WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing Website Detection Using Ensemble Machine Learning Methods

Atmakuri. Venkata Sai Mounika, Dasari. Vignan

Under Graduation of Computer Science
Engineering Institute of Technology,
Department of CSE, GMR Institute of Technology

A B S T R A C T

Phishing is a highly dangerous cybercrime that involves stealing user information and credentials through deceptive emails and websites that appear legitimate. Over the years, phishing attacks have seen a significant increase, impacting numerous internet users. In these attacks, the perpetrator chooses a target organization and creates a fraudulent website that closely resembles the legitimate one. They then distribute spam emails or share links via social media or other communication channels to reach a large number of potential victims. When users click on these links, they are redirected to the phishing website. Detecting phishing websites accurately poses a considerable challenge due to various dynamic factors. To address this issue, ensemble methods have emerged as the leading solution for classification tasks. Ensemble learning combines the predictions of multiple independent classifiers to achieve higher performance compared to individual classifiers. The timely detection of phishing attacks has become more critical than ever, and machine learning algorithms can play a crucial role in accurately identifying such attacks before users suffer any harm. To this end, a novel ensemble model is proposed for detecting phishing attacks on websites. The model incorporates several state-of-the-art classifiers, including random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM. To optimize the performance of these ensemble classifiers, a genetic algorithm (GA) is employed. The GA helps identify the best-performing classifiers, which are then used as base classifiers for the stacking ensemble method. This approach aims to enhance the overall detection capabilities and provide robust protection against phishing attacks.

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.

Methodology

An optimized stacking ensemble model for phishing websites detection

The training, ranking, and testing phases make up the methodology's three primary sections. Random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM were trained throughout the training phase without being optimised. This is done for two reasons: one, to gain a broad understanding of ensemble classifier performance prior to optimising it, and second, to determine which aspect of phishing websites is most helpful. The evolutionary algorithm was then used to improve the aforementioned classifiers. In this case, the genetic algorithm was utilised to choose the ideal model parameter values in order to increase the suggested model's overall accuracy. The optimised classifiers were ranked later during the ranking phase and used as a base classifier for the ensemble classifier—the stacking approach. New websites were gathered and used as testing data during the testing phase. These procedures will be used in the detection phase to identify if a website is benign or dangerous in the future.

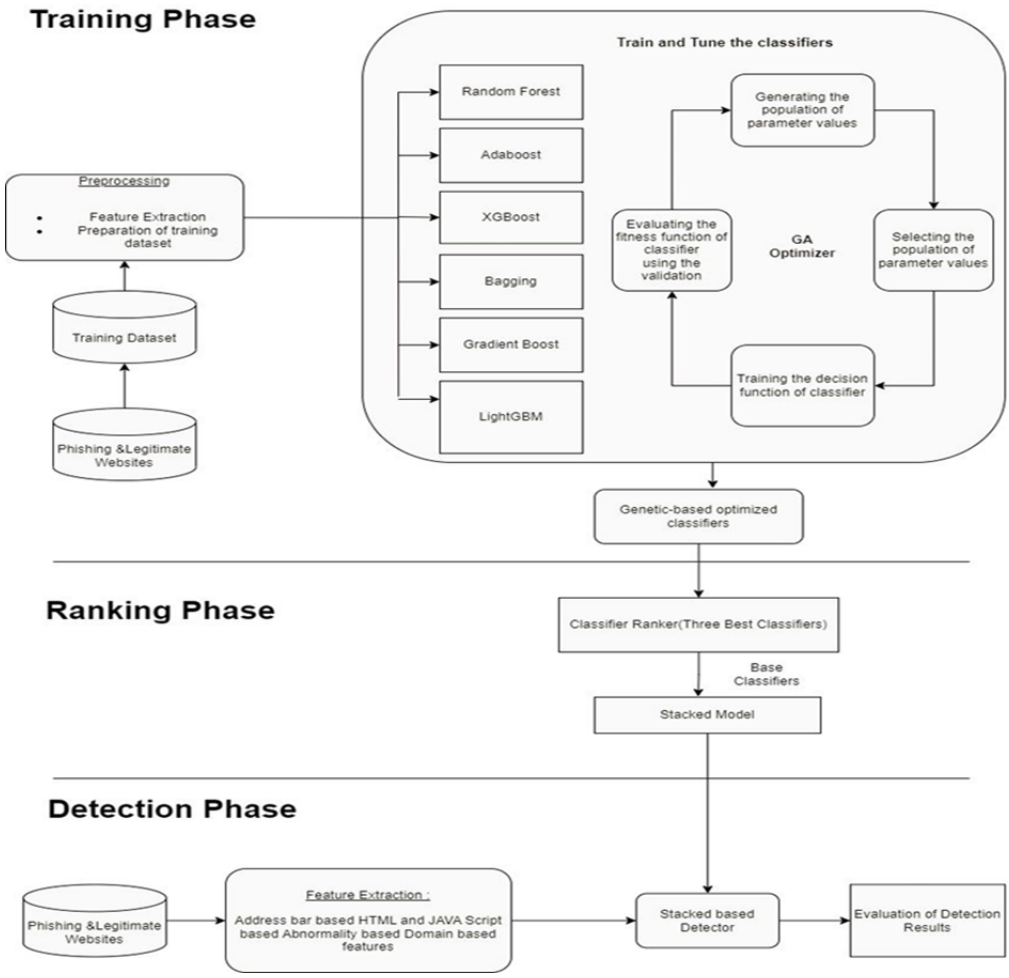


Fig.1 – Flowchart of processes

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

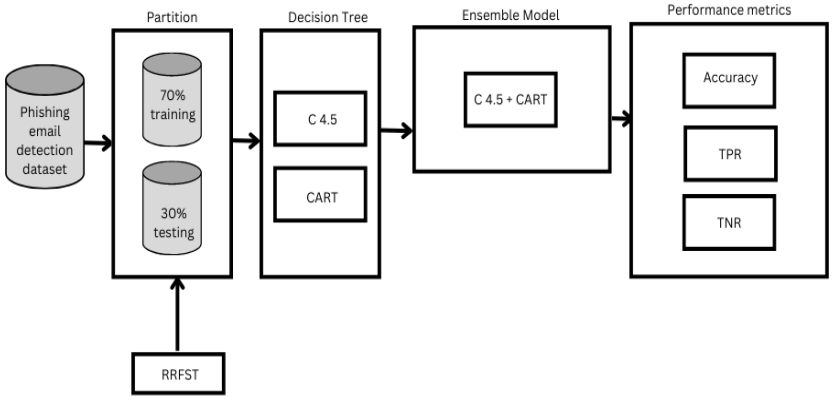


Fig. 2 - Proposed architecture of the method

2.3 A novel ensemble machine learning method to detect phishing attack

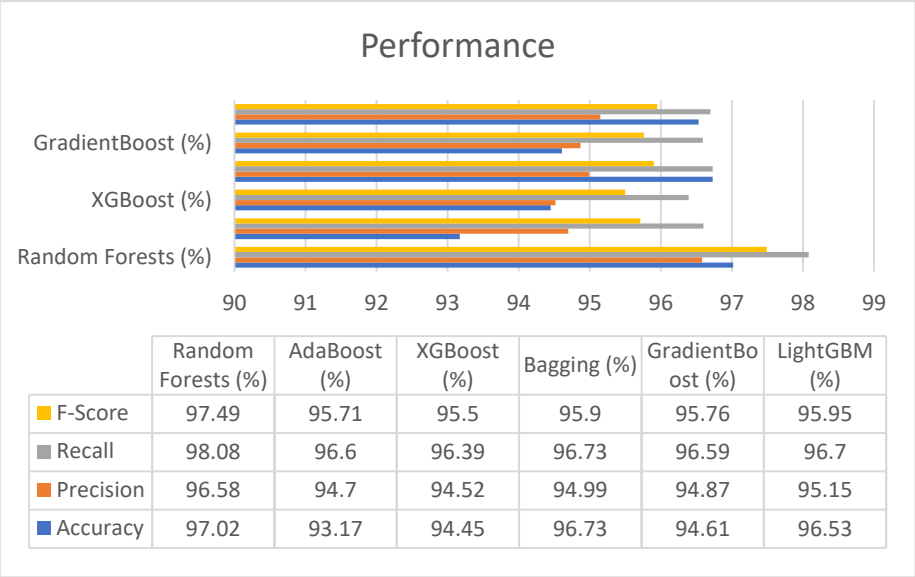
The propose model is made up of three major steps: Data Collection, Pre-Processing, and Ensemble Model. Dataset Collection: In this step, we take the dataset as input and identify features. Pre-Processing : This step includes placing URL instances in two categories as legitimate and phishing URLs. Ensemble 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



	Random Forests (%)	AdaBoost (%)	XGBoost (%)	Bagging (%)	GradientBoost (%)	LightGBM (%)
F-Score	97.49	95.71	95.5	95.9	95.76	95.95
Recall	98.08	96.6	96.39	96.73	96.59	96.7
Precision	96.58	94.7	94.52	94.99	94.87	95.15
Accuracy	97.02	93.17	94.45	96.73	94.61	96.53

Fig. 3.1- Comparison of accuracy metrics

Although they have shown good performance, they 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

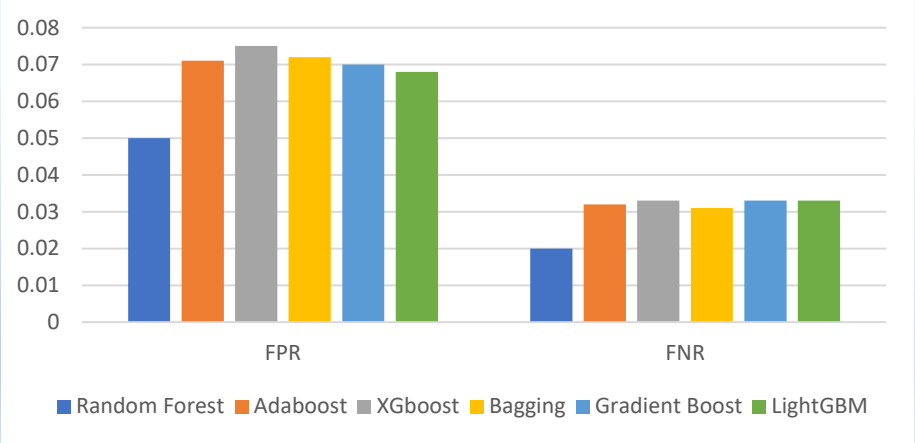


Fig. 3.2 - PNR and FNR rates

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

Classifier Name		**Without Optimization**		**With GA Optimization**	
Random Forests	Avg .	97.02 %		96.74 %	
	Variance	0.000		0.000	
AdaBoost	Avg .	93.17 %		93.63 %	
	Variance	0.000		0.000	
XGBoost	Avg .	94.45 %		97.01 %	
	Variance	0.000		0.000	
Bagging	Avg .	96.73 %		96.90 %	
	Variance	0.000		0.000	
GradientBoost	Avg .	94.61		97.13 %	
	Variance	0.000		0.000	
LightGBM	Avg .	96.53 %		96.42 %	
	Variance	0.000		0.000	

Fig 3.3 - Final observation result

3.2 The experiment was conducted using the Waikato Environment for Knowledge Analysis (WEKA) data mining software. As mentioned earlier, the original phishing email dataset was utilized to train and test binary classifiers, namely C4.5 and CART, along with their ensemble model. The training-testing dataset was split in a 70%-30% ratio.

The above 3.3 table presents the accuracy of the models. The ensemble of C4.5 and CART outperformed the individual classifiers, with C4.5 achieving an accuracy of 98.83% and CART achieving an accuracy of 98.95%. The ensemble model achieved an impressive accuracy of 99.11%, demonstrating its capability and robustness in accurately classifying phishing email data. The obtained results are visually represented in a comparative graph, as shown in the accompanying figure.

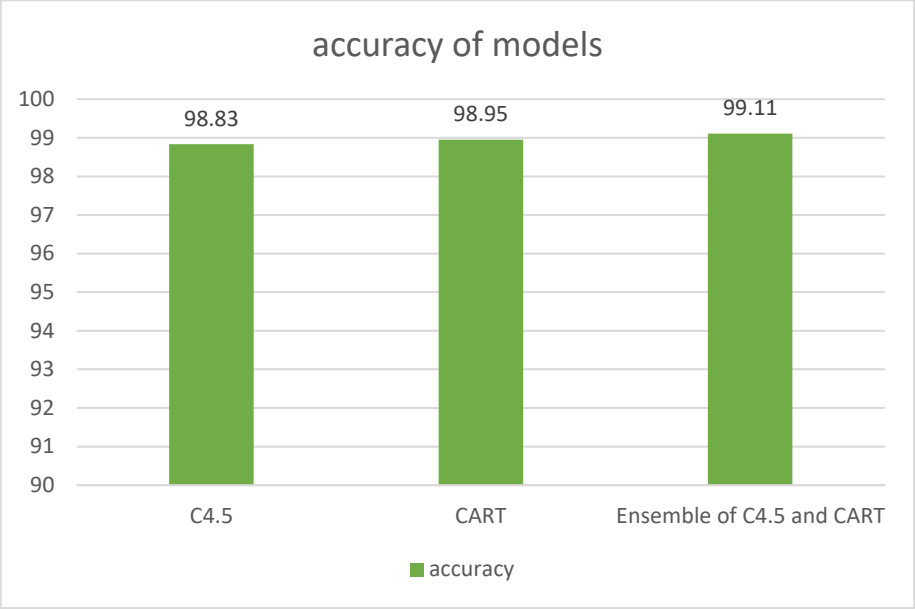


Fig .3.4 - Accuracy of models

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

Accuracy Metrics	**ANN+RFC**	**K-NN+RFC**	**C4.5+RFC**
TP Rate	0.981	0.983	0.977
FP Rate	0.041	0.038	0.054
Precision	0.968	0.97	0.958
Recall	0.981	0.983	0.977
ROC Area	0.997	0.996	0.996
F-measure	0.975	0.976	0.976
Accuracy	97.16	97.33	96.36

Fig .3.4 - Accuracy of models

4. Conclusion

Phishing is a deceptive technique where users are lured into fake websites or emails to steal their private data. Detecting phishing websites is crucial for protecting online transactions and preventing financial losses and data theft. In this context, we propose the use of ensemble learners to enhance phishing prevention and counteract such attacks. Our approach utilizes an ensemble classifier model that leverages a publicly available dataset for intelligent and automated identification of phishing websites.

Ensemble models significantly improve classifier performance in terms of accuracy, F-measure, and ROC area. Experimental results demonstrate that our model achieves a remarkable phishing page detection accuracy of 99.11%. By employing a suitable combination of ensemble classifiers, our model proves effective in identifying phishing webpages.

In future studies, we intend to incorporate feature selection methods into our model to reduce dependency on webpage content. Additionally, we plan to explore the implementation of deep learning techniques for phishing website detection. Furthermore, there is a need for further research to address phishing attacks targeting mobile devices. Given the widespread use of smartphones, they have become a prime target for attackers to carry out phishing attacks. Mobile phone users often prefer accessing their emails immediately, making it crucial to identify new features and develop efficient machine learning algorithms capable of detecting phishing attacks occurring on mobile devices.

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.