**A Hybrid Phishing Detection System: Machine Learning and NLP Based Analysis of URL Structures and Email Content**

Waleed Al-Mudaifea1, a), Haydar Alhassar1, b), Uzma Jafar1, 2, c) and Hafiz Adnan Hussain1, 2, d)

*1Faculty of Computing and Informatics, Multimedia University 63100, Cyberjaya, Selangor, Malaysia  
2Centre for Big Data and Blockchain Technologies, CoE for Advanced Cloud, Multimedia University 63100, Cyberjaya, Selangor, Malaysia*

*c) Corresponding author: uzma.jafar@mmu.edu.my*

*a) haider.a.alhassar@gmail.com*

*b) waleedmudiefa@gmail.com*

*d) hafiz.adnan@mmu.edu.my*

**Abstract.** Phishing attacks have evolved into one of the most pervasive cybersecurity threats, exploiting technical vulnerabilities and human behavior to compromise sensitive information. Traditional detection methods, such as blacklist- based filtering and heuristic rule systems, increasingly struggle to defend against sophisticated phishing techniques that employ obfuscated URLs and contextually manipulative emails. In response to these challenges, recent research has turned towards hybrid detection systems that combine machine learning (ML) and natural language processing (NLP) methodologies in datasets from Phishtank, Alexa, Kaggle, OpenPhish. This systematic literature review synthesizes two independent but complementary research studies: one focused on phishing URL analysis using ML-driven feature extraction and ensemble modeling, and the other centered on semantic analysis of phishing emails employing deep NLP architectures such as BiGRU and RoBERTa. The review evaluates their methodologies, empirical outcomes, and limitations through a comparative meta-analysis, highlighting the strengths and trade-offs of structural versus content-based detection approaches. Results demonstrate that hybrid systems, particularly those integrating lexical URL features with semantic email analysis, significantly outperform traditional single-method defenses, achieving accuracies exceeding 97% across diverse datasets. However, challenges related to dataset imbalance, computational cost, and adversarial evasion remain significant barriers to operational deployment. This review concludes by identifying key future research directions, including the need for lightweight NLP models, adversarial resilience, and multilingual phishing detection to create scalable and adaptive cybersecurity defenses.

## INTRODUCTION

The phishing threats continue to emerge as one of the most rampant and ever-evolving cybersecurity threats by preying on the vulnerability of humans by the use of fraudulent emails and replica sites to capture sensitive information. The first quarter of 2023 saw the number of phishing threats registered increase to nearly twice the amount, 1,624,144, from 888,585 in the last quarter, as reported by the Anti-Phishing Working Group (APWG) [1]. These numbers mirror the rising sophistication of the phishing operations, even though there are constant improvements to traditional security measures. Malicious URLs continue to be the main method used in phishing attacks. Bad actors continue to utilize evolving and anonymized URLs to trick users into giving up personal details, infecting their devices with malware, or running malicious scripts [2]. This evolving threat makes it crucial for organizations to detect and stop phishing attacks quickly. According to IBM, the average cost of a data breach is now around $4.88 million, highlighting the need for more proactive and effective security measures [2].

These traditional phishing defense techniques, such as blacklist-based filtering, rule-based heuristic systems, and end-user awareness training, offer weak resistance to these threats. The blacklist-based techniques depend reactively, according to pre-identified areas, and are, in essence, incapable of detecting newly synthesized phishing URLs. Heuristic systems that rely on predefined pattern matching for phishing detection consistently fall behind as attackers develop innovative techniques designed to bypass static rules. [4]. While user education remains a necessary component, it alone cannot effectively counteract the dynamic and evolving nature of social engineering tactics. Moreover, approaches that focus solely on URL analysis or email body scanning fail to account for the inherently multi-modal characteristics of modern phishing campaigns [5]. As phishing operations grow in complexity, there is a clear need for more advanced and adaptive detection mechanisms capable of interpreting contextual nuances and identifying subtle anomalies [6].

For these reasons, the existing research papers propose the deployment of a hybrid phishing detection system conceptualized based on machine learning and natural language processing (NLP) techniques to scan URL structures and email bodies [7]. Utilizing these techniques side by side, the planned system aims to enhance the accuracy of detection, adaptability, and offer improved resistance to evolving phishing schemes [8]. Based on previous studies, we synthesized a set of research questions we aim to answer from this study:

1. How effective are hybrid phishing detection systems that integrate machine learning and natural language processing compared to traditional methods?
2. What specific features and methodologies yield the highest detection accuracy in identifying phishing attempts?
3. What are the current limitations and challenges in implementing machine learning-based phishing detection systems?
4. How can natural language processing techniques enhance the analysis of email content and URL structures to improve phishing detection?

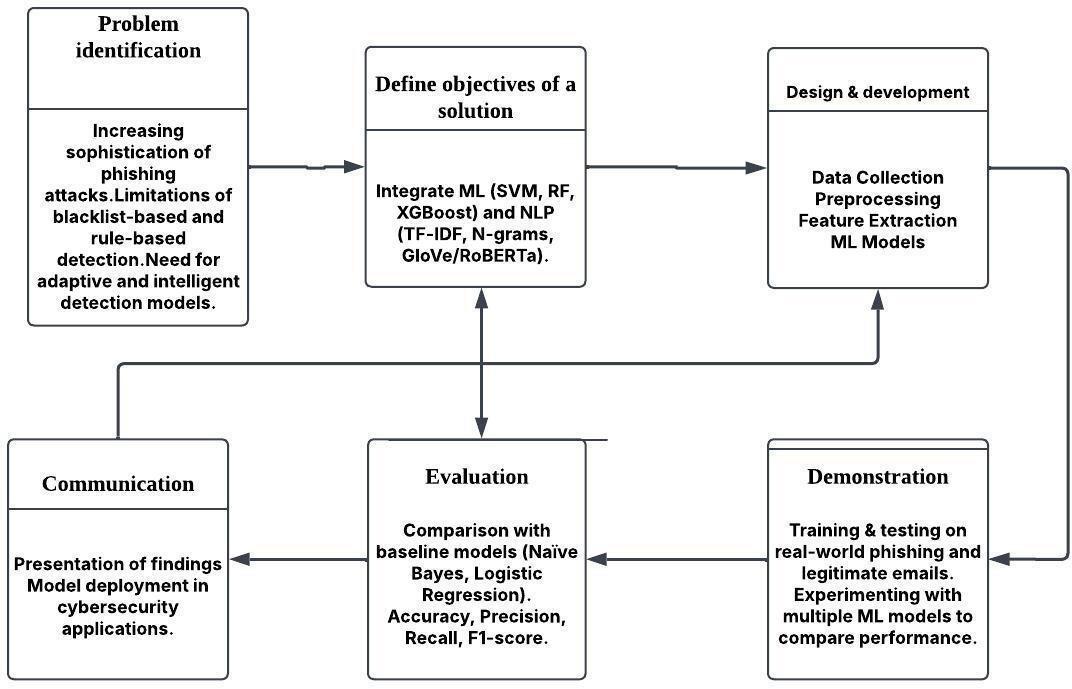
## METHODOLOGY

This research adopts the Design Science Research Methodology (DSRM) to develop a hybrid phishing detection system that integrates machine learning (ML) and natural language processing (NLP) techniques. The study begins with a comprehensive literature review across databases such as IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar, using keywords like “phishing detection,” “machine learning,” “deep learning,” and “natural language processing” to identify existing gaps in phishing prevention methods. The main objective is to build a comprehensive solution that analyzes URLs and email content to detect phishing threats more accurately. This includes evaluating the effectiveness of ML algorithms such as Random Forest, GLM, XGBoost, and ensemble methods in identifying malicious URL patterns like homograph attacks, typo-squatting, and suspicious domain structures.

NLP algorithms are coded to identify linguistic patterns, semantic anomalies, and contextual discrepancies in emails. Our super learner model is created by placing the selected NLP and ML models side by side to obtain the best possible accuracy of detection while also utilizing the URL structure in addition to text features like length, obfuscation, and suspicion indicators. The techniques of data imbalance management are incorporated to make the system train in an adequate time and generalize the model. The XGBOOST has to be tuned to give due weightage to the model. The system is exercised and tested by executing the actual phishing and legitimate examples, and the system's functionality is extensively tested by utilizing the measures of accuracy, precision, recall, and F1-score. The output is compared to the baseline model to determine the gains in terms of the detection capability, adaptability, and fault tolerance. Finally, the findings are documented for dissemination, with a view toward real-world deployment in cybersecurity applications. [Figure 1](#_k4nqufq43mky) illustrates how the Design Science Research Methodology (DSRM) is applied in developing the phishing detection system [11]. This methodology follows a structured six-phase approach, beginning with problem identification, where the need for an improved phishing detection system is established. The define objectives phase sets clear goals, including using ML and NLP techniques [12].

### Inclusion and Exclusion Criteria

For each included study, we extracted data on research objectives, methodologies, feature extraction techniques, machine learning algorithms, performance metrics, and key findings. A quality assessment was then carried out using a modified version of the Critical Appraisal Skills Programme (CASP) checklist, which evaluated each study’s methodological rigor, adequacy of sample size, relevance and appropriateness of performance metrics, and the extent of comparative analysis conducted against baseline methods. Studies were selected based on the following four filtering questions, explained in [Table 1.](#_8camw9qtzxhe)



**FIGURE 1.** DSRM methodology

**TABLE 1.** Inclusion criteria and filtering questions

|  |  |  |
| --- | --- | --- |
| **Filtering Question** | **Inclusion Criteria** | **Exclusion Criteria** |
| Does the study focus on phishing detection using computational methods? | Studies primarily focused on detecting phishing attempts using computational approaches | Studies focused solely on user education or policy measures without technical components |
| Does the study incorporate machine learning and/or natural language processing techniques? | Studies utilizing ML algorithms, NLP methods, or hybrid approaches | Studies using only traditional rule-based or blacklist approaches without ML/NLP components |
| Was the study published between 2019 and 2025? | Studies published within this timeframe to ensure relevance to current threats | Studies published before 2018, as they may not address current phishing techniques |
| Does the study provide empirical evaluation with performance metrics? | Studies with clear methodology and quantifiable results (accuracy, precision, recall, F1-score) | Conceptual papers without empirical validation |

### Data Extraction and Synthesis

This systematic literature review followed rigorous academic standards to ensure validity, reliability, and transparency. Research papers were selected based on clear inclusion and exclusion criteria, focusing on machine learning and natural language processing techniques for phishing detection. The methodological framework adhered to best practices, using a structured search across recognized academic databases and applying critical appraisal techniques to assess quality. For each study, we extracted details about methodology (features, algorithms, data sources), performance results (accuracy, precision, recall, F1-score), and limitations. We synthesized this information to address our research questions, first comparing the performance of various detection methods and then discussin URL versus email/content analysis. Finally, we consider the challenges and future directions identified in the literature.

The PRISMA flowchart in [Figure 2](#_iyp6rr59l1x0) outlines the systematic process for identifying, screening, assessing, and including studies in this review. Initially, 436 research papers were identified through database searches for their relevance to phishing detection, focusing on computational methods such as machine learning and natural language processing. However, 183 papers were removed during the Identification phase for being irrelevant or not meeting search criteria. In the Screening phase, the 253 remaining records were reviewed by examining their titles and abstracts. We excluded 127 papers that were unrelated to phishing detection or did not address the necessary computational methods. In the Eligibility phase, we performed a full-text review of the remaining 126 reports to ensure alignment with the focus of phishing detection research. 50 papers were excluded due to a lack of empirical data or failure to meet inclusion criteria.

A diagram of a paper

AI-generated content may be incorrect.

**FIGURE 2.** PRISMA flowchart

Our systematic literature study accounted for 76 papers, and they met our chosen criteria, as listed in Table 2. The papers formed the premises for the study in the present systematic literature study, providing insightful information about the performances of varied phishing detection techniques and revealing trends in the field. The following table has the year-wise distribution of the papers considered, providing insight into the spurt in interest in phishing detection during the last couple of years. The trend indicates surging interest in phishing detection techniques, i.e., sophisticated machine learning- and natural language processing-based techniques, in favor of the 2024 papers, counting 16, and those in 2023, counting 14. The result mirrors the intensifying value of these techniques in fighting modern-day cyber threats.

**TABLE 2.** Gathered paper’s yearly statistics

|  |  |
| --- | --- |
| **Database** | **Result** |
| 2024 | 16 |
| 2023 | 14 |
| 2022 | 11 |
| 2021 | 9 |
| 2020 | 12 |
| 2019 | 14 |
| **Total** | **76** |

### Data Extraction and Synthesis

The study developed two modular machine learning pipelines for phishing detection one for URLs and another for emails as shown in [Table 3.](#_hs32yoofsly7) In the URL-based system, a labeled dataset of phishing and benign URLs was cleaned and normalized, with malformed entries removed. Lexical and structural features such as URL length, subdomain size, presence of HTTPS, and symbol frequency were extracted using tools like tldextract and urllib. After standardizing features, models including Random Forest, Logistic Regression, and Ridge Classifier were trained and compared. A StackingClassifier ensemble with XGBoost as the meta-learner improved overall performance, especially phishing recall. The email-phishing pipeline ingested messages from six public datasets, applying cleaning steps like stripping headers, removing HTML, and lowercasing text. Text features were vectorized using TF-IDF on unigrams and bigrams, followed by dimensionality reduction via SVD to 300 components. SVM, Random Forest, and Logistic Regression were tuned and stacked, with XGBoost again serving as the meta-learner. Both pipelines were encapsulated into deployable scikit-learn-style objects using joblib, ensuring reproducibility, consistent transformations, and scalable deployment.

## RESULTS

To rigorously assess the efficacy of the proposed phishing URL detection framework, a series of supervised learning models were evaluated on a stratified test set using standard classification metrics: accuracy, precision, recall, and F1-score. The models tested include Logistic Regression, Ridge Classifier (a linear model formulated as a Generalized Linear Model), Random Forest, and a Super Learner ensemble composed of the aforementioned base learners with an XGBoost meta-learner. In a 60-40 testing split, the results were as shown in [Table 4.](#_blxs56pv1cn7)

**TABLE 3.** Overview of phishing detection projects with corresponding datasets, models, and feature engineering techniques

|  |  |  |  |
| --- | --- | --- | --- |
| **Project** | **Datasets** | **Models** | **Features** |
| URL Phishing | Kaggle Phishing URL dataset | Random Forest, Logistic Regression, GLM  XGBoost as Meta Learner | URL Structure Features, Unique Character-Based Heuristics, Presence of IP addresses, Keyword- Based Features, URL Obfuscation or Redirects. |
| Email Content | Kaggle, Enron, SpamAssasin,CEASE\_ 08. | SVM. Random Forest  And XGBoost as Meta Learner | TF-IDF, N-grams, Feature Embedding,LSA, Email-specific cues, Urgency words Semantic features |

**TABLE 4.** URL classification results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Super Learner | 0.9904 | 0.9914 | 0.9919 | 0.9916 |
| Random Forest | 0.9898 | 0.9903 | 0.9919 | 0.9911 |
| Logistic Regression | 0.8759 | 0.8334 | 0.9786 | 0.9002 |
| Ridge Classifier (GLM) | 0.8332 | 0.7757 | 0.9966 | 0.8723 |

The Super Learner, employing a stacking approach with passthrough enabled, demonstrated the highest overall performance across all metrics. It achieved an accuracy of 99.04%, with a precision of 99.14%, a recall of 99.19%, and an F1-score of 99.16%. These results indicate a highly balanced classifier capable of identifying phishing attempts with minimal compromise between Type I and Type II errors. As a 60-40 split, with Phishing URLs at 134,850 and Benign URLs totaling 100,945.

The confusion matrixof the Super Learner model further contextualizes this performance:

1. True Negatives (TN): 53,501 benign URLs correctly classified.
2. False Positives (FP): 439 benign URLs incorrectly labeled as phishing.
3. False Negatives (FN): 465 phishing URLs misclassified as benign.
4. True Positives (TP): 40,000 phishing URLs correctly identified.

This corresponds to a false positive rate of approximately 0.82% and a false negative rate of 1.15%, which are notably low and particularly critical in phishing detection scenarios, where both misclassifying benign content and failing to flag malicious URLs carry operational risks. While both the Super Learner and Random Forest classifiers yielded near-identical performance, the ensemble approach provided marginal yet consistent gains in both recall and F1-score, confirming the benefit of model diversity and meta-learning in complex classification tasks. In contrast, although Logistic Regression and Ridge Classifier achieved high recall scores (97.86% and 99.66%, respectively), their comparatively lower precision values resulted in reduced overall accuracy and F1-scores. This behavior indicates a higher rate of false positives, which, while preferable to false negatives in some security contexts, may lead to increased alert fatigue or unnecessary blocking in production systems.In summary, the ensemble methodology, particularly the stacking-based Super Learner, demonstrates superior robustness and generalization capability for phishing URL detection, effectively balancing detection sensitivity and specificity. These findings reinforce the viability of ensemble approaches in security-centric binary classification domains as shown in [Table 5.](#_fb5y2cxff1iz)

**TABLE 5.** Email classification results on the 60/40 split of the Kaggle phishing-email dataset (15 838 benign, 17 157 phishing in test)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Super Learner | 0.99005 | 0.99 | 0.99 | 0.99 |
| Random Forest | 0.9819 | 0.98 | 0.98 | 0.98 |
| SVM | 0.9893 | 0.99 | 0.99 | 0.99 |
| XGBoost | 0.9865 | 0.99 | 0.98 | 0.99 |

We incorporated Logistic Regression both as a standalone baseline and as the meta-learner in our stacking ensemble because of its strengths on high-dimensional, sparse feature spaces (e.g. TF-IDF): fast convergence, calibrated probabilistic outputs, and a linear decision boundary that complements tree-based and kernel methods. As a meta-learner, it effectively combines out-of-fold predictions from heterogeneous base models, improving overall calibration and minority-class recall. To ensure broad coverage of real-world email styles and phishing tactics, we trained and tested on the Kaggle phishing-email dataset, which merges six publicly available corpora (CEAS ’08, Enron, Ling, Nazario, Nigerian Fraud, SpamAssassin). This aggregation yields a large, diverse corpus of benign and malicious messages, mitigating dataset-specific biases and enabling our models to generalize across varied vocabulary, formatting, and obfuscation strategies. By leveraging this combined dataset, our Super Learner achieves

99.00 % accuracy and 99 % precision/recall on phishing, outperforming each standalone model and demonstrating the robustness of ensemble learning in email-level phishing detection.

The confusion matrix of the Super Learner model further contextualizes this performance:

1. True Negatives (TN): 15 676 benign emails correctly classified.
2. False Positives (FP): 162 benign emails incorrectly labeled as phishing.
3. False Negatives (FN): 166 phishing emails misclassified as benign.
4. True Positives (TP): 16 991 phishing emails correctly identified.

This corresponds to a false positive rate of approximately 1.02% and a false negative rate of approximately 0.97%. When it comes to phishing detection, both rates are extremely low, reducing missed attacks while preventing needless alarms. Our stacking technique decreases both error rates by about half, illustrating the resilience of model variety and meta-learning in email-level phishing categorization. In contrast, the Random Forest alone generated a 2.04% false positive rate (323/15,838) and a 1.59% false negative rate (273/17,157).

## CONCLUSION

Hybrid phishing detection systems combining machine learning and natural language processing significantly enhance defense against phishing attacks. By analyzing URL structures and the semantic content of emails/webpages, these systems address limitations of traditional single-facet methods. Our review of a URL-focused ML model and an email-focused NLP model showed each has strengths, and their integration boosts detection efficacy to over 97% accuracy. Ensemble models and deep learning techniques, especially with linguistic analysis, improve the identification of sophisticated phishing attempts. However, we noted challenges such as handling imbalanced data, reducing false positives, and resisting evasion tactics. While computational efficiency is a concern, new model optimization techniques help make real-time deployment feasible. In summary, neither URL nor content analysis alone suffices for today's phishing threats; a combination is most effective. Future efforts should aim to make these hybrid systems lightweight, adaptive, and generalizable across languages and attack variations. By integrating findings from various studies and regularly updating models with new phishing strategies, the cybersecurity field can better combat phishing as a significant threat.

**References**

1. Anti-Phishing Working Group (APWG), *Phishing Activity Trends Report, 1st Quarter 2023*, May 2023.
2. A. E.-S. El-Metwaly, et al., “Detection of Phishing URLs Based on Machine Learning and Cybersecurity,” in *2024 International Telecommunications Conference (ITC-Egypt)* (IEEE, Egypt, 2024).
3. N. Reyes-Dorta, P. Caballero-Gil, and C. Rosa-Remedios, “Detection of malicious URLs using machine learning,” Wireless Networks (2024).
4. M. Aljabri and S. Mirza, “Phishing attacks detection using machine learning and deep learning models,” in *Proceedings of the 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)* (IEEE, Riyadh, Saudi Arabia, 2022), pp. 175–180.
5. Y. A. Muhammed, “Detection of Malicious URL Using Machine Learning: A Review” (2023).
6. S. K. Garnayak, A. Kumar, S. Sahoo, B. Gouda, and V. Sharma, “Enhancing Cybersecurity: Machine Learning Techniques for Phishing URL Detection” (2024).
7. Hevapathige and K. Rathnayake, “Super Learner for Malicious URL Detection,” in *Proceedings of the 2022 2nd International Conference on Advanced Research in Computing (ICARC)* (IEEE, Belihuloya, Sri Lanka, 2022), pp. 114–119.
8. M. Karajgar, S. Sawardekar, S. Khamankar, N. Tiwari, M. Patil, V. Borate, Y. Mali, and A. Chaudhari, “Comparison of Machine Learning Models for Identifying Malicious URLs” (2024).
9. A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, “An effective phishing detection model based on character level convolutional neural network from URL,” Electronics **9**(9), 1–24 (2020).
10. D. Menaga, S. Vijay, V. Vignesh, and S. Ramalakshmi, “An Efficient Detection of Phishing Websites using Machine Learning,” in *Proceedings of the 2024 International Conference on Smart Electronics and Communication (ICOSEC)* (IEEE, 2024), pp. 1189–1194.
11. S. Raja, et al., “Structural Analysis of URL for Malicious URL Detection Using Machine Learning,” Journal of Advanced Applied Scientific Research (2023).
12. V. Ramanathan, et al., “Psychological tactics in phishing emails: A comprehensive review,” Journal of Cybersecurity (2022).