International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 9, Issue 7, July 2020 **ISSN 2319 - 4847**

URL Features Based Phishing Websites Identification Using Deep Learning Techniques

[1]B Aruna Kumari, [2]K Sushmitha Varma, [3]J Shreya and [4]M M S Sahithi

[1,2,3,4]Maharaja vijayaram gajapati raj college of engineering

ABSTRACT

Phishing is a deception that often attacks the users by stealing their personal information like bank account details, credit card details and other data through email, website and SMS. This exploitation can be controlled by understanding and identifying certain features and URLs. In this paper, we suggest a method to detect the phishing websites and minimize the error rate. A Machine Learning technique and a Deep Learning technique are implemented on the data of 2015 URLs and its features. These techniques are tested to observe which one of the two algorithms can give us the effective result. On evaluating, the Multilayer perceptron method achieves 83.3% accuracy by detecting a large number of phishing websites, whereas decision tree classifier achieves 81.1% accuracy. This gives the best algorithm that is suitable for detecting the phishing websites.

Keywords: Phishing, Legitimate, Decision tree algorithm, Multilayer perceptron (MLP).

1. INTRODUCTION :

Cybercrime is a violation that is committed against individuals with a motive to intentionally harm the victim directly or indirectly by using modern telecommunication networks such as Internet and mobile phones. Issues surrounding these types of crimes have become high-profile, particularly those regarding hacking, unwarranted mass-surveillance, copyright infringement, child pornography and phishing. Phishing is a cyberattack that uses disguised email as a weapon to steal users sensitive information. R.S Rao [1] elaborated the intention of phishers by saying phishers mainly trick the email recipients by making them believe that the message they received is something important more like a request from their bank or a note from someone in their company that needs to be seen and addressed right away. Any naïve user would immediately click the link and proceed to download the attachment. This would be an opportunity door for phishers to steal the data of the users. Several anti-phishing techniques emerge everyday but phishers find their way to break their mechanisms. Phishers use a set of features in creating fake websites which leads to the deception. These features can help in identifying and differentiating the phishing websites from the legitimate websites [2]. A Hybrid model proposed by Vaibhav Patil et al [3] uses all the three approaches like blacklist and whitelist, heuristics and visual similarity. It includes checking the http traffic, comparing the domain of each URL with the while list and black list of domains. Further analysis is done on the entire website by considering URL features. According to the Cyber Security Breaches Survey 2018, **43% of businesses** were victims of a cyber security breach. In the U.S, the state of California lost more than **$214 million** through cybercrime alone. It is estimated that the **biggest online data breach** compromised around 130 million user accounts. Many companies and businesses which are working online are either getting hacked or losing their data. The number of cyberattacks has only intensified but never lessened. There is the highest chance for the online brands of being targeted and attacked including payment providing applications too. In the history of cybercrimes, the first cybercrime was recorded in the year 1820 in a textile industry in France. 67% of responding businesses detected cybercrime in the year 2005. This has been recorded as the first report to provide data on monetary loss and system downtime resulting from cyber incidents. In 2017, **McAfee's Economic Impact of Cyber Crime** has become one of the victims to lose 780,000 records per day. In 2018, 93% of the files were found to be affected. The malware observed was polymorphic, meaning it has the ability to constantly change its code to evade detection. Over 50% of devices that got infected once were re-infected within the same year. According to Imperva Cyberthreat Defense Report 2019, a cyberattack became responsible for affecting **78% of surveyed organizations**. These cases are increasing day by day and researchers have come up with some widely used anti-phishing techniques like List-based, Heuristic based, Machine Learning based and Visual similarity based techniques which can detect and alert the phishing websites on the scene. The current paper focusses on Machine Learning and Deep Learning techniques and describes an effective approach against phishing scams based on URLs. The rest of the paper is structured as the related work regarding the paper, our methodology, results of our experiments and then concludes the future directions and remarks.

2. RELATED WORK:

Phishing detection is not only expensive but also it is getting tough for both humans and machines to detect the websites accurately. Users are getting deceived into giving away their data. The phishers on the other hand, are motivated to send their message bypassing automated detection systems and tricking users into interacting with the message. While the manipulation continues, there are a few aspects that are very challenging for them to fully hide. It is the destination of URLs. URLs are the key means that helps the users to locate information on the Internet. To detect phishing websites, classification models are used that can perform the analysis of features of URLs. Ranking of the sites are also considered as they are in great use to enhance the prediction of phishing websites [14]. According to Joby James et al [3], different classifying algorithms were analysed in Waikato Environment for Knowledge Analysis (WEKA) workbench and MATLAB. Kholoud Althobaiti et al [3] reviewed URL-based anti-phishing features that were aimed at both humans and automated systems. The main focus is on humans and computers to obtain a more comprehensive feature list. Parikh et al [4] stated, phishers started attacking users of online banking, payment services such as PayPal, and online e-commerce sites. The approach is a malicious URL is created and is then directs the user to the desired malicious page that the attacker wants. These can be recognized by utilizing URL identification strategy like Random Forest algorithm. This has three stages namely Parsing, Heuristic Classification of data, Performance Analysis. Many techniques have been proposed to detect phishing, and are classified into three categories such as blacklist, heuristics and machine learning [19][22][24][29]. Some use computer vision algorithms to detect phishing attacks [20] while some developed a real-time, language-independent anti-phishing classifier [21][27][30]. To prevent phishing, network operators need to protect their users from accessing such malicious sites Shima et al [5]. Since, the number of phishing sites is huge and growing, we need an automated and adaptive approach to defend customers from such activities. One of the best techniques is the hybrid approach combining both clustering and classification [11], [15][25]. It can be done ranking URLs and adapting to new characteristics of them. According to IBMs research team, Vazhayil et al [6][28] the number of spam mails is increasing rapidly and it is also noticed that more than half of the emails produced are scam. Phishers usually send emails with phishing URLs that direct them to fake websites. In the mail they mix authentic links and false links to make it look more appealing. Modern web browsers use add-ons and plugins to detect phishing URLs and we can also create a classifier which accepts a link from a user and checks the URL based on its URL features [12], [16][26]. Tools use blacklisting to detect dangerous websites but fails when an unknown phishing URL is encountered. This constantly requires updating the database for the mentioned techniques. Classical Machine Learning techniques like logistic regression using bigram, c4.5 decision tree to calculate heuristic values [13], naïve based classifier and support vector machines [17], RBF Network [18], Deep Learning techniques like CNN and CNN long short-term memory as architecture are also used. Most of the existing systems use only one Machine Learning algorithm to predict the accuracy. Sophiya Shikalgar [23] proved that using only one algorithm is not a good approach to improve the prediction accuracy. According to Neda Abdelhamid [7], to overcome the limitations of traditional anti-phishing approaches, computer security experts developed toolbar visualization techniques, such as EbayGuard, Netscape, Netcraft, McAfee webadvisor among others. This will help in revealing certain security information to the end-user about potential online risks, such as phishing attacks. A. Ahmed and N. Abdullah et al [8] stated that 70% of successful phishing attacks have begun through social networks and the lack of awareness and education on web spoofing attacks have caused the fall of the victims. They also averred that the inability to distinguish between the fake and legitimate web pages is still a challenge in the existing prevention solutions of web spoofing.

3. METHODOLOGY:

There are various methods involved to detect phishing and legitimate websites. The method we used focusses on identifying phishing and legitimate URLs based on its features. In this paper, the algorithms used are a Machine Learning technique known as Decision Tree (algorithm 1) and a Deep Learning technique known as MLP (algorithm 2). Machine learning techniques can construct classification models by analyzing a dataset. The classifiers like Decision Tree, KNN, Random forest, Naïve Bayes and Artificial Neural Networks can achieve very high classification accuracy. A Machine Learning classifier such as a Decision Tree classifier is imported from sklearn library. The data will be fitted into the model. A confusion matrix is created and computed which further results in the calculation of the metrics. Thereafter graphs will be plotted based on those metrics. The Neural Network activation functions like Tanh, Sigmoid function, Linear function and ReLU play a crucial role in Deep Learning as they introduce the non-linear property and determine the output of any model along with the accuracy. Similar to the Decision Tree classifier, an activation function is used, as it can predict the probability and gives an output. To use this function, Keras module can help in creating a Neural Network model where the hidden layers can be added to the model to reduce the error

occurrence. This will follow the same process as the Decision Tree classifier and provides the confusion matrix with its metrics. Henceforth the graphs will be plotted. The dataset that has been used is downloaded from Phish tank and it consists of 2015 URLs with selected features. With respect to the processed dataset, all the unnecessary columns have been removed while the rest of the dataset is identified by the label class 1 for phishing URLs and 0 for legitimate URLs. It is divided into 70% training data and 30% testing data. Decision tree classifier is fitted upon training data. Similarly the deep learning technique is also fitted upon the same training data. This helps to find the confusion matrix and then leads to the calculation of accuracy. The algorithm which gives the highest accuracy is considered as the effective algorithm.

Algorithm 1 (Decision Tree):

```
Calculating the accuracy of the training tuples of the data partition.
Input:    Data partition, a set of training tuples.
          Attribute list, the set of attributes.
Output:  Accuracy.
Method:
import the dataset, D
if attribute list is not empty then
apply attribute selection method; //remove unnecessary attributes
attribute list ← attribute list − splitting attribute;
split the dataset, D  into training tuples and testing tuples.
create the model(decision tree classifier);
fit the training tuples into the model;
predict the result for the testing tuples by using the model
create the confusion matrix;
calculate the accuracy //accuracy_score
          return accuracy_score;
```

Algorithm 2 (Multilayer perceptron):

```
Calculating the accuracy of the training tuples of the data partition.
Input:    Data partition, a set of training tuples.
          Attribute list, the set of attributes.
Output:  Accuracy.
Method:
import the dataset, D
if attribute list is not empty then
apply attribute selection method; //remove unnecessary attributes
attribute list ← attribute list − splitting attribute
split the dataset, D  into training tuples and testing tuples
create the model(keras model)
add hidden layers to the model
fit the training tuples into the model
predict the result for the testing tuples by using the model
create the confusion matrix
calculate the accuracy //accuracy_score
          return accuracy_score;
```

4. FLOWCHART:

The following flowchart (Figure (i)) outlines the process of learning which algorithm is best suitable for the detection of phishing websites. This flowchart gives a vivid view of how the procedure takes place. It states that out of the Decision

Tree and Neural Network algorithms, the algorithm that results in maximum value of accuracy is to be used. Initially, a dataset full of URLs and its features is downloaded and then split into training and testing data. Applying the addressed algorithms, the accuracies will be calculated. This helps in understanding which technique is efficient once the accuracy rates are displayed. In the end, the algorithm that gives the maximum accuracy is considered as the better approachable process to detect the phishing websites than the other algorithm.

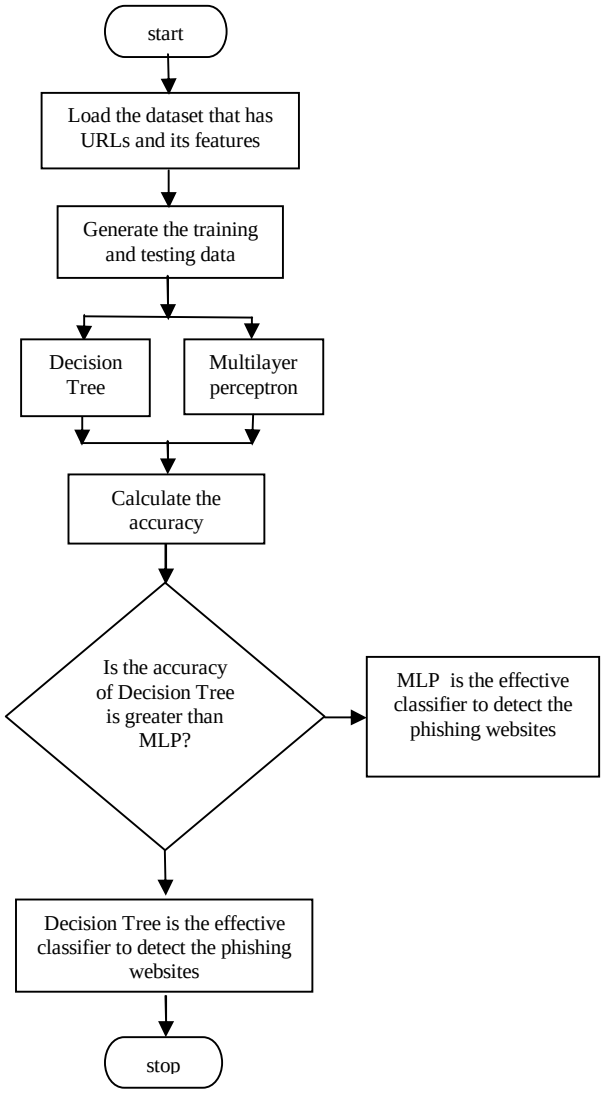

Figure (i) – process for phishing website detection

5. ENVIRONMENTAL SETUP:

To detect the phishing and legitimate websites, Machine Learning algorithm and Deep Learning algorithm run on Windows 7, 32 bit Operating system using core I3 processor and having Python version of 3.6.5. Pandas are used to read the dataset. The dataset will be split into training and testing data by using sklearn library. The same library will be used to import the decision tree classifier which can further give the confusion matrix. With the help of confusion matrix, the metrics like accuracy, recall, precision, f-measure, specificity and sensitivity are calculated. Once the result of the metrics is obtained, the graphs will be drawn using matplotlib library. In the final step, a three dimensional graph for run, test_size and accuracy is attained. To test the same problem using neural networks, import the keras library and create a neural network model. This library is used to add hidden layers to the model. The sklearn library is used to split the dataset into training and testing which further gives confusion matrix and its metrics. The graph will be drawn to visualize the metrics using matplotlib library.

6. EXPERIMENTAL EVALUATION:

The bar graph below displays the scores of the metrics (accuracy, recall, precision and f1-score). The scores can be of any value as they are entirely based on the confusion matrix. A method can be only called as a perfect method when the values are high. From both figures (ii), (iii), it is conveyed that the values indicate that the approach of using a Deep Learning model is better than the Decision Tree classifier.

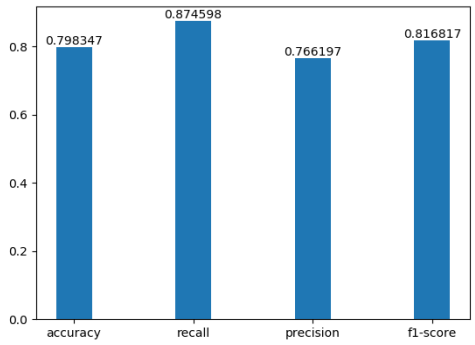

Figure (ii) - Decision Tree Classifier graph

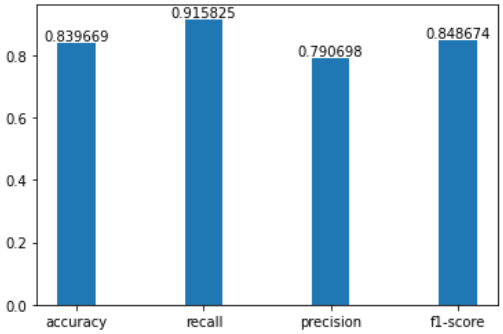

Figure (iii) - Neural Network graph

A 3-D graph (Figure (iv)) is drawn by taking the run, test_size and accuracy as three different dimensions. Run indicates the number of times we execute a code, test-size indicates the percentage of data that is taken from data set, accuracy denotes the quality of being correct.

run	test_size	accuracy
1	0.1	0.8465
2	0.2	0.8461
3	0.3	0.8413
4	0.4	0.8424
5	0.5	0.8313
6	0.6	0.8196
7	0.7	0.8065
8	0.8	0.7872
9	0.9	0.7502

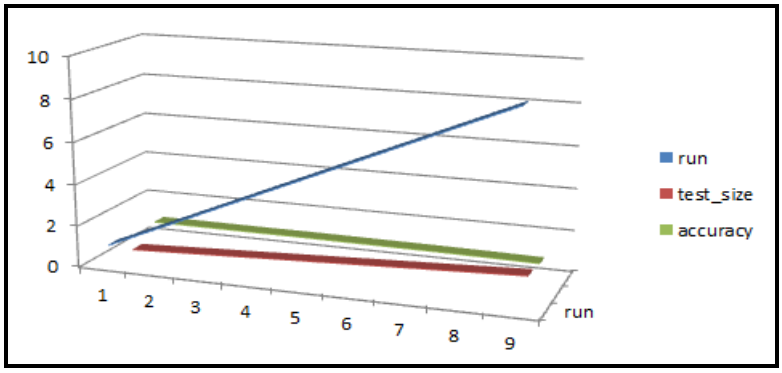

Figure (iv) - Run vs Accuracy

7. CONCLUSION AND FUTURE WORK:

Phishing attacks are continuing to grow and phishers are finding new ways to attack and commit crime. There are already some classifiers that detect phishing websites and give good prediction rates. Every model that has ever been developed spectacles the effective technique which can be used in identifying the phishing websites. The model we propose uses a Machine Learning and a Deep Learning technique. The Decision tree classifier of Machine Learning resulted in 79.83% of accuracy and Multilayer perceptron of Deep Learning achieved 83.96% of accuracy. It is clear that the Deep Learning approach showed a better performance and accuracy rate than the Machine Learning approach. This model was done by taking few selected features that distinguished both phishing and legitimate websites. In future, more features and URLs are added to improve the accuracy of the model and also other algorithms are considered that can possibly increase the efficiency of the proposed system.

REFERENCES:

[1] R S Rao, PhishShield: A desktop application to detect phishing Webpages through heuristic approach, 2015, ISBN: 978-1-4503-4956-7/17/04.
[2] https://github.com/chamanthmvs/Phishing-Website-Detection
[3] KholoudAlthobaiti, GhaidaaRummani, and Kami Vaniea, A review of human- and computer-facing URL phishing features, 2019, ISBN: 978-1-7281-3027-9
[4] S. Parikh, D. Parikh, S. Kotak, and P. S. Sankhe, A New Method for Detection of Phishing Websites: URL Detection, 2018, ISBN 978-1-5386-1974-2.
[5] K. Shima, Classification of URL bitstreams using bag of bytes, 2018.
[6] A. Vazhayil, R. Vinayakumar, and K. Soman, Comparative study of the detection of malicious URLs using shallow and deep networks, 2018, ISBN: 978-1-5386-4430-0.
[7] Neda Abdelhamid, Phishing detection: A recent intelligent machine learning comparison based on models content and features, 2017.
[8] A. Ahmed, N. Abdullah, Real time detection of phishing website, 2016, ISBN: 978-1-5090-0996-1.
[9] Joby James, Sandhya L, Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, 2013.
[10] Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Detection and Prevention of Phishing Websites using Machine Learning Approach, 2018.
[11] Mohammed Nazim Feroz, Susan Mengel, Phishing URL detection using URL Ranking, 2015.
[12] Ammar Yahya Daeef, R. Badlishah Ahmad, Yasmin Yacob, Ng Yen Phing, Wide scope and fast websites phishing detection using URLs lexical features, 2016.
[13] R. Kiruthiga, D. Akila, Phishing Websites Detection Using Machine Learning, 2019, ISSN: 2277-3878.
[14] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen, A novel approach for phishing detection using URL-based Heuristic, 2014.
[15] Bhawna Sharma, Parvinder Singh, PhishAlert: An Efficient Phishing URL Detection via Hybrid Methodology, 2019, ISSN: 2278-3075.
[16] Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon and Lee, Heuristic-based Approach for Phishing Site Detection Using URL Features, 2015, ISBN: 978-1-63248-056-9.
[17] Ankit Kumar Jain and B. B. Gupta, PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning, 2018.
[18] Ying Xue, Yang Li, Yuangang Yao, Xianghui Zhao, Jianyi Liu, Ru Zhang, Phishing Sites Detection Based On Url Correlation, 2016, ISBN: 978-1-5090-1256-5.
[19] Ram B. Basnet and Andrew H. Sung, Mining Web to Detect Phishing URLs, 2012.

[20] Rachel Greenstadt and Sadia Afroz, PhishZoo: Detecting Phishing Websites By Looking at Them, 2011, ISBN: 978-0-7695-4492-2.
[21] Jianyi Zhang and Yonghao Wang, A Real-time Automatic Detection of Phishing URLs, 2012, ISBN: 978-1-4673-2964-4.
[22] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen, Detecting Phishing Web sites: A Heuristic URL-Based Approach, 2013.
[23] Ms. Sophiya Shikalgar, Dr. S. D. Sawarkar and Mrs. Swati Narwane, Detection of URL based Phishing Attacks using Machine Learning, 2019, ISSN: 2278-0181.
[24] P Srinivasa Rao, MHM Krishna Prasad,K Thammi Reddy,"A Novel And Efficient Method For Protecting Internet Usage From Unauthorized Access Using Map Reduce" Published In IJITCS (MECS) Vol. 5, No. 3,Pp:49-55,February 2013.
[25] Bhawna Sharma, Parvinder Singh, PhishAlert: An Efficient Phishing URL Detection via Hybrid Methodology, 2019, ISSN: 2278-3075.
[26] Jin-Lee Lee, Dong-Hyun Kim, Chang-Hoon and Lee, Heuristic-based Approach for Phishing Site Detection Using URL Features, 2015, ISBN: 978-1-63248-056-9.
[27] P Srinivasa Rao, S Satyanarayana ,"Privacy Preserving Data Publishing Based On Sensitivity in Context of Big Data Using Hive",Journal of Bigdata(Springer), Volume:5,Issue:20, ISSN: 2196-1115, July 2018.
[28] Satish Muppdi, P Srinivasa Rao, M Rama Krishna Murthy,"Identification of Natural Disaster Affected Area Using Twitter ", 2nd International conference on Cyber Security, Image Processing, graphics,Mobility and analytics,NCCSIGMA-2019,Advances in Decision Sciences, Image Processing, Security and Computer Vision,Springer Nature,PP:792-801,2019.
[29] P Srinivasa Rao, MHM Krishna Prasad,K Thammi Reddy, "An Efficient Semantic Ranked Keyword Search Of Big Data Using Map Reduce", IJDTA,Vol.8, No.6 , Pp.47-56,2015.
[30] P. S. Latha Kalyampudi, P. Srinivasa Rao, and D. Swapna, "An Efficient Digit Recognition System with an Improved Preprocessing Technique", Springer Nature Singapore, ICICCT 2019 – System Reliability, Quality Control, Safety, Maintenance and Management, pp. 312–321, 2019