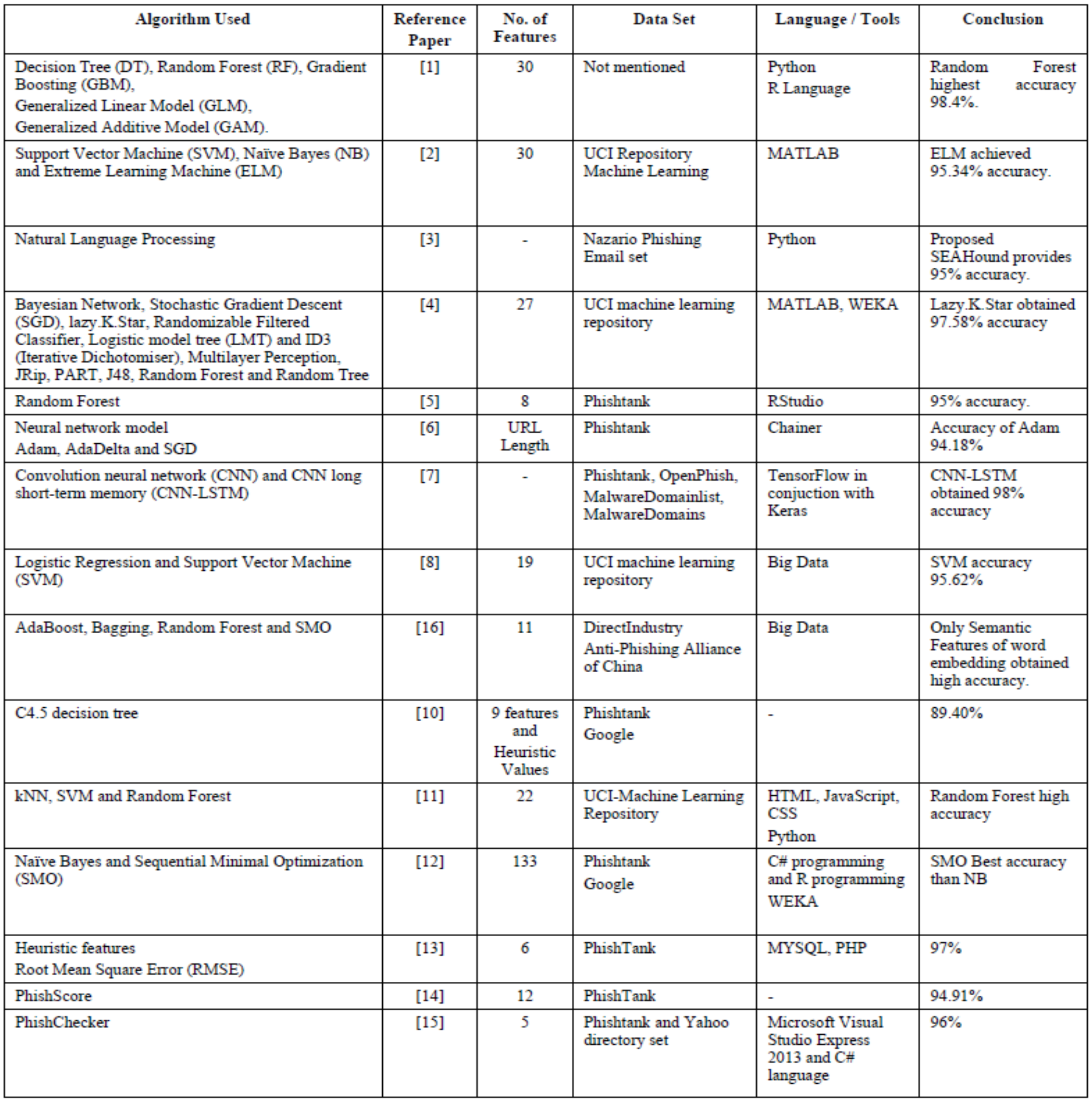
Another heuristic features detection method by authors [13] explains about the feature of URL such as PrimaryDomain, SubDomain, PathDomain and ranking of website such as PageRank, AlexaRank, AlexReputation to identify the phishing websites. Dataset used from PhishTank and experimental is splitted into 6 phases through MYSQL, PHP with 10 testing datasets. The proposed model contains two phases. In Phase I site features were extracted and in Phase II six values of heuristic are calculated. According to authors, if heuristic value is nearest to one, the site is considered as legitimate and if it is nearest to zero then the site is doubted as phishing site. Root Mean Square Error (RMSE)[13] is used to calculate accuracy and obtained 97% accuracy.

In this paper author [14] introduces a phishing URL detection system depends on URL lexical analysis named PhishScore. This approach is based on intra-URL relatedness[14][18]. This relatedness reflects the relationship into part of the URLRight around 12 site highlights removed from a solitary URL are utilized to include machine learning algorithms to identify phishing URLs. This experiment results accuracy of 94.91%.

RESULTS

This paper [15] focuses on detecting phishing website URLs with domain name features. Web spoofing attack categories content-based, heuristic-based and blacklist-based approaches[8][17] are explained and the proposed model PhishChecker is developed with the help of Microsoft Visual Studio Express 2013 and C# language[15]. Dataset used from Phishtank and Yahoo directory set and obtained an accuracy of 96%. This paper checks only the validity of URLs.

Table 1: Outline of Algorithms used to detect Phishing Website URLs



Algorithm Used	Reference Paper	No. of Features	Data Set	Language / Tools	Conclusion
Decision Tree (DT), Random Forest (RF), Gradient Boosting (GBM), Generalized Linear Model (GLM), Generalized Additive Model (GAM).	[1]	30	Not mentioned	Python R Language	Random Forest highest accuracy 98.4%.
Support Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM)	[2]	30	UCI Repository Machine Learning	MATLAB	ELM achieved 95.34% accuracy.
Natural Language Processing	[3]	-	Nazario Phishing Email set	Python	Proposed SEAHound provides 95% accuracy.
Bayesian Network, Stochastic Gradient Descent (SGD), lazy.K.Star, Randomizable Filtered Classifier, Logistic model tree (LMT) and ID3 (Iterative Dichotomiser), Multilayer Perception, JRip, PART, J48, Random Forest and Random Tree	[4]	27	UCI machine learning repository	MATLAB, WEKA	Lazy.K.Star obtained 97.58% accuracy
Random Forest	[5]	8	Phishtank	RStudio	95% accuracy.
Neural network model Adam, AdaDelta and SGD	[6]	URL Length	Phishtank	Chainer	Accuracy of Adam 94.18%
Convolution neural network (CNN) and CNN long short-term memory (CNN-LSTM)	[7]	-	Phishtank, OpenPhish, MalwareDomainlist, MalwareDomains	TensorFlow in conjuction with Keras	CNN-LSTM obtained 98% accuracy
Logistic Regression and Support Vector Machine (SVM)	[8]	19	UCI machine learning repository	Big Data	SVM accuracy 95.62%
AdaBoost, Bagging, Random Forest and SMO	[16]	11	DirectIndustry Anti-Phishing Alliance of China	Big Data	Only Semantic Features of word embedding obtained high accuracy.
C4.5 decision tree	[10]	9 features and Heuristic Values	Phishtank Google	-	89.40%
kNN, SVM and Random Forest	[11]	22	UCI-Machine Learning Repository	HTML, JavaScript, CSS Python	Random Forest high accuracy
Naïve Bayes and Sequential Minimal Optimization (SMO)	[12]	133	Phishtank Google	C# programming and R programming WEKA	SMO Best accuracy than NB
Heuristic features Root Mean Square Error (RMSE)	[13]	6	PhishTank	MYSQL, PHP	97%
PhishScore	[14]	12	PhishTank	-	94.91%
PhishChecker	[15]	5	Phishtank and Yahoo directory set	Microsoft Visual Studio Express 2013 and C# language	96%

III. CONCLUSION

This survey presented various algorithms and approaches to detect phishing websites by several researchers in Machine Learning. On reviewing the papers, we came to a conclusion that most of the work done by using familiar machine learning algorithms like Naïve Bayesian, SVM, Decision Tree and Random Forest. Some authors proposed a new system like PhishScore and PhishChecker for detection. The combinations of features with regards to accuracy, precision, recall etc. were used. Experimentally successful techniques in detecting phishing website URLs were summarized in Table 1. As phishing websites increases day by day, some features may be included or replaced with new ones to detect them.

REFERENCES

1. J. Shad and S. Sharma, “A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,” pp. 425–430, 2018.
2. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
3. T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.
4. M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.
5. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.
6. K. Shima et al., “Classification of URL bitstreams using bag of bytes,” in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.
7. A. Vazhayil, R. Vinayakumar, and K. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1–6.
8. W. Fadheel, M. Abusharkh, and I. Abdel-Qader, “On Feature Selection for the Prediction of Phishing Websites,” 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.
9. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017.
10. L. MacHado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.
11. A. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.
12. M. Aydin and N. Baykal, “Feature extraction and classification phishing websites based on URL,” 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp. 769–770, 2015.
13. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, “A novel approach for phishing detection using URL-based heuristic,” 2014 Int. Conf. Comput. Manag. Telecommun. ComManTel 2014, pp. 298–303, 2014.
14. S. Marchal, J. Francois, R. State, and T. Engel, “PhishScore: Hacking phishers’ minds,” Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 46–54, 2015.
15. A. A. Ahmed and N. A. Abdullah, “Real time detection of phishing websites,” 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016, 2016.
16. X. Zhang, Y. Zeng, X. B. Jin, Z. W. Yan, and G. G. Geng, “Boosting the phishing detection performance by semantic analysis,” in Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018, vol. 2018–Janua, pp. 1063–1070.
17. Dr.D.Akila, Dr.C. Jayakumar, "Acquiring Evolving Semantic Relationships for WordNet to Enhance Information Retrieval", International Journal of Engineering and Technology, Volume 6, November 5, pp. 2115-2128, 2014.
18. D.Akila,S.Sathya, G.Suseendran, “Survey on Query Expansion Techniques in Word Net Application”, Journal of Advanced Research in Dynamical and Control Systems, Vol.10(4), pp.119-124, 2018.