IPRC/24/153 Science, Technology, Mathematics & Medicine

A privacy-preserving explainable AI framework for phishing URL detection

M.K.P. Madushanka[1*], and S.R. Liyanage[2]

[1]Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka
[2]Department of Software Engineering, Faculty of Computing and Technology, University of Kelaniya, Sri Lanka

In the era of artificial intelligence (AI), achieving a balance between user privacy and the interpretability of AI systems remains challenging. This paper presents a novel framework that bridges explainable artificial intelligence (XAI) with privacy-preserving (PP) methods to enhance security and trust in phishing URL detection. Phishing continues to be a major cyber threat, with the anti-phishing working group (APWG) reporting 877,536 incidents in the second quarter of 2024. The framework presents PP techniques, such as differential privacy, federated learning, and homomorphic encryption, and the techniques made available through XAI, such as SHapley Additive exPlanations (SHAP). This strategy enables the protection of these sensitive details (i.e., URLs) from exposure in both model training and prediction, while still ensuring that the inner workings of machine learning (ML) methods remain interpretable to the users. It makes use of the University of California – Irvine (UCI) Phishing Websites Dataset, which consists of 11,055 records of authentic and phishing URLs and includes 30 attributes such as IP address inclusion, URL length, and Secure Sockets Layer (SSL) certificate status. A range of ML models such as XGBoost, Random Forest, GBDT and logistic regression were developedto classify the URLs. The model accuracy, precision, and recall metrics relative to one another, were used to determine the performance level of the models. SHAP explains what, while addressing levels of explanations both types as in the global or individual level. Visual aids including force graphs and SHAP summaries would assist cybersecurity professionals in making sense of these decisions, which would ultimately enhance the interpretability of the model. This proposed framework demonstrates how the growing phishing attacks can be mitigated within reasonable limits while maintaining transparency by integrating PP techniques and XAI thereby revolutionizing the cybersecurity space.

Keywords: Explainable AI (XAI), Machine learning (ML), Phishing detection, Privacy-preserving AI

*pavithram@kdu.ac.lk
ORCiD ID: https://orcid.org/0000-0002-9861-4179