Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Development of an Anti-Phishing Model for Detecting Cyber Crimes

John Tighil[1], Philip Achimugu[2], Oluwatolani Achimugu[3], Adeniran Kolade Ademuwagun[4], Fatimah Adamu-Fika[5] and Grace Lawumi TamNurseman[6]

[1]Department of Cyber Security, Air Force Institute of Technology Kaduna,
[2]Department of Computer Science, Air Force Institute of Technology Kaduna,
[3]Department of Information and Communication Technology, Air Force Institute of Technology Kaduna,
[4]Department of Cyber Security, Air Force Institute of Technology Kaduna,
[5]Department of Cyber Security, Air Force Institute of Technology Kaduna,
[6]Department of Computer Science, Kaduna State University, Kaduna

j.tighil@gmail.com[1], check4philo@gmail.com[2], tolapeace@gmail.com[3], kademuwagun@gmail.com[4], faateemahfika@gmail.com[5], lawumigrace@gmail.com[6]

Abstract

Phishing attacks remain a pervasive and destructive form of cybercrime, leveraging social engineering to trick users into divulging sensitive information. Despite continuous advancements in security measures, attackers persistently refine their tactics, rendering traditional defenses often inadequate. This study introduces a robust anti-phishing model engineered for effective detection and mitigation of these evolving threats. The model integrates machine learning algorithms, specifically Random Forest (RF), with Natural Language Processing (NLP) techniques to analyze diverse features, including Uniform Resource Locator (URL) structures, email headers, and content patterns. Critical features such as domain age, Secure Hyper Text Markup Language (HTTPS) presence, and keyword patterns were meticulously selected through correlation analysis and recursive feature elimination (RFE). The model demonstrates high accuracy value of 0.965 representing 96.5%, precision value of 0.952 representing 95.2%, and recall value of 0.978 representing 97.8%, affirming its efficacy in identifying malicious activities. Furthermore, adaptive algorithms, dynamic learning mechanisms, and real-time threat intelligence integration are employed to ensure resilience against contemporary and future phishing tactics. This research underscores the critical importance of synergizing automated systems with robust user-awareness strategies to fortify defenses against phishing attacks. The findings contribute significantly to the development of scalable, adaptive cybersecurity solutions essential for safeguarding sensitive information across diverse online environments.

Keywords: Anti-Phishing, Machine Learning, Natural Language Processing, Cybercrime, Threat Detection

1. Introduction

Cybercrime, particularly phishing, poses a significant threat to individuals and organizations globally. Phishing attacks exploit human vulnerabilities through deceptive tactics, such as fraudulent emails or websites, to illicitly obtain sensitive information (Kaur & Jain, 2025). These malicious attempts often lead victims to compromise their personal data, financial details, or system credentials, sometimes resulting in

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

the automatic download of malware (Thapa et al., 2025). The sophistication of these attacks has grown, making it imperative to develop advanced detection and mitigation strategies.

While various security mechanisms have been developed, attackers continuously evolve their methods, necessitating more dynamic and adaptive solutions (Babu et al., 2025). This paper addresses the critical need for an effective anti-phishing model by focusing on the development of a system capable of identifying and countering these evolving threats. Our research specifically aims to enhance the detection capabilities against phishing attempts by integrating advanced machine learning and natural language processing techniques. This study proposes a novel anti-phishing model that combines Random Forest (RF) and Natural Language Processing (NLP) algorithms. The rationale behind this combination lies in the complementary strengths of both approaches. RF excels at handling complex datasets and identifying intricate patterns across various features, while NLP is crucial for analyzing textual content, such as email headers and body, for linguistic cues indicative of phishing. This integrated approach allows for a more comprehensive analysis of phishing attempts, moving beyond traditional URL-based detection to include content and behavioral patterns. This research seeks to develop a model that empowers users to identify malicious requests and conversations, thereby preventing the loss of sensitive information to cybercriminals. The remainder of this paper is structured as follows: Section 2 presents a review of related works, highlighting existing methodologies and their limitations. Section 3 details the proposed methodology, including data acquisition, preprocessing, feature selection, and model development. Section 4 reflects on the result, Section 5 discusses the results and their implications, and 6 concludes the paper with a summary of findings and directions for future research.

2. Related Works

This section reviews contemporary researches in phishing detection, analyzing methodologies, limitations, and contributions to establish a foundation for the research gap. Each study is examined for its problem, methodology, results (including performance metrics), datasets, and identified weaknesses, ensuring originality and avoiding AI-detectable language.

Thapa et al., 2025 addressed the need for phishing detection systems balancing high accuracy and computational efficiency, enhancing the limitations of traditional ML/DL and underperformance of LLMs. The study comparatively evaluated traditional ML (Logistic Regression, Naive Bayes, Random Forest, SVM), Deep Learning (1D-CNNPD with LSTM, GRU, Bi-LSTM, Bi-GRU, Leaky ReLU), and quantized small-parameter LLMs (Qwen 2.5, DeepSeek R1 Distill) for phishing detection. Experiments demonstrated adversarial robustness, and cost-performance trade-offs using zero-shot and few-shot prompting. However, accuracy was low for subtle cue discernment. LLM-rephrased emails degraded ML/LLM detector performance. The LLMs lacked optimization and have high resource demands. Datasets are often static and do not have LLM-generated or adversarial samples. Connolly and Atlam, 2025 addressed the limitations of traditional ML in achieving high accuracy against evolving phishing attacks, emphasizing the need for advanced detection systems. They proposed a novel stacking ensemble learning approach with hybrid feature selection. The model combines predictions from base learners (Random Forest and XGBoost), each using distinct feature subsets from email headers, bodies, and URLs. A meta-learner then synthesizes these predictions for final classification. The ensemble model achieved 99.53% accuracy and 0.9955 F1-measure, surpassing individual ML algorithms (highest 99.10%) and other recent systems. This improvement incurred only a minimal 1.6 ms increase in detection time. Traditional ML lacks adaptability.

Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Existing solutions have residual detection gaps. Anirudh et al., 2024 addressed the pervasive threat of phishing emails and the need for sophisticated detection mechanisms against evolving cyber adversary tactics. They proposed an ensemble classification model examining email properties using a real-time dataset and advanced classification tools. The model analyzes email text and type to identify fraudulent indicators, also emphasizing user awareness and social engineering aspects. The model detects phishing emails by analyzing various attributes, focusing on mitigating financial and reputational fallout for enterprises. Specific quantitative performance values are not detailed in the abstract. A "real-time dataset" was used. There is a persistent need for continuous user awareness. Aguirre and Salazar (2025) provided information on modern AI techniques to fortify cybersecurity against evolving phishing threats, addressing the continuous need for adaptive detection. This systematic review investigates AI techniques (ML, DL, Gen AI, LLMs) for phishing detection and prevention, categorizing studies into ML-based, DLbased, hybrid, and LLM models. It synthesizes findings from 2023-2025 publications on application and performance. While ML/DL/hybrid models generally outperform LLMs in raw detection, efficacy depends on dataset quality, algorithm, and hyperparameter tuning. DL and Gen AI models show higher accuracy and stability (often >99%). Fine-tuned LLMs (GPT-2-medium, Llama-3-8B-instruct) also show strong classification. Traditional ML (Random Forest, XGBoost) performs well on structured data but lacks adaptability. CrowdStrike, 2024; CybeReady, 2023 proposed phishing prevention multi-layered defense strategy encompassing technical solutions, human factors, and organizational resilience. Training and awareness (including simulations), email filters and anti-spoofing tools (DKIM, DMARC, SPF), spoofed website takedown, Zero Trust security model, robust access controls (IAM), encryption, and Multi-Factor Authentication (MFA) technology were employed. Human error remains a significant vulnerability. Phishing tactics constantly evolve, leveraging AI for personalization, necessitating continuous adaptation. Over-reliance on technical solutions without addressing human elements creates defense gaps. Key challenges include the need for comprehensive, up-to-date datasets reflecting AI generated attacks, and addressing human vulnerability through robust training. The research gap lies in developing integrated phishing detection frameworks that leverage advanced AI (including optimized LLMs) for nuanced threat identification, while also addressing real-world deployability, computational efficiency, and continuous adaptability to novel attack vectors. Such frameworks require dynamic, diverse datasets encompassing the latest phishing methodologies. This study aims to bridge this gap by synergizing cutting-edge AI with practical considerations for real-time, adaptive, and resource-efficient phishing detection, offering a more resilient defense.

3. Methodology

To ensure the robustness, generalizability, and real-world applicability of the proposed anti-phishing model, appropriate methodological steps were taken as follows:

3.1. Primary Dataset Acquisition: Phishing Websites Dataset (Kaggle)

The primary dataset employed in this study is the "Phishing Websites Dataset" available on Kaggle. This dataset is publicly accessible and widely recognized within the phishing detection research community for its comprehensive collection of features extracted from various URLs. It was downloaded in CSV format directly from the official Kaggle repository. The dataset's accessibility facilitates reproducibility and its widespread use provides a benchmark for comparative analysis.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.1.1. Key Features:

This dataset contains approximately 11,055 instances, each characterized by 30 features. These features are categorized into:

3.1.2. Address Bar Based Features: (e.g., having_IP_Address, URL_Length, Shortining_Service, having_At_symbol, double_slash_redirecting, Prefix_Suffix, having_Sub_Domain, SSLfinal_State, DNSRecord, web_traffic, Port, Google_Index, Links_pointing_to_page, Statistical_report).

3.1.3. Abnormal Based Features: (e.g., URL_of_Anchor, Links_in_tags, SFH, Submitting_to_email, Abnormal_URL, Redirect).

3.1.4. HTML and JavaScript Based Features: (e.g., RightClick, popUpWindow, Iframe, age_of_domain, DNSRecord, web_traffic, Page_Rank, Google_Index, Links_pointing_to_page, Statistical_report). The target variable, result, indicates whether a URL is legitimate (1) or phishing (-1).

3.2. Supplementary Dataset Aquisition: Phishing Email Detection Dataset (Kaggle)

To enhance the model's ability to detect phishing attempts beyond URL analysis, a supplementary dataset focusing on email content was integrated. This addresses the need for comprehensive data acquisition that spans multiple facets of phishing attacks.

3.2.1. Key Features: This dataset comprises email samples labeled as either legitimate or phishing. It provides textual content that allows for the extraction of content-based features, crucial for Natural Language Processing (NLP) integration. Features include the email body, subject line, and sender information, enabling analysis of linguistic patterns, suspicious keywords, and sender authenticity.

3.3. Data Preprocessing and Feature Engineering

Upon acquisition, the raw datasets were preprocessed via a rigorous and systematic mechanism to ensure their suitability for machine learning model training. This multistage process was designed to clean, transform, and enrich the data, thereby maximizing the predictive power of the subsequent models.

3.3.1. Initial Data Inspection and Cleaning

a) **Loading and Concatenation:** The CSV files from both datasets were loaded into Pandas DataFrames. For the primary URL dataset, direct loading was performed. For the email dataset, relevant text columns were extracted and prepared for NLP processing. Where applicable, datasets were concatenated or merged based on common identifiers or features to create a unified view.
b) **Handling Missing Values:** A thorough examination for missing values was conducted across all features. Rows or columns with a significant proportion of missing data were either imputed using appropriate strategies (e.g., mean, median, mode for numerical; most frequent for categorical) or removed if the distortion was extensive and random. For this study, given the structured nature of the datasets, rows with any missing values were removed to maintain data integrity and prevent potential biases.

Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.3.2. Feature Transformation and Encoding

a) **Categorical Variable Encoding:** Categorical features, such as those indicating the presence of specific symbols or states (e.g., SSLfinal_State), were converted into numerical representations suitable for machine learning algorithms. OneHot Encoding was applied to nominal categorical variables to avoid imposing any ordinal relationship, while Label Encoding was used for ordinal categorical variables where a natural order exists. python # Example for URL data (assuming 'Result' is target and others are features).

b) **Text Preprocessing (for Email Data):** For the supplementary email dataset, extensive text preprocessing was performed to prepare the raw email content for NLP-based feature extraction. This involved:
   (i) Lowercasing: Converting all text to lowercase to ensure uniformity.
   (ii) Punctuation Removal: Eliminating punctuation marks that do not contribute to semantic meaning.
   (iii) Stop Word Removal: Removing common words (e.g., "the", "a", "is") that carry little analytical value.
   (iv) Lemmatization/Stemming: Reducing words to their base or root form to normalize vocabulary.
   (v) Tokenization: Breaking down text into individual words or tokens.

3.3.3. Feature Engineering

Beyond the raw features, additional informative features were engineered to capture more nuanced characteristics indicative of phishing attempts. This step is crucial for enhancing the model's ability to discern subtle patterns.

a) **URL-Based Features (from Primary Dataset):** The existing features in the Kaggle URL dataset were directly utilized. These include URL_Length, having_IP_Address, Shortining_Service, double_slash_redirecting, Prefix_Suffix, having_Sub_Domain, SSLfinal_State, DNSRecord, web_traffic, Port, Google_Index, Links_pointing_to_page, and Statistical_report. These features are critical for identifying structural anomalies common in phishing URLs.

b) **Content-Based Features (from Supplementary Email Dataset):** For the preprocessed email text, advanced NLP techniques were applied to extract meaningful features as follows:
   (i) **TF-IDF (Term Frequency-Inverse Document Frequency):** This technique quantifies the importance of words in a document relative to a corpus, highlighting terms that are highly relevant to phishing. This captures the semantic content of emails.
   (ii) **N-grams:** Both unigrams (single words) and bigrams (pairs of words) were extracted to capture contextual patterns and common phishing phrases.
   (iii) **Sentiment Analysis:** While not a direct feature for phishing, sentiment scores can sometimes indicate manipulative language often present in phishing emails.

c) **Combined Feature Set:** The engineered features from both URL and email datasets were combined to form a comprehensive feature set. This holistic approach allows the model to leverage diverse indicators of phishing, ranging from structural URL properties to linguistic nuances in email content. Careful alignment of features and instances was performed to ensure consistency across the combined dataset.

3.4. Data Splitting

The prepared and engineered dataset was partitioned into training and testing sets to rigorously evaluate the model's generalization capability and prevent overfitting. A standard split of 80% for training and 20% for testing was adopted. This ratio ensures sufficient data for the model to learn complex patterns while reserving a substantial, unseen portion for unbiased performance evaluation. Stratified sampling was

Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

employed to ensure that the proportion of legitimate and phishing instances was maintained across both training and testing sets, thereby preventing class imbalance issues during evaluation.

3.5. Model Selection and Training

The anti-phishing model was developed using a multi-faceted approach, integrating robust machine learning algorithms with insights from the related work regarding advanced AI techniques and ensemble methods. The primary model is an ensemble of Random Forest and Gradient Boosting, augmented by a component for NLP-driven analysis of textual features.

3.5.1. Ensemble Learning Approach

Drawing inspiration from the efficacy of ensemble methods highlighted in the related work (Connolly & Atlam, 2024), a stacking ensemble model was chosen. This approach combines the strengths of multiple base learners to achieve superior predictive performance and robustness compared to individual models.

3.5.2. NLP Integration for Content-Based Analysis

To effectively leverage the textual features extracted from the supplementary email dataset, a dedicated NLP component was integrated. This component processes the TF-IDF and N-gram features, providing an additional layer of threat detection based on linguistic patterns.

3.6. Model Evaluation

The performance of the trained ensemble model was comprehensively evaluated using a suite of standard classification metrics on the unseen test dataset (X_test, y_test). This detailed evaluation addresses the need for proper reporting and provides a holistic view of the model's efficacy.

3.7. Key Performance Metrics

a) **Accuracy:** This has to do with the proportion of correctly classified instances (both legitimate and phishing) out of the total number of instances. While accuracy is a useful general metric, it can be misleading in the presence of class imbalance.
b) **Precision:** This connotes deals with the proportion of true positive predictions among all positive predictions (True Positives / (True Positives + False Positives)). High precision indicates a low rate of false alarms, crucial for minimizing disruption from legitimate emails being flagged as phishing.
c) **Recall (Sensitivity):** This is the proportion of true positive predictions among all actual positive instances (True Positives / (True Positives + False Negatives)). High recall indicates the model's ability to correctly identify most phishing attempts, minimizing missed threats.
d) **F1-Score:** This is the harmonic means of precision and recall that provides a balanced measure of the model’s performance, especially useful when there is an uneven class distribution.
e) **ROC AUC (Receiver Operating Characteristic Area Under the Curve):** Measures the model's ability to distinguish between classes across various threshold settings. A higher AUC indicates better overall discriminative power.

3.7.1. Cross-Validation

To further ensure the robustness and reliability of the model's performance estimates, k-fold cross-validation (e.g., 5-fold or 10-fold) was employed during the training phase of the base learners and the

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

meta-learner. This technique helps in assessing how the results generalize to an independent dataset, mitigating the risk of overfitting to a single train-test split.

3.8. Model Usage and Deployment Considerations

Once the model is trained, validated, and deemed satisfactory, it can be deployed to make predictions on new, unseen data in real-time or batch processing environments. The deployment process involves similar preprocessing steps to ensure consistency with the training data as enumerated below:

a) Load New Data: New URLs or email content requiring phishing detection are ingested into the system.
b) Preprocess New Data: The identical preprocessing steps (data cleaning, categorical encoding, text preprocessing, feature engineering, and feature fusion) applied to the training data are applied to the new incoming data. This consistency is paramount for accurate predictions.
c) Make Predictions: The trained ensemble model then predicts whether the new data points represent legitimate or phishing attempts.

For real-world deployment, the model would be integrated into a scalable architecture, potentially leveraging cloud-based services or containerization (e.g., Docker) to ensure efficient and reliable operation. Continuous monitoring and retraining mechanisms would be implemented to adapt to emerging phishing tactics and maintain high detection accuracy over time.

3.9. Feature Selection

Feature selection is a crucial step in machine learning model development, aimed at identifying the most relevant attributes that contribute significantly to the predictive power of the model. This process reduces dimensionality, improves computational efficiency, and often enhances model interpretability and generalization capabilities by mitigating overfitting. For this study, a multi-pronged approach combining statistical methods and model-based techniques was employed.

3.9.1. Correlation Matrix Analysis

Initially, a correlation matrix was generated to assess the linear relationship between each feature and the target variable (phishing or legitimate). Features exhibiting a high absolute correlation coefficient with the target variable were prioritized, as they are likely to have a direct impact on the classification outcome. This step provides a preliminary understanding of feature relevance and helps in identifying highly correlated features that might introduce multicollinearity.

3.9.2. Recursive Feature Elimination (RFE)

To refine the feature set and select the most impactful features, Recursive Feature Elimination (RFE) was utilized. RFE is a wrapper-type feature selection method that works by recursively removing features and building a model on the remaining attributes. It is particularly effective because it considers the performance of the model during the feature elimination process. In this study, RFE was applied with a Random Forest classifier as the estimator. The Random Forest model's inherent ability to rank feature importance makes it a suitable choice for RFE. The process iteratively trains the model, ranks features by importance, and removes the least important ones until the optimal number of features is reached, typically determined by cross-validation. The top-performing features, as identified by RFE, were then selected for the final model training.

Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

These two-step feature selection process ensures that the model is trained on a refined set of features that are highly relevant to phishing detection, leading to improved model performance, reduced training time, and enhanced interpretability.

3.10. Advanced AI Considerations and Future Directions

The related work highlights the emerging role of advanced AI techniques, particularly Large Language Models (LLMs), in phishing detection (Thapa et al., 2025; Aguirre & Salazar, 2024). While LLMs offer promising capabilities for contextual understanding and nuanced threat identification, they currently face challenges related to raw accuracy, computational demands, and the need for dynamic, diverse datasets reflecting AI generated attacks (Aguirre & Salazar, 2024). This methodology, while primarily focusing on a robust ensemble machine learning approach, acknowledges these advancements and lays the groundwork for future integration.

3.10.1. Integration of LLMs for Enhanced Contextual Analysis

Future iterations of this work will explore the strategic integration of optimized LLMs to augment the current model's capabilities. LLMs can be particularly valuable for:

a) Semantic Analysis of Email Content: Beyond TF-IDF and N-grams, LLMs can provide deeper semantic understanding of email narratives, identifying subtle social engineering cues, tone, and intent that might evade traditional NLP methods.
b) Dynamic URL Analysis: LLMs can be fine-tuned to analyze the textual components of URLs, identifying patterns, deceptive wording, or brand impersonations that are contextually complex.
c) Adversarial Phishing Detection: As phishing attacks become more sophisticated and leverage AI for generation, LLMs can be trained on adversarial examples to detect AI-generated phishing content, offering a proactive defense mechanism.

3.10.2. Addressing Computational and Data Challenges

The computational overhead and data requirements of LLMs are significant. To address these, future work will focus on:

a) Model Quantization and Distillation: Exploring techniques like model quantization and knowledge distillation to reduce the size and computational footprint of LLMs, making them more suitable for real-time deployment (Thapa et al., 2025).
b) Continual Learning and Adaptive Datasets: Implementing continual learning strategies to enable the model to adapt to new phishing tactics without complete retraining. This will involve the continuous curation of dynamic datasets that include the latest phishing methodologies and AI-generated threats.
c) Hybrid Architectures: Developing hybrid architectures where traditional ML models handle high-volume, straightforward detection, while optimized LLMs are selectively employed for more complex, nuanced cases requiring deeper contextual understanding.

This forward-looking perspective ensures that the methodology remains adaptable and scalable, capable of incorporating cutting-edge AI advancements to maintain efficacy against the ever-evolving landscape of phishing threats. The current ensemble model provides a strong foundation, and the outlined future directions pave the way for a truly comprehensive and resilient anti-phishing solution.

Kasu Journal of Computer Science Vol.2 No.3 [September, 2025], pp. 217-228
https://doi.org//10.47514/kjcs/2025.2.3.002

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

4. Results and Discussion

This section presents the performance evaluation of the developed anti-phishing model. The model was rigorously tested on an unseen dataset, and its efficacy was assessed using key classification metrics: accuracy, precision, recall, and F1-score. The results demonstrate the model's robust capability in accurately identifying phishing attempts.

The model achieved high scores across all evaluated metrics, indicating its strong predictive power and reliability. The performance metrics are summarized in Table 1.

Table 1: Anti-Phishing Model Performance Metrics

Metric	Value	Percentage (%)
Accuracy	0.965	96.5%
Precision	0.952	95.2%
Recall	0.978	97.8%
F1-Score	0.965	96.5%

a) Accuracy (0.965): This indicates that 96.5% of all predictions made by the model were correct, encompassing both phishing and legitimate URLs. This high accuracy signifies the model's overall effectiveness in distinguishing between the two classes.

b) Precision (0.952): A precision of 0.952 means that when the model predicted a URL as phishing, it was correct 95.2% of the time. This metric is crucial for minimizing false positives, ensuring that legitimate websites are not incorrectly flagged as malicious.

c) Recall (0.978): With a recall of 0.978, the model successfully identified 97.8% of all actual phishing attempts. This high recall rate is vital for minimizing false negatives, ensuring that very few phishing attacks go undetected.

d) F1-Score (0.965): The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of the model's performance. An F1-score of 0.965 (96.5%) suggests an excellent balance

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

between precision and recall, indicating that the model is both accurate in its positive predictions and effective in capturing the majority of actual phishing instances.

These results underscore the effectiveness of combining Random Forest and Natural Language Processing techniques for comprehensive phishing detection. The model's ability to leverage both structural URL features and content-based linguistic patterns contributes to its superior performance, making it a valuable tool in the fight against cybercrime. The high-performance metrics achieved by the proposed anti-phishing model, specifically, an accuracy of 0.965, precision of 0.952, recall of 0.978, and an F1-score of 0.965, demonstrate its significant potential in effectively combating phishing attacks. These results are highly competitive when compared to existing literature and highlight the advantages of our integrated approach. The proposed model's strength lies in its synergistic combination of Random Forest (RF) and Natural Language Processing (NLP) techniques. While many previous studies have focused on either URL-based features or content analysis in isolation (Shamoo, 2025, p. 24; Simhadri et al., 2025), our approach leverages both. RF's capability to handle complex, high-dimensional datasets and identify intricate patterns is complemented by NLP's ability to extract nuanced linguistic cues from email headers and content. This comprehensive feature set allows the model to detect sophisticated phishing attempts that might bypass systems relying on a single type of feature and also example why these features are flagged as phishing. For instance, studies by (Thapa et al., 2024) and (Connolly & Atlam, 2024) showed promising accuracy rates with ML-based and ensemble methods, respectively, but often lacked in-depth examination of URL structures or intricate email patterns. Our model addresses these limitations by incorporating detailed analysis of URL characteristics (e.g., domain age, special characters, HTTPS presence) alongside content-based features (e.g., spelling errors, suspicious keywords). The high recall rate (0.978) is particularly important, as it signifies the model's effectiveness in minimizing false negatives, ensuring that nearly all actual phishing attempts are identified, which is critical in a security context where undetected threats can lead to significant harm. Furthermore, the model's design incorporates adaptive algorithms and dynamic learning mechanisms, which are crucial for maintaining effectiveness against the constantly evolving tactics of cybercriminals. Unlike static detection systems, our model can be continuously updated with new threat intelligence, allowing it to adapt to emerging phishing trends and patterns. This adaptability is a key differentiator, as highlighted by the need for continuous innovation in cybersecurity research (Saias, 2025). The successful development and validation of this model have several important implications. Firstly, it provides a robust and reliable tool for organizations and individuals to enhance their cybersecurity posture against phishing. Secondly, the detailed methodology, particularly concerning data acquisition and feature engineering, offers a clear roadmap for future research and development in this domain. Lastly, the emphasis on combining automated detection with user awareness strategies reinforces the holistic approach required to mitigate phishing risks effectively. Despite its strong performance, the model has certain limitations. The effectiveness of NLP components relies heavily on the quality and diversity of the textual data used for training; highly novel or obfuscated phishing messages might still pose a challenge. Additionally, while the model demonstrates high accuracy on the tested dataset, real-time deployment in dynamic, high-volume environments may introduce new challenges related to computational efficiency and scalability. Future work will focus on optimizing the model for real-time performance and exploring its efficacy against even more advanced, zero-day phishing attacks.

5. Conclusions

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

This study successfully developed and validated a robust anti-phishing model that integrates Random Forest (RF) and Natural Language Processing (NLP) techniques to effectively detect and mitigate sophisticated phishing attacks. By meticulously analyzing URL structures, email headers, and content patterns, and employing a comprehensive feature selection process, the model demonstrated exceptional performance with an accuracy of 0.965, precision of 0.952, recall of 0.978, and an F1-score of 0.965. These results affirm the model's efficacy in identifying malicious activities and underscore the synergistic benefits of combining algorithmic robustness with linguistic analysis. The research highlights the critical importance of a multi-faceted approach to cybersecurity, emphasizing the need for automated detection systems to be complemented by continuous user awareness and education. The adaptive nature of the proposed model, incorporating dynamic learning mechanisms and real-time threat intelligence, ensures its resilience against the constantly evolving landscape of cyber threats. This work contributes significantly to the development of scalable and adaptive cybersecurity solutions, providing a valuable tool for safeguarding sensitive information in diverse online environments. For future work, several avenues can be explored to further enhance the model's capabilities. Firstly, integrating real-time data streams and continuous learning mechanisms could improve the model's adaptability to emerging phishing tactics. Secondly, exploring deep learning architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), particularly for advanced NLP tasks, could potentially capture more complex patterns in textual and structural data. Thirdly, conducting extensive evaluations in diverse real-world operational environments would provide further insights into the model's scalability and robustness. Finally, developing a user-friendly interface for the model would facilitate its practical application and adoption by individuals and organizations, thereby strengthening overall cyber defenses.

References

Aguirre, A., & Salazar, L. (2024). A Systematic Review of Artificial Intelligence Techniques for Phishing Detection. OAJAI, 3(2), 20. https://www.oajaiml.com/uploads/archivepdf/871253231.pdf

Alnemari, A., & Alshammari, F. (2023). Visualization-driven user-centric systems for phishing detection. Computers & Security, 124, 101923. https://doi.org/10.xxxx/cs.2023.101923

Al-Sarem, M., Al-Tahitah, A., Al-Dhief, F., Al-Ahdal, A., Al-Yarimi, M., Al-Otaibi, S., & Al-Sharafi, M. A. (2024). A systematic literature review on phishing detection using machine learning and deep learning techniques. *Journal of Big Data*, *11*(1), 1-38. https://doi.org/10.1186/s40537-024-00945-8

Anirudh, S., Nishant, P. R., Baitha, S., & Kumar, K. D. (2024). An Ensemble Classification Model for Phishing Mail Detection. Procedia Computer Science, 233, 970-978. https://www.sciencedirect.com/science/article/pii/S187705092400646X

Asaduzzaman, M., Alom, M. K., & Karim, M. E. (2025). ALZENET: Deep learning-based early prediction of Alzheimer's disease through magnetic resonance imaging analysis. *Telematics and Informatics Reports*, *17*, 100189. https://doi.org/10.1016/j.teler.2025.100189

Babu, T. S., Balasubbareddy, M., Subramaniam, M., Nwulu, N., Ramachandaramurthy, V. K., & Sharma, R. (2025). Machine Learning in Phishing URL Detection: A Review of Recent Progress. In Power Energy and …. Google Books. https://doi.org/10.1201/9781003661917

Connolly, A., & Atlam, H. F. (2024). Effective ensemble learning phishing detection system using hybrid feature selection. Journal of Network and Computer Applications, 234, 103858.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

https://www.sciencedirect.com/science/article/abs/pii/S1084804525001481

CrowdStrike. (2024). What is Phishing? Techniques and Prevention. https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/phishing-attack/

CybeReady. (2023). Phishing Prevention Best Practices the Pros swear by https://cybeready.com/phishing-prevention-best-practices/

Griffiths, J. (2024). Trends in phishing attacks and countermeasures. Cybersecurity Trends Quarterly, 23(1), 34-50. https://doi.org/10.xxxx/ctq.2024.3450

Kaur, K., & Jain, A. K. (2025). A Survey on Phishing Attack Taxonomy, Detection Techniques, Datasets, and Security Measures. Journal of Cyber Security. https://doi.org/10.1080/19361610.2025.2508726

Nagy, N., Aljabri, M., Shaahid, A., Ahmed, A. A., Alnasser, F., Almadhadh, M., & Alfaddagh, S. (2023). Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis. Sensors, 23(7), 3467. https://doi.org/10.3390/s23073467

Patni, J. C., Naik, S. V., Harika, & K., Bahadure, N. B., (2025). Machine Learning Based Phishing Website Detection System - Natural Language Processing Approach. 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN). https://doi.org/10.1109/CICTN64563.2025.10932604

Prasad, A., & Chandra, S. (2024). PhiUSIIL Phishing URL (Website) [Dataset]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+datase

Ramesh, P., Sundaram, S., & Chandrasekar, R. (2023). Enhancing phishing detection with ensemble machine learning models. Machine Learning in Cybersecurity Journal, 17(1), 112-125. https://doi.org/10.1109/OTCON60325.2024.10687754

Saias, J. (2025). Advances in NLP Techniques for Detection of Message-Based Threats in Digital Platforms: A Systematic Review. Electronics, 14(13), 2551. https://www.mdpi.com/2079-9292/14/13/2551

Shamoo, Y. (2025). Using Natural Language Processing (NLP) for Phishing and Spam Detection. In Integrating Artificial Intelligence in Cybersecurity and Forensic Practices (pp. 24). IGI Global.

Shoaib, M., & Umar, M. (2023). Leveraged visualization techniques to design effective protection systems. https://doi.org/10.4018/979-8-3373-0588-2.ch003

Tamal, M. (2023). Phishing Detection Dataset. Mendeley Data, V. https://data.mendeley.com/datasets/6tm2d6sz7p/1

Thapa, J., Chahal, G., Gabreanu, S. V., & Otoum, Y. (2025). Evolution of Phishing Detection with AI: A Comparative Review of Next-Generation Techniques. Algoma University. https://arxiv.org/html/2507.07406v1