PhishFence: Integrating Explainable AI with Probabilistic Classifiers for Phishing Detection

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

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

All authors contributed equally to this work.

Abstract Phishing attacks remain a prevalent and financially damaging threat in digital communications, affecting both vulnerable individuals and large companies. 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 propose PhishFence, a phishing email detection pipeline that pairs Sentence-BERT (SBERT) contextual text embeddings with a machine learning classifier to achieve high detection accuracy. To address the lack of transparency in traditional classifiers, we integrate SHAP (SHapley Additive exPlanations) to generate feature-level explanations. These SHAP outputs are interpreted by the OpenAI API into natural language, making classification rationale understandable to end-users. We combine six publicly available datasets into a unified dataset of approximately eighty thousand emails, representing a diverse distribution of phishing emails. We evaluate five models—logistic regression, random forest, multinomial naive Bayes, multi-layer perceptron, and BERT—in terms of accuracy, precision, recall, and F1-score. Our BERT-based model achieved an accuracy of 99.27%, outperforming the other four models. We further present a web application that enables users to submit emails for quick and reliable classification. Overall, PhishFence matches the accuracy of leading phishing detection models while providing feature-level explanations for classification decisions, demonstrating the effectiveness of pairing XAI with contextual embeddings and probabilistic classifiers in improving phishing detection.

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