INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 3 - V8I3-1275)
Available online at: https://www.ijariit.com

A comparative analysis of Machine Learning Algorithms on malicious URL prediction

B. Naveen Kumar
basettynaveen18@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

S. Viswa Teja Reddy
syamakuruviswatejareddy@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

A. Yashwanth Sai
yashwanthsai560@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

G. Umesh Kumar
gollaumeshyadav73@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

M. Srikanth
srikanth461335@gmail.com
Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh

Abstract*— Phishing is a type of deception in which a person or entity poses as a legitimate user. Phishing is a technique for fooling consumers that has grown more prevalent in cyberspace. The majority of phishing texts are cryptic. Many strategies plan has now been intended to deal with the phishing problem in the literature. There is currently no solid remedy in place to prevent such assaults. This article proposes a To detect phishing assaults, a human having to learn forecasting engine is used taking this into account. Logical regression beats the other methods in in both of precision and failure rate, according to the experimental investigation. With logistic regression, you can predict URLs with accuracy.

***Keywords*—** *Machine learning, Logistic Regression, Threats, Phishing, Attacks, DecisionTrees*

1. INTRODUCTION

Phishing is really the criminally unethical act of impersonating a trustworthy source in an online transaction to get personal information such as usernames, passwords, and credit card numbers. A phishing portal is a commonly chosen attack that attempts to steal personal credit card details digits, checks account details, Social Security numbers, and passwords in intention of committing fraud. Phishing has a long history. significant detrimental impact on revenue, customer relationships, marketing operations, and a company's entire image. Messages that are commonly used attract unsuspecting members of the public.

Phishing often is usually carried out by e-mail or online chatting, and it usually asks victims to submit confidential info on a bogus site that looks and feels exactly like the real one. Phishing is a sort of social engineering that takes use of the inaccessibility of present online safety technologies to deceive people. Phishing is a type of cybercrime that employs the use of a fake email as a tool. The object is to convince the email reader that the communication is exactly what you're looking for. A coworker's letter or a bank request, for instance, and the need to examine or receive an application. Phishing is distinguished by the way the message is delivered. The offenders pose as a credible law enforcement agency Usually a real or seemingly real person or organization with whom the victim might do business It was one of the older sorts of cyber-attacks, with advanced spoofing techniques and approaches dating back to the 1990s.

It is also one of the most common and deadly diseases. Phishing was responsible for 29percent of all fraud assaults in the first quarter of 2019, with India ranking second behind the United States on the list of countries that host big phishing attacks.

Phishing is indeed the illegal use of an email medium to impersonate a trusting guy in order to gain access of private details such as identities, credentials, and credit card details. According to a survey, phishing and exploit code attacks are the most popular online fraud methods worldwide, with India being one of the top three affected countries. Deception accounts for over 40 percent of all cyber-attacks, according to the RSA Quarterly Fraud Survey. As per the poll, Canada, the U.s., India, and Brazil have the most scam attacks and criminal fraud statistics and analysis.

2. RELATED WORK

Phishing has become a major problem in the internet age. In this age of the internet, the safety of our data on the internet is becoming increasingly crucial. Happy et al. [6] describe phishing as one of the most harmful methods for hackers to obtain users' accounts such as usernames, passwords, and account numbers without their awareness. Users are ignorant of this type of hazard, and they will fall prey to a Phishing scam at some point. This might be due to a lack of a combination of financial assistance and personal experience, a lack of market understanding, or a lack of brand trust. As a result, phishing site detection is required. In this paper, we offer a technique for phishing detection based on URLs.

The internet has become the centre of everyone's attention these days. Everyone used the internet for online shopping as well as other activities like online banking, online booking, online recharge, and more. Phishing is a type of online risk that occurs when a user clicks on a link that leads to a fake website. Login ids, passwords, and credit card numbers are all examples of sensitive information. Junaid et al. [7] proposed a machine-learning-based phishing detection method. Overall, the proposed technique has the maximum efficiency when paired with the Support vector machine classifier, properly segregating 95.66 percent of the time. With only 22 per cent of the breakthrough features, there is a risk of phishing and acceptable websites.

A lack of funding, personal observations, business reputation, or brand confidence might all be factors. As a result, the identification of phishing sites is required. In this work, Mehmet et al. [8] suggested a technique for phishing identification based on URLs. Researchers employed eight different methods to evaluate the URLs of three separate datasets in order to compare the findings. The first strategy investigates various features of the URL; the second investigates the website's authenticity by determining where it is hosted and who manages it; and the third way investigates the website's graphic presence. Machine Learning methods and algorithms are used to analyse these many properties of URLs and webpages.

Phishing has posed a severe danger to information protection because it employs social engineering and other sophisticated techniques to get personal information from consumers. Despite the different defences proposed by academics, phishing perpetrators will ultimately figure out a way to get past them because such precautions require extensive effort. Because manual feature engineering is inefficient at identifying newly developing phishing assaults, a low-cost, reliable phishing detection system is required. In a capsule neural net with several parallel divisions, one convolution operation extracts the shallow characteristics of URL, while other layers of the capsule create proper features of URL and validity of URL [11-15].

In this research, Yong et al. [9] created a unique approach for identifying phishing websites that focuses on detecting a website's Uniform Resource Locator (URL), which has been demonstrated to be an effective and precise way of detection. To offer you a better picture, our new capsule-based neural network is separated into several parallel components. The first eliminates shallow features from URLs, while the other two create correct visual features of URLs and use the shallow features to evaluate URL authenticity. The ultimate output of our system is calculated by adding the outputs of all divisions. Extensive testing on a dataset collected from the Internet indicate that our system can compete with other cutting-edge detection methods while consuming a fair amount of time.

It's critical to recognise and protect against phishing website assaults. The term "neural networks" refers to a type of They are an excellent heuristic machine learning tool for phishing website identification and prevention due to their active learning skills from large-scale datasets. [10] Erzhon et al. However, certain irrelevant variables may lead the machine learning system to overfit during the data training process, resulting in the training model's failure to effectively anticipate and identify phishing websites. This paper demonstrates a neural network-based paradigm for identifying pushing websites (OFS- NN), which is based on an efficient feature selection approach. The initial stage in this approach is to propose a function discovery method that is responsive towards phish URLs' severe housing.

This algorithm picks characteristics for identifying phishing websites depending on the estimated effective value of each function and sets a threshold to exclude any irrelevant features. The neural net subsequently trains an effective classifier for detecting and forecasting the websites that are being pushed using the optimal feature set that was specified. According to the results of the studies, the recommended OFS-NN is a good approach for predicting as well as forecasting. It has a low risk of false positives and a strong potential for generalisation. Moreover, during the sample training phase, the optimum feature selection technique improves the accuracy of machine learning methods.

3. PROPOSED WORK

This section gives a high-level overview of the proposed community detection technique. The proposed model focuses on identifying phishing attacks by allowing users to check phishing websites choices, blacklists, and information. This is in accordance with a few options that will be used to distinguish between authentic and faked websites. There are numerous alternatives available, such as URLs and domain identities. This research is limited to URLs and domain name properties. For locating the most effective outcome, we usually employ two algorithms. As a result, the logistic regression and decision tree classification algorithms are being used.

(a) Data collection

One of the most important jobs in the development of a machine learning model is data collection. It is the collection of task-related data, together with some specified factors, in order to evaluate and provide some useful consequence. However, some of the data is beginning to show signs of wear and tear, such as erroneous, incomplete, or wrong statistics. As a result, processing the data is required before evaluating it and returning to the findings.

(i) Data cleaning: Fill in missing numbers, smooth out creaking data, detect and delete outliers, and repair anomalies to clean up the data.
(ii) Data Transformation: Data modifications that improve the data's norm include flattening, grouping, and generalization.
(iii) Data Selection: It refers to a set of algorithms or routines that enable us to choose the best data for our system.

(b) Data pre-processing

Data pre-processing is a cleaning operation that gets converted unstructured raw information into a neat, well-structured

dataset that may be used for future study. Because data is often received from a number of sources, it must be of adequate quality and in a certain format before the model can learn or be educated with it. This will make it easier to get more and more accurate results with meaningful data. The essential phases in pre-processing include filling in missing values and null values, removing possible outliers, and normalizing data.

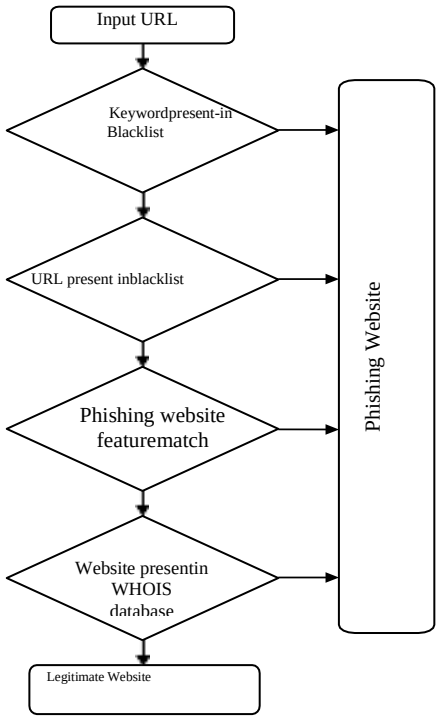

Fig. 1: System workflow

(c) Splitting of Data

In machine learning, each dataset is usually divided into two categories: training data and testing data. The performance variable, as well as other variables, are included in the training range. The model examines the data and looks for patterns. The remaining dataset serves as a validation set to confirm that our model's predictions are accurate. We utilised the scikit library's train test split method to separate our files. The test size parameter determines how much data may be utilised in the test sample. The remainder is kept in the freezer The amount of data that can be used in the test sample is determined by this parameter. The rest is stored in the freezer. Each of them may be specified, as can the training dataset by train size. A function that creates random numbers is known as randomized state. The train and trial sets were split 80:20 in our dataset, with the random state set to 0.

(d) Data Clustering

Cluster analysis, often known as data clustering, is a machine learning approach for grouping unlabeled data into categories. The following is a description of it: "A method for grouping data points into distinct clusters depending on how similar they are. Objects having possible resemblances are grouped together in a category with few or no resemblances to another."

.

It does so by looking for comparable patterns in the unlabeled dataset, such as form, scale, color, and behavior, and categorizing them based on the presence or absence of such patterns. It's an unsupervised learning method, which implies the algorithm isn't supervised and operates on a dataset with no labels. After employing this clustering approach, each cluster or group is given a cluster-ID, which ML systems may utilize to facilitate the study of huge and complicated datasets. Clustering is a prominent mathematical data analysis approach.

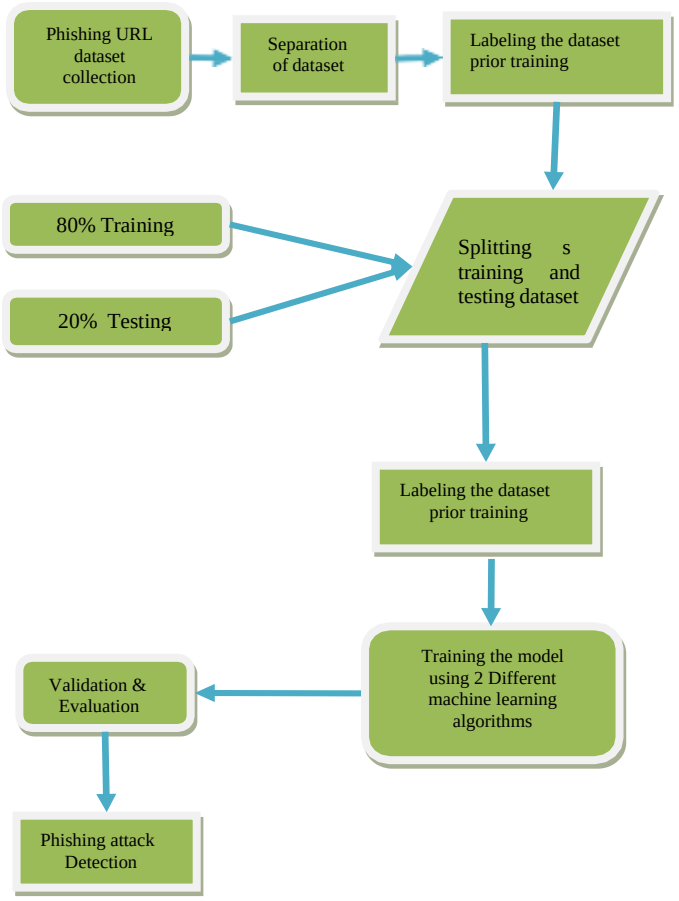

Fig. 2: System Architecture

(e) Analysis

After they've been taught, the approach assesses the efficiency of two interrelated machine learning algorithms, logistic regression and decision tree classifier.

4. SCHEMES USED

(a) Logistic Regression

The odds ratio construct is used in logistic regression to calculate the likelihood. This is the mathematical connection between the chance of an event happening and the probability of it not happening. To estimate the likelihood of a src attribute, the supervised training classification technique regression analysis is utilised. Characteristics of the target or variable quantity is shattered Regression seems to be the simplest method of regression analysis where the vector quantity is separated (binary). Using a big number of notional, arbitrary, interim, or ratio-level freelancing variables, logistic regression may be used to see the relationship between one dependent binary variable and a large number of nominal, ordinal, interval, or ratio-level freelance variables.

It is a statistical model that models contingent probability using logistic logistics. It is used to investigate the relationship between one split variable quantity and one or more independent variables (categorical or continuous). This is in contrast to linear multiple regression, which uses a continuous variable quantity. Regression is easier to use, analyse, and train than other methods.

(b) Decision tree

A Logistic Regression is indeed a supervised classifier learning program that may be used for classification or regression, with classification being the most popular application. In a tree-structured classifier, internal nodes reflect dataset attributes, branches represent selection laws, and leaf nodes provide the

output. The decision tree algorithm and the node are the two nodes that make up a decision tree. Every decision is made up of decision nodes, each of which has several branches, and leaf nodes, which have no extra branches.

The dataset's functioning influences the judgments or tests that are made. The idea behind the decision tree is also easily understood due to the fact that it was created in the shape of a tree. When utilizing Indecision Trees, keep the following in mind: We usually start at the bottom of the tree to forecast the class mark to register. We usually follow the branch that resembles the value and jump to the next logical form on the comparison.

Each continuous load edge from a node correlate to the test case's probable solutions, and each node acts as a test case for a few characteristics. Each subtree rooted at new nodes is changed on a regular basis, and the process is algorithmic.
The number the amount the quantity the quantity of records and characteristics affects the temporal complexity of decision trees among the provided data. Elevated data may be handled by decision forests with reasonable accuracy.

5. PERFOMANCE ANALYSIS

We employed precision, remember, F1 score, and F1 measure i n logistic regression and decision tree classifier to evaluate the accuracy and detect spam websites.As a consequence, the decis ion tree classifier was 85 percent accurate, and logistic regressi on was 97 percent accurate.The projected model's output show s if the URL is dangerous or not.

Using data collection, we analyse the data to panda bears and NumPy libraries, which is then preprocessed into a proper data set, which we then implement inside the SKlearn (Scikit) which consists of data, from which the data is classified into th e training set and testing set, and both sets are transmitted to th e model. In this case, the model goes over both the positive and negative aspects of the situation.

The logistic regression and decision tree classification algorithms are used. The effectiveness is then assessed and quantified using various measures.

This page displays the results of all of the classifiers listed on the phishing dataset. We tested these algorithms on over 35000 samples and a variety of performance indicators, as well as the results are presented in the form of graphs below.

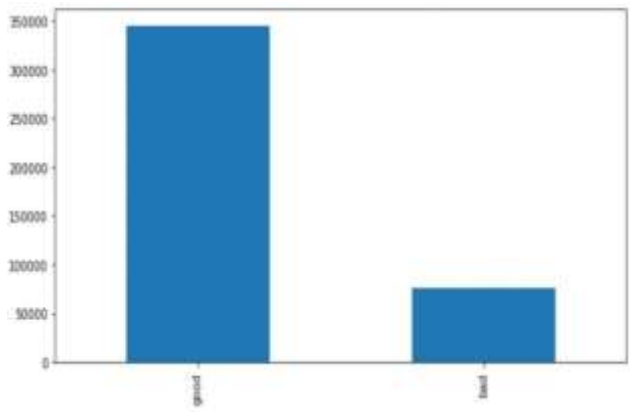

Fig. 3: Phishing Vs Non-phishing

The performance analysis of the dataset is shown in Figure 3. To detect phishing websites, several machine learning experts have created algorithms and methodologies. Following a thorough examination, we have determined that the great bulk of the research is carried out using well-known machine learning techniques.

A variety of approaches were tried in terms of accuracy, precision, memory, and other factors. phishing website URLs are growing thanks to trial-and-error detection tactics. End-users may easily prepare our phishing detection model with the help of the plugin. We propose to construct the phishing detection system as a scalable home internet that will contain on-line training so that new phishing assault patterns may be learnt quickly and our models' accuracy can be improved with greater extracting features in the future.

We estimate the Tree based classifier's efficiency to be 85 cent using several metrics such as precision, recall, and F Ranking.

Table 1: Decision Tree –accuracy

	Precision	**Recall**	**F1 score**	**Support**
Bad	0.82	0.26	0.40	14964
Good	0.86	0.99	0.92	69129
Accuracy			0.86	84093
Macro avg	0.84	0.62	0.66	84093
WeightedAvg	0.85	0.86	0.83	84093

Various factors such as specificity, recall, and F Score are used to calculate the accuracy of the Logistic regression.

Table 2: Logistic regression-accuracy

	Precision	**Recall**	**F1 score**	**Support**
Bad	0.98	0.88	0.93	14964
Good	0.97	1.00	0.98	69129
Accuracy			0.97	84093
Macro avg	0.98	0.94	0.96	84093
Weightedavg	0.97	0.97	0.97	84093

Each classifier's sensitivity is computed using test samples. False positives and false negatives, on the other hand, are taken into account when calculating a classifier's overall performance in detecting attacks. We don't want visitors to be able to browse phishing URLs, thus false positives are a critical consideration when choosing the best classifier.

6. CONCLUSION

A comparison of machine learning algorithms for URL prediction is offered in this research. The major goal is to ensure security and prevent the user from gaining access to their sensitive data. It is possible to determine if a website is authentic or not using machine learning algorithms. The accuracy level of logistics regression and classifier was found to be approximately 97 percent using performance metrics and our literature review, thus these modules were picked for classification. As the number of phishing websites grows every day, this is something that everyone should be aware of. Other parameters may be included for examination in the future, allowing the accuracy to be enhanced even further.

REFERENCES

[1] Li, Q., Cheng, M., Wang, J., & Sun, B. (2020). LSTM based Phishing Detection for Big Email Data. *IEEE Transactions onBig Data*.

[2] Mao, J., Tian, W., Li, P., Wei, T., & Liang, Z. (2017). Phishing-alarm: robust and efficient phishing detection via page component similarity. *IEEE Access*, *5*, 17020-17030.

[3] Raza, M. Q., Mithulananthan, N., Li, J., & Lee, K. Y. (2018). Multivariate ensemble forecast framework for

demand prediction of anomalous days. IEEE Transactions onSustainable Energy, 11(1), 27-36.

[4] Shyni, C. E., Sundar, A. D., & Ebby, G. E. (2018, February). Phishing Detection in Websites using Parse Tree Validation.
In *2018 Recent Advances on Engineering, Technology and Computational Sciences (RAETCS)* (pp. 1-4). IEEE.

[5] Tan, M., Yuan, S., Li, S., Su, Y., Li, H., & He, F. (2019). Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning. *IEEE transactions onpower systems*, *35*(4), 2937-2948.

[6] Chapla, H., Kotak, R., & Joiser, M. (2019, July). A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier. In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 383-388). IEEE.

[7] Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing Detection Using Machine Learning Techniques. *arXiv preprint arXiv:2009.11116*.

[8] Korkmaz, M., Sahingoz, O. K., & Diri, B. (2020, July). Detection of Phishing Websites by Using Machine Learning- Based URL Analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies(ICCCNT)* (pp. 1-7). IEEE.

[9] Huang, Y., Qin, J., & Wen, W. (2019, October). Phishing URL Detection Via Capsule-Based Neural Network. In 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID) (pp. 22-26). IEEE.

[10] Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: an effective phishing websites detection model based on optimal feature selection and neural network. *IEEE Access*, *7*,73271-73284.

[11] Neha Roy, Y. Bevish Jinila (2015), "A survey on security challenges and malicious vehicle detection in Vehicular ad hoc networks", Contemporary Engineering Sciences, Vol.8, No. 5 – 8, pp. 235-240.

[12] Y. Bevish Jinila (2015), "Anonymization based location privacy preservation in Vehicular ad hoc networks", Vol.8, No.1-4, pp.109-114

[13] Shyry, S.P, "Efficient identification of bots by K-means clustering", Advances in Intelligent Systems and Computing, 2016, 398, pp. 307–318,2016.

[14] Prayla Shyry, S,"Hybrid Trio detection Approach: A Framework for Intrusion Detection", Lecture Notes on Data Engineering and Communications Technologies, 2019, 26, pp. 1032–1038.

[15] Karunakaran, P. "Deep Learning Approach to DGA Classification for Effective Cyber Security. "Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 04 (2020): 203-213.

