International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A HYBRID MACHINE LEARNING BASED APPROACH OF URL FRAUD DETECTION

Rampam Pavan Kalyan[1], Maddirala Mukesh[2], Chagalamarri Hemamber[3] ,Dinesh Thunga[4]

[123]Department of Computer Science and Engineering & CHRIST (DEEMED TO BE UNIVERSITY)
[4]Department of Mechanical Engineering & CHRIST (DEEMED TO BE UNIVERSITY)

Abstract - Phishing is widely used to scam people aroundthe world. This attack is mainly used to steal the user's information such as Debit/Credit card, passwords and other sensitive data. This attack can be enforced on large mob at a time so, anyone can be target. This attack is initiated by sending a fake URLs through email, social media but link appears to be so genuine sent right from original organization Thus, user tempts to enter the information. In this article our team is going to handle this kind of attacks. Using Machine learning techniques, our web app is to going to detect whether it is a phishing link or not

Key Words: Phishing, Support Vector Machine, Decision Tree classifier, Machine Learning, CyberSecurity**.**

1.INTRODUCTION

Internet-connected devices and their applications are becoming increasingly widespread all over the world as a result of technical advancement. This however garner more attention for other internet-connected computer security challenges as well. Several efforts have been made to address these difficulties and machine learning methods are frequently used in their implementation [3-5]. A cyberattack known as phishing tricks victims into accessing malicious content and disclosing personal information. The majority of phishing use the same domain and website interface as trustworthy websites. There is a great need for an intelligent plan to protect consumers from cyberattacks. In this study, we propose a method for detecting URLs that employs machine learning techniques. Our work aims to increase cyber-attack detection rates by offering high detection accuracy and low false-negative and false-positive rates as phishing attempts grow more prevalent. False-negative sites are those that are misidentified as authentic websites, whereas false-positive sites are those that are misinterpreted as legitimate websites.

The start of the attack was specifically intended to target few people in particular area but, now general public are also victims. In Starting days, user needs to manually enter the information. But one clicks on to the link and providing some access levels to attacker is enough to steal information.The fig .1 depicts as simple explanation of phishing. When a victim visits a website and clicks on an external link, such as an advertisement or another pop-up link, phishing has already started.

The user gets directed to that website when they click on a phishing link. The attacker utilizes the victim's information to get access to other reliable websites after acquiring it. Numerous alternative detection procedures are developed and used in the literature to identify this kind of phishing attempt.

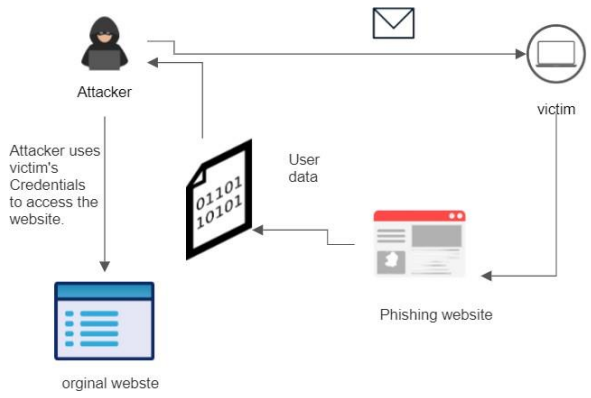

Fig. 1 General Process of phishing link

A. Phishing Process

When Attacker sends a Phishing link to victim via email, social media or through any other platform. Using link masking attacker can change the link into original domain. By seeing that domain user believes link is from legitimate organization. The user tempts click on the link. The Phishing site seems to be look like an original website. In most of the cases only front end of the website is prepared and backend won't be done. The user does not recognize that part. The user feels safe in the website by looking that front end of the website and tries to fill up credentials required. This could be user's username and password or credit/debit card's details with OTP. After gathering the information from the victim. The attacker tries to get on to the original website. Enters the gathered information from victim. This us how Phishing happens in today's world. Sometimes knowing about the loss, the user can't even map the entry point of the fraud. Because he/she doesn't even know that is a trap.

B. Statistics

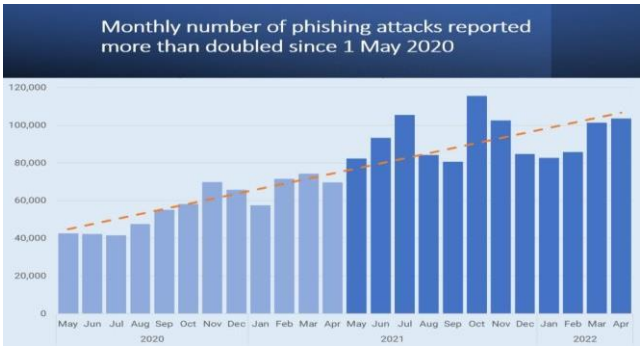

Fig. 2 Statistics of Phishing Attacks

The Fig. 2 clearing states that Phishing attacks are prone to cyber space. Risk of getting exploited increasing day by day. These are very few numbers in which there is a different situation in ground level. These numbers show only those who register case and in efficient investigation they found phishing is the cause of attack

2. Related work

The research on this field is to carried for long time since attackers change their methods accordingly new techniques need to be developed to save public from these attacks. A study done by Mr. Anklet Kote and colleagues titled as "Detection of Phishing Websites Using Data Mining" published in 2020. The model extracts the features of the website via URL when user use it. The features will be running against test data. The main objective is to detect phishing links. It uses Random Forest Algorithm be used to attain the desired goal. The study by Amani Alswailem and colleagues titled as "Detection of Phishing Website Using Data Mining" Published in 2019. In which global set offeatures are extracted including hard-to-forge features from

URLs. It uses Phish Tank as a repository for finding blacklisted links and for legitimate links it uses top one million links. And another by Manish Jain and colleagues titled as "Phishing Website Detection System using Machine Learning "published in 2020. Machine learning techniques such as Naïve Bayes classifier, support vector machine and random forest. This study by Hemali Sampat and colleagues titled as "Detection of Phishing Website using Machine Learning" published in 2018. In this paper, System uses different methods such as extracting URL feature, Black List, and WHOIS database is implemented. The education awareness is also provided from this article. In order to identify phishing websites in the year 2020, Alsariera,Adeyemo, Balogun, and Alazzawi used a total of fourdifferent meta-learner models. These models were referred to individually as AdaBoostExtra Tree (ABET), Bagging-Extra Tree (BET), Rotation Forest - Extra Tree (RoFBET), andLogitBoost-Extra Tree (LBET) [10]. Phishing Detection Using Multidimensional Characteristics was the title of the paper. This technique was inspired by the numerous guisesin which phishing assaults can take place [11].

The study by Nandini Patil and colleagues named known as "Phishing Website Detection based on Machine Learning" published in 2019.The created neuro-fuzzy framework,known as Fi-NFN, was conceptualized by Pham, Nguyen,Tran, Huh, and Hong in 2018. This architecture includes ananti-phishing model that guards users of fog from becomingthe intended victims of phishing attacks [12].

In this article they used data mining algorithms rather than traditional algorithms which proved to be very efficient. The model scans the phishing link and by unsupervised learning techniques it adds new suspicious keywords into database. The secondary model detects the phishing links based on checking links in Black List, WHOIS database. Another study by Dhananjay Merat and colleagues titled as "A Machine Learning Approach for Phishing and Its Detection techniques "published in 2019. Dynamic machine methods are used such as single layer artificial intelligence. It has fouroutcomes TP (True Positive, FN (False Negative), TN (True Negative) and FP (False Positive). The Study by Sachin Nandrajog and colleagues titled as "Phishing website Detection:" came out in 2021. It used hybrid approach which increased accuracy. 70.18% of accuracy for naïve bayes, 89.44% accuracy for Random Forest and at last 91.43% of accuracy when using hybrid algorithm. A Study by Ammar Odch and colleagues gave title as "Machine Learning Techniques for Detection of Website Phishing: A Review forPromises and Challenges" in 2021. The framework Tang andMahmoud introduced in 2021 for recognizing phishing websites was built on deep learning. They put their solution into practice by creating a browser extension for Google Chrome. A warning is displayed whenever a phishing website is detected, and the user receives real-time results about the risk of phishing websites through this plugin [13].

3.PROPOSED WORK

This study proposed a method to predict phishing links using machine learning. When compared to other systems we are not using any dataset because at long run it might damage the efficiency. We are fully focused on benchmarking and are sourced from third-party services like domain registers, search engines, WHOIS records. This record is used for extracting feature vectors from the given link listed vocabulary, host and word. The vocabulary feature is for textURLs that includes: host-name length, URL length, URL tokens and so on. The host-name length, URL length, the frequency of dots in the URL. A weka function called "String to Word Vector" is employed to change the URL into carrier specific words such as Index, Using IP, Long URL, Short URL, Symbol @, Redirecting/prefix and prefix, Sub Domains, HTTPS, Domain Reg length, Favicon, Non- Standard port and HTTPS Domain Port, Req URL, Anchor URL, link in script tags, server form handler, website forwarding, website traffic, Page Rank, Google Index, Iframe Redirection, Age of domain, Links pointing to page, Stats Report. In this we use a dataset to test the algorithms for accuracy and time efficiency. Also, this dataset is collected from the Kaggle which is a data company that hosts open-source datasets for various research purposes.

A. Implementation Strategy

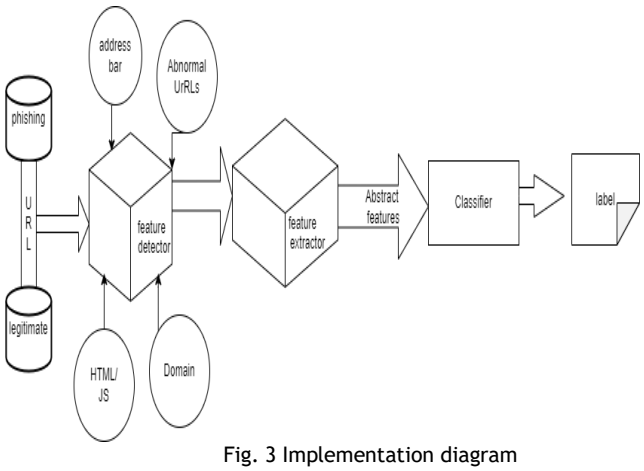

Fig. 3 Implementation diagram

In Fig 3. We can see the block diagram for the implementation proposed model. In this the process first stepis started at the URLs phase, then it is passed to the module to start extracting information from the URLs to check the attributes of the link and the sent further process and at this step the extracted attributes is made into an abstract feature packet and the sent for classification in which the algorithms do their work and classify the features into their perspective divisions, which is then graded as per the algorithm to procced to the final step i.e. is to label the link that has been processed by the algorithms and then label is given to the final value to URL as phishing link or a legitimate link. In this stage we evaluate the models by

entering the link with predictor variables to each model, then the models will predict the targeting variable according to the prediction results and we will compare it with real values to get results.

B. *Algorithm*

In recent research, the algorithm used for detection has many limitations. Support Vector Machine: It is a supervised machine learning algorithm. It is used for classifying involving two classes. Usage of SVM gives improved outcome. Decision Tree: It is another supervised machine learning algorithm where they are pre- defined outcomes. Decision tree is frequently in classification adversities. Outcome will be out in binary format. The rules decision tree has can easily determine the target variable where large-scale database has lot of similarities. Decision tree can analyze trained data and classify missing data.

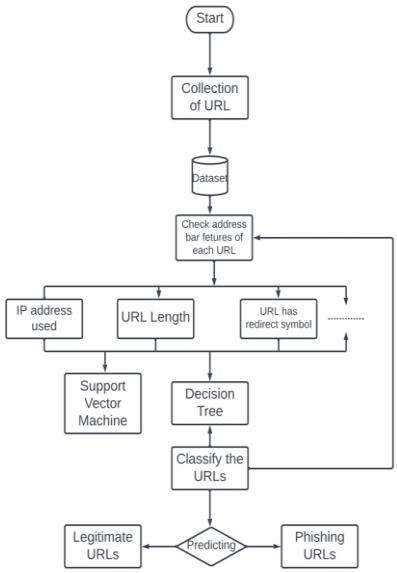

Fig 4. System architecture

C. Algorithm basic steps

If in each and every case belonging to similar feature class. The tree generates leaf node and this leaf node returned with marking of same class. Calculate the potential data which is done for every given attribute for test mode. After this, generated information is a result that is require.

D. Methodology

Data Collection: Collect a large dataset of URLs from various sources including legitimate and fraudulent URLs. Data Preprocessing: Preprocess the data by removing any duplicates, invalid URLs, and URLs with missing features. Feature extraction: The next step is to extract features from the URLs that can be used as inputs to the machine learning algorithms. This can include features such as the length of the URL, the presence of certain keywords or characters, the number of slashes, and other relevant information.

Implement and train the hybrid model: The next step is to implement the hybrid model of SVM and DT classifiers. The dataset should be split into a training set and a testing set. The training set should be used to train the SVM and DT classifiers, and the testing set should be used to evaluate the performance of the hybrid model. The hyperparameters of the SVM and DT classifiers should be tuned to optimize the performance of the hybrid model.

Evaluate the performance: The final step is to evaluate the performance of the hybrid model. This can be done by calculating metrics such as accuracy, precision, recall, and F1 score. The results should be compared to the performance of other state-of-the-art URL fraud detection methods to determine the effectiveness of the hybrid model.

The Support Vector Machine (SVM) algorithm is used to find a hyperplane in an N-dimensional space. The entire number of features is referred to as N in this situation. The identified hyperplanes serve as representations of the decision points that categories the data points. The size of the hyperplane is defined by the number of features, which in turnis determined by the overall number of features. The scenario is pretty difficult because there are 87 different aspects. Support vectors are the data points that are situated closer to the hyperplane. These are very important in the creation of the SVM since they help determine where the hyperplane will be. The objective of the SVM is to achieve a margin that is as large as possible between the data points and the hyperplane [6].

Decision tree algorithm can be used for URL fraud detection by building a decision tree model that classifies URLs as either fraudulent or legitimate based on their features. The decision tree model is trained on a dataset of labeled URLs, where each URL is represented by a set of features such as the length of the URL, the number of subdomains, presence of special characters, etc.

During training, the decision tree algorithm recursively partitions the dataset based on the values of the features to create a tree-like structure of decision rules. Each internal node of the tree represents a decision rule based on a feature value, and each leaf node represents a class label (fraudulent or legitimate). To classify a new URL, the decision tree model traverses the tree from the root node to a leaf node, following the decision rules based on the feature values of the URL. The class label assigned to the URL is the class label of the leaf node it reaches.

The performance of the decision tree model can be evaluated using metrics such as accuracy, precision, recall, and F1-score. The model can also be further optimized by pruning the tree to prevent overfitting or by using an ensemble of decision trees such as a random forest. Overall, the decision tree algorithm is a powerful and interpretable method for URL fraud detection that can capture complex interactions between URL features and accurately classify URLs as fraudulent or legitimate.

Hybrid models are used in URL fraud detection to improve the accuracy of the classification process. A hybrid model is a combination of multiple machine learning algorithms that work together to provide a more accurate prediction than any single algorithm could. One common approach to building a hybrid model for URL fraud detection is to combine the strengths of decision trees and support vector machines (SVMs). Decision trees are good at handling categorical data and can capture complex relationships between the input features, while SVMs are good at handling continuous data and can separate data into distinct categories using a hyperplane.

The hybrid model works by first using the decision tree algorithm to generate a set of rules that can be used to classify URLs as either legitimate or fraudulent. These rules are then used to filter out the most obvious cases of fraud. The remaining

URLs are then passed to the SVM algorithm, which uses a more sophisticated technique to classify them as either legitimate or fraudulent. The output of the SVM algorithm is then combined with the output of the decision tree algorithm to produce a final prediction. This approach can result in a more accurate prediction than either algorithm used on its own. Additionally, hybrid models can be tuned to optimize their performance, making them a powerful tool for detecting URL fraud.

4. IMPLEMENTATION SPECIFIES AND RESULTS

Accuracy is probably the most well-known method for validating a machine learning model so that it may be applied for classification tasks. Its extensive use can be partially attributable to the fact that it can be executed with a respectable level of simplicity. It is simple to comprehend and simple to put into practice. For assessing a model's performance for straightforward scenarios, accuracy is a useful criterion to use, and it can be calculated using Equations 1 and 2 below. The proportion of accurate predictions that can be made is determined by a metric called accuracy, which is used when dealing with classification challenges. It needs to be calculated we must first take the total number of forecasts and divide that number by the number of predictions that turned out to be accurate; the accuracy of the suggested model can be determined by following the steps in Table I.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Several points of view that are connected to the classificationtechnique have been used in this research to apply the prediction model of the proposed phishing website.
The objective experimental evaluation is the purpose of this application. This study tested various algorithms to determine how accurate they were at identifying phishing websites by their URL. The highest rate of testing accuracy was provided by the Gradient boosted classifier, which was 97.4 percent. There is a possibility of four outcomes:

- TP (True Positive)
- TN (True Negative)
- FP (False Positive)
- FN (False Negative)

Depending upon these four parameters outcome is decided. In simple words, the above outcome decides whether the linkis legitimate or not.

TABLE I

ML Model	Accuracy	F1-score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
XGBoost Classifier	0.969	0.973	0.993	0.984
Random Forest	0.967	0.971	0.993	0.990
Svm	0..964	0.968	0.980	0.965
Decision tree	0.960	0.964	0.991	0.993
K-Nearest Neighbours	0.956	0.961	0.991	0.993
Logistic Regression	0.934	0.941	0.943	0.927
Naïve Bayes Classifier	0.605	0.454	0.293	0.997
Proposed model	0.947	0.950	0.965	0.934

5. COMPARISON OF METHODS

The Support Vector Machine (SVM) algorithm is used to find a hyperplane in an N-dimensional space. The entire number of features is referred to as N in this situation. Consequently, in this project, the study employed machine learning techniques to identify phishing websites and reduce their success rates. With 11,053 URLs, 32 features, and a 50:50 split between training and testing, and chosen Kaggle dataset. The huge number of features that distinguish the dataset apart from other datasets helped boost accuracy. This study trained our dataset using 9 machine learning techniques. Gradient Boosting Classifier, XGBoost Classifier, Random Forest, Svm, Decision tree, K-Nearest Neighbours, Logistic Regression, Naïve Bayes Classifier, Proposed model were among the techniques utilized. In terms of test accuracy, recall, and precision among the employed algorithms, proposed model given 94.7, 95.0, 96.5, 93.4.

The study indicate that the best algorithm for identifying phishing websites using hybrid model is decision tree and support vector machine. This methods for identifying phishing websites have such high accuracy, making our project superior to earlier ones. In future study, this is to improve accuracy by enlarging the dataset. Compare to other algorithms svm and decision tree support some characterics that can help for detection of url.

Thus, the implementation may make use of certain big data. However, using Big Data necessitates the application of deep

learning techniques, as described in [8], and parallel code execution is required to reduce training time, particularly when using GPU technologies [7].

6. CONCLUSION AND FUTURE WORK

It is crucial to detect the Phishing links because at later it can make you pay a lot. It reduces the risk of getting exploited.

However, there is very little awareness in the society about this issue and technical persons too. It is so critical to protect our privacy and security. In this paper we presented an approach for Phishing link detection. Analysis based on URL enhances speed of detection. Based on results we can conclude that inclusion of decision tree gave an advantage to classify missing data. Our web app can be installed in any device it gives us added advantage. In upcoming years, by applying feature selection algorithm and dimensionality reduction methods can reduce the number of features and omitting unwanted stuff. There are many technologies like hybrid methods and dynamic methods can be used in future projects.

7. REFERENCES

[1] Huang. H, Zhong. S, and Tan. J (2009, August). "Browser-side countermeasures for deceptive phishing attack." In 2009 Fifth international conference on Information assurance and Security (pp. 352-355).\

[2] U. Ozker and O. K. Sahingoz, "Content based Phishing Detection with Machine Learning." 2020 i=International Conference on Electrical Engineering (ICEE), 2020, pp1-6 doi: 10.1109/ICEE49691.2020.9.\

[3] Huang. H, Zhong. S, and Tan. J (2009, August). "Browser-side countermeasures for deceptive phishing attack." In 2009 Fifth international conference on Information assurance and Security (pp. 352-355).\

[4] O. K. Sahingoz, U. Cekmez and A. Buldu," Internet of Things (IoTs) Security: Intrusion Detection using Deep Learning" 2021, Jour□nal of Web Engineering, 2021, pp. 1721–1760, vol. 20, iss. 6, doi: 10.13052/jwe1540-9589.2062.

[5] R. Yetis and O. K. Sahingoz," Blockchain Based Secure Communication for IoT Devices in Smart Cities," 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 2019, pp. 134- 138, doi: 10.1109/SGCF.2019.8782285.

[6] Gandhi, R. (2018, July 5). Support Vector Machine -intro duction to machine learning algorithms. Medium. Retrieved May 18, 2022, from https://towardsdatascience.com/support-vector-machine introduction to- machine-learning-algorithms-934a444f.

[7] S. I. Baykal, D. Bulut and O. K. Sahingoz," Comparing deep learning performance on Bigdata by using CPUs and GPUs," 2018 Electric Electronics,Computer Science, Biomedical Engineering' Meeting (EBBT), 2018,pp.1-6,doi:10.1109/EBBT.2018.8391429

[8] S. C. Kalkan and O. K. Sahingoz," Deep Learning Based Classification of Malaria from Slide Images," 2019 Scientific Meeting on Electrical Electronics Biomedical Engineering and Computer Science (EBBT),2019,pp. 1-4, doi:10.1109/EBBT.2019.8741702

[9] Multilayer Perceptron. Multilayer Perceptron - an overview — ScienceDirect Topics. (n.d.). Retrieved May18,2022, fromhttps://www.sciencedirect.com/topics/computer scien ce/multilayer perceptron.

[10] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun and A. K. Alazzawi," AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," in IEEE Access, vol. 8, pp. 142532-142542, 2020, doi: 10.1109/ACCESS.2020.3013699.

[11] P. Yang, G. Zhao and P. Zeng," Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019,doi:10.1109/ACCESS.2019.2892066.

[12] C. Pham, L. A. T. Nguyen, N. H. Tran, E. -N. Huh and C. S. Hong," Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks," in IEEE Transactions on Network and Ser vice Management, vol. 15, no. 3, pp. 1076-1089, Sept. 2018, doi: 10.1109/TNSM.2018.2831197.

[13] L. Tang and Q. H. Mahmoud," A Deep Learning-Based Framework for Phishing Website Detection," inIEEE Access, vol. 10, pp. 1509-1521, 2022, doi: 10.1109/ACCESS.2021.3137636.

[14] Machine learning random forest algorithm - javatpoint. www.javatpoint.com. (n.d.). Retrieved May 18, 2022, from https://www.javatpoint.com/machine-learning-random-forestalgorithm.\

[15] A.V. Ramana, Rao, K.L. Rao, R.S. Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Soc. Netw. Anal. Min. 11, 110 (2021). https://ezproxy.iku.edu.tr:2444/10.1007/s13278-021-00829-w.

[16] Abdullateef O. et al., "Improving the phishing website detection using empirical analysis of Function Tree and its variants", Heliyon, vol 7, Issue 7, 2021, e07437, https://doi.org/10.1016/j.heliyon.2021.e07437.

E. Kocyigit, M. Korkmaz, O.K. Sahingoz, B. Diri,"Real-Time Content Based Cyber Threat Detection with Machine Learning". In: Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., Madureira, A. (eds) Intelligent Systems Design and Application, 2021, ISDA 2020. Advances in Intelligent Systems