International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparative Study of KNN and LR Approaches of Machine Learning with Respect to the Identification of Phishing Websites

Dr. Sachin Kadam[1], Mrs. Nidhi[2], Dr. Pratibha Deshmukh[3], Mrs. Nidhi Khare[4], Mr. Irfan Khatik[5]

Submitted: 17/07/2023 **Revised**: 07/09/2023 **Accepted**: 22/09/2023

Abstract: With the advent of Internet and growth in the field of Information and Communication technology, phishing attacks are becoming very common source for finding users personal or confidential information. These types of attacks are executed through email, websites, instant messaging services etc. This type of attack is very common and is also considered as one of the major threats to the organization. Therefore, it becomes very important for an individual to check if the message has been received from the trusted sender, as it fools the victim by pretending to be the original user and asking them to share their personal and confidential information. There are lots of techniques which are used to detect phishing websites. In this paper, the two machine learning classification algorithms: K-Nearest Neighbors (KNN) and Logistic Regression (LR) are applied to the phishing and non-phishing website URLs dataset. The performance of classification algorithms KNN and LR are compared by using the classification report accuracy, precision, confusion matrix, sensitivity, f-score and time required for its execution. Hence, this paper will compare the accuracy of KNN and LR models in order to find phishing websites. The major objective of this paper is to use key features to detect phishing websites with higher accuracy and also lower rate of error.

Keywords: *KNN modal ML, LR model ML, Phishing and Non Phishing Websites identification using KNN, Phishing and Non Phishing Websites identification using LR*

1. Introduction

The internet evolution has changed our lives; it has been widely used in all the fields for easing out our tedious jobs. The various domains like banking, healthcare, education, retail etc. have been massively benefitted and the way the tasks are automated as well as the facilities provided in turn of that has made lives of individuals easy. As the Internet has penetrated in all domains, therefore it becomes necessary to look at the security measures too. Using appropriate security mechanisms leads to prevention of unauthorized access and misuse of data related to any person or organizations sensitive or confidential information. The attackers use various techniques to exploit the vulnerabilities and penetrate the structure which helps them to gain illegal right of entry to private information.

Phishing is a cybercrime which involves lure the user into given that their private, secret and susceptible information. This may include usernames, passwords, banking details etc. Phishing attacks are appropriate familiar as false websites are on the rise; these attacks are mainly through email, SMS etc. These attacks may cause low to high level losses for an individual or any organization by misuse or leak the data. Example: A investigator established that the email link was forwarding it to a forged website that seem to have the correct URL in the browser, but the user was trick by using some characters in flanked by which resemble the lawful URL and domain name. Various Machine Learning techniques are used to detect phishing websites. Various classification algorithms can be used to analyse phishing websites accurately and efficiently. The main goal of this paper is to find out the key features for detecting phishing websites by applying KNN and LR machine learning algorithms on website URL datasets. These websites can easily be detected and effectively in less time with high accuracy.

2. Literature Review

Chaudhry et al. [1] have discussed different phishing attacks, how it could be detected and prevented, in their paper they talk about different modes from which these attacks could happen like email, malicious sites. Basit et al. [2], paper presented a literature review of different techniques in AI to detect phishing websites which are happening because of user data available online like their login credentials, cards detail. In the paper they compare different techniques for detecting phishing attacks like Deep learning, machine learning etc. Odeh et al. [3] have briefly

[1] *Institute of Management and Entrepreneurship Development, Bharati Vidyapeeth (Deemed to be University), Pune (India)*
sachin.a.kadam@bharatividyapeeth.edu
[2] *Bharati Vidyapeeth's Institute of Management & Information Technology, Navi Mumbai (India)*
nidhi.poonia@bharatividyapeeth.edu
[3] *Bharati Vidyapeeth's Institute of Management & Information Technology, Navi Mumbai (India) pratibha.deshmukh@bharatividyapeeth.edu*
[4]*Symbiosis Skills and Professional University, Pune (India) nidhikhare89@gmail.com*
[5] *Fergusson College, Pune (India)*
irfan.khatik@fergusson.edu

discussed how phishing attacks are getting information of sensitive data of users and in the paper discussed and surveyed about different machine learning techniques like SVM, RF, NB. In the paper different deep learning techniques are also compared for detecting phishing websites. Wu et al. [4] have shown that the toolbar plays very important role in detecting phishing websites, in the paper he discussed the user should pay attention to toolbars warnings if website looks fake and also analysed in his study that the toolbars security is not able to protect phishing attacks. Song et al. [5] talks about classifiers based on machine learning that are not able to detect phishing pages. Athulya et al. [6], in their paper discussed different phishing risks, attacks and prevention by using a hybrid approach to finding it in less time with more accuracy. A study by Gupta et al [7] provides a different taxonomy of phishing attacks history and its solutions and in the paper they also discussed the effect and impact of these attacks. Fette et al. [8], in their paper communicates about what is phishing how the data which is important for the user is steeled by sending email of spoofed websites and in their study the evaluated more than 7000 emails of having phishing and non-phishing email to detect them using machine learning techniques. Basit et al. [9], this paper discussed about how phishing attacks has been increased in pandemic phase and cybercriminals have taken benefit as online access of everything increased, in the paper author compared three ML algorithms ANN, Decision Tree and KNN with RFC for detecting phishing attack websites. Chen et al. [10] used Machine learning highly developed methods in their paper, the authors veteran the URLs and verified whether it was a phishing URL or not, and in this paper, they used a programmed approach for detecting lawful and false URL phishing attacks. Apruzzese et al. [11], in their paper compared Random Forest and Decision Tree to compare the dataset of phishing websites and Random Forest gives better results in detecting. In the paper by Aljabri et al. [12], they compared the machine learning and deep learning techniques for detecting phishing websites on two data sets one based on lexical and other on domain and in result random forest give highest accuracy. Christobel, A., et al. [13] compares different classification methods for datasets. Apruzzese et al. [14], paper discussed about ML and its application for detecting phishing websites and they used their ML-PWD Phishing Website Detection software for evasion attacks. In paper by Bajpai et al. [15] performance of KNN algorithm is measured for the dataset

3. Research Problem Statements

Now days where we all use internet for our day to day activities and do all kind of work online attackers took benefit of it and try to lift our secret information through phishing websites so we really required detecting these websites. Here we are using KNN and LR models to detect whether the website is legitimate or not.

3.1. Research Objectives:

- to study the KNN approach with respect to its identification capability of phishing websites when respective URLs are provided
- to study the LR approach with respect to its identification capability of phishing websites when respective URLs are provided
- compare the KNN and LR approaches with respect to their identification capability of phishing websites with respective website URLs are provided

3.2. Research Methodology:

Experimental research methodology in the Machine Learning domain is applied on the dataset in the Fig. 1, the system flow of our experimental research is been shown. First, we collected the datasets (phishing and non-phishing website URLs) from Kaggle as input after that preprocessing of the dataset took place and selected different features of extraction and processing further. We splitted the data in training and testing datasets. Then we applied KNN and LR models and We observed the accuracy and time taken for giving results and compare our models

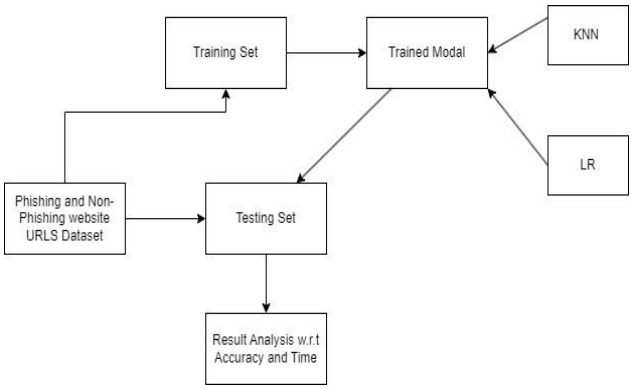

Fig. 1. System flow of Research.

3.3. Design of the Experiment:

- Select a sample of website URLs corresponding to phishing and non-phishing categories.
- Train and test a KNN model with the objective to measure its capability to identify phishing websites.
- Train and test an LR model with the objective to measure its capability to identify phishing websites.
- Compare the efficiency of KNN and LR models with respect to the identification of phishing websites.

3.4. Selection of Samples:

- Sample size (n): 200000 records
- Discussion about the dataset: In this study phishing and non-phishing websites datasets were processed to compare KNN and LR models. The dataset was gathered from Kaggle. The dataset used has more than 200000 records of phishing and non-phishing websites.
- Labels include two categories good and bad. Good - This is not a phishing site or malicious website and Bad- This is a phishing site or malicious website. Among the websites or URLs non phishing (76%) and phishing (24%).
- Sampling Methodology: Secondary Data from Kaggle.

4. Experiment and Result

In the experiment using Machine learning model KNN and LR the phishing website's URLs is being evaluated on the basis of higher prediction accuracy in detecting bad websites or phishing websites and fastest convergence time for classification.

4.1. Experiment using KNN Model

Here in the setup for the analysis and classification of non-phishing websites and phishing websites KNN and LR machine learning classification algorithm is used.

4.1.1. KNN Model:

K-Nearest Neighbor (KNN) classification algorithm is one of the most used techniques in the pattern recognition field it belongs to supervised learning and it is used for both classification and regression models [13]. The K-nearest neighbor's algorithm is being used as it is the mainly used because its robust and make extremely correct predictions on the data set [16].

Fig.2. show a confusion matrix of the prediction and results of an actual good (non-phishing website) / actual bad (phishing website) and predicted good (non-phishing website)/ predicted bad (phishing website) prediction and results of a classification problem using the KNN model

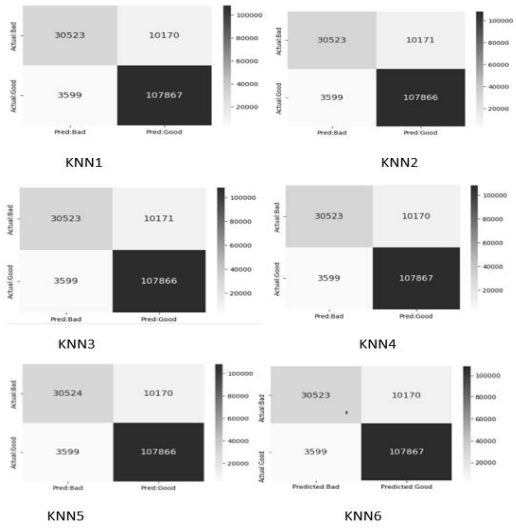

Fig. 2. Confusion matrix results using the KNN algorithm on website URLs dataset

Table 1. Performance Table for KNN Model for The Dataset of Website URLs

Approach	Accuracy	Average	Precision	F-Score	Sensitivity
KNN1	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%
KNN2	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%

KNN3	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%
KNN4	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%
KNN5	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%
KNN6	91%	Macro Average	90%	88%	86%
		Weighted Average	91%	91%	91%

Table 1. shows the measurement of Accuracy, Precision, F-Score, and Sensitivity for K-Nearest Neighbors (KNN) on phishing and non-phishing website datasets. The performance of these methods is measured using a confusion matrix. From this table, we can determine that KNN achieves an accuracy of 91% which is a great accuracy for a diagnosing phishing website. KNN achieve the Macro average precision (90%), sensitivity (86%), and F-score (88%) and Weighted Average precision (90%), sensitivity (86%), and F-score (88%) for the dataset. This performance is measured using cross-validation.

4.1.2. Using LR Model

Logistic regression is used to calculate or predict the probability of a website phishing or not using binary classification [17]. It is a Machine Learning algorithm where Linear Regression model is used for the classification of websites it uses cost function.

Fig.3. show a confusion matrix of the prediction and results of an actual god (non-phishing website) / actual bad (phishing website) and predicted good (non-phishing website)/ predicted bad (phishing website) prediction and results of a classification problem using LR model.

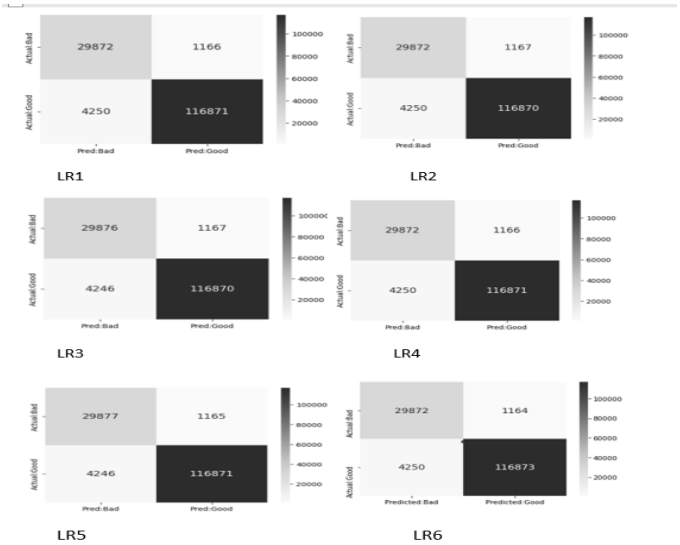

Fig. 3. Confusion matrix results using the LR algorithm on website URLs dataset

Table 2. Performance Table for LR Model for The Dataset of Website URLs

Approach	**Accuracy**	**Average**	**Precision**	**F-Score**	**Sensitivity**
LR1	96%	Macro Average	93%	95%	96%
		Weighted Average	97%	97%	96%
LR2	96%	Macro Average	93%	95%	96%
		Weighted Average	97%	97%	96%
LR3	96%	Macro Average	93%	95%	96%

		Weighted Average	97%	97%	96%
LR4	96%	Macro Average	93%	95%	96%
		Weighted Average	97%	97%	96%
LR5	96%	Macro Average	93%	95%	96%
		Weighted Average	97%	97%	96%
LR6	96%	Macro Average	93%	95%	96%
		Weighted Average	97%	97%	96%

Table 2. shows the measurement of Accuracy, Precision, F-Score, and Sensitivity for Logistic Regression on phishing and non-phishing website datasets. The performance of these methods is measured using a confusion matrix. From this table, we can determine that LR achieves an accuracy of 96% which is a really great accuracy for diagnosis phishing websites ULR achieves the Macro avg precision (93%), sensitivity (96%), and F-score (95%), and Weighted Average precision (97%), sensitivity (96%) and F-score (97%) for the dataset. This performance is measured using cross-validation.

5. Model Performance Comparisons

Table 3. Comparative Analysis of LR and KNN modal Identification of Phishing Websites

Comparative Analysis of LR and KNN Identification of Phishing Websites				
No. of Experiments	**LR (Accuracy of bad Websites detected)**	**Time Taken(min)**	**KNN (%Accuracy of bad Websites detected)**	**Time Taken(min)**
1	96%	04:30	91%	25:00
2	96%	05:00	91%	27:00
3	96%	06:00	91%	32:00
4	96%	04:20	91%	24:00
5	96%	05:30	91%	30:00
6	96%	04:20	91%	27:00
Average	**96%**	**04:56**	**91%**	**27:30**

Table 3. shows the Comparative Analysis of LR and KNN models in the Identification of Phishing Websites. Here results analysis of the LR and KNN models was tested six times and its average is compared according to the accuracy of detecting bad Websites (phishing websites) and time consumption for the execution of different models.

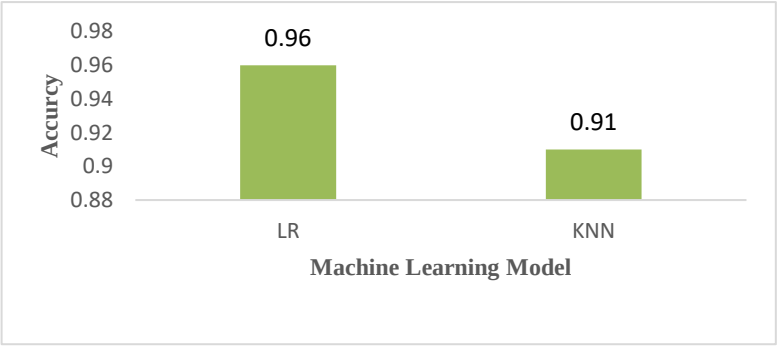

Fig. 4. Accuracy Graph of the dataset of websites for LR and KNN model

In Fig 4. we compare the performance of ML classifiers KNN and LR by accuracy graph and to determine the performance we apply both the model on the datasets six times and the average result from the accuracy graph shows that Logistic Regression has achieved better accuracy for the website URLs dataset

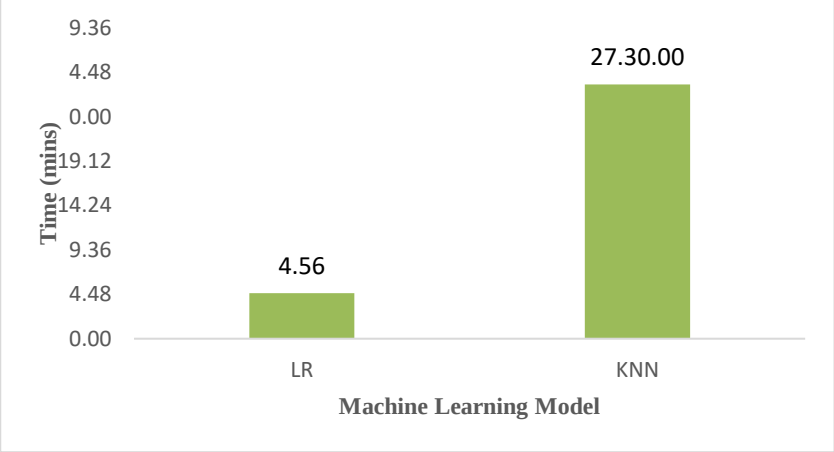

Fig 5. Average Time Graph for LR and KNN model for website URLs dataset

In Fig. 5. we compare the performance of ML classifier KNN and LR according to the time taken for processing the dataset six times the average performance result from the graph shows that Logistic Regression has achieved less time than KNN for detecting bad or phishing website URLs dataset.

6. Conclusions

On phishing and non-phishing websites dataset, we apply two main machine learning classification algorithms six times, which are KNN and LR to classify phishing website URLs give the outputs and different results.

Among LR and to K-Nearest Neighbor model LR model have higher prediction accuracy in detecting bad websites or phishing websites. LR model has very fast convergence time as compared to KNN model in detecting bad websites. Hence, the LR model is the appropriate algorithm to use for the classification of phishing website URLs. Through this paper, we can conclude that the LR model is much better than the k-nearest neighbor KNN model in terms of the classification of phishing and non-phishing website URLs.

References and Footnotes

Author contributions

Dr. Sachin Kadam1: Conceptualization, Methodology, Data study

Mrs. Nidhi2: Writing-Original draft preparation, algorithm application and comparison of model result

Dr. Pratibha Deshmukh3: Grammar & Proofreading, Formatting, Similarity check,

Mrs. Nidhi Khare4: Writing-Reviewing, Editing, References

Mr.Irfan Khatik5: Conclusion, Abstract

Conflicts of interest

The authors declare no conflicts of interest

References

[1] Chaudhry, J. A., Chaudhry, S. A., & Rittenhouse, R. G. (2016). Phishing attacks and defenses. *International journal of security and its applications*, *10*(1), 247-256.

[2] 2. Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*, 139-154.

[3] Odeh, A., Keshta, I., & Abdelfattah, E. (2021, January). Machine learning techniquesfor detection of website phishing: A review for promises and challenges. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0813-0818). IEEE.

[4] Wu, M., Miller, R. C., & Garfinkel, S. L. (2006, April). Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 601-610).

[5] Song, F., Lei, Y., Chen, S., Fan, L., & Liu, Y. (2021). Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers. *International Journal of Intelligent Systems*, *36*(9), 5210-5240.

[6] Athulya, A. A., & Praveen, K. (2020, June). Towards the detection of phishing attacks. In *2020 4th international conference on trends in electronics and informatics (ICOEI)(48184)* (pp. 337-343). IEEE.

[7] Gupta, B. B., Arachchilage, N. A., & Psannis, K. E. (2018). Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems*, *67*, 247-267.

[8] Fette, I., Sadeh, N., & Tomasic, A. (2007, May). Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web* (pp. 649-656).

[9] Basit, A., Zafar, M., Javed, A. R., & Jalil, Z. (2020, November). A novel ensemble machine learning method to detect phishing attack. In *2020 IEEE 23rd International Multitopic Conference (INMIC)* (pp. 1-5). IEEE.

[10] Chen, Y. S., Yu, Y. H., Liu, H. S., & Wang, P. C. (2014, August). Detect phishing by checking content consistency. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)* (pp. 109-119). IEEE.

[11] Apruzzese, G., Conti, M., & Yuan, Y. (2022, December). SpacePhish: The Evasion-space of Adversarial Attacks against Phishing Website Detectors using Machine Learning. In *Proceedings of the 38th Annual Computer Security Applications Conference* (pp. 171-185).

[12] Aljabri, M., & Mirza, S. (2022, March). Phishing attacks detection using machine learning and deep learning models. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (pp. 175-180). IEEE.

[13] Christobel, A., & Sivaprakasam, Y. (2011). An empirical comparison of data mining classification methods. *International Journal of Computer Information Systems*, *3*(2), 24-28.

[14] Apruzzese, G., Conti, M., & Yuan, Y. (2022, December). SpacePhish: The Evasion-space of Adversarial Attacks against Phishing Website Detectors using Machine Learning. In *Proceedings of the 38th Annual Computer Security Applications Conference* (pp. 171-185).

[15] Bajpai, D., & He, L. (2020, September). Evaluating KNN performance on WESAD dataset. In 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 60-62). IEEE. Wu, X., Zhu, F., Zhou, M., Sabri, M. M. S., & Huang, J. (2022). Intelligent Design of Construction Materials: A Comparative Study of AI Approaches for Predicting the Strength of Concrete with Blast Furnace Slag. Materials, 15(13), 4582

[16] Page, A., Turner, J. T., Mohsenin, T., & Oates, T. (2014, May). Comparing raw data and feature extraction for seizure detection with deep learning methods. In The twenty-seventh international flairs conference.

[17] Mahesh, T. R., Vivek, V., Kumar, V. V., Natarajan, R., Sathya, S., & Kanimozhi, S. (2022, January). A comparative performance analysis of machine learning approaches for the early prediction of diabetes disease. In 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE

[18] Ramasamy, J. ., Doshi, R. ., & Hiran, K. K. . (2023). Three Step Authentication of Brain Tumour Segmentation Using Hybrid Active Contour Model and Discrete Wavelet Transform. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 56–64. https://doi.org/10.17762/ijritcc.v11i3s.6155

[19] Waheeb , M. Q. ., SANGEETHA, D., & Raj , R. . (2021). Detection of Various Plant Disease Stages and Its Prevention Method Based on Deep Learning Technique. Research Journal of Computer Systems and Engineering, 2(2), 33:37. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/30