Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 1 (2025)

Enhancing user protection by identifying phishing attacks using Explainable AI

Rohan Kumar Sinha[1], Prof. Pradeeep Kumar[2]

Indian Institute of Management Lucknow[1]
Indian Institute of Management Lucknow[2]

Abstract:- In today's world, artificial intelligence (AI) is solving complex problems for various domains, and cyber security is one of them. A phishing attack is one of the most important cyber security attacks. Researchers have found that AI can detect phishing attacks much better than humans but do not give the reasons that were important in detecting it, which could be very useful to domain experts and users. So, we are using explainable artificial intelligence (XAI) to understand the reasons for Artificial Intelligence-based phishing attack detection. We have used a standard phishing attack dataset from Kaggle and applied four different AI techniques to detect the phishing attack and then applied SHAP on it to understand the reasons for Artificial Intelligence-based phishing attack detection.

Keywords: Explainable AI, Phishing Attack, Feature Identification

Introduction and Literature Review:

The domain of artificial intelligence (AI) has seen exponential growth, exerting a profound impact on several sectors and instigating fundamental changes in conventional approaches. Cybersecurity is one of the sectors which has been significantly affected by artificial intelligence. Researchers have found that AI techniques give good accuracy in detecting phishing attacks [1,12]. For a better understanding, let us look at the AI in cyber security.

The basic understanding of cyber security is that it is the process of giving security or protecting digital assets from cyber-attacks. Researchers have given a formal definition as "cyber security can be defined as the protection of cyberspace itself, the electronic information, the ICTs that support cyberspace, and the users of cyberspace in their personal, societal and national capacity, including any of their interests, either tangible or intangible, that are vulnerable to attacks originating in cyberspace"[2]. There are many types of cyber-attacks. Phishing is one of the most popular.

Phishing is a type of Cyberattack that involves tricking individuals. A formal definition is “Phishing is a scalable act of deception whereby impersonation is used to obtain information from a target”[3]. It is one of the contributors to “various cyber incidents such as data breaches and ransomware attacks, financial frauds, and denial of service attacks”[4]. So, it becomes very important to identify phishing attacks.
The identification/detection of phishing attacks is a crucial task. “The conventional approaches for phishing attack detection give low accuracy and can recognize only about 20% of phishing attacks. Machine learning approaches give good outcomes for phishing detection” [5]. There are several AI-based phishing detection approaches, “among which deep learning algorithms provided promising results” [6]. However, these AI techniques do not give explanations which

will be helpful to users in identifying phishing attacks. So, XAI becomes important in identifying phishing attacks.

Explainable AI will help users in detecting a phishing attack by identifying the various features of the phishing attack. As AI techniques give good accuracy in detecting phishing attacks, XAI will help users identify the correct features responsible for the phishing attacks. However, there are limited studies on using explainable AI in identifying phishing attacks [7,8,9]. So, it is important to study XAI's application to identify features responsible for phishing attacks. The research objective and research questions are given below:

Research Objective: Utilizing the powerful Explainable Artificial Intelligence technique to understand the reasons of Artificial Intelligence based phishing attack detections. As Artificial Intelligence techniques give good accuracy in detecting phishing attacks, understanding the reasons/ features used by Artificial Intelligence would be useful to domain experts and users.

Research Question 1: Using Explainable Artificial Intelligence, find out the top n important features that has the highest contribution in Artificial Intelligence based phishing attack detection?

Research Question 2: Using Explainable Artificial Intelligence, find out whether a particular feature have significant contribution in Artificial Intelligence based phishing attack detection?

Research Question 3: Using Explainable Artificial Intelligence, find out the features which have more contribution than a user defined threshold in Artificial Intelligence based phishing attack detection?

In this paper, we have answered these questions using a very powerful and popular explainable AI technique, SHAP (SHapley Additive exPlanations). We have taken SHAP,as it is one of the most popular XAI which gives global explanation, where as other popular XAI technique like (Local Interpretable Model-Agnostic Explanations) LIME gives local explanation.
For the purpose of analysis, we have used a standard phishing attack dataset from Kaggle and applied four different AI techniques to detect the phishing attack. Based on their performance, we selected an AI technique and applied SHAP to identify the features responsible for detecting phishing attacks. This can be useful for users and industry experts when detecting phishing attacks in different situations.
The remaining paper has three sections- Experiment, Results, and Conclusion.

Experiment:

We have done rigorous comparative experiments on standard phishing attack Kaggle dataset using four different AI algorithms. We evaluated the performance of these algorithms using various performance parameters. Finally, based on these performance parameters, we chose an AI algorithm and applied SHAP on it to identify features important in detecting phishing attacks.
The details of various components of the experiments are as follows. We have taken that dataset from Kaggle. This is a balanced dataset that has 5000 normal and 5000 phishing data, and it contains 50 columns [10]. We have used four powerful and popular AI algorithms: Logistic

Regression, XGBoost, Random Forest, and ANN. We have used the popular XAI algorithm-SHAP. It is a game theory-based XAI algorithm [11].

Before running the experiments, we pre-processed the data, and we had to analyze the data using various statistical tests. Based on it, we found that the feature “HttpsInHostname” has only one value for all. We deleted it as it did not add value. Similarly, we have tried to understand the relationship between various features, for which we have used places between the features. We found that there is no major correlation between any two features, so we can go ahead with all the features.

Results:

We have evaluated the performance of various AI algorithms using various parameters such as accuracy, F1 score, precision, recall, area under curve, and others. Figure 1 shows the performance of all four algorithms using the confusion matrix, and Figure 2 shows the accuracy of all four algorithms. From these matrices, we can say that all four algorithms are performing decently well. But XGBsoost has the best accuracy.

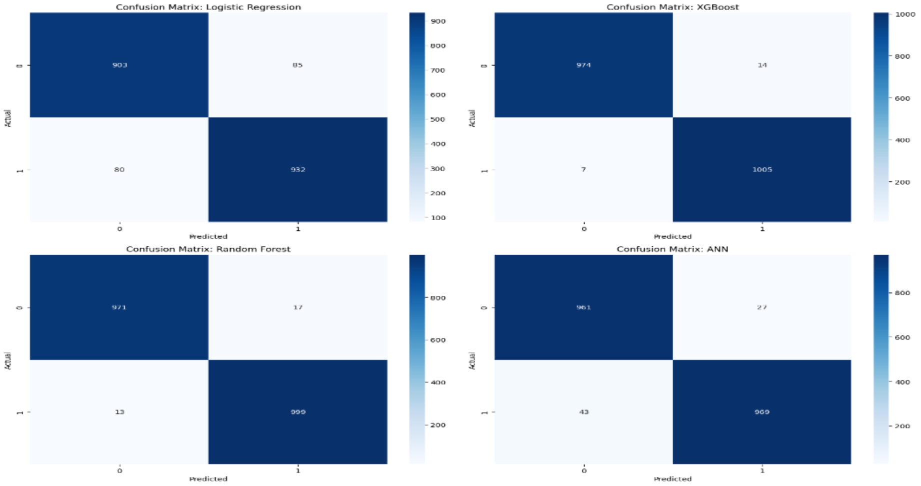

Fig. 1. Confusion Matrix of all the four algorithms

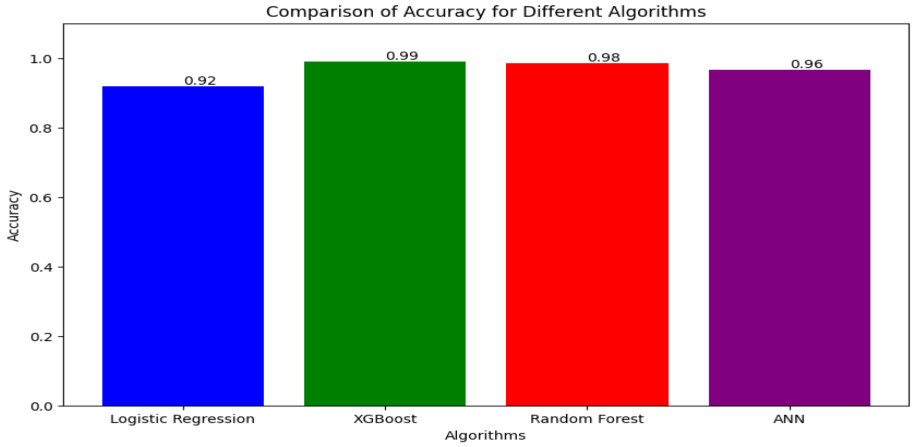

Fig. 2. Accuracy of all the four algorithms

Figure 3 shows the ROC curve for four algorithms. It also gives information about the area under the curve of all four algorithms. XGBoost and Random Forest have the perfect area under a curve score of 1. Similarly, we have analyzed it from the perspective of precision-recall. Figure 4 shows the precision-recall curve for all the four algorithms. Here, XGBoost and random forest also have the perfect area under a curve score of 1.

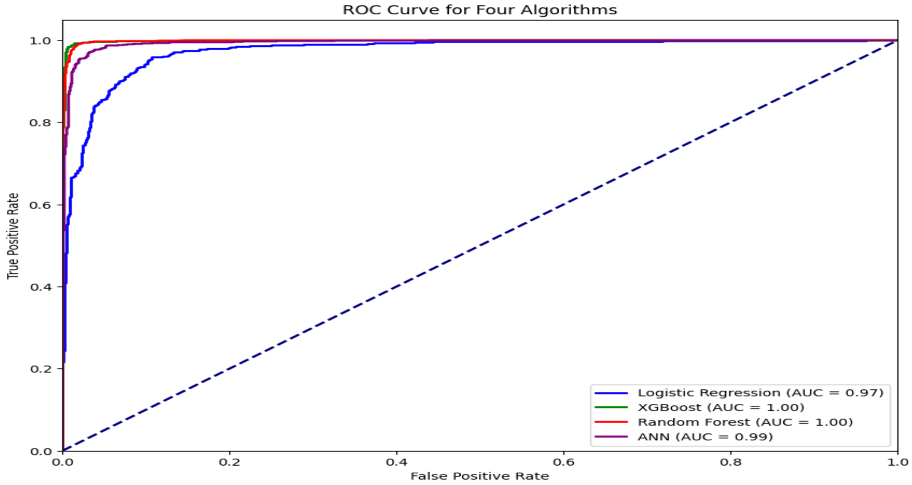

Fig. 3. ROC curves of all the four algorithms

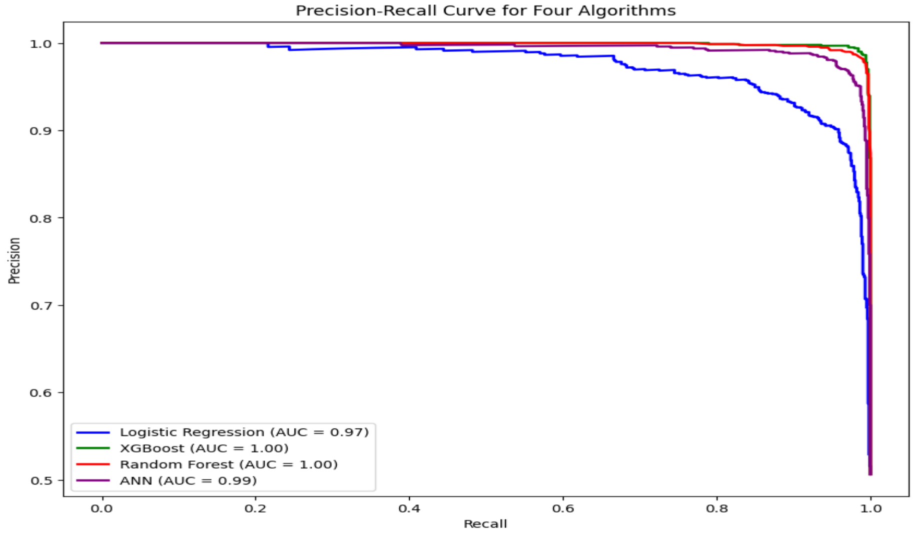

Fig. 4. Precision recall curves of all the four algorithms

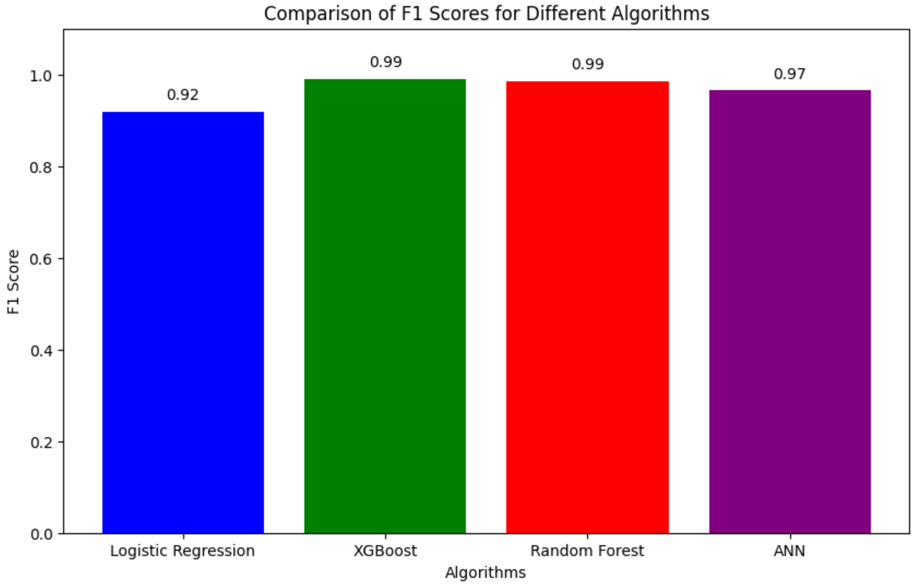

Fig. 5. F1 Scores of all four algorithms

Figure 5 compares the F1 Scores of all the algorithms. As expected, XGBoost and random forest have the best F1 scores. We can see that all four algorithms are working well. Based on all these performance matrices, we can say that it is XGBoost to be the best among them. So, we have chosen it for further analysis using the SHAP algorithm.

XGboost algorithm has high accuracy and F1 score. So, it can be said that it is good for detecting phishing attacks. However, it is difficult for users to identify which features we should use as we boost the algorithm to detect phishing attacks. To solve this problem, we have applied SHAP and identified the features that are helpful in detecting phishing attacks. Figure 6 shows all the features and their contribution to detecting phishing attacks in descending order, i.e., the first feature is the most important feature, and the last feature is the least important feature.

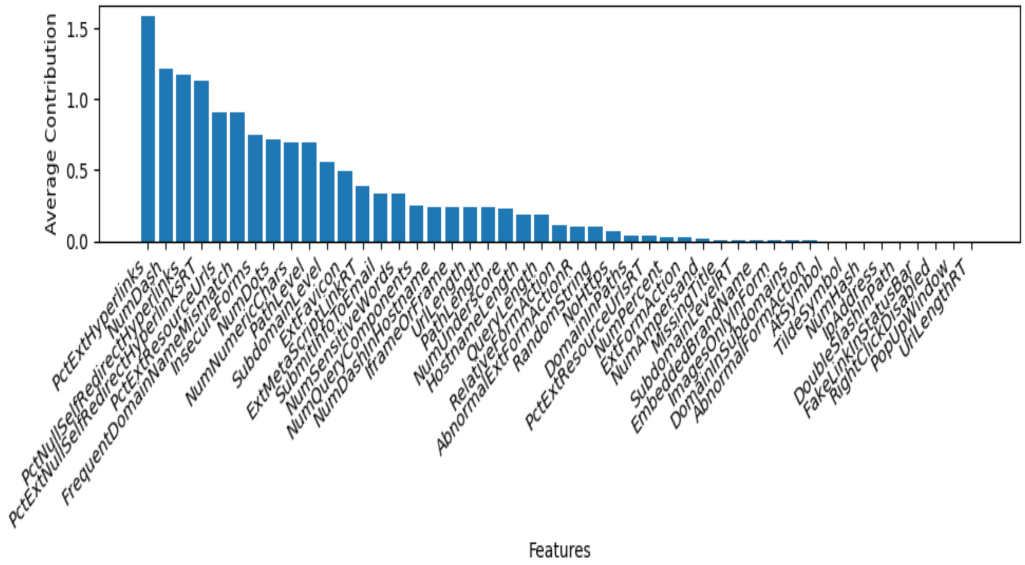

Fig. 6. Average contribution of all the features in descending order

Conclusion:

Generally, phishing attack detection is limited to feature consider important to the domain expert. Although some more features may be very helpful in identifying a phishing attack, finding such features is a difficult task from a human perspective. The combination of a good AI algorithm with high accuracy and a popular XAI algorithm like SHAP can solve this problem. It can act as an expert with high accuracy in predicting the details of the features responsible for it. So, in our case of identifying Phishing attacks, we have used XGBoost as the AI algorithm as it has high accuracy and SHAP as the XAI algorithm. Combining both helps identify features that are difficult for humans to analyze and identify and answer some important questions. Some important questions that we are answering are **(**I) find out the top n important features that has the highest contribution in Artificial Intelligence based phishing attack detection? (II) find out whether a particular feature have significant contribution in Artificial Intelligence based phishing attack detection? (III)find out the features which have more contribution than a user defined threshold in Artificial Intelligence based phishing attack detection?

Answers the first question, **(**I) find out the top n important features that has the highest contribution in Artificial Intelligence based phishing attack detection?. We are trying to find top n features, where n is user defined. For understanding purpose, we have taken n to be five. The top five features identified by the combination are PctExtResourceUrls, NumDash, PctNullSelfRedirectHyperlinks, PctExtNullSelfRedirectHyperlinksRT, and PctExtResourceUrls. Users/domain experts might not have considered these features when detecting a phishing attack. However, as these suggested features are based on an AI algorithm, which has almost 100% accuracy, these suggested features are likely to have a high impact on detecting phishing attacks.

Similarly answering the second question, find out whether a particular feature have significant contribution in Artificial Intelligence based phishing attack detection?. We have taken “MissingTitle” as the feature whose significant contribution needs to be checked. Although “MissingTitle” may look important to some user/domain expert. Our SHAP based analysis shows that has a minuscule contribution in detecting phishing attacks. This might be an useful insight for the domain experts/user to recheck the importance of MissingTitle” contribution in detecting phishing attacks.

Answering the last question, find out the features which have more contribution than a user defined threshold in Artificial Intelligence based phishing attack detection? We have taken 0.5 as threshold contribution which a user defined value. Based on 0.5 threshold, we have the important features. There are 11 features that satisfy the criteria. These features are: PctExtResourceUrls, NumDash, PctNullSelfRedirectHyperlinks, PctExtNullSelfRedirectHyperlinksRT, PctExtResourceUrls, FrequentDomainNameMismatch, InsecureForms, NumDots, NumNumericChars, PathLevel, and SubdomainLevel. these set of features can give a new perspective to look at the set of important feature for detecting phishing attacks. It will be useful to the domain expert and user to look the detection of phishing attack from a non human perspective which can have their biases.

So, this paper helps in understanding and analyzing the AI based phishing attack detection using the XAI technique. This XAI based analysis is free from human bias and can help answering research questions some of which have been discussed above.

References

1. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, *10*, 93104-93139.
2. Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *computers & security*, *38*, 97-102.
3. Lastdrager, E. E. (2014). Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, *3*, 1-10.
4. Biswas, B., Mukhopadhyay, A., Kumar, A., & Delen, D. (2024). A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks. *Decision Support Systems*, *177*, 114102.
5. Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*, 139-154.
6. Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. *Knowledge and Information Systems*, *64*(6), 1457-1500.
7. Fan, Z., Li, W., Laskey, K. B., & Chang, K. C. (2024). Investigation of phishing susceptibility with explainable artificial intelligence. *Future Internet*, *16*(1), 31.
8. Calzarossa, M. C., Giudici, P., & Zieni, R. (2023, July). Explainable Machine Learning for Bag of Words-Based Phishing Detection. In *World Conference on Explainable Artificial Intelligence* (pp. 531-543). Cham: Springer Nature Switzerland.
9. Fan, Z., Li, W., Laskey, K. B., & Chang, K. C. (2024). Investigation of phishing susceptibility with explainable artificial intelligence. *Future Internet*, *16*(1), 31.
10. https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning
11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
12. Alhogail, Areej, and Afrah Alsabih. "Applying machine learning and natural language processing to detect phishing email." Computers & Security 110 (2021): 102414.