**Title: Unveiling suspicious phishing attacks: enhancing detection with an optimal feature vectorization algorithm and supervised machine learning**

**Author Names:** Maruf A. Tamal ,Md K. Islam Md K. Islam, Touhid Bhuiyan, Touhid Bhuiyan Abdus Sattar Abdus Sattar1Nayem Uddin Prince Nayem Uddin Prince

**Abstract: Introduction:** The dynamic and sophisticated nature of phishing attacks, coupled with the relatively weak anti-phishing tools, has made phishing detection a pressing challenge. In light of this, new gaps have emerged in phishing detection, including the challenges and pitfalls of existing phishing detection techniques. To bridge these gaps, this study aims to develop a more robust, effective, sophisticated, and reliable solution for phishing detection through the optimal feature vectorization algorithm (OFVA) and supervised machine learning (SML) classifiers.

**Methods:** Initially, the OFVA was utilized to extract the 41 optimal intra-URL features from a novel large dataset comprising 2,74,446 raw URLs (134,500 phishing and 139,946 legitimate URLs). Subsequently, data cleansing, curation, and dimensionality reduction were performed to remove outliers, handle missing values, and exclude less predictive features. To identify the optimal model, the study evaluated and compared 15 SML algorithms arising from different machine learning (ML) families, including Bayesian, nearest-neighbors, decision trees, neural networks, quadratic discriminant analysis, logistic regression, bagging, boosting, random forests, and ensembles. The evaluation was performed based on various metrics such as confusion matrix, accuracy, precision, recall, F-1 score, ROC curve, and precision-recall curve analysis. Furthermore, hyperparameter tuning (using Grid-search) and k-fold cross-validation were performed to optimize the detection accuracy.

**Results and discussion:** The findings indicate that random forests (RF) outperformed the other classifiers, achieving a greater accuracy rate of 97.52%, followed by 97.50% precision, and an AUC value of 97%. Finally, a more robust and lightweight anti-phishing model was introduced, which can serve as an effective tool for security experts, practitioners, and policymakers to combat phishing attacks.

**Keywords:** Phishing detection Optimal Feature Vectorization Algorithm (OFVA) Supervised Machine Learning (SML) URL-based features Random Forest classifier Cybersecurity

https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1428013/full