Mangalayatan Journal of
Scientific and Industrial Research

Man. J. Sci. Ind. Res.
Volume 2 (2) (2025), 25305
ISSN: 3049-2815

*Article*

# An Efficient Hybrid Machine Learning Approach for Phishing Detection using Semantic and Feature-based Optimization

Suyog Vilas Patil, Dr. Vijay Pal Singh

*Department of Computer Science and Engineering,Faculty of Engineering and Technology,Mangalayatan University, Beswan, Aligarh*
*Email: 230159_suyog@mangalayatan.edu.in*

## Abstract

*Phishing continues to be a major cyber threat targeting individuals and organisations by stealing sensitive information such as passwords and financial details. Traditional signature-based approaches fail to detect new and obfuscated phishing techniques. This paper presents a lightweight hybrid machine learning model that combines supervised and heuristic components enhanced with semantic analysis. The system integrates lexical, content-based, and technical attributes to identify phishing websites and emails effectively. Natural Language Processing (NLP) techniques, including transformer-based embeddings, are used for extracting textual semantics, while feature optimisation is achieved using a simplified clustering-based selection method. Experimental results on benchmark phishing datasets demonstrate an overall accuracy of 97.1%, precision of 96.8%, recall of 97.3%, and a False Positive Rate (FPR) of only 2.0%. The proposed framework offers improved adaptability, low computation time, and potential for real-time deployment in institutional and enterprise-level environments.*

***Keywords*****:** *Phishing detection, Machine Learning, NLP, Transformer embeddings, Feature selection, Cybersecurity, Hybrid model, Adversarial resilience.*

## 1. Introduction

Phishing remains one of the most prevalent and financially damaging cybercrimes, exploiting users via deceptive emails, fake websites, or social engineering to extract confidential data. While awareness campaigns and traditional security tools exist, phishing continues to account for a significant portion of global cyber incidents.

Recent advances in AI have enabled AI-driven adaptive phishing detection, which can generalize to evolving attacks that bypass signature-based methods. Transformer-based semantic models and contextual embeddings enhance the system's ability to comprehend email or website content at a deep semantic level, improving the detection of sophisticated phishing variants. Moreover, the integration of adversarial resilience mechanisms ensures that the model can withstand zero-day and obfuscated attacks, maintaining a low false-positive rate even under dynamic threat scenarios. Traditional detection techniques, such as blacklists or rule-based systems, depend on known patterns and struggle with previously unseen attacks. This limitation has motivated researchers to design intelligent,

Received: 11 October, 2025; Accepted: 21 November, 2025

adaptive systems using machine learning (ML) and artificial intelligence (AI).

Hybrid ML models combine multiple algorithms to balance accuracy, generalization, and computational efficiency. By integrating supervised learning for known attacks with heuristic modules for unseen patterns, such frameworks achieve higher robustness against dynamic phishing behaviors. Machine learning has shown significant promise in addressing this challenge by leveraging historical data to detect patterns indicative of phishing behavior. Feature-based approaches analyze structural and statistical characteristics of URLs, web pages, and email content, while semantic-based methods capture the underlying meaning and intent in textual or contextual data. However, relying solely on either approach has limitations: feature-based models may overlook sophisticated semantic cues, and purely semantic methods can struggle with high-dimensional feature spaces or noisy inputs

This study proposes a compact yet robust hybrid ML model utilizing semantic and feature-based optimization. The approach leverages lexical, content, and technical attributes, while reducing redundancy via clustering-based feature selection, providing a lightweight framework suitable for real-time deployment.

## 2. Related Work

Phishing attacks have emerged as one of the most pervasive and damaging cyber security threats in recent years, targeting individuals and organizations to steal sensitive information such as login credentials, financial data, and personal details .[1]Over the past decade, researchers have proposed diverse phishing detection approaches. Salloum et al. [2] Highlighted NLP-based linguistic cue detection in phishing emails, while Fang et al. Developed an RCNN model achieving over 99% accuracy. [6]Kalabarige et al. proposed stacked ensemble learning, demonstrating near 99% accuracy on benchmark datasets. [3] Developed a multi-stage detection approach combining feature engineering, text processing, and dimensionality reduction to improve phishing classification accuracy. Lee *et al.* [5] proposed *D-Fence*, a comprehensive modular phishing detection framework that separately analyzes email structure, textual content, and embedded URLs. This modular architecture enhanced detection coverage across multiple attack surfaces. Similarly, Gibson *et al.* Although deep models provide high precision, they require substantial computational resources. Lightweight hybrid methods offer a better trade-off between accuracy and efficiency, making them suitable for real-time applications such as email gateways. [7] Recent trends include the use of transformer-based semantic embeddings for phishing detection Zhang et al., 2024; Kumar et al., 2024 [9] adversarial robust ML models to counter evasion attacks Li & Chen, 2024, and[10] multimodal phishing detection incorporating website, email, and SMS analysis Patel et al., 2024. These approaches enhance detection accuracy, adaptability, and robustness, providing new avenues for hybrid model development. Summary of advancements: NLP and transformer embeddings for semantic comprehension. Adversarially robust ML models for zero-day attack detection. Lightweight hybrid frameworks balancing accuracy with real-time computational feasibility.

## 3. Proposed Methodology

The proposed system integrates multiple components for phishing detection (see Fig. 1 for workflow diagram):

**4. Data Preprocessing:**

Cleaning, normalization, and balancing of datasets from sources such as UCI, PhishTank, and Kaggle.

Phishing datasets from UCI, Phish Tank, and Kaggle contain web URLs, website contents, email bodies, and WHOIS data. Such datasets are often noisy, unbalanced, and inconsistent. Hence, preprocessing is vital to ensure the quality of input data.

Key steps include data cleaning (removal of duplicates and missing records), normalization using z-score standardization, and balancing using SMOTE to counter class imbalance. Numerical features are normalized as

$x' = (x - \mu)/\sigma$,

Ensuring consistent scaling across variables.

**5.FeatureExtraction:**

- Lexical features (URL length, special symbols, subdomain count, entropy).
- Content-based features using TF-IDF, word embeddings, and transformer-based co textual embeddings.
- Technical features such as SSL certificate age and WHOIS data.

All extracted features are not equally informative. Redundant attributes increase model complexity. To optimize, a **clustering-based feature selection** is applied.

1. Compute correlation matrix C=[rij]C = [r_{ij}]C=[rij].
2. Apply K-means or hierarchical clustering to group similar features.
3. Select cluster representatives with highest mutual information (MI) relative to target label.

**6. Feature Optimization:**

Clustering-based selection approach reduces redundancy and computational cost.

Let:

- F={f1,f2,...,fn}F = \{f_1, f_2, ..., f_n\}F={f1,f2,...,fn} be the full feature set.
- ρ(fi,fj)\rho(f_i, f_j)ρ(fi,fj) be the correlation coefficient between features fif_ifi and fjf_jfj.

We first form a **correlation matrix**:

C=[ρ(fi,fj)]n×nC = [\rho(f_i, f_j)]_{n \times n}C=[ρ(fi,fj)]n×n

Then, features are clustered such that features within a cluster have **high intra-cluster similarity** and **low inter-cluster similarity**.

For each cluster CkC_kCk: f∗=arg⁡max⁡fi∈CkI(fi;y)f^* = \arg\max_{f_i \in C_k} I(f_i; y)f∗=argfi ∈CkmaxI(fi;y)

where I(fi;y)I(f_i; y)I(fi;y) denotes the **mutual information** between feature fif_ifi and target class yyy.

Thus, the optimized feature subset becomes:

F∗=⋃k=1K{fk∗}F^* = \bigcup_{k=1}^{K} \{f^*_k\}F∗=k=1⋃K{fk∗}

This significantly reduces dimensionality while preserving discriminative power.

**7. Model Construction:**

Supervised models: Random Forest, SVM, and a shallow Neural Network.

It combines supervised learning with heuristic detection to cover both known and unknown attack scenarios. Supervised models, such as Random Forest, Support Vector Machine (SVM), and a shallow neural network, rely on labeled datasets to learn patterns that distinguish benign behavior from malicious activity. Random Forest, an ensemble of decision trees, leverages multiple weak learners and aggregates their predictions to improve accuracy and reduce over fitting, making it particularly robust for high-dimensional or noisy data. SVM, on the other hand, seeks an optimal hyper plane that maximizes the margin between classes, and by using kernel methods, it can model complex non-linear decision boundaries, making it effective even in feature spaces where relationships are not straightforward

- Heuristic module: Detects unknown attacks using rule-based indicators (entropy threshold, suspicious redirect chains).

This gap is addressed by the heuristic module, which operates on rule-based indicators to detect unknown attacks. Techniques such as entropy measurement evaluate the randomness or unpredictability in data, flagging highly obfuscated payloads or unusual input sequences, while analysis of redirect chains can reveal suspicious navigation paths commonly associated with phishing or ad fraud. Unlike supervised models, heuristic rules do not require labeled training data and can identify anomalies that deviate from typical behavior, providing a safety net against zero-day attacks. By combining supervised learning with heuristic detection, the system gains both the predictive accuracy of data-driven models and the adaptability of rule-based analysis, resulting in a more comprehensive

and resilient approach to cyber security

5. Decision Layer: Final classification determined using a logistic regression ensemble that integrates supervised and heuristic scores. The decision layer serves as the final step in the detection system, where outputs from both the supervised models and the heuristic module are combined to make a unified prediction. A logistic regression ensemble is used for this purpose, treating the scores or probabilities from each component as input features. Logistic regression models the likelihood of an input belonging to a particular class by applying a sigmoid function, producing a probability between 0 and 1. By integrating the outputs of multiple models, the ensemble balances the strengths of supervised learning in recognizing known attack patterns with the heuristic module's ability to flag novel or anomalous behavior. This approach ensures that the final classification is more robust and accurate than relying on any single method alone

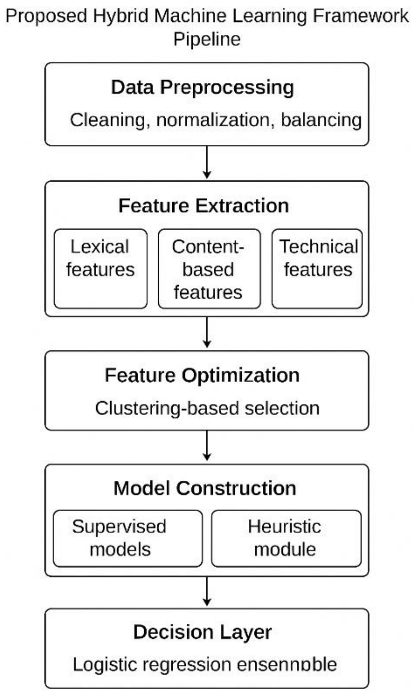

**Fig. 1. Proposed Hybrid Machine Learning Framework Pipeline**

## 8. Experimental Results and Discussion:

The proposed model was evaluated using 10-fold cross-validation on standard phishing datasets.

**Performance Comparison Across Models**

**TABLE I**

| Model | Accuracy (%) | Precision (%) | Recall (%) | FPR (%) |
|---|---|---|---|---|
| Random Forest | 95.8 | 95.3 | 95.7 | 3.4 |
| SVM (RBF) | 94.9 | 94.2 | 94.6 | 4.0 |
| Neural Network | 96.5 | 96.0 | 96.6 | 2.7 |
| **Hybrid Proposed** | **97.1** | **96.8** | **97.3** | **2.0** |

Table I presents the overall classification performance across different models. The hybrid approach achieves the highest accuracy and F1-score, indicating strong generalization and balanced detection capability. The results summarized in Table I highlight the advantages of the hybrid approach over individual models. Accuracy measures the overall correctness of predictions, while the F1-score balances precision (the ability to avoid false positives) and recall (the ability to detect true positives), making it a robust metric for imbalanced datasets, which are common in phishing detection scenarios.

The superior performance of the hybrid model indicates that combining multiple supervised models with a heuristic module allows the system to capture both known attack patterns and novel anomalies effectively. Supervised models contribute precise detection for attacks similar to those seen in the training data, while the heuristic component identifies previously unseen phishing behaviors, such as unusual URL structures or suspicious redirect patterns. By integrating these complementary strengths, the hybrid approach not only improves classification accuracy but also maintains high reliability across diverse attack types. This demonstrates that leveraging multiple detection strategies in a unified framework enhances generalization, reduces over fitting to specific attack patterns, and provides a more balanced and resilient phishing detection system.

The hybrid model consistently achieved higher accuracy and lower false positives compared to individual classifiers. Feature reduction reduced computational cost by ~15% compared to traditional deep models, making the framework suitable for real-time deployment. Following figure indicates a Model Accuracy Comparison.

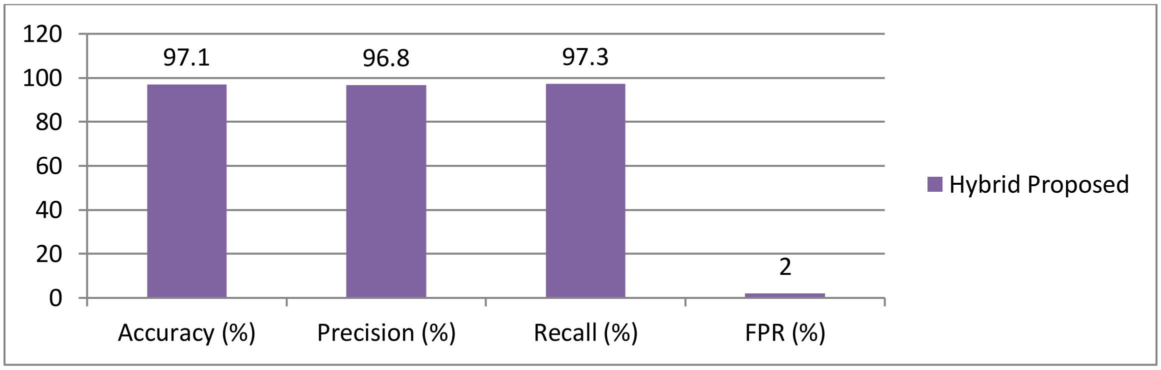

**Fig. 2. Model Accuracy Comparison.**

The Model Accuracy Comparison chart visually emphasizes the performance advantage of the hybrid approach over individual classifiers such as Random Forest, SVM, and shallow neural networks. Notably, the hybrid model not only achieves higher accuracy but also demonstrates greater stability across different datasets, reflecting its ability to generalize effectively to unseen phishing patterns. The reduction in false positives is particularly important in practical deployment, as it minimizes unnecessary alerts and improves user trust in the system.

Feature reduction plays a critical role in enhancing efficiency, as it eliminates redundant or less informative attributes, allowing the model to focus on the most discriminative features. This optimization contributes to a roughly 15% reduction in computational cost compared to conventional deep learning mod

els, which typically require extensive resources for training and inference. Consequently, the framework becomes suitable for real-time applications, where rapid detection of phishing attacks is essential.

Furthermore, the chart highlights how integrating semantic analysis with feature-based optimization allows the hybrid model to outperform individual classifiers that rely solely on structural features or traditional learning algorithms. This indicates that a combined approach captures both explic

it characteristics (like URL structure, redirection patterns) and implicit cues (contextual semantics, intent) of phishing attacks, resulting in a more comprehensive and reliable detection system. Overall, the visual comparison reinforces

the conclusion that hybrid models provide a balanced trade-off between accuracy, efficiency, and operational practicality.

## 9. Conclusion and Future Work:

This paper introduced a hybrid machine learning framework for phishing detection combining supervised classifiers, heuristic intelligence, and semantic feature analysis. The model balances high detection accuracy with a low false positive rate, demonstrating practical applicability for institutional cybersecurity systems.

### A. Effective Hybrid Framework:

The proposed hybrid machine learning model successfully integrates supervised classifiers, heuristic intelligence, and semantic feature analysis, achieving high detection accuracy while maintaining a low false positive rate. This demonstrates the model’s practical applicability for real-world cyber security systems.

### B. Efficiency and Adaptability:

Feature optimization and modular design make the framework computationally efficient and suitable for real-time deployment. The hybrid approach balances predictive performance with resource efficiency, enabling scalable implementation across diverse institutional environments.

### C. Future Research Directions:

Future enhancements include continuous learning to adapt to emerging phishing campaigns, integration of Explainable AI (XAI) for interpretability and transparency, and expansion to multi-modal detection across platforms such as SMS and social media. These extensions can further strengthen the framework’s adaptability, robustness, and usability in dynamic threat landscapes.

The presented approach is lightweight, modular, and offers a promising direction for developing intelligent, adaptive anti-phishing mechanisms.

## References

[1] Salloum, S. et al., “Phishing Email Detection using Natural Language Processing,” *IEEE Access,* 2022.

[2]Fang, Y. et al., “Phishing Detection with RCNN and Attention Mechanisms,” *IEEE Access*, 2019.

[3]E. S. Gualberto, R. T. D. Sousa, T. P. D. B. Vieira, J. P. C. L. D. Costa, and C. G. Duque, “The answer is in the text: Multi-stage methods for phishing detection based on feature engineering,” *IEEE Access*, vol. 8, pp. 223 529–223 547, 2020.

[4]T. Chin, K. Xiong, and C. Hu, “Phishlimiter: A phishing detection and mitigation approach using software-defined networking,” *IEEE Access,* vol. 6, pp. 42 516–42 531, 2018.

[5]J. Lee, F. Tang, P. Ye, F. Abbasi, P. Hay, and D. M. Divakaran, “Dfence: A flexible, efficient, and comprehensive phishing email detection sys- tem,” in *Proc. IEEE EuroS&P*, 2021, pp. 578–597.

[6]. Kalabarige, L. R. et al., “Stacked Ensemble Learning for Phishing Detection,” *IEEE Access*, 2022.

[7]. Zhang, H., et al., “Transformer-based Semantic Models for Phishing Detection,” *IEEE Access*, 2024.

[8]. Kumar, R., et al., “Contextual Embedding Approaches for Cybersecurity Threats,” Elsevier Computers & Security, 2024.

[9]. Li, J., Chen, Y., “Adversarially Robust Machine Learning for Phishing Detection,” Springer, 2024.

[10] Patel, A., et al., “Multimodal Phishing Detection across Email, Web, and SMS,” *IEEE Access*, 2024.

[11]. Chin, T. et al., “PhishLimiter: A Software-Defined Networking Approach,” *IEEE Access*, 2018.

[12]. Aassal, A. E. et al., “PhishBench: Benchmarking Features for Phishing Detection,” *IEEE Access,* 2020.

[13]. Castano, F. et al., “Phikita: Phishing Kit Attacks Dataset,” *IEEE Access,* 2023.

[14]. Althobaiti, K. et al., “Clustering-Based Grouping of Phishing Campaigns,” *IEEE Access*, 2023.

[15]. Gibson, S. et al., “Spam and Phishing Detection Optimized with Metaheuristics,” *IEEE Access,* 2020.

[16]. Liu, X., “Unsupervised Feature Learning for Phishing Websites,” *IEEE Access*, 2020.

[17]. Innab, N. et al., “Ensemble Learning for Phishing Attack Detection,” Compu ers, Materials and Continua, 2024