Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation

Smita Sindhu
USN:1BM16CS107
Department: CSE
B.M.S. College of Engineering
Bengaluru, India
1bm16cs107@bmsce.ac.in

Sunil Parameshwar Patil
USN:1BM16CS112
Department: CSE
B.M.S. College of Engineering
Bengaluru, India
1bm16cs112@bmsce.ac.in

Arya Sreevalsan
USN:1BM16CS133
Department: CSE
B.M.S. College of Engineering
Bengaluru, India
1bm16cs133@bmsce.ac.in

Faiz Rahman
USN:1BM16CS135
Department: CSE
B.M.S. College of Engineering
Bengaluru, India
1bm16cs135@bmsce.ac.in

Guide Name: Ms. Saritha A. N.
Assistant Professor, CSE
B.M.S. College of Engineering
Bengaluru, India
saritha.cse@bmsce.ac.in

***Abstract*— Phishing is a common attack used to obtain sensitive information using visually similar websites to that of legitimate websites. With the growing technology, phishing attacks are on the rise. Machine Learning is a very popular approach to detect phishing websites. This paper explains the existing machine learning methods that are used to detect phishing websites. The paper explains the improved Random Forest classification method, SVM classification algorithm and Neural Network with backpropagation classification methods which have been implemented with accuracies of 97.369%, 97.451% and 97.259% respectively.**

Keywords*: *Phishing, Phishing attacks, Machine Learning, Random Forest, SVM, Neural Network, Backpropagation

I. INTRODUCTION

Phishing is a fraudulent practice in which an attacker tries to obtain sensitive information by impersonating someone else to benefit himself/herself in a malicious way. Today, most of the users are accessing the services online, so it has become very easy for phishers to obtain user's confidential information. The website contents of phishing websites look very similar to that of legitimate websites and hence prompts people to provide their sensitive information. Phishing attacks can be prevented by making users distinguish between phishing and legitimate websites. Most of the phishers use images rather than text which are difficult to detect. Various tools and mechanisms have been developed to detect phishing websites and to prevent attacker from obtaining sensitive information. Blacklisting is one the easy way to detect phishing websites but can't be used to find new phishing websites. It is also a time consuming process.

In this paper, improved Random Forest model, SVM classification method and Neural Network with backpropagation model have been discussed which were implemented with accuracies of 97.369%, 97.451% and 97.259% respectively.

II. RELATED WORK

In the paper "URL Phishing Data Analysis using Random Forest" [4], SVM and Random Forest classification methods are discussed. The datasets are obtained from UCI Machine Learning Repository. The first dataset comprises of 30 features and second dataset comprises of 10 features. In Random Forest method, a number of classification trees are used which are created randomly by making use of different subsets of dataset to ensure that overfitting does not happen. The accuracy obtained using Random Forest method is 95.1%.SVM, a supervised Machine Learning algorithm makes use of hyperplane to do the classification. The data point is assigned to a particular class based on where it lies. The accuracy obtained with SVM classification algorithm is 92.62%.

In the paper "Phishing Detection from URLs by using Neural Networks" [9], two classification methods namely Artificial Neural Network and Deep Neural Network are discussed. A total of 27 features are considered which include length of URL

978-1-7281-7213-2/20/$31.00 ©2020 IEEE

and subdomain number. 10-fold cross validation was used where original dataset was divided into 10 parts. Nine datasets were used for training and one dataset was used for testing. In Artificial Neural Network, one hidden layer framework was used. The accuracy achieved using this classification algorithm is 91%. In Deep Neural Network, two hidden layers were used. This classification algorithm was implemented using Tensorflow. The accuracy achieved using this method is 96%.

In the paper "Extraction of Features and Classification on Phishing Websites using Web Mining Techniques" [10], the method used is web mining technique. This method, BOG (Bag of Words) representation model is used which is used to extract information from documents. Application of data mining techniques on text present in documents to extract useful information is called web content mining. BOG is used to classify documents. The document is classified and put into topic hierarchy where it best fits in. In this method, web phishing dataset is taken, pre-processed and features are selected. Then, various classification algorithms like Naive Bayes, Random Forest, KNN, SVM are used and their performances are assessed. Using Naive Bayes algorithm, 92.9806% of phishing data instances were classified correctly, using KNN,97.1777% of phishing data instances were classified correctly, using Random Forest,97.2592% of phishing data instances were classified correctly and using SVM,93.8037% of phishing data instances are classified correctly. Hence, Random Forest method achieves better performance than remaining algorithms.

In the paper "Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection" [11], the method used is Wrappers Feature Selection that uses a classifier to predict significant features in predicting phishing websites. It is practically not possible to include all the features to train classifier. So, only the most distinguished features are included to train the classifier to detect phishing websites. In this method, inductive classifier is used. The basic idea is to remove redundant features by training the classifier. For each features subset, a score is assigned depending on classification error rate of model. It provides most distinguished features set and improves the performance of Machine Learning classifier. This method uses N-fold cross validation technique to predict phishing websites. The small dataset is partitioned into 'n' equal datasets and the model is trained using remaining datasets. This process is repeated n times. The final accuracy achieved is the average of n-accuracies obtained after running the classifier model n times. The TPR obtained using this method is 0.971 and FPR is 0.969. The advantages of this method is that it provides most important features used for classifier and also improves the performance of phishing website detection. The disadvantages of this method is that it is more time-consuming and involves extra computational overhead.

In a paper published by IEEE [18], the model uses the method of lexical analysis of URL to extract features that helps in detecting phished webpage. The training data is obtained from the lexical feature or the surface level features of a URL. This is then fed to a confidence-weighted learning algorithm. This algorithm then classifies or matches each binary vector from the URL to the binary vectors that it has been trained to detect a malicious website. A URL is split into three units: the protocol (e.g. http), the domain (the parent site or the one that follows the protocol) and path of the object being accessed. These are then converted to tokens. A domain token usually helps in classifying an input URL as malicious or not and hence is considered as- fuzzy blacklist. Also there are certain rules maintained that can be used for classification based on the surface level features. Any confidence-weighted level algorithm can be used here, only difference is that instead host based features, the algorithm uses lexical features of a URL. This method has produced a result with error rates lower than 3%.

A paper published in WORM'07, November 2, 2007, Alexandria, Virginia, USA [19], identifies the different ways a site can be phished and the algorithms to identify these. The methods for obfuscating (making it a phishing website) a URL are: replacing a hostname with IP address, replacing a domain name with a fake but valid looking name, appending extra letters and numbers after the domain name and misspelling host and domain names. The model was trained with the dataset containing whitelist and blacklist. The model is trained with features that are categorized into four types Page Based, Domain Based, Type Based and Word Based features. Then a logistic regression is used to classify the input into phishing or benign URL. Using this technique around 777 unique webpages per day were found as phished website using with this model.

III. PROPOSED APPROACH

Three methods to detect phishing websites have been implemented. The dataset is taken from UCI Machine Learning repository. It consists of several parameters namely IP Address, URL Length, sub-domain, domain registration length, @ symbol and request URL. Some of the python libraries used for implementation are sklearn, utils, numpy and pandas. The libray "sklearn" was used to implement Random Forest, SVM and Neural Network with backpropagation classification algorithms. The library "sys" was used to extract information about constants, functions and methods. The features of URL are extracted using Lexical Feature Extraction and Random Forest classification method, SVM classification method and Neural Network with backpropagation classification algorithm are run to classify the websites as phishing website or legitimate website. The classifier algorithm giving the best accuracy score is selected as the final classifier algorithm.

Initially the classifiers when used produced accuracy rates of 87.34%, 89.63% and 89.84% for neural networks, random forest and SVM classifiers. To improve the accuracy rates, lexical feature extraction was used. This is implemented in utils.py.

Using the above results, which states that SVM is the best classifier, is used to implement a chrome extension. The chrome extension gives an alert popup stating whether the loaded site is prone to phishing or not.

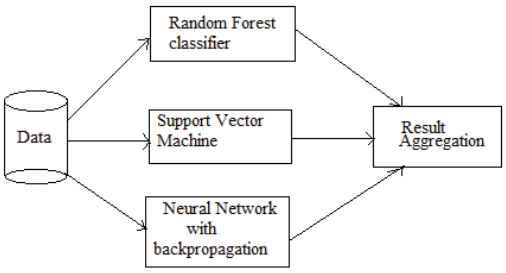

Fig. 1. High level design

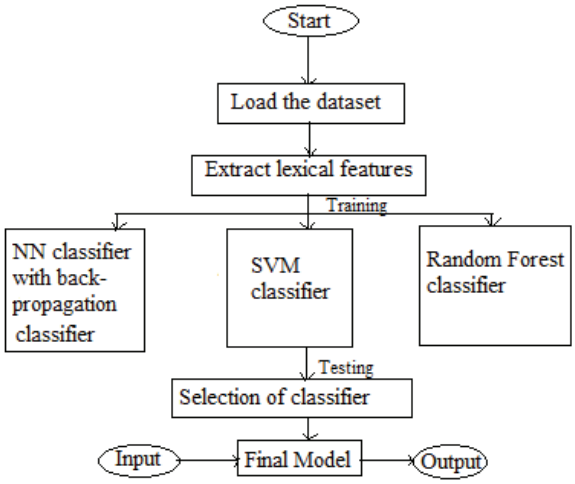

Fig. 2. System architecture

IV. DATASET

The dataset is from UCI Machine learning repository. It consists of 11,055 URLs with 6157 phishing instances and 4898 legitimate instances. Each instance contains 30 features. The result can have either a value of 1 (not phishing) and -1 (phished URL). Each column represents a feature and holds the values 1, -1 or 0. '1' if the URL is completely phished, '0' if the URL is partially phished, '-1' if the URL is benign.

The dataset includes following features: IP address (malicious IP addresses have extra characters or change in certain letters from the original one, e.g.: www.129.B7.fake.html), Long and Short URL (long and short URLs can indicate a phished website), use of @ symbol, '//' symbol can be an indication of redirection, presence of an anchor tag or not, having forms submitting to a blank page, functions related to HTML tags such as onMouseOver(), any pop ups arising from the page, disabled right click to stop users from viewing page source and presence of iframe tags resulting into redirection.

V. MACHINE LEARNING ALGORITHM

The algorithms used are Random Forest, SVM and Neural Networks with back-propagation. Initially lexical feature extraction is performed first on the dataset and then passed to these algorithms.

Random Forest classifier involves combining the results of various stump trees (with single hierarchy) to reach to a conclusion. Each of these trees' results are calculated separately and then combined to give a prediction. This algorithm is implemented using sklearn's RandomForestClassifier module. Since this classifier is irreproducible, the results from this classifier varies from time to time, hence is not constant.

Neural Network includes a series of input layer, number of hidden layers and an output layer. The usage of backpropagation is to reduce the error in final result as the error gets back propagated and the weight given to each hidden layer neurons changes with each iteration. The classifier has been implemented using sklearn's MLPClassifier. Neural networks can be used to perform complex computations; however, it slows down with large datasets.

SVM (support vector machine) involves having a hyperplane that separates the two categories (here: phished or benign website). This hyperplane is also called as the margin. On training with the dataset, the classifier places benign and phished websites on either of the plane, hence classifying the websites. This classifier has been implemented using sklearns's svm.SVC() using the kernel mode as polynomial with a degree of 9.

VI. CHROME EXTENSION

The browser extension is written in JavaScript with python trained model. From the experiment results, SVM is chosen as the final classifier as it classifies better than neural networks and provides better accuracy results unlike random forest classifier.

The extension contains content.js file that extracts the features from the URL and applies the SVM algorithm to find out whether the URL is phished or not, background.js that connects the content.js to frontend and manifest.json that contains the meta data for the extension. The chrome extension has a front end as an alert pop up that says whether the URL is phished or not with an 'OK' button.

Functions implemented in the content.js are:

Is_IPIn_URL(): to check IP address in the URL

Is_Long_URL(): to check if length of the given URL is above 75 characters.

Is_Tiny_URL(): to check if URL has less than 20 characters.

Is_AlphaNumeric_URL(): presence of any '@' symbols or achor tags.

Is_Redirecting_URL(): to check if there is any '//' symbols.

Is_IllegalHttps_URL():check whether there is multiple 'https' in the given URL.

Is_ScLnk_From_Different_Domain(): presence of scripts in URL.

Is_Form_Action_Invalid(): check for blank submit forms.

Is_Iframe_Present(): is there presence of iframe tags redirecting the link to another page.

Once the feature selection is done then, the data is sent to SVM classifier. If the result is 1, then the website is benign, else if the result is -1, the website is phished.

VII. EXPERIMENTAL RESULTS

The accuracies achieved using Random Forest, SVM and Neural Network with backpropagation classification algorithms are 97.369%, 97.451% and 97.259% respectively. The algorithms were improved using lexical feature extraction from the given URL. SVM is found to be the best classifier among the three as it gives better frequency than Neural Networks. Even though Random Forest classifier gives greater accuracy rates than SVM classifier most of the time, the accuracy rates are not constant hence proving SVM to better than random forest classifier.

The chrome extension identifies the phished URLs up to an accuracy of 97.451%.

Table 1. Comparison of accuracies of Machine Learning algorithms

ML Algorithm	Old Result Accuracy	New Result Accuracy (improved using lexical feature analysis on each algorithm)
Random Forest	87.34%	97.369%
Support Vector Machine	89.63%	97.451%
Neural Network with Backpropagation	89.84%	97.259%

VIII. CONCLUSION

Phishing detection is an important step towards phishing attacks and the algorithm used needs to reliable to ensure maximum protection. Various existing methods to detect phishing websites have been mentioned in this paper. Random Forest classification method, SVM classification algorithm and Neural Network classification algorithm were improved. The accuracies obtained using improved Random Forest classification, SVM classification method and Neural Network classification algorithm are 97.835%, 97.89% and 95.444% respectively.

Since, SVM classification algorithm gave better accuracy as compared to that of Random Forest and Neural Network classification algorithms, SVM is chosen as final classifier algorithm for classification of websites as phishing or legitimate.

REFERENCES

[1] Che-Yu Wu, Cheng-Chung Kuo , Chu-Sing Yang,"A Phishing Detection System Based on Machine Learning", International conference on Intelligent Computing and its Emerging Applications (ICEA) 2019

[2] J. Jagadessan, Akshat Shrivastava, Arman Ansari, Laxmi Kanta, Mukul Kumar," Detection and Prevention Approach to SQLi and Phishing Attack using Machine Learning", International Journal of Engineering and Advanced Technology(IJEAT) ISSN:2249-8958, Issue-A:2019

[3] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh,Aram Alsedrani ,"Detecting Phishing Websites Using Machine Learning" ,2nd International Conference on Computer Applications & Information Security 2019.

[4] Dr. G. Ravi Kumar, Dr. S. Gunasekaran, Nivetha R., Sangeetha Prabha K, Shanthini G., Vignesh A. S., "URL Phishing Data Analysis and Detecting Phishing Attacks using Machine Learning in NLP", International Journal of Engineering Applied Sciences and Technology(IJEAST) Vol. 3, Issue 8, ISSN No.2455-2143, Pages 70-75, Published: December 2018

[5] Muhammet Baykara, Zahit Ziya Gürel "Detection of Phishing Attacks", 6th International Symposium on Digital Forensic and Security (ISDFS), 22-25 March, 2018

[6] Ishant Tyagi, Jatin Shad, Shubham Sharma, Sidharth Gaur & Gagandeep Kaur," A Novel Machine Learning Approach to Detect Phishing Websites".5th International Conference on Signal Processing and Integrated Networks (SPIN) Feb 2018.

[7] Tianrui Peng, Ian G. Harris, Yuki Sawa , "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning" 12th IEEE International Conference on Semantic Computing 2018.

[8] S. Jagadeesan, Anchit Chaturvedi, Shashank Kumar, "URL Phishing Analysis using Random Forest", International Journal of Pure and Applied Mathematics, 2018

[9] Ozgur Koray Sahingoz, Saide Işilay Baykal and Deniz Bulut, " Phishing Detection from URLs by using Neural Networks", International Conference on Computer Science engineering and Applications, 2018

[10] Nandhini S., Dr. V. Vasanthi," Extraction of Features and Classification on Phishing Websites using Web Mining Techniques", IJEDR Volume 5, Issue 4, ISSN:2321-9939, Published:2017

[11] Waleed Ali" Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection", IJACSA (International Journal of Advanced Computer Science and Applications, Vol. 8 No. 9, Issue:2017

[12] Fadi Thabtah, Neda Abdelhamid "Deriving Correlated Sets of Website Features for Phishing Detection: A Computational Intelligence Approach", Journal of Information & Knowledge Management Vol. 15, No. 4 (2016) 1650042 (17 pages) World Scientific Publishing Co., 25 November 2016

[13] K.N. Manoj Kumar,K. Alekhya-"Detecting Phishing Websites using Fuzzy Logic", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 10, October 2016.

[14] Priyanka Singh, Yogendra P.S. Maravi , Sanjeev Sharma,"Phishing Websites Detection through Supervised Learning Networks", International Conference on Computing and Communications Technologies (ICCCT) 2015.

[15] Vinnarasi Tharania. I, R. Sangareswari, M. Saleembabu," Web Phishing Detection in Machine Learning Using Heuristic Image Based Method", International journal of Engineering Research and Applications(IJERA) ISSN:2248-9622 Vol. 2, Issue: October 2012

[16] A. Belabed, E. Aïmeur, A. Chikh "A personalized whitelist approach for phishing webpage detection" Seventh International Conference on Availability, Reliability and Security 2012.

[17] Mona Ghotaish Alkhozae,Omar Abdullah Batarfi "Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code", International Journal of Information and Communication Technology Research, Volume 1 No. 6, October 2011

[18] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner "Lexical Feature Based Phishing URL Detection Using Online Learning", Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence, AISec 2010, Chicago, Illinois, USA, October 8, 2010

[19] Sujata Garera, Niels Provos, Monica Chew, Aviel D. Rubin "A framework for detection and measurement of phishing attacks", WORM '07 Proceedings of the 2007 ACM workshop on Recurring malcode Pages 1-8, Alexandria, Virginia, USA, November 02 - 02, 2007