**Phishing Website Detection Using Ensemble Machine**

**Learning Methods**

A.V.S. Mounika ,Dr. K. Lakshmana Rao

Under Graduation of Computer Science

Engineering Institute of Technology,

Associate Professor ,Dr. k. Lakshmana Rao,

Department of CSE, GMR Institute of Technology

**Abstract:** One of the most dangerous cybercrimes is phishing, where the user’s information and credentials are stolen using fake emails or websites that are sent to the target and look like legitimate ones. Phishing attacks have been increasing over the years, and affect many internet users. In this type of attack, the phisher selects any organisation as a target, and then develops a phishing website that is similar to the organisation’s legitimate website. The phisher then sends several spam emails or posts these links using social media or any communication medium to many internet users, who may click on these links and be redirected to the phishing website Accurate detection of phishing websites is a challenging problem because it depends on several dynamic factors. Ensemble methods are considered the state-of-the-art solution for many classification tasks. Ensemble learning combines the predictions of several separate classifiers to obtain a higher performance than a single classifier. Timely detection of phishing attacks has become most crucial than ever. Machine learning algorithms can be used to accurately detect phishing attacks before a user is harmed. This learning observes- a novel ensemble model to detect phishing attacks on the website. We select random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM. A genetic algorithm (GA) is used to optimize the performance of several ensemble classifiers. Then, the best optimized classifiers are used as base classifiers for the stacking ensemble method.

**Keywords:**Cyberattacks,Web threat, Phishing Website Detection, Machine Learning, Ensemble Learning

# Introduction

Phishing attacks are the practice of sending fraudulent communications that appear to come from a reputable source. It is usually done through email. The goal is to steal sensitive data like credit card and login information, or to install malware on the victim’s machine. Phishing is a common type of cyber-attack that everyone should learn about to protect themselves. This has been a great threat for humans right now. These phishers are different from hackers as these people does not need to access the files or any other data from the disk or from the computer, they just trap people by making identical and duplicate interface of a website. These are mostly used to trap the credentials of users who has access to a huge amount of money or assets. these look alike in appearance but in different in URL

In order to receive confidential data, criminals develop unauthorized replicas of a real website and email, typically from a financial institution or other organization dealing with financial data

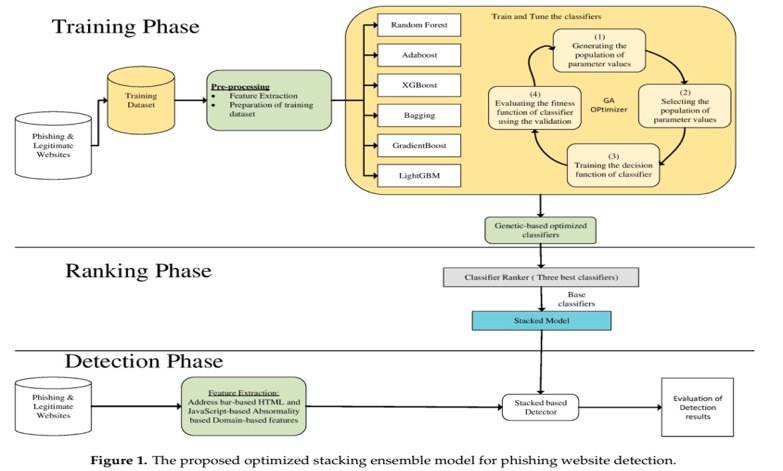
Scientists and researchers have implemented so many algorithms like recurrent neural network (RNN), Artificial Neural Network (ANN), K-Nearest (KNN),etc. but random forest has shown the best accuracy

Email spoofing is one of the techniques used when phishing someone, email spoofing is where you will receive a professional kind of Email from a specific site or user where you can brief into the sender’s information. While you take a close look at that, the sender info would be made in the much-hidden format so that it doesn’t create a frightening kind of feeling for the person in the receiving end to trust and do the further reply. They can also reach out to you via IM technique, IM is generally referred to as instant messaging where the experience of live chatting could be promoted. It creates a space for both the persons in the opposite end to have a smooth communication between them, it is little similar to text messaging but instant messaging as per the word it differs in a certain aspect possible. All these techniques are mainly used for manipulating the people and make them believe in the authenticity to collect the information that is needed by them. Proposals like detection through html , url and license related have made a good help for the detection of phishing websites

1. **METHODOLOGY**

**2.1** **An optimized stacking ensemble model for phishing websites detection**

The methodology consists of three main phases: the training, ranking, and testing phases. In the training phase, random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM were trained without optimization. The reason behind this is twofold: on the one hand, to obtain a general insight into the performance of ensemble classifiers before optimizing them, and on the other hand, to explore which of the phishing websites’ characteristics is most useful. The aforementioned classifiers were then optimized using the genetic algorithm. Here, the genetic algorithm was used for selecting the optimal values of model parameters in order to improve the overall accuracy of the proposed model. Later, in the ranking phase, the optimized classifiers were ranked and used as a base classifier for the ensemble classifier— the stacking method. In the testing phase, new websites were collected and used as testing data. In the detection phase, these steps will be applied to any website in future in order to detect whether it is a benign or malicious website



**2.2 An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique.**

RRFST reduces the features from phishing E-mail data set based on simple remove and replace policy.

**Input:**

A[ ] = Phishing E-mail or phishing website data set with ‘n’ number of features

Ac= Accuracy of ensemble model (C4.5+CART) with ‘n’ number of features

fr = Removed any one feature from A[ ] .

**Output :**

Ar[ ]= Reduced feature subsets

Md= Computationally efficient model

RRFST (A[ ] , Ar[ ] )

* Start
* Ar [ ]=Remove any one feature fr from feature set A[ ]
* A’c = Apply Ar[ ] feature set to ensemble of C4.5 and CART and calculate the accuracy
* if (A’c > =Ac)

{

Don’t replace the removed feature fr into A[ ]

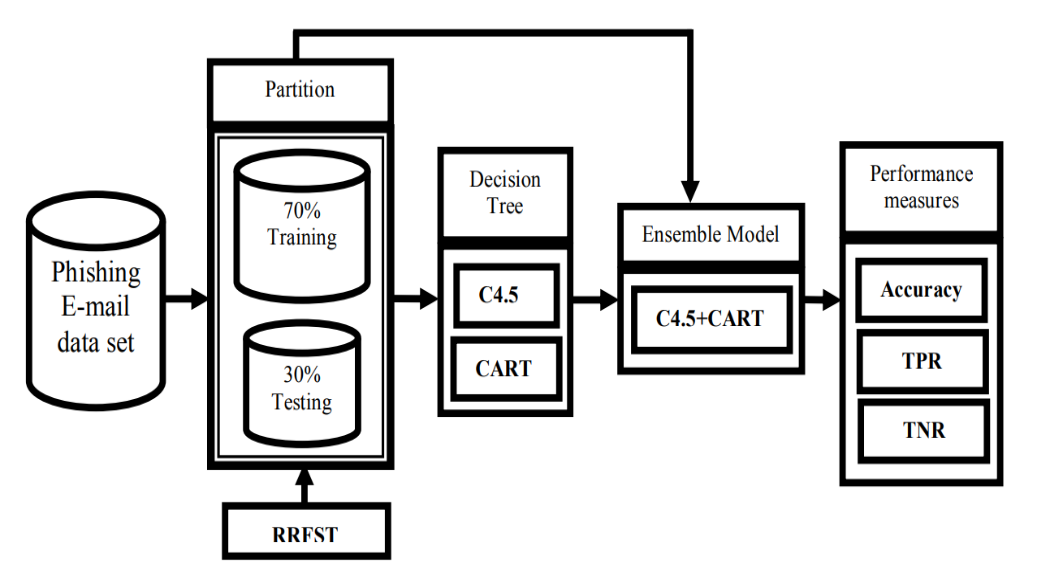
Else

Replace the removed feature fr into A[ ]

}

Ar [ ]= A [ ]

* Repeat step 2 to 4 until completion of all features of A[ ]
* Md = Apply the Ar[ ] to the ensemble of C4.5 and CART
* Recommended Md as the best model in case of ensemble of C4.5 and CART with reduced feature set
* Stop



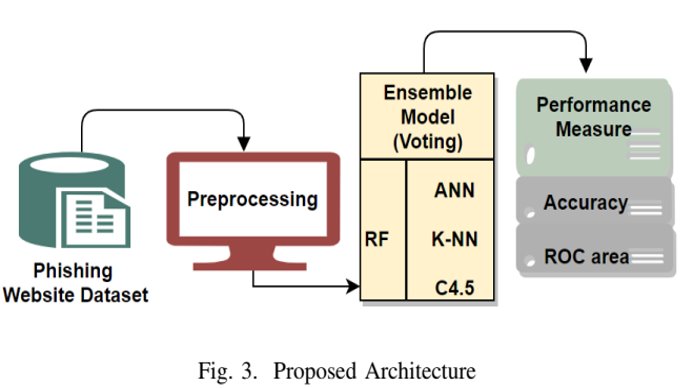
**Fig 2. Proposed architecture of the method**

**2.3 A novel ensemble machine learning method to detect phishing attack**

The propose model is made up of three major steps: Data Collection, Pre-Processing, and Ensemble Model. Dataset Collection: In this step, we take the dataset as input and identify features. Pre-Processing : This step includes placing URL instances in two categories as legitimate and phishing URLs. Ensemble Models: We use a novel ensemble model to detect phishing attacks over a website as ensemble techniques showed better performance in the past.

Voting algorithm is used to combine two classifiers taking RFC as a base classifier with ANN, KNN, and C4.5 algorithms. All the algorithms are used with a batch size of 100 and 10-fold cross-validation is used to validate the efficiency of the classifier.

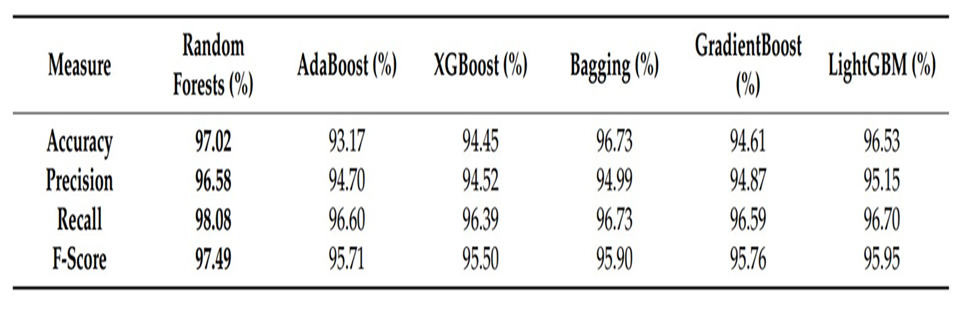
Performance of different combinations of ensemble methods to identify the best combination in detection.



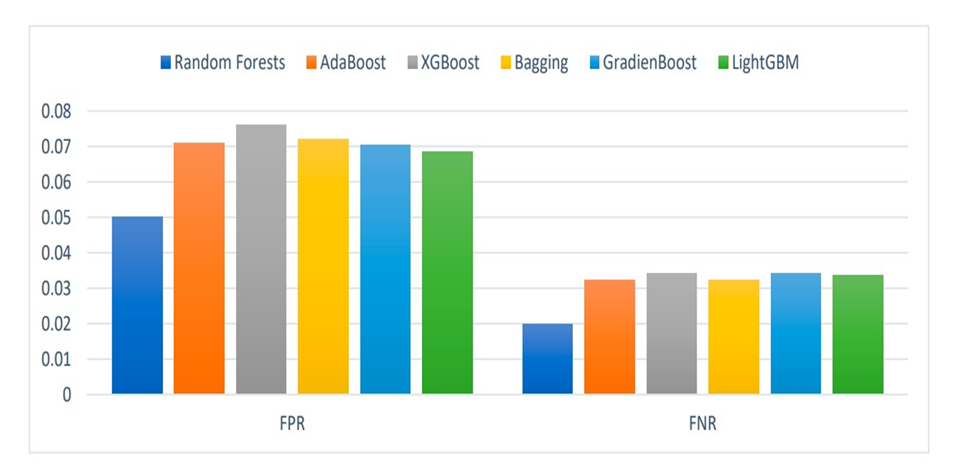
## 3 RESULTS AND DISCUSSIONS

**3.1** In this section we are going to check the results and have a discussion about the model. At first a set of ensemble classifiers are trained using 10 fold cross validation without using genetic algorithm to find out the difference .

For this they have taken three datasets and applied various algorithms like adaboost, XG boost ,random forest etc .Apart from all the classifiers n, the random forests classifier yielded the best performance compared with the other classifiers in terms of accuracy, precision, recall, and F-score; it achieved 97.02% accuracy. The following table shows the detailed results of different classifiers



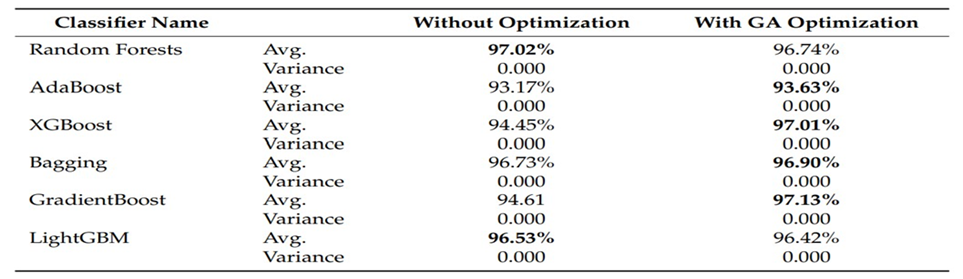
**Fig 3.1 Comparision of accuracy metrics**



**Fig. 3.2 PNR and FNR rates**

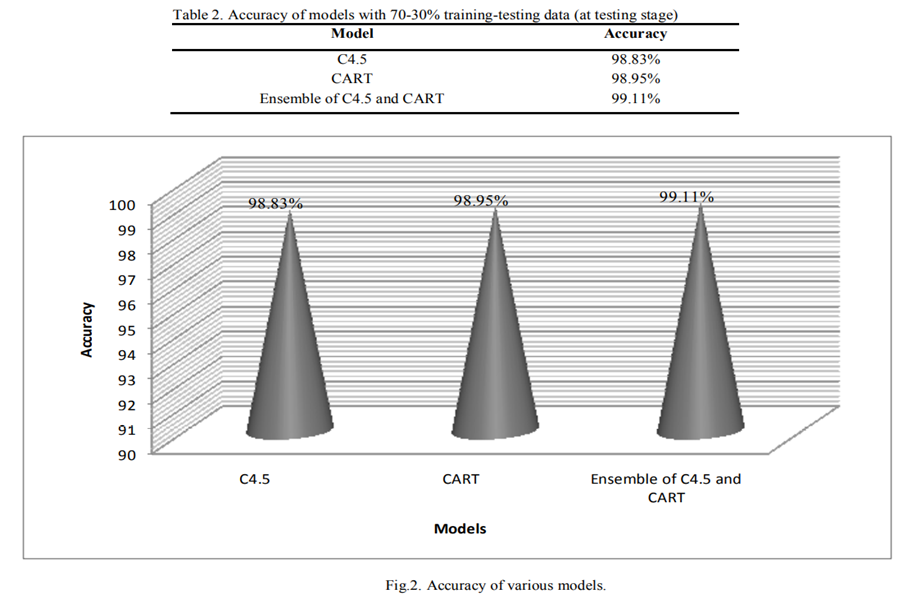
Although they have shown good performance, we have used GA classifier to enhance the accuracy .They have adjusted some parameters by tuning them for this the GA algorithm .Except for random forest classifier all the classifiers have shown better results after enhancing with GA

The average accuracy and variance values of all of the classifiers, before and after conducting GA optimization:



**Fig 3.3 final observation result**

**3.2** Experiment is carried out using Waikato Environment for Knowledge Analysis (WEKA) data mining software .As stated above first of all original phishing E-mail data set is presented to C4.5, CART and its ensemble model with 70%-30% ratio of training-testing data set to develop binary classifier. Table 2 show that accuracy of models, where ensemble of C4.5 and CART was found better than individual of C4.5 (98.83% accuracy) and CART (98.95% accuracy) with 99.11% accuracy and is enough capable and robust for classification of phishing E-mail data more accurately. A comparative graph of obtained results is also shown in figure

****

**3.3** In this methodology they have used ANN KNN and decision c4.5 on Random Forest algorithm. They evaluated the performance of the proposed ensemble model using an ensemble of multiple classifiers. The common classifier is random forest classifier and It can be seen that the TP rate improves in an ensemble of KNN and RFC. The precision of RFC and KNN have the highest among all that is 0. 970.The ensemble of KNN with RFC gives an accuracy of 97.33%.

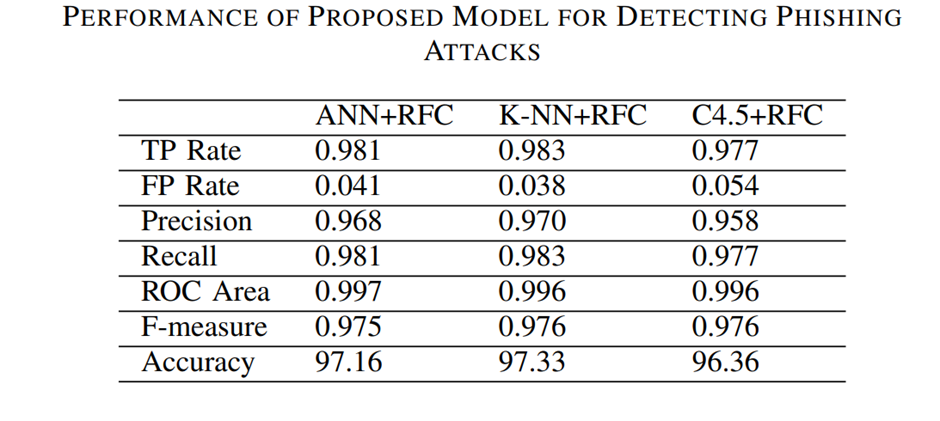
Finally, we made a conclusion that the algorithms which are infused with Random Forest have gained accuracy

**True positive rate (TPR)**

True Positive Rate is also called sensitivity, this shows positively tested subjects under examination. It can also be defined as the total number of true positives by the total number of true positives and false negatives

**True negative rate (TNR)**

True negative rate is also called specificity shows negative tested identified correctly. It can also be defined as total number of true negatives by the total number of false positives and true negatives



**3.3 Comparison Table:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F-Score** | **ROC-Area** | **TP Rate** | **TF Rate** |
| Ensemble of RF, AD-B, XG-B, Light- GBM | 98.39 | 98.46 | 98.13 | 98.43 | - | - | - |
| ANN+RFC | 97.16 | 96.8 | 98.1 | 97.5 | 99.7 | 98.1 | 4 |
| KNN+RFC | 97.33 | 97 | 98.3 | 97.6 | 99.6 | 98.3 | 3.8 |
| C4.5+RFC | 96.36 | 95.8 | 97.7 | 97.6 | 99.6 | 97.7 | 5.4 |
| C4.5 | 98.83 | - | - | - | - | - | - |
| CART | 98.95 | - | - | - | - | - | - |
| Ensemble of C4.5 and CART | 99.11 | - | - | - | - | - | - |

**4 CONCLUSION**

Phishing is a technique that allows the user to deceive i.e. fake e-mails and to lure another user onto fake websites in order to steal someone’s private data. Website detection for phishing is an important aspect as it has impact on online transactions . The world needs to advance protection techniques and anticipate and prevent financial losses and the stealing of data. we have proposed ensemble learners that will advance phishing prevention and work against attacks. What’s more, we have adopted an ensemble classifier model that uses a publicly available dataset to offset phishing websites in an intelligent and automated manner. Ensemble models improve the performance of the classifiers in terms of classification accuracy, F-measure and ROC area. Experimental results reveal that by using ensemble models it is possible to detect phishing pages with an accuracy of %99.11, given the presented model. If a proper combination of ensemble classifiers is employed, our model is able to detect phishing webpages.

In the future, with future studies, feature selection methods can be planned in our present model that will take away the need for the dependency on webpage content. Deep learning techniques can also be implemented for detection of phishing websites. Furthermore, we need further study to detect phishing attacks via mobile devices. At present, ‘smartphones’ are a very popular technological offspring. These smartphones are also a common point where attackers merge, where phishing attacks occur. Users of mobile phones prefer to read their e-mails on their phones immediately. Thus, what is needed is to find possible new features as well as potent, fast machine learning algorithms that will detect phishing attacks that occur in mobile devices

**5 REFERENCES**

* Al-Sarem, M., Saeed, F., Al-Mekhlafi, Z. G., Mohammed, B. A., Al-Hadhrami, T., Alshammari, M. T., ... & Alshammari, T. S. (2021). An optimized stacking ensemble model for phishing websites detection. Electronics, 10(11), 1285.
* Dutta, A. K. (2021). Detecting phishing websites using machine learning technique. PloS one, 16(10), e0258361
* Parmar, S. R. (2020). Detection of Phishing URL using Ensemble Learning Techniques (Doctoral dissertation, Dublin, National College of Ireland).
* Basit, A., Zafar, M.,Javed, A. R., & Jalil, Z. (2020, November). A novel ensemble machine learning method to detect phishing attack. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE
* Kiruthiga, R., & Akila, D. (2019). Phishing websites detection using machine learning. International Journal of Recent Technology and Engineering, 8(2), 111-114..
* Basit, A., Zafar, M.,Javed, A. R., & Jalil, Z. (2020, November). A novel ensemble machine learning method to detect phishing attack. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE
* Hota, H. S., Shrivas, A. K., & Hota, R. (2019). An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique. Procedia computer science, 132, 900-907.
* Dharani, M., Badkul, S., Gharat, K., Vidhate, A., & Bhosale, D. (2021). Detection of Phishing Websites Using Ensemble Machine Learning Approach. In ITM Web of Conferences (Vol. 40, p. 03012). EDP Sciences.
* Miao, M., & Wu, B. (2020, June). A flexible phishing detection approach based on software-defined networking using ensemble learning method. In Proceedings of the 2020 4th International Conference on High Performance Compilation, Computing and Communications (pp. 70-73).
* Ubing, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. International Journal of Advanced Computer Science and Applications, 10(1).
* Nagaraj, K., Bhattacharjee, B., Sridhar, A., & Sharvani, G. S. (2018). Detection of phishing websites using a novel twofold ensemble model. Journal of Systems and Information Technology.
* Taha, A. (2021). Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting. Mathematics, 9(21), 2799.
* Zamir, A., Khan, H. U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., & Hamdani, M. (2020). Phishing web site detection using diverse machine learning algorithms. The Electronic Library, 38(1), 65-80.
* Subasi, A., & Kremic, E. (2020). Comparison of adaboost with multibooting for phishing website detection. Procedia Computer Science, 168, 272-278.
* Niranjan, A., Haripriya, D. K., Pooja, R., Sarah, S., Deepa Shenoy, P., & Venugopal, K. R. (2019). Ekrv: Ensemble of knn and random committee using voting for efficient classification of phishing. In Progress in advanced computing and intelligent engineering (pp. 403-414). Springer, Singapore.