A Machine Learning-Based Framework for Phishing Website and URL Detection with Feature Importance Analysis

Ahmad Ghassan Ata ,Mohammad Eyad Samada
Department of Cybersecurity
[Zarqa University], Jordan

Abstract

The increasing prevalence of web applications has been accompanied by a rise in phishing attacks that traditional rule-based and blacklist-based detection methods often fail to detect. This study proposes a unified machine learning framework for phishing website and URL detection using supervised classification models, following a consistent pipeline of feature extraction, preprocessing, training, and evaluation across two datasets. To address class imbalance, models were evaluated using the F1-score, where Neural Networks achieved the best performance on the first dataset, while several classifiers showed near-perfect results on the second dataset. Given the negligible performance differences and the need for interpretability, Random Forest was selected for feature importance analysis, which revealed that a small subset of discriminative features plays a critical role in effective phishing detection, highlighting the value of combining robust evaluation with explainable machine learning techniques.

Keywords

Machine Learning, Phishing Detection, Phishing Websites, Phishing URLs, Feature Importance, Random Forest, F1-score

1.Introduction

Phishing attacks represent a major cybersecurity threat, exploiting deceptive websites and malicious URLs to steal sensitive user information. Due to their constantly evolving nature, traditional detection mechanisms based on blacklists, handcrafted rules, and signatures often fail to identify newly generated or obfuscated phishing content. As a result, machine learning approaches have gained attention

for their ability to learn discriminative patterns from website and URL data.

Machine learning models leverage structural, lexical, and behavioral features to differentiate phishing attempts from legitimate instances. However, many existing studies emphasize detection accuracy while overlooking experimental consistency and model interpretability, which limits trust in their outcomes. To address these challenges, this study proposes a unified machine learning framework for phishing website and URL detection, applying a consistent experimental pipeline across multiple datasets. The framework adopts the F1-score to handle class imbalance and integrates feature importance analysis to enhance interpretability, demonstrating that Random Forest provides a balanced trade-off between strong detection performance and explainability.

1.1 Contributions

The main contributions of this work are summarized as follows:

A unified machine learning pipeline for phishing website and URL detection is proposed to ensure experimental consistency.

Multiple supervised classification models are evaluated using the F1-score to address class imbalance in phishing datasets.

A comparative analysis is conducted across two datasets representing phishing websites and phishing URLs.

Feature importance analysis is performed using Random Forest to interpret model decisions and identify influential phishing-related features.

2. Related Work

Phishing detection has received significant attention in recent years due to the rapid growth of fraudulent websites and malicious URLs targeting online users. Early approaches primarily relied on blacklist-based and rule-based mechanisms; however, these techniques were limited in detecting newly generated or obfuscated phishing attacks.

Several studies have investigated the use of machine learning techniques for phishing website detection. Zamir et al. [2] evaluated multiple classifiers, including Support Vector Machines and Random Forest, using URL-based and structural features. Although their results demonstrated promising detection accuracy, the study

relied on a single dataset and primarily reported accuracy as the main evaluation metric, which limits its suitability for imbalanced phishing scenarios.

Other works focused specifically on phishing URL detection using lexical features extracted from URLs. Verma and Das [1] explored machine learning-based phishing URL detection by analyzing URL characteristics and applying traditional classifiers. While effective, this approach was restricted to URL-level analysis and did not consider phishing website detection as a complementary task.

Comparative studies have also been proposed to evaluate different machine learning algorithms for phishing detection. Mosa et al. [3] compared multiple classifiers for both phishing website and URL detection. However, variations in preprocessing pipelines, feature engineering strategies, and evaluation metrics made fair comparison challenging. Additionally, these studies primarily emphasized performance metrics without analyzing the contribution of individual features to the classification decision.

More recently, real-time phishing detection approaches have been explored. Alotaibi et al. [4] proposed a machine learning-based framework for real-time phishing URL detection using large-scale datasets. Despite achieving high detection performance, the study focused mainly on operational efficiency and did not address model interpretability or cross-scenario evaluation.

Several survey and large-scale assessment studies have further reviewed phishing detection techniques using machine learning and deep learning approaches [5]–[7]. These works highlighted key challenges such as class imbalance, dataset dependency, and lack of interpretability. In contrast, this study proposes a unified and interpretable framework evaluated consistently across two phishing scenarios.

3. Methodology

This section describes the proposed machine learning framework used for phishing website and phishing URL detection. An overview of the complete methodology is illustrated in **Figure 1**, highlighting the unified experimental pipeline applied across both datasets.

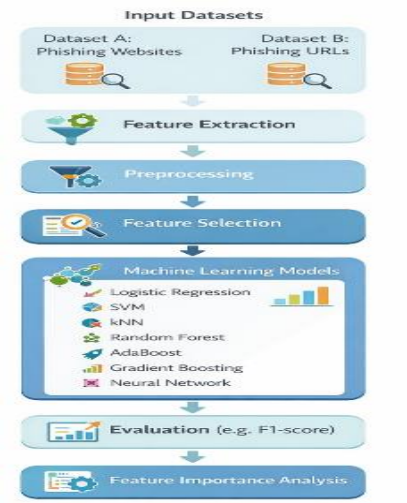

Figure 1. Overview of the proposed machine learning framework for phishing website and phishing URL detection, illustrating data preprocessing, feature selection, model training, evaluation using the F1-score, and feature importance analysis.

3.1 Datasets

The experiments were conducted using two distinct datasets. The first dataset focuses on phishing website detection, while the second dataset targets phishing URL classification. Each dataset contains both phishing and legitimate samples and was processed independently while following the same experimental pipeline to ensure fair and consistent comparison.

The datasets differ in terms of feature complexity and classification difficulty, allowing evaluation of model performance under varying phishing detection scenarios.

3.2 Feature Extraction and Preprocessing

A comprehensive set of features was extracted to capture characteristics commonly associated with phishing behavior, including lexical properties, structural patterns, and abnormal URL or website attributes. Preprocessing steps included data cleaning, normalization, and feature selection to ensure data quality and consistency.

Feature ranking techniques were applied to reduce redundancy and emphasize the most discriminative phishing-related features, improving model robustness and interpretability.

3.3 Machine Learning Models

Several supervised machine learning classifiers were evaluated, including Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes, Decision Tree, Random Forest, Gradient Boosting,

AdaBoost, and Neural Networks.

All models were trained and evaluated using identical experimental settings to ensure that observed performance differences are attributable solely to model characteristics rather than experimental variations.

3.4 Implementation Workflow

This subsection presents the implementation workflow of the proposed methodology. The workflow illustrates the sequence of data preprocessing, feature selection, model training, and evaluation steps as executed using a visual data mining environment.

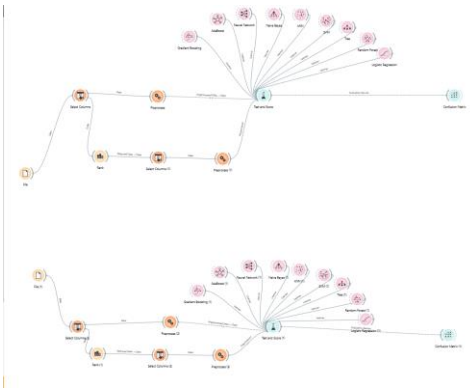

Figure 2. Implementation workflow of the proposed machine learning framework for both phishing website and phishing URL datasets.

The workflow is applied consistently across both datasets to ensure fair and reproducible evaluation.

4.Experimental Setup

This section describes the experimental setup adopted to ensure fair, reliable, and reproducible evaluation of the proposed framework. It outlines the evaluation strategy, performance metrics, and measures taken to maintain experimental consistency across all conducted experiments.

4.1 Evaluation Strategy

A stratified 10-fold cross-validation strategy was employed to ensure reliable and unbiased performance estimation while preserving the class distribution in phishing and legitimate samples.

4.2 Evaluation Metrics

The F1-score was selected as the primary evaluation metric due to its balanced consideration of precision and recall, which is particularly important in phishing detection tasks characterized by class imbalance. Additional metrics, including

accuracy and confusion matrices, were analyzed to further assess model behavior.

4.3 Experimental Consistency

All experiments were conducted using the same preprocessing steps, feature sets, and validation strategy to ensure that observed performance differences are attributable to model characteristics rather than experimental variations.

The F1-score was selected as the primary evaluation metric due to its balanced consideration of precision and recall, which is particularly important in phishing detection tasks characterized by class imbalance. Additional metrics, including accuracy and confusion matrices, were analyzed to further assess model behavior.

4.3 Experimental Consistency

All experiments were conducted using the same preprocessing steps, feature sets, and validation strategy to ensure that observed performance differences are attributable to model characteristics rather than experimental variations.

5.Results

5.1 Model Performance

The experimental results demonstrate varying levels of classification difficulty across the two datasets. On the phishing website dataset, Neural Networks achieved the highest F1-score, while Random Forest and other ensemble-based models demonstrated competitive performance with negligible differences. On the phishing URL dataset, most classifiers achieved near-perfect F1-scores, indicating high feature separability.

S.No	Model	F1-score
1	Gradient Boosting	0.949
2	AdaBoost	0.968
3	Neural Network	0.970
4	Naive Bayes	0.929
5	kNN	0.944
6	SVM	0.749
7	Decision Tree	0.960
8	Random Forest	0.968
9	Logistic Regression	0.928

Table 1 F1-score comparison of different machine learning models on the phishing website dataset.

S.No	Model	F1-score
1	**Gradient Boosting**	**1.000**
2	**AdaBoost**	**1.000**
3	**Neural Network**	**1.000**
4	**Naive Bayes**	**0.998**
5	**kNN**	**1.000**
6	**SVM**	**1.000**
7	**Decision Tree**	**0.992**
8	**Random Forest**	**1.000**
9	**Logistic Regression**	**1.000**

Table 2 F1-score comparison of different machine learning models on the phishing URL dataset.

5.2 Feature Importance Analysis

Given the negligible performance differences and the importance of model interpretability, Random Forest was selected for feature importance analysis. The analysis aims to identify the most influential features contributing to phishing website detection.

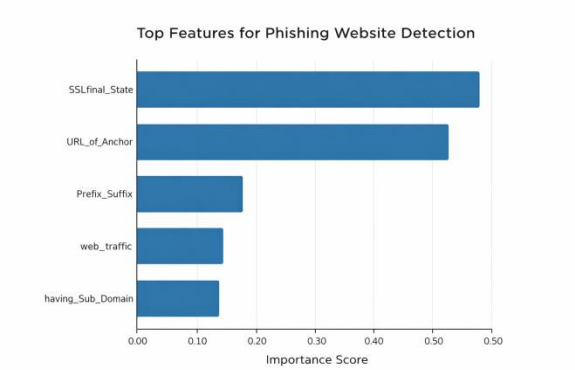

Figure 3 illustrates the relative importance of the top-ranked features identified by the Random Forest model on the phishing website dataset.

6.Discussion

The results highlight the impact of dataset characteristics on phishing detection performance. The phishing website dataset represents a more challenging classification problem, where slight performance differences among models are observable. In contrast, the phishing URL dataset exhibits high separability, allowing most classifiers to achieve optimal performance.

Although Neural Networks achieved marginally higher F1-scores on the phishing website dataset, Random Forest was selected for further analysis due to its robustness and interpretability. The feature importance analysis confirms that the extracted features effectively capture phishing-related patterns,

reinforcing the validity of the proposed feature engineering approach.

7.Conclusion and Future Work

This study presented a machine learning-based framework for phishing website and phishing URL detection using a consistent and interpretable experimental pipeline. Multiple classifiers were evaluated across two datasets, demonstrating the influence of dataset complexity on model performance.

While some models achieved slightly higher detection performance, Random Forest provided a strong balance between competitive detection performance and explainability, making it suitable for feature importance analysis. Future work may investigate cross-dataset generalization, advanced feature selection techniques, and explainable deep learning models to further enhance phishing detection capabilities.

8.References

[1] R. Verma and A. Das, “Phishing URL Detection Using Machine Learning Methods,” in Proc. 8th Int. Conf. on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, Mar. 2022, pp. 1–6, doi: 10.1109/ICACCS54159.2022.9785123.

[2] M. Zamir, H. U. Khan, B. Iqbal, and F. Ahmed, “Phishing Website Detection Using Machine Learning Algorithms,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 1, pp. 1–8, 2019, doi: 10.14569/IJACSA.2019.0100101.

[3] D. T. Mosa, M. Aburrous, and S. M. Darwish, “Machine Learning Techniques for Phishing Website and URL Detection,” Journal of Information Security and Applications, vol. 72, Art. no. 103429, 2023, doi: 10.1016/j.jisa.2023.103429.

[4] M. Alotaibi, A. Alabdulkarim, and S. Alshammari, “Real-Time Phishing URL Detection Using Machine Learning,” Engineering Proceedings, vol. 107, no. 1, pp. 108–115, 2025, doi: 10.3390/engproc2025107108.

[5] S. Marchal, J. François, R. State, and T. Engel, “PhishStorm: Detecting Phishing with Streaming Analytics,” IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458–471, 2014, doi: 10.1109/TNSM.2014.2363949.

[6] A. Rao and R. Verma, “A Comprehensive Survey on Phishing Detection Techniques,” Computer Science Review, vol. 34, pp. 100–118, 2019, doi: 10.1016/j.cosrev.2019.100189.

[7] H. Shirazi, S. Simpson, N. Sadeh, A. T. Thomas, and A. Crandall, “Large-Scale Assessment of Phishing Detection Using Machine Learning,” IEEE Security & Privacy, vol. 18, no. 2, pp. 18–27, 2020, doi: 10.1109/MSEC.2019.2952121.