PhishFence: Integrating Explainable AI with Probabilistic Classifiers for Phishing Detection

Thomas Ha[1,4], Andrew Lee[2,4], Connor Lee[3,4], Dr. Eugene Pinsky[4]

[1]*Sharon High School, Sharon, MA 02067*
[2]*Yongsan International School of Seoul, Seoul, South Korea 04347*
[3]*Palos Verdes Peninsula High School, Rolling Hills Estates, CA 90274*
[4]*Boston University, Boston, MA 02215*

Abstract Phishing attacks remain a prevalent and financially damaging threat in digital communications, affecting both vulnerable individuals and large companies and corporations. Attacks are increasingly leveraging obfuscation techniques to deceive end-users and evade traditional spam filters. Furthermore, conventional "black box" phishing detection techniques fail to provide clear explanations to end-users regarding classification decisions, resulting in poor transparency and trust. We present PhishFence, a phishing email detection method that combines probabilistic classification models with explainable artificial intelligence (XAI) techniques to enhance both phishing detection accuracy and interpretability. We utilize six free and publicly available phishing email datasets and combine them to form a large dataset, representing a diverse sample of phishing emails. We evaluate a variety of probabilistic machine learning classifiers for accuracy and several other metrics, determining that the best models achieve a similar accuracy compared to leading phishing detection models. When applied to a completely new validation dataset and a handful of anonymized real-world sample emails, PhishFence achieved similarly high results, indicating robustness against overfitting. We further integrate SHAP, a model-agnostic XAI method, to quantify the impact of specific words on the classification and determine the primary indicator words for phishing and legitimate emails. This analysis demonstrates the effectiveness of pairing explainable AI with probabilistic machine learning classifiers for warning users of potentially harmful emails.

Keywords: Cybersecurity; phishing detection; explainable artificial intelligence; machine learning; probabilistic classification; email security.