International Journal of Scientific Research and Engineering Development-– Volume 6 Issue 3, May-June 2023

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Classification Based Web Service Selection Apporach

[1]Jeya Kumar, [2]S. J. Vivekanandan,[3]Balu Yogeshwar, [4]Pathiputturu Dinesh.

[1,2]Assistant Professor, Dhanalakshmi College of Engineering, Chennai

[3,4] Students, Dhanalakshmi College of Engineering

Abstract **–**Phishing is a type of social engineering attack that is frequently used to steal customer data, including login credentials and credit card details. Due to the improvements Websites are the main source of cyberattacks in internet technology. Even though there are a number of defences against phishing attempts, the attackers frequently alter their strategies. Machine learning is one of the methods that is most frequently utilised to resolve cybersecurity challenges. Machine learning and deep learning techniques have been useful for solving security-related problems in recent years.Because most phishing assaults share a few basic traits, machine learning is best suited for detecting them. Many machine learning approaches have been used in this study to identify phishing assaults. Two priority-based algorithms are suggested here.The ultimate fusion classifier is chosen based on the output of these algorithms. We applied a novel fusion classifier to a dataset from UCI and attained a 97% accuracy. Python was used to carry out our experiments.

Keywords:Phishing, Cyber Security, Machine learning, Priority based algorithms, Fusion, UCI, Python.

1.INTRODUCTION

The phrase used most often today is "social engineering." Threats from the internet are causing many issues for everyone. Phishing is one of the social engineering techniques that is most frequently utilised. When an attacker poses as a reliable source and dupes a victim into opening an email, SMS, or instant message, it takes place. There are numerous techniques to conduct phishing. As an illustration, many faculty members receive a spam email from a certain university. The user may be prompted by the email to click the link. The link opens a copy of the internet page upon clicking it. The hacker keeps track of and uses the new password.In a phishing attack, individuals are coerced into visiting unlawful websites and disclosing sensitive information such as passwords, credit card numbers, and bank account information. Using an antivirus or firewall is one of the most popular defences against cyberattacks. However, antivirus protection can't completely shield users from phishing attempts.The users are being directed by the phishers to a fake/dummy webserver, which is the cause of this. Secure browser connections are also used by attackers to carry out their illicit activities. Because businesses are unable to train their personnel in this area, phishing attempts are on the rise due to a lack of methods for preventing them. The typical defences employed by the businesses include phishing attack simulation training, modernising all of their systems with the most recent security procedures, or encrypting sensitive data. Careless surfing is one of the most common causes of falling victim to this phishing assault Legitimate websites and phishing websites share similarities. The fake website shares the same aesthetic as the real one.For instance, a user might get an email from PayPal (even though it's not truly from PayPal) informing them that their account has been restricted. In picture 1, a sample phishing email is displayed. User credentials are stolen if the user responds and takes action. Arthur Samuel first used the phrase "machine learning" in 1959, and it now permeates every industry. There are many methods used in machine learning, including supervised learning model, unsupervised learning models, and reinforcement learning models.There are two samples of data used in supervised learning: train data and test data. The learning step, which follows training, involves building a model using the train-data.During the testing step, the built model is utilised to assess the model's performance using test data. In unsupervised learning, the data does not have any labels.An interactive learning model is reinforcement learning.Regression or classification are two types of supervised learning. Regression results are anticipated as numbers, whereas classification results are predicted as labels. Most difficulties in real life can be resolved through supervised learning. For phishing detection, supervised learning algorithms perform well because the classification of suspect URLs in phishing attempts may be viewed as a classification problem. Many classifications exist.There are many algorithms accessible, but selecting the appropriate one is essential to addressing the given problem. The remaining parts of the essay are arranged as follows. Literature review is presented in Part 2, research technique is described in Section 3, experiment findings are presented in Section 4, and the conclusion is presented in Section 5.

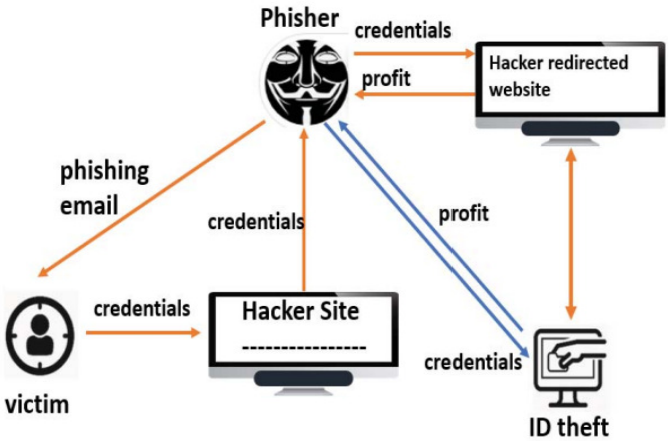

Fig. 1. Phishing attack diagram

2. LITERATURE SURVEY

It's not a novel idea to use machine learning to stop phishing assaults. Some researchers used phishing assaults to test machine learning algorithms. et al., Vahid Shahrivari

For the purpose of phishing detection, [1] used machine learning techniques.

They used the random forest algorithm, together with the logistic regression classification algorithm, SVM, Adaboost algorithm, KNN, and ANN, to obtain high accuracy. To detect phishing assaults, Dr.G. Ravi Kumar [2] et al. used a variety of machine learning techniques. For improved performance, they used Natural Language Processing algorithms. They used a Support Vector Machine to generate excellent accuracy from data that had already undergone NLP processing. With their model, Venkateshwara Rao [3] et. al. successfully detected phishing assaults using decision trees, support vector classifiers, and random forest models.Amani Alswailem[4] et. al used a random forest approach to successfully identify phishing assaults using a variety of machine learning models. Meenu[5] et. al used Artificial Neural Networks, Support Vector Classifiers, Decision Tree Classifiers, and Logistic Regression to predict phishing emails and found that the logistic regression classifier provided the best results. For the purpose of phishing attack identification, Abdul Basit[6] et al. analysed a number of strategies, trends, opportunities, and difficulties.

A deep learning model called an artificial neural network can address classification and regression issues. There is no requirement to use a feature selection strategy in ANN. Neural networks are capable of automatically extracting features.For the purpose of detecting phishing websites, some researchers used ANN. With the use of naive Bayes classifier and ANN, Sandeep Kumar [7] et. al., who used machine learning techniques, were able to detect phishing websites with an accuracy of 89.3%.

Manish Jain [8] et. al use machine learning approaches to detect phishing. Support vector machines and random forest classifiers were used by Jagadeesan[9] et al. to detect phishing.

Two separate datasets from the UCI repository were used. Using a real-world dataset of 1,353 URLs, Arun Kulkarni[10] et. al. developed four categorization techniques and attained an accuracy of more than 90%. Machine learning is an effective method for managing phishing website identification, as demonstrated by R. Kiruthiga[11] et. al's comparison of various machine learning techniques.Preeti[12] et. al used a variety of classification algorithms to identify phishing websites. They used Logistic Regression, Random Forest, Decision Tree, and SVM. They used Logistic Regression and had high accuracy with 1200 URLs. Nevertheless, when they used Logistic Regression on 12,000 sites, they got less accurate results. They deduced from the experiments that Decision Trees function effectively regardless of the size of the dataset. By utilising various phases, such as the convolutional layer, activation function, pooling step, and flattening, convolutional neural networks are able to handle the classification problem. It resembles ANN after flattening. Deep learning models were proposed by Ali Aljofey[13] et al. for the detection of phishing websites. They used a convolutional neural network model at the character level and got good results. Convolutional neural networks and attention-based hierarchical recurrent neural networks were used by Y. Huang et al. to present a model for phishing Website identification.

3. RESEARCH METHODOLOGY

First, we obtained a data set of phishing websites from the UCI [15] repository. The dataset was then subjected to a number of data preparation techniques. Both categorical features and missing values are absent from the dataset. After that, we used several feature selection techniques, and lastly, machine learning methods like support vector machines, decision trees, and random forests were utilised. When every classification technique had been used, the best model for phishing website detection was chosen. In Figure 2, the suggested model was displayed. We started by gathering a dataset from the UCI repository. Afterwards, we used data preparation strategies. ANOVA and Mutual information were two feature selection techniques that were later used. We then used a variety of machine learning classification algorithms.

Following that, we used two priority algorithms. Final fusion is based on these two algorithms.

A. Dataset:

Any machine Learning algorithm performance is basically depending on the selected dataset.

Number of Training samples	Number of Testing Samples	Total
7738	3317	11055

TABLE I. DATASET DETAILS

We collected a Phishing Websites Data Set from UCI ML repository. The collected data from UCI is in weka arff file format. We converted the arff dataset into csv file format. The

dataset-1 contains 30 features/attributes. In 30 features, last feature (Result) is a dependent feature and remaining are independent features (like port, HTTPS token, URL_of_Anchor, Abnormal_UR, web_traffic etc.).

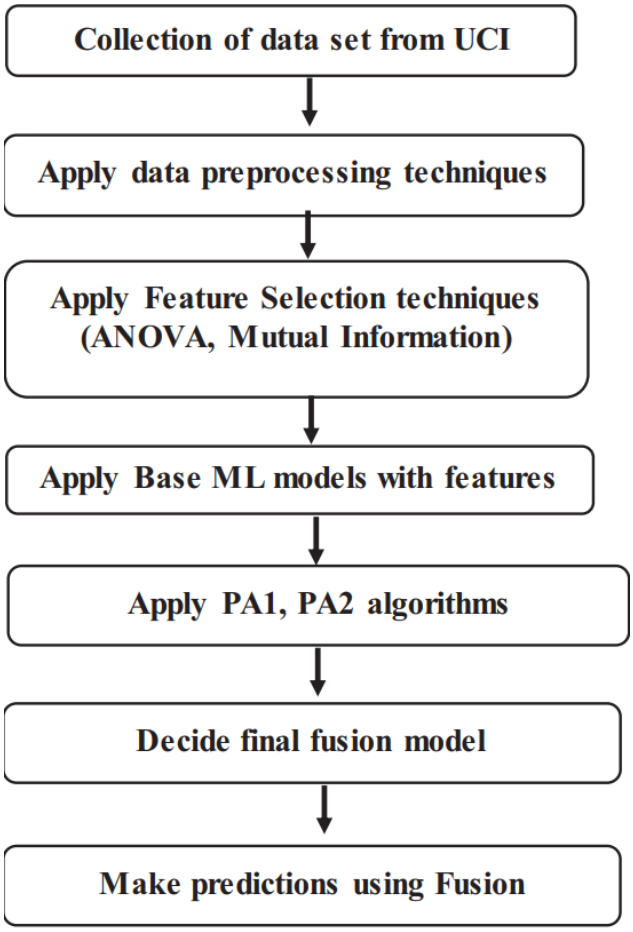

Fig. 2. Proposed Framework for SMS Spam Detection

The dataset's key characteristics are IP address, website URL length, double-slash redirection, any prefixes or suffixes, and if a subdomain exists or not. SSL final state, port, anchor URL, SFH, anomalous URL, and whether or not there are any pop-up messages on the website? Iframe, page rank, Google index, various links going to the page, shortening service, the age of the domain asking for Links, Does the domain have any subdomains? Having at least one symbol, various Links in Tags, an HTTPS Token, the length of the Domain Registration, etc. There are 11055 samples in the dataset. There are 7738 samples total that were used for training. 3317 samples total were used for testing.

B. Feature Selection

Feature selection process reduces the complexity of machine learning algorithm. There are various number of feature selection techniques available. We used two feature selection techniques namely ANOVA F-value,Mutual Information.Based on these two methods, we identified best features.We calculated ANOVA F-value for all the features in the dataset.Three features namely 'Iframe', 'Fevicon', 'popUpWidnow' got zero value of F-value.So these features are eliminated.Next,we calculated Mutual Information between every feature anddependant variable.Seven features namely 'having_At_Symbol', 'double_slash_redirecting', 'port', 'Abnormal_URL', 'Right-Click', 'popUpWidnow', DNSRecord got zero value,so these features also eliminated.So we removed 9 features from the original dataset.

C. Machine Learning Algorithm:

We applied several machine learning algorithms for phishing website detection.We applied following classification algorithms.

Logistic Regression

Logistic regression is a supervised learning algorithm for classification problems. It a regression technique in the background so its name includes regression.It two functions namely logit and sigmoid for processing the datasets.As it uses regression type technique in the background processit is named as logistic regression.

Support Vector Machine

SVM generates a hyperplane for classifying the dataset into various groups. SVM works well with nonlinear data as well. The hyperplane allows for a certain error rate that is not attainable with conventional classification models. A classifier called Support Vector Machine uses more complex mathematical notations. Support vector machines are helpful in the solution of both classification and regression issues.

K-Nearest Neighbors

K-NN is a simple but efficient classification algorithm. In KNN, nearest neighbors are identified and based on the count of neighbors. A new data point is assigned to a particular class based on this count.

Decision Tree Algorithm

Decision tree is a tree-based algorithm in which Gini index/gain measure is used to identify the root. This procedure applied recursively to build the whole tree.There is a threshold for splitting the tree.

Random Forest Algorithm

Random forest is an ensemble learning model. It combines various decision tree to assign a new data point to a class.As it

uses the decision of several decision trees, it is considered as a powerful model.

Ensemble learning

Ensemble Learning combines various classification algorithms into one. Random Forest is also ensemble model. Ensemble learners can be created using stacking classifiers and voting classifiers.

D. Fusion

In this research, we used a novel fusion classifier. We used the two priority-based algorithms PA1,PA2 to develop the fusion classifier. True positive rate and True negative rate are the foundations of these two algorithms. We used base classifiers and calculated TPR and TNR before applying PA1 and PA2. We used all of the fundamental classifiers, including logistic regression, Naive Bayes, decision trees, random forests, and gradient boosting. Classifier with support vectors. We used two priority-based methods after applying the base classifiers.

Priority algoithm1(PA1):

PA2 gives high priority to the classifier that is good in both categories(classes) .Here priority for base classifiers calculated using CDN(class difference number). CDN=Avg PA1 gives equal priority to True positive rate and True Negative rate. So it averages the values of both TPR and TNR(Avg-Value).Based on the average value ,we assigned priorities to base classifiers. -Value/|TNR-TPR|

IV. EXPERIMENTATION AND RESULTS

A. Evaluation Metric

Precision, Recall, Accuracy are the three measures used for comparing performance evaluation of classifiers. Recall=TP/ (TP + FN) Accuracy= (TP + TN) /(TP+TN+FP+FN) Precision=TP/ (TP + FP) Here, TP-True Positive, FP-False Positive TN-True Negative ,FN-False Negative.

B. Applying classification algorithms

We applied six Machine Learning base classification algorithms. After applying algorithms, True Positive Rate, True Negative Rate values are noted. TPR & TNR values are shown in table2. We achieved good True Positive Rate for of 97% with Random Forest.Gradient Boosting also given 96.7%. We achieved 95.9% True Negative Rate with Random Forest.

Algorithm	True Positive Rate	True Negative Rate
Logistic Regression	94.1%	89.5%
Naive Bayes	84.3%	93.7%
DTR	95.6%	93.9%
RF	97%	95.9%
Adaboost	95.3%	90.9%
Gradient Boosting	96.7%	94.8%

TABLE II. RESULTS OF EXPERIMENTS WITH ML CLASSIFIE

C. Applying PA1 algorithm

After applying base classification models,we applied proposed priority based algorithm PA1.The results of PA1 are shown in table3. PA1 algorithm considers both the True Positive Rate and True Negative Rate, so equal priority given to both of them. The results are shown in table-3. From table-3, it is shown that Random Forest is assigned with highest priority. The reason for this is that, the average value of TPR and TNR is more for random forest. Similarly, Gradient Boosting is assigned with p2(second highest priority), as the average of TPR and TNR is 95.7%. In this way, all the base classifiers are assigned with priorities p1, p2, p3, p4, p5, p6.

D. Applying PA2 Algorithm

After applying PA1,we applied second proposed priority based algorithm PA2.The results of PA2 are shown in table4.

Algorithm	TNR	TPR	Avg.Value	PA1 Priority
LR	94.1%	89.5%	91.8%	p5
Naive Bayes	84.3%	93.7%	89%	p6
DTR	95.6%	93.9%	94.75%	p3

TABLE III. RESULTS OF EXPERIMENTS WITH PA1-ALG

RF	97%	95.9%	96.4%	p1
Adaboost	95.3%	90.9%	93.1%	p4
Gradient Boosting	96.7%	94.8%	95.7%	p2

TABLE IV. RESULTS OF EXPERIMENTS WITH PA2-ALG

Alg	TNR	TPR	Avg	Diff	CDN	PA2 priority
LG	94.1	89.5	91.8	4.6	19.9	p5
NB	84.3	93.7	89	9.4	9.4	p6
DT	95.6	93.9	94.75	1.7	55.7	p2
RF	97	95.9	96.4	1.1	87.6	p1
AB	95.3	90.9	93.1	4.4	21.1	p4
GB	96.7	94.8	95.7	1.9	50.3	p3

ICDN=Avg-Value/|TNR-TPR|

From table-4,it is observed that Random Forest is assigned with highest priority. The reason for this is that, the value of CDN is more for random forest. Similarly, Decision Tree Classifier is assigned with p2(second highest priority), as DTC achieved second highest CDN value. Similarly, all the base classifiers are assigned with priorities p1, p2, p3, p4, p5,p6.PA2 algorithm assigns highest priority to a classifier that is performing good in both the classes.

E. FINAL FUSION

The final fusion is based on the priorties achieved from PA1,PA2 algorithms.In PA1 algorithm Random Forest achieved highest priority and Gradient Boosting achieved second highest priority.In PA2 algorithm,Random Forest achieved highest prioity.Next two priorities are assigned to Decision Tree and Gradient Boosting.Based oh these results,we created a fusion model with Random Forest,Decision Tree classifier and Gradient Boosting.We applied fusion technique with with stacking classifier in the final model and achieved an accuracy of 97%.

F. Comparison with Previous Work

We compared our proposed model wth previous works for phishing website detection.Table-5 and Figure-3 shows the comparison of the proposed model with previous work. In [5],authors achieved an accuracy of 95% with logistic regression. In [7],they achieved ana ccuracy of 89.3% with ELM.In this paper, we applied a novel fusion classifier with two prioity algorithms and achieved an accuracy of 97%.

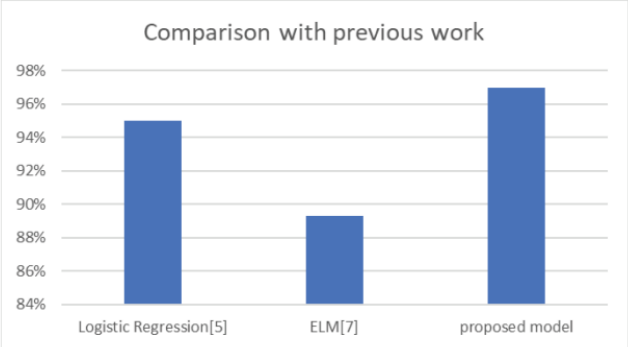

Fig. 3. Comaprison woth previous work

TABLE V. COMPARISON WITH PREVIOUS WORK

Model	Accuracy
Logistic Regression [5]	95%
ELM [7]	89.3%
proposed model	97%

V. CONCLUSION & FUTURE WORK

In this paper, we applied various machine learning algorithms logistic regression, decision tree classifier, random forest classifier, AdaBoost, gradient boosting classifier for the phishing detection. We used a dataset from the UCI machine learning repository for our experiments. Later, we applied two priority algorithms PA1, PA2. Based on the results of prioritybased algorithms final fusion model was decided. Later, we applied a fusion classifier and achieved an accuracy of 97%.The proposed model was tested on one dataset only.In future,we will test several fusion models on more datasets.

REFERENCES

[1] Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques", arXiv:2009.11116v1 [cs.CR] 20 Sep 2020.

[2] G.Ravi Kumar, Dr.S.Gunasekaran, Nivetha.R, "URL Phishing Data Analysis and Detecting Phishing Attacks Using Machine Learning In NLP," in International Journal of Engineering Applied Sciences and Technology-2019,Vol. 3, Issue 10, ISSN No. 2455-2143.

[3] K.Venkateswara Rao, Jagan Mohan Reddy D, G L Vara Prasad, "An Approach for Detecting Phishing Attacks Using Machine Learning Techniques," in Journal of Critical Reviews,vol-7,issue-18,2020.

[4] Amani Alswailem, Norah Alrumayh, Bashayr

Alabdullah, Dr. Aram Alsedrani, "Detecting Phishing Websites Using Machine Learning," in International Conference on Computer Applications & Information Security (ICCAIS), 978-1-7281-0108-8/19- 2019 IEEE.

[5] Meenu, Sunila godara, "Phishing Detection using Machine Learning Techniques," in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958,vol-9,issue-2,Dec-2019.

[6] Abdul Basit,Maham Zafar.Xuan Liu,Abdul Rehman Javed,Zunera Jalil.Kashif Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques.: Telecommunication Systems", https://doi.org/10.1007/s11235-020-00733-2,Springeroct,2020ISBN:978-1-5386-0965-1.

[7] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana GutthaKhalilian and Nikravanshalmani, Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques," in International Journal of Innovative Technology and Exploring Engineering (IJITEE),volume-2,issue-8S2,June 2019.

[8] Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, Nidhi Gupta, "Phishing Website Detection System Using Machine Learning.," in International Research Journal of Engineering and Technology (IRJET),Voume-7,Issue-5,May-2020.

[9] Jagadeesan, S., Chaturvedi, "URL phishing analysis using random forest", International Journal of Pure and Applied Mathematics, 118(20), 4159–4163.

[10] Arun Kulkarni,Leonard L.Brown, "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications,Volume-10,No-7,2019.

[11] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning",International Journal of Recent Technology and Engineering,Volume-8, Issue-2S11, ISSN: 2277-3878,September 2019.

[12] Preeti, Rainu Nandal, and Kamaldeep Joshi "Phishing URL Detection Using Machine Learning", International Conference on Advanced Communication and Computational Technology,Lecturer Notes in Electrical EngineeringVolume-668,,pages-547-560,2019.

[13] Ali Aljofey,Qiang Qu,J-P Niyigena ," An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL", Electronics, Electronics 2020, 9, 1514; doi:10.3390/electronics9091514,MDPI.2020.

[14] Huang, Y. Yang, Qin.Q, J Wen W," Phishing URL Detection via CNN and Attention-Based Hierarchical RNN Proceedings of the IEEE International Conference On Trust, Security And Privacy in Computing And Communications,IEEE International Conference On Big Data Science& Engineering-2019.

[15]https://archive.ics.uci.edu/ml/datasets/Phishing+Websites.

