https://doi.org/10.59476/ilpmt2026.v2i1.794

A HYBRID MACHIN ELEARNING AND DEEP LEARNING SYSTEM FOR PHISHING EMAIL DETECTION IN WEBMAIL PLATFORMS

Elda Xhumari, Rivalda Bedini
University of Tirana, Albania

Abstract

Phishing emails continue to represent a major cybersecurity threat, exploiting social engineering techniques and deceptive language to compromise user credentials, financial data, and organizational systems. As phishing campaigns become increasingly sophisticated and automated, traditional rule-based detection methods struggle to identify new attack patterns. In response to this challenge, this research proposes a hybrid phishing email detection framework that combines classical machine learning and deep learning techniques for integration within webmail platforms. The objective of the study is not only to improve phishing detection accuracy but also to demonstrate the operational feasibility of deploying intelligent detection systems directly in a real-world email environment.

The proposed framework is trained on a multi-source dataset consisting of more than11,000 labeled emails balanced between legitimate and phishing messages. The dataset combines legitimate emails from the Enron Email Corpus with phishing samples obtained from public phishing repositories and AI-generated phishing emails. This diverse dataset allows the model to capture both traditional phishing patterns and emerging AI-generated attack styles. To enhance classification performance, the system integrates textual representations extracted from email subject and body with engineered structural indicators such as suspicious domain patterns, URL entropy, keyword flags, subdomain counts, and other metadata-based features.

Two modeling pipelines were implemented and evaluated. The first employs classical machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machine, trained using TF-IDF textual features combined with engineered attributes. The second pipeline utilizes deep learning architectures, specifically Bidirectional Long Short-Term Memory (Bi-LSTM) networks and the DistilBERT transformer model, to capture contextual language patterns and semantic relationships present in phishing messages. DistilBERT was selected due to its balance between strong predictive capability and relatively low computational cost, enabling near real-time email analysis.

Experimental results demonstrate strong classification performance across all evaluated models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The best-performing model was integrated into the Roundcube webmail platform to enable real-time phishing detection. Incoming emails are automatically analyzed and suspicious messages can be redirected to a dedicated phishing folder, demonstrating the practical applicability of hybrid AI-based detection systems for strengthening operational email security.

Keywords: *phishing detection, machine learning, deep learning, email security, hybrid classification.*

Introduction

Email phishing represents one of the most critical and escalating threats in modern cybersecurity. Through social engineering and psychological manipulation, attackers impersonate trusted entities, such as financial institutions or business associates, to deceive victims into disclosing sensitive credentials, clicking malicious links, or executing actions that result in financial or data loss. Phishing attacks exploit users trust and remain one of the most commonly used techniques in cybercrime, targeting both individuals and organizations (A. S. & A. H., 2023). The increasing prevalence of these attacks and their potential impact on organizations and individuals underline the need to develop effective detection and prevention methods, thus reducing the risk and costs derived from these incidents (Anti-Phishing Working Group, 2024). The scale of this threat continues to grow: phishing attacks reached 989,123 reported incidents in the fourth quarter of 2024 alone, representing the highest quarterly figure recorded that year, up from 932,923 in the third quarter and 877,536 in the second quarter, with a monthly average of approximately 330,000 attacks (Anti-Phishing Working Group, 2024).

Despite the existence of spam filters and rule-based defenses, such attacks continue to bypass conventional detection mechanisms. Traditional detection techniques often rely on predefined rules or signature-based approaches, which struggle to identify newly emerging phishing strategies or AI-generated messages (A. S. & A. H., 2023). As phishing campaigns evolve and become more sophisticated, more advanced approaches are required to detect malicious emails effectively.

The aim of this research is to develop and evaluate an intelligent phishing email detection system using hybrid machine learning (ML) and deep learning (DL) approaches, and to demonstrate its practical applicability

through deployment in a real webmail environment. The object of the study is email-based phishing detection using textual and structural email features. To achieve this aim, the following objectives were defined: (1) constructing a comprehensive multi-source labeled dataset of over 11,000 emails from sources including Enron, Public Phishing Email Corpus, OpenPhish, and AI-generated samples; (2) implementing classical ML classifiers, namely Logistic Regression, Random Forest, and Support Vector Machine, with hybrid TF-IDF and engineered feature inputs; (3) developing DL architectures including a Bidirectional LSTM and a fine-tuned DistilBERT transformer; (4) comparatively evaluating all models using accuracy, precision, recall, F1-score, and ROC-AUC metrics; and (5) integrating the best-performing model into the Roundcube webmail platform for real-time phishing detection.

The research employs an empirical methodology combining quantitative model evaluation with system integration testing, using scikit-learn for ML models and TensorFlow/Keras for DL architectures, applied across both real-world and synthetically generated email datasets.

Methodology and equipment

This study follows an empirical research design aimed at developing and evaluating a hybrid phishing email detection system. The central hypothesis is that combining textual features with engineered structural and URL-based attributes produces superior classification performance compared to single-modality approaches.

The dataset was constructed from multiple sources to ensure diversity and coverage. Legitimate emails were drawn from the Enron Email Corpus, a well-established collection of real corporate communications (cs.cmu.edu/~enron). Phishing samples were sourced from a public phishing email corpus containing real attack emails (academictorrents.com), and supplemented with approximately 2,800 AI-generated phishing emails (huggingface.co/datasets/kxm1k4m1/generate_phishing_email_final). The final dataset contains 11,388 labeled email samples, with 6,002 legitimate (52.7%) and 5,386 phishing (47.3%), achieving a near-balanced class distribution suitable for model training without resampling.

Data preprocessing involved text normalization including lowercasing, removal of non-ASCII characters, stripping of reply chains, signature blocks, and legal disclaimers, as well as filtering of empty or invalid entries. Hash-based deduplication was applied to eliminate redundant samples across sources.

Feature engineering produced 28 numeric and boolean attributes per email, capturing URL complexity indicators such as entropy, path depth,

domain count, and maximum URL length, alongside content-based signals including subject length, special character count, and keyword flags. Textual features for classical ML models were represented using TF-IDF vectorization applied separately to the subject and body fields, while deep learning models used tokenized sequences with padding for uniform input length.

Five models were developed and evaluated: Logistic Regression, Random Forest, and Support Vector Machine using scikit-learn, alongside a Bidirectional LSTM and a fine-tuned DistilBERT transformer implemented in TensorFlow/Keras and HuggingFace Transformers respectively. All models used a hybrid input combining textual and engineered numeric features. ML models were validated using k-fold cross-validation, while DL models used train/validation/test splits with binary cross-entropy loss, dropout regularization, and early stopping.

The complete workflow was implemented in Python 3.10, with Pandas and NumPy for data processing, Matplotlib and Seaborn for visualization, and the Roundcube webmail platform for real-world system integration and deployment testing.

Presentation of research results (Analysis)

System architecture and modeling approach

The phishing detection system is built on a hybrid input architecture that combines textual features extracted from email content with engineered numerical and boolean indicators. The overall framework operates across three stages: preprocessing, model training, and deployment. Two parallel modeling pipelines were developed and evaluated: a classical machine learning pipeline comprising Logistic Regression, Random Forest, and Support Vector Machine, and a deep learning pipeline comprising a Bidirectional LSTM and a fine-tuned DistilBERT transformer. Both pipelines share a common preprocessing phase and are trained on the same dataset to ensure a fair comparison.

For the ML pipeline, email subject and body text were vectorized using TF-IDF with unigram and bigram ranges, and combined with standardized engineered features through scikit-learn's ColumnTransformer. For the DL pipeline, the Bi-LSTM used a Keras tokenizer with a vocabulary of 20,000 words and sequences padded to 256 tokens, while DistilBERT used the HuggingFace WordPiece tokenizer with identical sequence length. In both DL architectures, a dual-branch design was adopted: a text branch producing contextual embeddings and a numerical branch processing engineered features, with outputs concatenated at a fusion layer before final classification. Training used binary

cross-entropy loss, Adam optimizer, dropout regularization, early stopping, and class weighting to address label imbalance.

Model performance results

All five models were evaluated on a held-out test set of 1,703 emails using accuracy, precision, recall, F1-score, and ROC-AUC. Results are summarized in Table 1.

Table 1. Final Performance Metrics for Each Model (Test Set)

Model	Accuracy	Precisio	Recall	F1-	ROC-AUC
Logistic Regression	99.47%	99.26%	99.63%	99.44%	0.9999
Random	99.59%	99.26%	99.88%	99.57%	1.0000
SVM	99.30%	99.13%	99.38%	99.25%	0.9999
BiLSTM	99.76%	99.66%	99.88%	99.77%	1.0000
DistilBERT	99.94%	99.87%	100.00%	99.94%	0.9999

All models surpassed 99% across every metric, confirming the effectiveness of the hybrid feature design. DistilBERT achieved the highest overall performance, producing zero false positives and missing only a single phishing email across the entire test set, an error rate of just 0.06%. BiLSTM followed closely with an error rate of 0.23%, misclassifying only 4 emails in total. Among classical ML models, Random Forest was the strongest performer with 99.59% accuracy and only 7 total misclassifications, while SVM produced the most errors at 12 misclassifications, though still achieving 99.30% accuracy. ROC-AUC values of 0.9999 or above across all models indicate near-perfect class separation regardless of threshold.

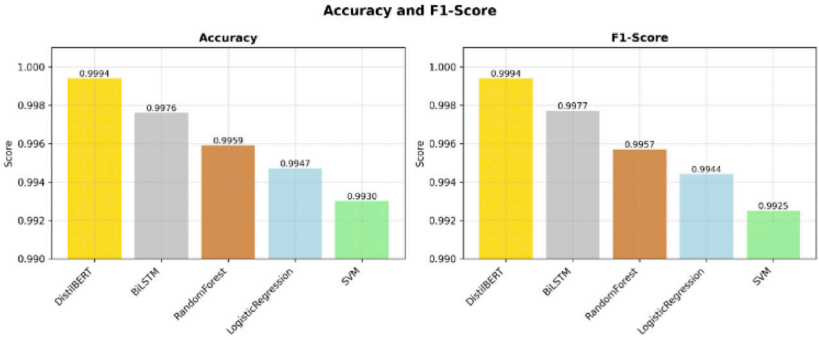

Fig. 1. Comparison of Model Accuracy and F1-Score Across All Models

ML vs. DL comparison

Grouping models by approach reveals a clear advantage for deep learning in minimizing false positives. The ML group produced 19 total false positives across all three classifiers, compared to only 1 for the DL group. Mean accuracy for DL models was 99.85% compared to 99.45% for ML models, and mean F1-score was 99.86% versus 99.42%, as shown in Table 2.

Table 2. *Comparison of ML vs. DL Models*

Group	Mean Acc.	Mean Prec.	Mean Recall	Mean F1	FP	FN	Best Model
Machine Learning	99.45 %	99.22 %	99.63 %	99.42 %	19	9	Random Forest
Deep Learning	99.85 %	99.76 %	99.94 %	99.86 %	1	4	DistilBER T

Random Forest closely approached DL-level recall, missing only one phishing email, and offers the additional advantage of feature interpretability, where URL entropy, number of domains, and keyword flags emerged as the most influential predictors. ML models are significantly lighter in computational requirements, making them more suitable for resource-constrained deployment environments. Deep learning is the preferred choice when accuracy and reliability are paramount, while machine learning remains attractive in constrained environments due to speed, interpretability, and simplicity (Uddin et al., 2024).

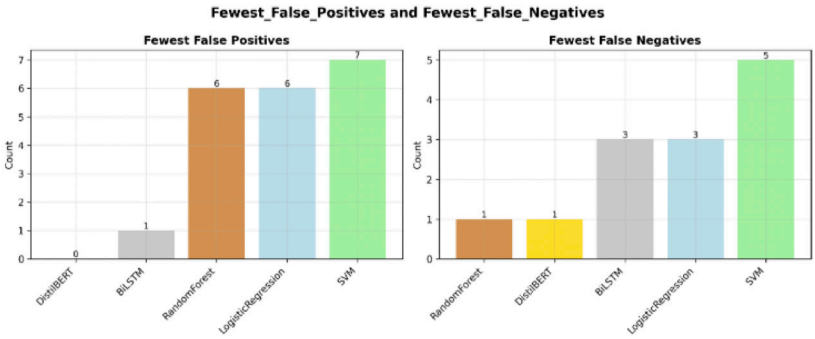

Fig. 2. *False Positive and False Negative Counts Across All Models*

Production testing

The best-performing model was integrated into the Roundcube webmail platform for real-time phishing detection. Two deployment configurations were tested: a DistilBERT-only branch and a hybrid RF + DistilBERT branch implementing a two-stage pipeline where Random Forest pre-filters emails before passing borderline cases to DistilBERT for deeper semantic analysis. During live testing with a balanced mix of real and simulated phishing emails, both branches successfully detected all phishing attempts. The DistilBERT-only branch occasionally produced isolated false positives on legitimate emails containing multiple embedded links. The hybrid branch demonstrated greater consistency, maintaining zero false positives across repeated test sessions while preserving accurate phishing detection, confirming its robustness and suitability for long-term production deployment.

Error Analysis

Despite near-perfect performance, a small number of misclassifications occurred across all models, summarized in Table 3. Phishing emails that were missed typically employed subtle social engineering without overt URL manipulation, closely mimicking legitimate communication styles. False positives were primarily legitimate emails containing multiple hyperlinks or urgent transactional language, which inadvertently triggered phishing indicators such as high URL entropy or keyword flags.

Table 3. Breakdown of Prediction Outcomes (TN, FP, FN, TP) by Model

Model	TN	FP	FN	TP	Total	Error Rate
Logistic Regression	893	6	3	801	1703	0.53%
Random Forest	893	6	1	803	1703	0.41%
SVM	892	7	5	799	1703	0.70%
BiLSTM	898	1	3	801	1703	0.23%
DistilBERT	898	0	1	804	1703	0.06%

These findings suggest that future improvements should incorporate URL reputation checking, adversarial training samples, and larger transformer architectures to address edge cases where semantic subtlety outweighs structural signals.

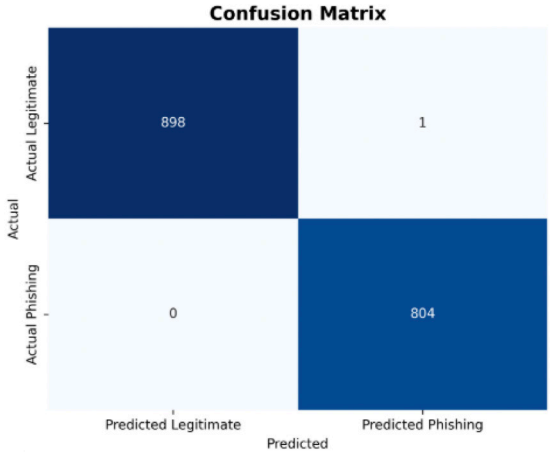

Fig. 3. DistilBERT Confusion Matrix (Test Set)

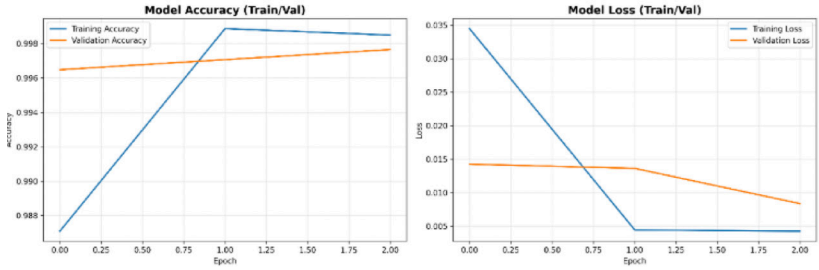

Fig. 4. DistilBERT Training and Validation Accuracy and Loss Curves

Conclusions

This study developed and evaluated an intelligent phishing email detection system combining classical machine learning and deep learning approaches within a unified hybrid architecture. The research confirmed that integrating engineered structural and URL-based features with contextual textual embeddings produces superior detection performance compared to single-modality approaches.

All five models achieved over 99% accuracy on the test set. DistilBERT delivered the strongest results with 99.94% accuracy, zero false positives, and only one missed phishing email, confirming the advantage of transformer-based architectures in capturing semantic nuance (Songailaite et

al., 2023). BiLSTM followed closely with an error rate of 0.23%, while Random Forest proved the most competitive classical model, approaching deep learning performance while offering interpretability and lower computational cost.

Real-world deployment in the Roundcube webmail platform demonstrated the system's practical feasibility. The hybrid RF + DistilBERT production configuration maintained zero false positives across repeated live testing sessions, confirming its stability and suitability for operational email security environments.

For practical deployment, the following recommendations are proposed. First, a risk-based classification approach, categorizing emails into severity levels rather than binary outcomes, enables more nuanced organizational response. Second, detection thresholds should be calibrated conservatively to minimize false positives and preserve user trust. Third, continuous retraining with recent phishing campaigns is essential given the rapid evolution of attack strategies, particularly AI-generated phishing. Fourth, the system should complement rather than replace user awareness, with clear explanations of flagged emails reinforcing human judgment alongside automated detection.

Future work should address multilingual detection to extend applicability beyond English datasets, adversarial robustness against evasion tactics such as obfuscated URLs and modified phishing text, and the integration of larger transformer architectures balanced with distillation techniques for efficient deployment. Incorporating multimodal signals including images, sender metadata, and logos, alongside integration with enterprise security infrastructures such as SIEM platforms, represents a promising direction for comprehensive next-generation phishing defense.

List of references

1. A. S., & A. H. (2023). Phishing emails detection model using deep learning.
2. Anti-Phishing Working Group. (2024). Phishing activity trends report: 4th quarter 2024. Anti-Phishing Working Group.
3. Songailaitė, M., Kankevičiūtė, E., Zhyhun, B., & Mandravickaitė, J. (2023). BERT-based models for phishing detection. In Proceedings of the 28th International Conference on Information Society and University Studies (IVUS 2023).
4. Uddin, M. A., Islam, M. M., Hossain, M. S., & others. (2024). An explainable transformer-based model for phishing email detection: A large language model approach.