Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-1 Issue-2, April 2021

K-Nearest Neighbor Based URL Identification Model for Phishing Attack Detection

Tsehay Admassu Assegie

Abstract*: Phishing causes many problems in business industry. The electronic commerce and electronic banking such as mobile banking involves a number of online transaction. In such online transactions, we have to discriminate features related to legitimate and phishing websites in order to ensure security of the online transaction. In this study, we have collected data form phish tank public data repository and proposed K-Nearest Neighbors (KNN) based model for phishing attack detection. The proposed model detects phishing attack through URL classification. The performance of the proposed model is tested empirically and result is analyzed. Experimental result on test set reveals that the model is efficient on phishing attack detection. Furthermore, the K value that gives better accuracy is determined to achieve better performance on phishing attack detection. Overall, the average accuracy of the proposed model is 85.08%.

Keywords: Phishing attack, Machine learning, KNN Network security, Phishing detection.

I. INTRODUCTION

Phishing is a type of social engineering attack where the attacker attempts to gain access to a system by collecting sensitive information such as user name, password and credit card details [1]. The attacker uses forged websites to collect the sensitive information from online users such as mobile bank customers. Once, the sensitive information is gathered though forged websites, then the attacker gets access to a system and such access causes financial loss to the legitimate users.

Numerous methods and countermeasures have been proposed to safeguard users from phishing attack [2]. One of phishing attack detection approaches is content-based anti-phishing. Content-based anti-phishing approach is visual similarity employed to identify the contents of phishing websites form legitimate website by analyzing the similarity of contents. Another approach to phishing attack detection is classification of URL into malicious and legitimate classes by employing machine-learning algorithm and proposing model that automatically detects malicious URL and takes an action to stop access to such URLs.Although, different approaches have been proposed for detecting phishing attack, phishing attack remained a major challenge to the business industry involving online transaction. One of the major challenges of phishing attack detection is that attackers are adapted to different phishing attack detection approaches. Consequently, an effective and efficient phishing detection approach is important to tackle the problem of phishing attack. Hence, we are motivated in designing and implementing an alternative approach to detect phishing attack more efficiently with machine learning. Overall, the contribution of this study is to provide an effective model to phishing attack detection by proposing machine-learning model that classifies URL into phishing or legitimate classes automatically. Therefore, the objectives of this study is to explore the answers to the following questions:

- How to create a classification model by employing KNN algorithm to detect phishing attack?
- What is the accuracy of KNN algorithm on phishing attack detection?
- What is the value of K that gives optimal accuracy on phishing attack detection?

II. LITERATURE REVIEW

Numerous studies has been conducted on phishing attack detection problem and a number of approaches have been proposed although phishing attack detection still remains a major challenge and much work is required to overcome the challenges of phishing attack. This section presents a review of recently published studies related to phishing attack detection using automated predictive model for decision support on identifying whether a given URL is suspicious or legitimate.In [3], the authors proposed decision tree based model for phishing attack detection. The authors employed 11,055 observation of suspicious or phishing and nonsuspicious or legitimate websites with 30 features. The performance of the proposed model is evaluated on test set and result shows that decision tree algorithm is effective on phishing attack detection, although the accuracy is promising, there is still larger scope for improving the performance to get better results.In another study [4], conducted on phishing attack detection problem, random forest algorithm is employed to phish tank dataset and model for phishing attack detection is proposed to classify RUL into malicious or phishing and legitimate classes. The performance of the proposed is evaluated and the result shows that the proposed model has acceptable accuracy on phishing attack detection. In [5], convolutional neural network based intelligent phishing attack detection model is proposed.

Manuscript received on 31 March 2021 | Revised Manuscript received on 05 April 2021 | Manuscript Accepted on 15 April 2021 | Manuscript published on 30 April 2021.

***** Correspondence Author

Tsehay Admassu Assegie*, Department of Computer Science, Faculty of Computing Technology, Aksum Institute of Technology, Aksum University, Axum, Ethiopia. Email: tsehayadmassu2006@gmail.com

The proposed model is evaluated against performance on phishing attack detection and the accuracy of the model is acceptable.

In [6], deep learning framework is proposed for website phishing attack detection. The proposed model is effective on detecting phishing attack. The accuracy of the proposed model is tested empirically and result shows that the model has better accuracy. The model is helpful on automatic identification of URL as phishing attack or legitimate.

In [7] support vector machine based phishing attack detection model is proposed for automatic and intelligent classification of URL into legitimate and malicious classes. The proposed model has 60% accuracy on test set as shown in experimental result analysis on the performance of the model on phishing attack detection.

In [8], hybrid approach using an ensemble voting classifier is used to implement an intelligent phishing attack detection. In the implementation, the authors applied Naive Bayes and Decision tree algorithm as base classifier and voting ensemble classifier to achieve better accuracy on phishing attack detection.

In [9], random forest algorithm is employed on phish tank dataset and a model for phishing attack detection is proposed. The authors conducted experimental test on the model the accuracy of the model against phishing identification is analyzed. Experimental result shows better accuracy is achieved on phishing attack detection using the proposed method. A comparative study conducted on an ensemble classification [10] namely, random forest and xboost model for phishing attack detection shows that an ensemble model performed well on phishing detection as compared to the random forest model.

In [11], random forest is applied on phishing tank data repository and an automated phishing attack detection model is proposed. The performance of the proposed model is evaluated against time complexity on test set and result shows that the model is effective on detection of suspicious emails.

In another study [12], hybrid approach is proposed to phishing attack detection problem. A machine-learning model is proposed for automatic identification of phishing attack by employing K-nearest neighbor (KNN) and Support vector machine (SVM). In [13], random forest and K-nearest neighbor (KNN) algorithm is employed for phishing attack detection. Comparative result on the performance of the KNN and random forest model shows that random forest is better on phishing attack detection as compared to KNN.

In another study [14], Naïve Bayes and logistic regression algorithm is employed to implement a model for phishing attack detection. The proposed model detects phishing attack ased on the prior experience provided to learning algorithm during training phase. The comparison on the performance of logistic regression and Naïve Bayes shows that logistic regression performed better compared to Naïve Bayes. In [15], hybrid support vector machine and neural network based phishing attack detection model is proposed to automatically classify a URL into suspicious and legitimate classes. The analysis of the performance of the model on identification of phishing attack shows that hybrid approach is better than an individual algorithm on detecting phishing attack.

III. METHODOLOGY

This study employed data collected from kaggle Website phishing data repository. We have employed K-Nearest Neighbor (KNN) with Python programming language to implement the proposed model in Jupyter Notebook environment. The data repository consists of 1,353 observations of malicious and legitimate classes and 10 features describing each observation in the dataset. The dataset features are summarized in table 1.

Table- I: Phishing attack dataset feature description

Feature	Decription
URL_Length	Length of URLs.
popUpWidnow	HTML source code contains pop-up
age_of_domain	Age of the domain name
having_IP_Address	URL host name contains IP address
web_traffic	Number of web pages visited by the visitors
Request_URL	Request URL
SSLfinal_State	The existence of HTTPS in URL
SFH	An empty string or about: blank
URL_of_Anchor	Tags and website domain name

The correlation matrix for the features shown in table 1 are demonstrated in figure 1, which shows the relationship between each features in the dataset. As shown in figure 1, URL-length is strongly related to the result feature or the class variable. This is an indication of phishing attack. Thus, if the URL length is larger in length, than the URL is more likely phishing attack or illegitimate.

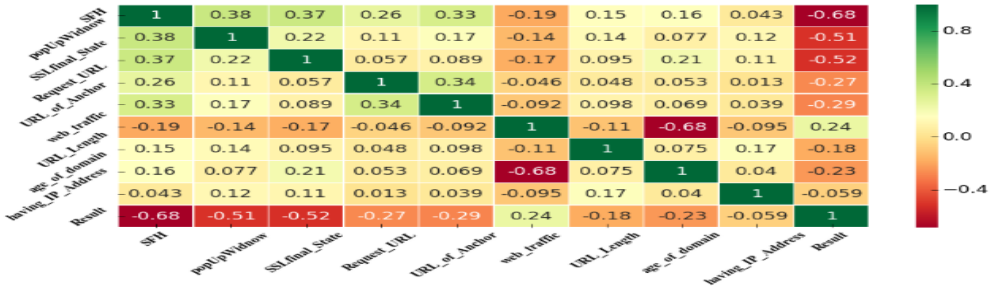

Fig. 1.Correlation of phishing attack dataset features

IV. EXPERIMENTAL RESULT AND DISCUSSIONS

In this section, experimental results of the study are explained. We have evaluated the proposed model on test set and the performance of the model is analyzed. To conduct the experiment on the effectiveness of the proposed model, accuracy metric is used as performance measure.

A. Accuracy of the proposed model

The accuracy of the proposed model is evaluated against different values of K to find K value that gives better accuracy on detecting phishing attack through URL classification. Table 2 demonstrates accuracy of the proposed against different values of K.

Table- 2: The effect of neighbors on accuracy

Neighbors (1-10)	Accuracy of the model in percentage
1	82.28%
2	83.02%
3	84.13%
4	84.50%
5	84.87%
6	85.97%
7	85.97%
8	87.08%
9	85.23%
10	87.82%

As illustrated in table 2, the accuracy of the model on phishing attack detection has slightly increased with the increase in K value with the highest accuracy at k=10. Thus, the highest accuracy is achieved when k value is 10. Furthermore, the accuracy of the proposed model varies with different values of k is shown in figure 2. As demonstrated in figure 2, training and test accuracy has decreased with higher values of k. This shows that determining better k value is vital to achieve possible highest accuracy with KNN model on phishing detection.

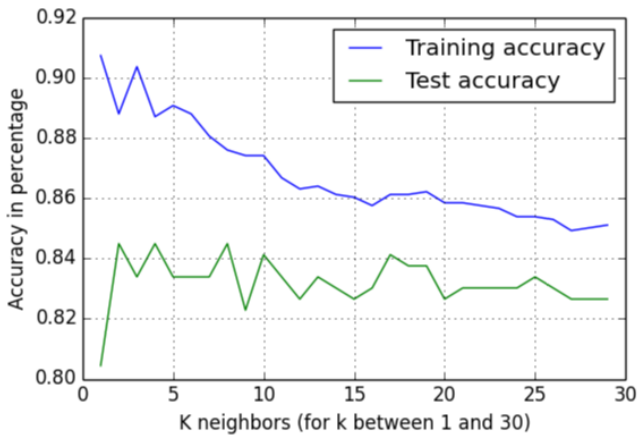

Fig. 2.The effect of neighbors on performance

Figure 2 shows how the accuracy of the proposed model changes with different values of K. As shown in figure 2, the test accuracy is highest when the values of K is between 3 and below 20. The optimal number of neighbors is 10, which is used for training for achieving highest possible accuracy on phishing attack detection.

B. Confusion matrix

A confusion matrix is used as performance measurement to evaluate the proposed phishing attack detection model. We have analyzed the performance of the model on test data consisting of 106 observations and the confusion matrix of the proposed model for phishing attack detection is shown in figure 3.

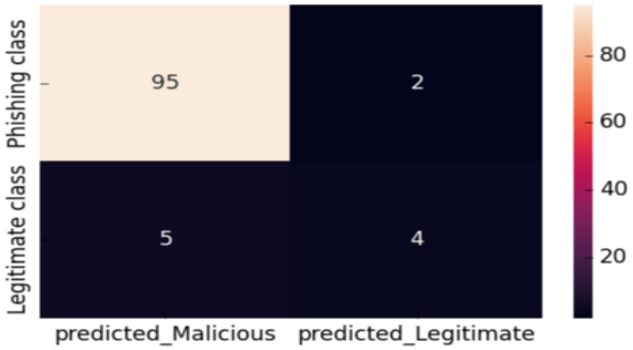

Fig. 3.Confusion matrix

The confusion matrix of the proposed model shown in figure 3 reveals that the model is efficient on phishing attack detection. However, there are misclassified observations among the 106 observations used in testing. In 106 observations of testing set containing suspicious or phishing and legitimate observations, 7 observations are misclassified Thus, the result shows that the performance of the proposed model is promising.

V. CONCLUSION

In this study, K-Nearest Neighbor (KNN) based model is proposed for phishing attack detection through classification URL into malicious or phishing and legitimate class. We have analyzed the performance of the proposed model using 106 observations for testing the model on phishing detection. Experimental result using accuracy metrics as performance measure shows that proposed model is effective on detecting phishing attack. Overall, the accuracy of the proposed model is 85.08%. In conclusion, we found that the proposed model is useful for combating the escalating cases of phishing attack in the business industry

As a future work, we recommend the researchers to conduct a study on phishing attack detection suing different machine learning algorithms such as support vector machine, decision tree and random forest. Furthermore, more dataset is required to design and implement an effective and robust model to overcome the problem of phishing attack in the business industry.

REFERENCES

1. AlMaha Abu Zuraiq, Mouhammd Alkasassbeh, Review: Phishing Detection Approaches, IEEE, 2019. [CrossRef]

2. 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, IEEE, 2019.
3. A. K. Shrivas, Ramkishun Suryawanshi, Decision Tree Classifier for Classification of Phishing Website with Info Gain Feature Selection, International Journal for Research in Applied Science & Engineering, 2017.
4. Adwan Yasin, Abdelmunem Abuhasan, An intelligent classification model for phishing email detection, International Journal of Network Security & Its Applications, 2016. [CrossRef]
5. Weiping Wang, Feng Zhang, Xi Luo, Shigeng Zhang, PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks, Hindawi Security and Communication Networks Volume 2019. [CrossRef]
6. Ping Yi, Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu Web Phishing Detection Using a Deep Learning Framework, Hindawi Wireless Communications and Mobile Computing Volume 2018. [CrossRef]
7. Mouad Zouina, Benaceur Outtaj, A novel lightweight URL phishing detection system using SVM and similarity index, Zouina and Outtaj Hum. Cent. Comput. Inf. Sci, 2017. [CrossRef]
8. Ch. Chakradhara Rao, A. V. Ramana, Detection of Phishing Websites Using Hybrid Model, JOURNAL OF ADVANCEMENT IN ENGINEERING AND TECHNOLOGY, 2018.
9. Vamsee Muppavarapu, Archanaa Rajendran, Shriram Vasudevan, Phishing Detection using RDF and Random Forests, The International Arab Journal of Information Technology, Vol. 15, No. 5, September 2018.
10. T. A. Assegie, Bizuneh H. D., Improving network performance with an integrated priority queue and weighted fair queue scheduling, Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 1, pp. 241-247, July 2020, doi: 10.11591/ijeecs.v19.i1.pp241-247. [CrossRef]
11. Bayu Adhi Tama, Kyung Hyune Rhee, A Comparative Study of Phishing Websites Classification Based on Classifier Ensembles, Journal of Multimedia Information System VOL. 5, NO. 2, June 2018.
12. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, Intelligent Phishing Website Detection using Random Forest Classifier, International Conference on Electrical and Computing Technologies and Applications, IEEE, 2017.
13. Assegie, T.A, Software defined network emulation with OpenFlow protocol, International Journal of Advances in Applied Sciences (IJAAS), vol. 9, no. 1, pp. 70-76, March 2020, doi: 10.11591/ijaas.v9.i1.pp70-76. [CrossRef]
14. Altyeb Altaher, Phishing Websites Classification using Hybrid SVM and KNN Approach, International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017. [CrossRef]
15. Nur Sholihah Zaini, Deris Stiawan, Mohd Faizal Ab Razak, Ahmad Firdaus, Wan Isni Sofiah Wan Din, Shahreen Kasim, Tole Sutikno, Phishing detection system using machine learning classifiers, Indonesian Journal of Electrical Engineering and Computer Science Vol. 17, No. 3, March 2020. [CrossRef]
16. Nur'Ain Maulat Samsudin, Cik Feresa binti Mohd Foozy, Nabilah Alias, Palaniappan Shamala, Nur Fadzilah Othman, Wan Isni Sofiah Wan Din, Youtube spam detection framework using naïve Bayes and logistic regression, Indonesian Journal of Electrical Engineering and Computer Science Vol. 14, No. 3, June 2019. [CrossRef]
17. Assegie, T.A., Nair, P.S, A review on software defined network security risks and challenges, TELKOMNIKA, Vol.17, No.6, December 2019, pp.3168~3174. [CrossRef]
18. Abhishek Kumar, Jyotir Moy Chatterjee, Vicente García Díaz, A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing, International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 1, February 2020. [CrossRef]
19. Assegie, T.A., Nair, P.S, Comparative Study on Methods Used in Prevention and Detection Against Address Resolution Protocol Spoofing Attack," Journal of Theoretical and Applied Information Technology, vol. 97, no. 16, August 2019.

AUTHORS PROFILE

Tsehay Admassu Assegie has completed his M.Sc. in Computer Science at Andhra University, Department of Computer Science and Systems Engineering, Vishakhapatnam (India) in 2016 and B.Sc. in Computer Science in 2013 at Dilla University Dilla, SNNP (Ethiopia). Tsehay is currently working as full time lecturer at Department of Computer Science, Aksum University, Axum (Ethiopia). Tsehay has published 20 research articles in international journals and participated in international conference. His research interest include Artificial Intelligence, Machine Learning, Network Security and Software Defined Network. Tsehay has authored two books published with ISBN number. He has 3 years of research experience and over 7 years of teaching experience in academia.