Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X

Volume 4, Issue 1, 2026

CROSS DATASET GENERALIZATION OF MACHINE LEARNING MODELS FOR PHISHING URL DETECTION

Yusra Arooj[*1], **Dr. Jawaid Iqbal**[2], **Azeem Akram**[3]

[*1]*Master of Science in Computer Science, from Riphah International University, Islamabad.*
[2]*Associate Professor, Faculty of Computing, Riphah International University, Islamabad.*
[3]*Master of Science in Software Engineering, Riphah International University, Islamabad*

[*1]yusraarooj5@gmail.com, [2]jawaid.iqbal@riphah.edu.pk, [3]akramazeem947@gmail.com

DOI: https://doi.org/10.5281/zenodo.18276928

Keywords
Phishing detection, Machine learning, URL features, Cybersecurity

Article History
Received: 25 November 2025
Accepted: 01 January 2025
Published: 16 January 2026

Copyright @Author
Corresponding Author: *
Yusra Arooj

Abstract

Phishing attacks continue to pose a serious and persistent challenge in the domain of cyber security, resulting in substantial financial losses and the compromise of sensitive user information. With the increasing sophistication of phishing techniques, machine learning methods have become a widely adopted solution for phishing URL detection due to their ability to automatically learn distinguishing patterns from data. However, a notable weakness in existing research is the frequent reliance on single dataset evaluation, which doesn't accurately reflect real-world operating conditions. This study addresses this limitation by examining the cross-dataset generalization capability of machine learning models for phishing URL detection. Two publicly available phishing URL datasets containing both lexical and structural URL features are utilized. Supervised learning models, namely Logistic regression, Support Vector Machine and Random Forest are trained on one dataset and evaluated on an independent dataset. Experimental results demonstrate that the Random Forest classifier consistently outperforms other models, achieving high detection accuracy while maintaining balanced precision and recall across both evaluation settings. These findings indicate that cross-dataset evaluation provides a more realistic and reliable assessment of model robustness. Overall, the study highlights the importance of moving beyond single dataset testing and offers practical insights for developing more dependable and deployable phishing detection systems.

1. INTRODUCTION

The widespread adoption of internet-based technologies has profoundly reshaped modern society, influencing how individuals, businesses and institutions communicate, conduct financial transactions and access information. Often services from the backbone of e-commerce, digital banking, education and social interaction. While these advancements have improved efficiency and accessibility, they have simultaneously introduced serious cyber security challenges. Among these threats, phishing attacks have emerged as one of the most common and damaging forms of cybercrime.

Phishing attacks typically exploit user trust by impersonating legitimate organizations through deceptive websites or malicious URLs. Victims are often tricked into disclosing confidential information such as login credentials, banking details and personal data. Over time, phishing techniques have become increasingly sophisticated. Attackers now use automated tools to rapidly generate fraudulent URLs and frequently modify domain names to evade traditional detection mechanisms. These malicious URLs often closely resemble legitimate one's through subtle changes in spelling, domain structure or the inclusion of security indicators such as HTTPS, making them difficult for users to identify.

Conventional phishing detection techniques, including blacklist based and rule-based approaches, have been widely used to counter these threats. However, blacklist-based methods depend on previously identified malicious URLs and require continuous updates, which limits their effectiveness against newly generated or zero- day phishing attacks. Rule based systems, while useful in constrained environments, lack flexibility and struggle to adapt to the evolving nature of phishing strategies. These limitations highlight the need for more intelligent and adaptive detection mechanisms.

In response, machine learning based approaches have gained significant attention in phishing detection research. By learning patterns from historical data, machine learning models can detect previously unseen phishing URLs. These models commonly analyze lexical and structural features such as URL length, the presence of IP addresses, special characters and domain related attributes. Numerous studies have reported high detection accuracy using supervised learning algorithms including Logistic Regression, Support Vector Machines, Decision Trees and ensemble methods.

Despite these promising outcomes, a critical limitation persists in the evaluation practices adopted by most existing studies. Typically, models are trained and tested on a single dataset or on randomly split subsets derived from the same source. Although such evaluations may produce impressive accuracy figures, they do not adequately reflect real-world deployment scenarios. In practical settings, phishing URLs originate from diverse sources and evolve overtime, leading to significant variation in URL structures and characteristics. Consequently, models evaluated using a single dataset may over fit dataset specific patterns and fail to generalize to unseen data.

This lack of robust generalization assessment raises concerns regarding the real-world reliability of phishing detection systems. Cross-dataset evaluation, in which a model trained on one dataset is tested on a completely independent dataset, provides a more rigorous and realistic measure of performance. Addressing this gap, the present study focuses on evaluating the cross-dataset generalization capability of machine learning models for phishing URL detection. Two publicly available datasets are employed and multiple supervised classifiers are assessed under cross-dataset conditions. Experimental results indicate that ensemble-based methods particularly, Random Forest demonstrate superior robustness and maintain high detection accuracy even when applied to unseen data. These findings emphasize the importance of cross-dataset evaluation in developing reliable and deployable phishing detection solutions.

2. LITERATURE REVIEW

Phishing detection has attracted substantial research interest due to the increasing prevalence and sophistication of phishing attacks. Early approaches primarily relied on blacklist-based techniques however, these methods proved insufficient for detecting newly generated phishing URLs. As a result, research gradually shifted toward machine learning based solutions that analyze URL characteristics to distinguish phishing websites from legitimate one's.

Ma et al. [1] proposed one of the earliest machine learning approaches for malicious website detection using lexical features extracted from URLs. Their study demonstrated that machine learning models could outperform traditional blacklist-based systems by identifying previously unseen malicious URLs. Nevertheless, the evaluation was limited to a single dataset, restricting insights into real-world generalization.

Whittaker et al. [2] introduced a largescale phishing detection system that utilized URL based features and machine learning classifiers to identify phishing campaigns automatically. The framework emphasized scalability and achieved high detection accuracy. However, the reliance on data from a single source raised concerns regarding the robustness of the model when deployed in diverse environments.

Marchal et al. [3] presented a lightweight phishing detection approach based solely on URL analysis, avoiding webpage content inspection and third-party services. While this method reduced computational overhead and latency, it did not examine performance across multiple datasets, leaving generalization capability unexplored.

Aburrous et al. [4] proposed a hybrid phishing detection framework that combined multiple feature sets with different classifiers to improve accuracy. Although the approach yielded improved results, the evaluation remained confined to a single dataset, limiting conclusions about model robustness. However, the assessment was limited to a single dataset and the potential impact of dataset variability was not explored.

More recently, deep learning techniques have been applied to phishing detection tasks. In [5], deep learning models were employed to automatically learn complex patterns from phishing website data, achieving higher detection accuracy than traditional machine learning approaches. Still, the increased computational complexity and lack of cross-dataset assessment limited the practical connection of the proposed approach.

A study [6] delved into phishing website discovery using deep literacy infrastructures and reported high discovery rates. Although the results were emotional, the study primarily concentrated on model delicacy within a controlled experimental setup and didn't assess robustness across different datasets. Also, a study [7] proposed a mongrel frame for automated phishing discovery by combining multiple machine literacy ways. While the frame bettered bracket performance, assessment was again limited to a single dataset.

Other studies, similar to [8], emphasized ensemble literacy ways to enhance phishing discovery delicacy. These styles validated bettered stability and performance; still, they frequently reckoned on dataset-specific patterns. Research presented in [9] stressed the significance of point selection in phishing discovery but didn't consider conception across datasets. Also, [10] explored URL grounded phishing discovery using statistical literacy approaches but estimated performance within a single dataset environment.

Overall, being literate validates that machine literacy and deep literacy ways can effectively descry phishing URLs. Still, a common limitation across most studies is the reliance on single dataset assessment strategies. Veritably many workshops probe how well models generalize when applied to independent datasets. This gap motivates the present study, which totally assesses phishing detection models using cross-dataset trials to give a more realistic assessment of model robustness and real-world connection.

3. RESEARCH PROBLEM

Phishing website detection has been widely investigated using machine learning and deep learning techniques, with many studies reporting high detection accuracy under controlled experimental settings. Despite these encouraging results, a careful analysis of existing literature review reveals several fundamental limitations that significantly reduce the practical applicability of these approaches in real world environments. One of the most critical and recurring issues identified across prior research is the lack of robust evaluation strategies that account for dataset diversity and the continuously evolving nature of phishing attacks.

The majority of existing phishing detection studies assess their proposed models using a single dataset or random train test splits derived from the same data source. Although this evaluation strategy simplifies experimentation and often results in high accuracy, it fails to represent realistic deployment scenarios. In practice, phishing URLs vary considerably across datasets due to differences in data collection time frames, sources, geographical regions and attack methodologies adopted by adversaries. As a result, models trained and evaluated on a single dataset may exhibit inflated performance by learning dataset specific patterns rather than capturing generalized phishing characteristics.

Another important challenge highlighted in the literature is the extensive reliance on complex deep learning models without sufficient analysis of their generalization capability. While deep learning approaches can achieve impressive performance, they typically require large training datasets and substantial computational resources. Furthermore, their effectiveness is rarely evaluated on entirely unseen datasets, raising concerns regarding over fitting and real world deployability. Lightweight machine learning models, although computationally efficient, also suffer from similar limitations when evaluated under narrowly defined experimental conditions.

Based on these observations, the central research problem addressed in this study is formulated as follows:

How effectively do machine learning models for phishing URL detection generalize when trained on one dataset and evaluated on a different, independent dataset that reflects real world conditions?

This research problem emphasizes the necessity of evaluating phishing detection models beyond conventional single dataset testing. Addressing this issue is essential for developing reliable, deployable and future ready phishing detection systems capable of operating effectively in dynamic cyber security environments.

4. PROPOSED SOLUTION

To address the identified research problem, this study proposes a structured machine learning based framework designed to evaluate the generalization capability of phishing URL detection models through cross-dataset experimentation. The proposed framework focuses on robustness, reproducibility and practical applicability while maintaining computational efficiency.

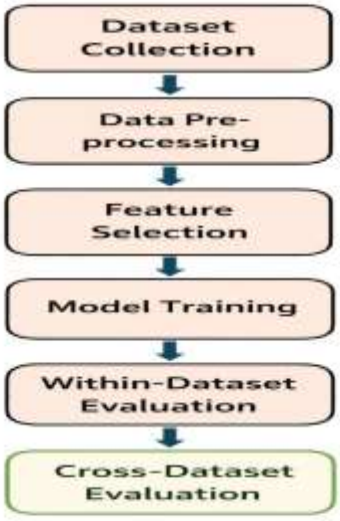

Figure 1. Overall Phishing Detection Framework

The overall workflow of the proposed phishing detection framework is illustrated in figure 1. The process begins with the selection of publicly available phishing URL datasets, followed by data preprocessing and feature extraction. The extracted feature set is then used to train supervised machine learning classifiers. Finally, model performance is evaluated using both within dataset and cross-dataset testing strategies to assess generalization capability under realistic conditions.

4.1 Dataset Selection and Description

Dataset	UCI Phishing Websites Dataset	Kaggle Phishing URLs Dataset
Number of URLs	11,055	˜10,000
Time Period	Not specified	2022-2024
Feature Type	URL Lexical and Structural features	URL Lexical and Structural features
Class Labels	Phishing (-1) Legitimate (1)	Phishing (-1) Legitimate (1)

Table 1. Description of Phishing URL Dataset

Two publicly available phishing URL datasets are utilized to ensure dataset diversity and independence. The first dataset is used for model training and within dataset evaluation, while the second dataset is exclusively employed for cross-dataset testing. Both datasets consist of URL based features that capture lexical and structural characteristics commonly associated with phishing behavior.

The datasets include attributes such as URL length, presence of IP addresses, use of special characters, domain registration information and security indicators. Each URL is labeled as phishing (-1) or legitimate (1), enabling supervised learning. The use of two independent datasets ensures that the evaluation reflects real world deployment scenarios rather than performance limited to a single dataset.

4.2 Data Preprocessing and Point Handling

Data preprocessing plays a crucial role in ensuring model reliability and consistency across datasets. This stage involves cleaning missing values, standardizing feature formats and ensuring compatibility between datasets to enable fair and meaningful evaluation. The ARFF datasets are first converted into a structured irregular format using Python libraries. Byte-decoded class markers are decrypted and converted into integer values to insure comity with machine literacy algorithms.

Missing values and inconsistencies are examined, and all features are retained in their original categorical or numerical form to save dataset integrity. Point normalization is applied where necessary to ameliorate model confluence, particularly for distance-grounded classifiers. The final dataset is resolved into point vectors (X) and target markers (y) previous to model training.

4.3 Machine Learning Models and Methodology

This study employs three extensively used supervised machine learning classifiers to estimate phishing discovery performance.

1. Logistic Regression (LR): A direct birth classifier used to model the probability of phishing URLs grounded on point benefactions.
2. Support Vector Machine (SVM): A periphery-grounded classifier capable of handling high-dimensional point spaces.
3. Random Forest (RF): An ensemble literacy system that combines multiple decision trees to ameliorate robustness and reduce overfitting.

Each model is trained using the same point set, and hyperparameters are named to balance performance and computational cost. The models are originally estimated using within- dataset testing to establish birth performance before cross-dataset evaluation is conducted.

4.4 Evaluation Strategy and Cross-Dataset Testing

To assess conception capability, the trained models are estimated on an entirely independent dataset not seen during training. This cross-dataset assessment strategy offers a realistic assessment of model robustness and stability. Model performance is evaluated using standard performance metrics, including accuracy, precision, recall and F1 score.

The experimental results show that while all evaluated models perform well when tested on data from the same dataset, ensemble-based methods demonstrate noticeably stronger generalization when applied to unseen data. In particular, the Random Forest classifier consistently outperforms the other models under cross-dataset evaluation. It achieves an accuracy of approximately 98% on an independent dataset, indicating a high degree of robustness and dependable performance in realistic deployment scenarios.

3. RESULTS AND ANALYSIS

This section presents a comprehensive evaluation of the proposal phishing detection framework. The experimental analysis is divided into two phases: within dataset assessment and cross-dataset generalization testing. The primary objective is to evaluate not only classification accuracy but also the robustness and generalization capability of machine learning models under realistic operational conditions.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.95	0.95	0.94	0.94
Support Vector Machine (SVM)	0.96	0.96	0.96	0.96
Random Forest	0.98	0.98	0.98	0.98

Table 2. Performance Comparison of Machine Learning Models

A.Experimental Setup

All experiments were conducted using Google Collaboration with Python based machine learning libraries, including Scikit-learn, NumPy and Pandas. The Random Forest classifier was selected as the primary model due to its ensemble learning capability and resistance to overfitting. For within dataset assessment, the dataset was partitioned into training and testing sets using an 80:20 split.

5.2 Within-Dataset Performance

The Random Forest model demonstrated strong baseline performance when both training and testing were performed on the same dataset, achieving high accuracy and balanced classification metrics.

5.3 Cross-Dataset Generalization Results

To assess real world applicability, the trained model was subsequently evaluated on a completely independent dataset. This experiment simulates real deployment scenarios where phishing patterns differ from those observed during training. The Random Forest model maintained high detection accuracy during cross-dataset testing, indicating effective generalization.

Metric	Value
Accuracy	98.47%
Precision	0.98
Recall	0.98
F1-score	0.98

Table.3 Cross-Dataset Generalization Performance of Random Forest Classifier

Table 3 presents the cross-dataset performance of the Random Forest classifier. The model achieved an accuracy of 98.47%, accompanied by consistently high precision, recall and F1 score values. These results confirm the robustness of the proposed approach when applied to previously unseen datasets.

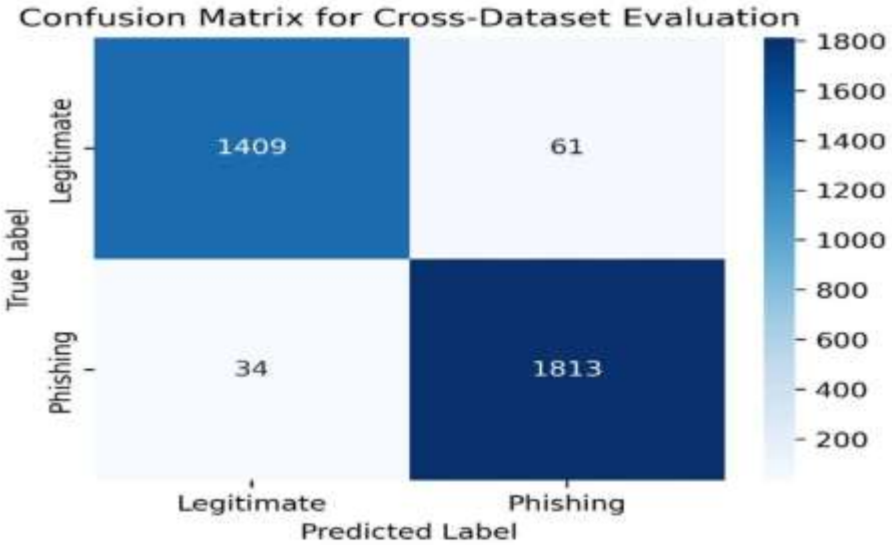

Figure.2 Confusion Matrix for Cross-Dataset Evaluation

Figure 2 illustrates the confusion matrix obtained from cross-dataset evaluation. The majority of phishing and legitimate URLs are correctly classified, demonstrating the effectiveness of the proposed model in practical deployment environments.

5.4 Comparative Analysis

The experimental findings clearly demonstrate that cross-dataset evaluation provides a more realistic and credible assessment of phishing detection models. While many prior studies report high accuracy using single dataset evaluations, their generalization capability remains uncertain. In contrast, this study confirms that well- designed machine learning models can achieve strong generalization performance without relying on complex deep learning architectures.

4. CONCLUSIONS

This study investigated the cross-dataset generalization capability of machine learning models for phishing URL detection by training models on one dataset and evaluating them on a completely independent dataset. This approach addresses a major limitation commonly observed in existing phishing detection research.

Experimental results demonstrate that the Random Forest classifier achieves high detection accuracy in both within dataset and cross-dataset scenarios, confirming its robustness and stability. The findings further indicate that URL based features, when combined with ensemble learning techniques, are effective capturing generalized phishing patterns.

The primary contribution of this research lies in its cross-dataset evaluation methodology, which provides a more realistic assessment of model performance and helps bridge the gap between academic research and real-world cyber security deployment.

Future Work

Future research may explore:

1. Incorporation of content-based and visual features
2. Assessment on larger, continuously streamlined datasets

3. Integration with real-time browser-based phishing detection systems
4. Comparison with advanced deep learning and hybrid models

5. REFERENCES

[1] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: Learning to detect malicious web sites from suspicious URLs,” in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009, pp. 1245–1254.

[2] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 2010, pp. 1–12.

[3] S. Marchal, J. Francois, R. State, and T. Engel, "Phish Storm: Detecting phishing with streaming analytics," IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458–471, Dec. 2014.

[4] M. Aburrous, M. A. Hossain, F. Thabtah, and K. Dahal, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert Systems with Applications, vol. 37, no. 12, pp. 7913–7921, 2010.

[5] S. Al-Ahmadi, M. Alqarni, and A. Alzahrani, "Phishing website detection using deep learning models," IEEE Access, vol. 8, pp. 173500–173512, 2020.

[6] M. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[7] A. M. Alshamrani, A. Chowdhary, and D. Huang, "The applicability of a hybrid framework for automated phishing detection," Computers & Security, vol. 95, p. 101869, 2020.

[8] R. Verma and K. Dyer, "On the character of phishing URLs: Accurate and robust statistical learning classifiers," in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 2015, pp. 111–122.

[9] A. Jain and B. Gupta, "Comparative analysis of features-based machine learning approaches for phishing detection," in Proceedings of the International Conference on Computing, Communication and Automation, Greater Noida, India, 2017, pp. 1–6.

[10] J. Zhang, C. Seifert, and E. H. Spafford, "Phishing detection using statistical learning methods," International Journal of Information Security, vol. 11, no. 5, pp. 357–371, 2012.