Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

Phishing website detection using novel machine learning fusion approach

A. Lakshmanarao
Department of Information Technology
Aditya Engineering College
Surampalem, A.P, India
laxman1216@gmail.com

P.Surya Prabhakara Rao
Department of Information Technology
Anurag University,Ghatkesar,
Hyderabad, Telangana, India
suryaprabhakar.p@gmail.com

M M Bala Krishna
Department of Computer Science& Engineering
St Mary's Womens Engineering College
Guntur, A.P, India
balakrishna2508@gmail.com

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395810

***Abstract*—The Phishing is a sort of social designing assault regularly used to take client information, including login accreditations and credit card numbers. With the enhancements in internet technology, websites are the major resource for the cyber-attacks. There are several counter measures available for avoiding phishing attacks, but phishers are changing their attacking methods from time to time. One of the most widely used techniques for solving cybersecurity issues is machine learning. From last several years, Machine Learning and Deep Learning Techniques are suitable for solving security related issues. Machine Learning is most suitable for detecting phishing attacks because most of the phishing attacks have common characteristics. This paper has applied several machine learning techniques for detecting the phishing attacks. Here, two priority-based algorithms are proposed. Based on the results of these algorithms, the final fusion classifier is decided. We used a dataset from UCI and applied a novel fusion classifier and achieved an accuracy of 97%. We used Python for implementing our experiments.**

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

I. INTRODUCTION

Social Engineering is the most widely used term today. Every individual facing lots of problems with cyber threats. One of the most widely used attacks in social engineering is phishing. It happens when an attacker behaves like a trusted source and hoodwinks a casualty into opening an email, text, or instant message. Phishing can be done in different ways. For example, a spam email from some university is distributed to many faculty members. The email may ask the user to click on the link. On clicking the link, it opens a duplicate website page. The attacker monitors and hijacks the new password. In a phishing attack, the users are forced to link to illegal websites and revealed their critical information like bank-related information, credit card details, passwords, etc. One of the most widespread solutions for cyber-attacks is using an antivirus or firewall. But unfortunately, antivirus software is unable to fully prevent phishing attacks.

The reason for this is that the phishers are diverting the users into a dummy/fake webserver. Attackers also using secure browser connections for making their illegal activities. The reason for an increase in phishing attacks is not having the correct tools for preventing these attacks, so companies are unable to train their employees in this field. The general countermeasures used by the companies are educating their employees with mock phishing attacks, updating all their systems with the latest security measures, or encrypting sensitive information. One of the most reasons for becoming a victim of this phishing attack is browsing without care. Phishing websites are similar to legitimate websites. The appearance of spoofed website is same as actual site. For example, a user may receive an email from PayPal (but not actually from PayPal) that the account of the user is limited. The sample phishing email is shown in figure-1. If the user responds and do some action, then user credentials are stolen. Machine Learning is a term coined by Arthur Samuel in 1959, which dominated all areas today. In machine learning, there are several techniques like supervised learning models, unsupervised learning models, reinforcement learning models. In supervised learning there are two samples of data, one is train data and the other is test-data. The train-data is used for the learning phase, where a model is constructed after training is done. The constructed model is used in the testing phase with test data for evaluating the performance of the model. In unsupervised learning, there are no labels attached to the data. Reinforcement learning is an interactive learning model. Supervised learning can be regression or classification. In regression, the predicted result is a number whereas in classification the predicted result is a label. Supervised learning solves most of the real-life problems. Classification of suspected URLs in phishing attacks can be treated as a classification problem, so supervised learning algorithms work well for phishing detection. There are number of classification

algorithms available, but choosing the right algorithm is a crucial part for solving the given problem. The rest of the paper is organized as follows. Section 2 contains literature survey,

Section 3 explains about proposed research methodology, Section fours shows the results of the experiments and finally Section 5 is conclusion.

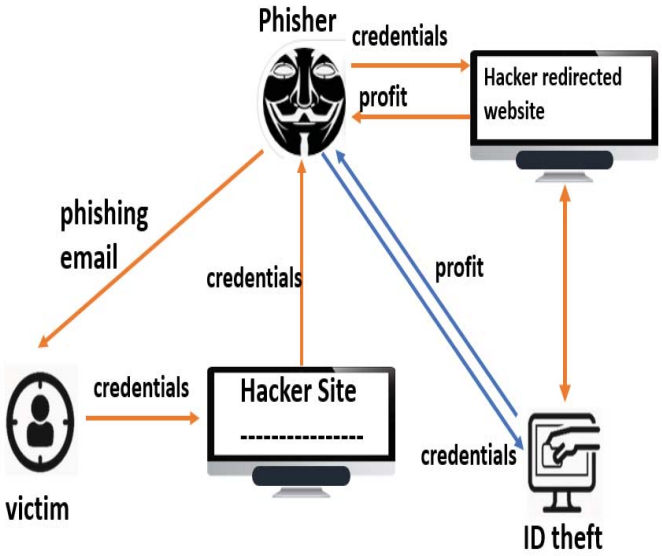

Fig. 1. Phishing attack diagram

II. LITERATURE SURVEY

Applying machine learning techniques to prevent phishing attacks is not a new era. Several researchers applied machine learning techniques to phishing attacks. Vahid Shahrivari et.al [1] applied machine learning techniques for phishing detection. They applied logistic regression classification algorithm, SVM, Adaboost algorithm, KNN, random forest, ANN and achieved good accuracy with the random forest algorithm. Dr.G. Ravi Kumar [2] et.al applied various machine learning algorithms for the detection of phishing attacks. They applied Natural Language Processing techniques for better performance. They achieved good accuracy with Support Vector Machine with preprocessed data by NLP techniques. Venkateshwara Rao [3] et.al applied decision trees, Support vector classifiers, and random forest models for detecting phishing attacks and achieved good accuracy with their model. Amani Alswailem[4] et.al applied several machine learning models to phishing attacks and achieved good accuracy with a random forest algorithm. Meenu[5] et.al applied Logistic Regression, Support Vector classifiers Decision Tree classifiers, and Artificial Neural Networks for predicting phishing emails and achieved more accuracy with logistic regression classifier. Abdul Basit[6] et.al reviewed various techniques, trends, opportunities, and challenges for phishing attack detection. Artificial Neural Network is a deep learning model which can solve classification and regression problems. In ANN, there is no need to apply any feature selection technique. Feature extraction can be automatically done by neural networks. Some researchers applied ANN for detecting phishing website detection. Sandeep Kumar [7] et.al applied machine learning techniques for the detection of phishing websites and achieved an accuracy of 89.3% with naïve Bayes classifier and ANN. Manish Jain [8] et.al surreys phishing detection using machine learning techniques. Jagadeesan[9] et al. applied support vector machines and random forest classifiers for phishing detection. They used two different datasets from the UCI repository. Arun Kulkarni[10] et.al proposed four classification algorithms on a real world dataset with 1,353 URLs and achieved an accuracy above 90%.R. Kiruthiga[11] et.al compared various machine learning techniques and shown that machine learning is a good tool for handling phishing website detection. Preeti[12] et.al applied various classification models for phishing website detection.They applied Decision Tree,Random Forest,SVM,Logistic Regression.With 1200 URLs,they achieved good accuracy with Logistic Regression.But,when they applied Logistic Regression to 12000 sites,they achieved less accuracy with Logistic Regression.From the experiments, they concluded that Decision Tree performs well irrespective of size of the dataset. Convolutional Neural Networks solves classification problem by using different steps like convolutional layer, activation function, pooling step and flattening. After flattening, it is similar to ANN. Ali Aljofey[13] et.al proposed deep learning model for phishing website detection. They applied character level convolutional neural network model and achieved good results.Y.Huang[14] et.al proposed phishing URL detection model using convolutional neural networks attention based hierarchical Recurrent Neural Networks.

III. RESEARCH METHODOLOGY

First, we collected a phishing Websites Data Set from UCI [15] repository. After that, we applied various data preprocessing techniques to the dataset. The dataset has no missing values and no categorical features. Later we applied different feature selection methods and finally, we applied various machine learning techniques like support vector machine, decision tree classification, random forest classification. After applying all classification techniques, we selected the best model for phishing website detection. The proposed model was shown in Figure-2. First, we collected a dataset from UCI repository. After that, we applied data preprocessing techniques. Later we applied two feature selection techniques ANOVA, Mutual information. Next, we applied several machine learning classification algorithms. Later, we applied two priority algorithms. Based on these two algorithms, final fusion classification model was decided.

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.).

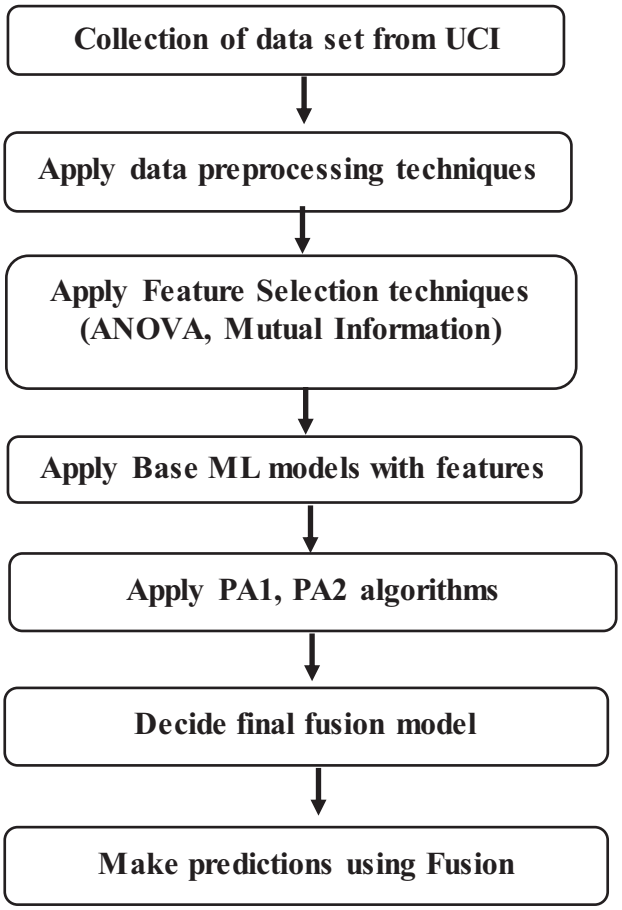

Fig. 2. Proposed Framework for SMS Spam Detection

The main features of the dataset are IP_address,URL length of website,double slashing redirecting,any prefix or suffix exists,having sub domain or not?,SSL final state,port,URL of the anchor,SFH,abnormal URL,any pop up notifications exisits in the wbisite or not?,Iframe,The age of the domain,page_rank,Google index,different links pointing to the webpage,shorting service, Requesting URLs,Is domain is having sub domain or not?,having_at_symbol, different Links_in_tags HTTPS-token,Domain-registeration-length etc.The dataset contains 11055 samples.The number of samples used for training are 7738.The number of samples used for testing are 3317.

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 and dependant 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 Algorithms

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 process,it is named as logistic regression.

Support Vector Machine

SVM create a hyperplane for dividing the dataset into different classes.SVM is well suited for nonlinear data also.Hyperplane provides for some error rate which is not possible in normal classfication models.Support Vector Machine is a classifier which involoves more mathematical notations.Suport Vector Machines are useful for solving both classification problems ans regresison problems.

K-Nearest Neighbors

K-NN is a simple but efficient classification algorithm. In K-NN,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

We applied novel fusion classifier in this paper. We created

fusion classifer using two priority based algorithms PA1,PA2.These two algorithms are based on True positive rate and True negative rate.Before applying PA1&PA2,we applied base classifiers and TPR and TNR are calculated.We applied all base classifiers like logistic regression,Naive Bayes,Decision Tree Classfier,Random Forest,Gradient Boosting classifier.Support Vector Classifier.After applying base classifiers,we applied two priority based algorithms.

Priority algoithm1(PA1):

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.

Priority algoithm2(PA2):

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 Metrics

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.

TABLE II. RESULTS OF EXPERIMENTS WITH ML CLASSIFIERS

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%

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.

TABLE III. RESULTS OF EXPERIMENTS WITH PA1-ALG

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

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%.

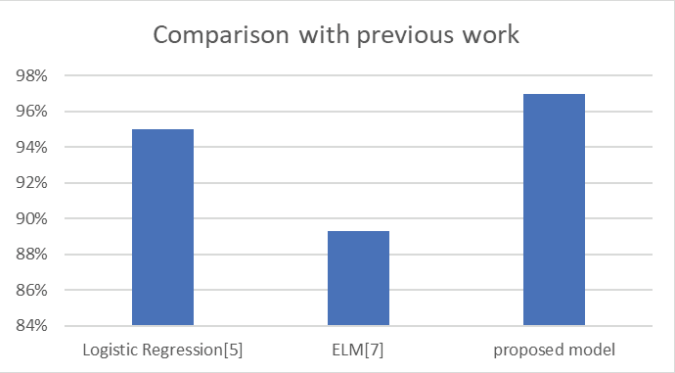

Fig. 3. Comaprison woth previous work

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 priority-based 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,Springer-oct,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.