HPDnet: A Hybrid Deep Learning Architecture Combining Bidirectional LSTMs and Heuristic Features for Phishing Email Classification

Parul Patel
Department of ICT, VNSGU

ABSTRACT
Phishing emails are one of the most common cybersecurity threats and continue to cause significant financial losses. Traditional rule-based email filters are becoming less effective because attackers constantly change the content and structure of phishing emails. This paper presents HPDnet, a hybrid deep learning model for phishing email detection that combines BiLSTM-based text analysis with handcrafted features such as URL count, urgency words, generic greetings, and formatting patterns. The model merges these features to improve phishing detection accuracy while maintaining interpretability. The paper also describes the data preprocessing, model architecture, training process, and performance evaluation using a labeled phishing email dataset. The proposed approach demonstrates that combining deep learning with domain-specific features provides an effective and practical solution for phishing email classification.

Keywords
Phishing detection, deep learning, bidirectional LSTM, natural language processing, cybersecurity, email classification

1. INTRODUCTION
Phishing is one of the most common initial-access techniques used in cyberattacks. It relies on social engineering rather than software vulnerabilities to compromise victims. A phishing email typically impersonates a trusted sender like a bank, a workplace IT department, or a shipping carrier and pressures the recipient into clicking a malicious link, entering credentials on a spoofed page, or transferring funds. Phishing attacks target human judgment instead of technical weaknesses. Since attackers frequently change their email content, website domains, and formatting to avoid detection, phishing detection systems need to continuously adapt to recognize new phishing attempts. Initially, phishing email detection has been approached as a supervised binary classification problem: given the text and structure of an email, predict whether it is phishing or legitimate. Early systems relied on manually engineered features like the presence of IP-based URLs, mismatched anchor text and hyperlink targets, or the count of embedded forms, feeding a classical classifier such as a support vector machine or random forest [1]. These systems were effective and interpretable but required continual manual feature engineering as attacker behavior shifted. The growth of deep learning has provided a new way to detect phishing. These models learn patterns directly from data instead of depending on manually created rules, although they can be harder to understand and explain [2][3].

This paper presents HPD-Net (Hybrid Phishing Detection Network), a deep learning model for phishing email detection that combines two approaches. Instead of relying only on deep learning or only on manually designed features, HPD-Net uses both. A bidirectional LSTM (BiLSTM) learns patterns from the email text, while a set of simple heuristic features captures common phishing signs such as urgent language and generic greetings. This approach has two main advantages. First, these heuristic features are easy to compute and are still effective in identifying phishing emails, especially when the amount of training data is limited. Second, the model becomes easier to understand because the system can show which phishing indicators were detected, along with the prediction score. This helps security analysts understand why an email was classified as phishing.

The remainder of this paper is organized as follows. Section 2 reviews related work in phishing email detection, spanning classical feature-based methods, deep learning approaches, and recent transformer-based systems. Section 3 describes the proposed methodology, including preprocessing, feature extraction, and model architecture. Section 4 details the experimental setup and results. Section 5 concludes the paper and outlines directions for future work.

2. RELATED WORK
Phishing emails remain one of the most effective cyberattack vectors due to their reliance on social engineering rather than technical vulnerabilities. Over the past two decades, phishing email detection has evolved from rule-based systems and handcrafted feature engineering to machine learning, deep learning, and, more recently, large language model (LLM)-based approaches.

Early phishing detection systems relied on manually designed features. One of the most influential systems was PILFER, which used only ten carefully selected features, such as IP-based URLs, mismatched hyperlinks, and domain age, to detect phishing emails with high accuracy and a low false-positive rate [1]. Building on this approach, Akinyelu and Adewumi [2] proposed a Random Forest classifier using a reduced set of phishing-specific email features. Their model achieved high classification accuracy while minimizing false positives, confirming the effectiveness of ensemble learning methods for phishing detection. However, both studies depended heavily on handcrafted features and external information such as domain-related attributes, limiting their adaptability to evolving phishing strategies [2]. As phishing attacks became increasingly sophisticated, researchers shifted toward deep learning models capable of learning textual representations directly from email content. Halgaš et al. [3] introduced an RNN-based phishing email classifier that analyzed the linguistic structure of email text without requiring manually engineered features. Their work demonstrated that sequential neural networks can automatically learn discriminative textual patterns and outperform conventional feature-based approaches.

Further advances in natural language processing encouraged the adoption of LSTM and CNN architectures. Bagui et al. compared several machine learning and deep learning models using one-hot encoding and word embeddings for phishing email classification. Their experiments showed that deep learning models outperformed traditional machine learning techniques in classification accuracy, while CNN combined with word embeddings achieved the best overall performance [4]. Recognizing the growing role of NLP, Salloum et al. (2022) conducted a systematic literature review of 100 studies on phishing email detection using natural language processing. Their review identified feature extraction and classification as the dominant research areas, with Support Vector Machines being the most widely used traditional classifier and TF-IDF and word embeddings serving as the most common text representation techniques. The review also highlighted the increasing adoption of deep learning while emphasizing the need for more robust semantic understanding and multilingual phishing detection methods [5].

Recent studies have explored privacy-preserving and distributed learning approaches. Sun et al. (2021) proposed Federated Phish Bowl, an LSTM-based federated learning framework that enables collaborative phishing email detection without sharing private email data. Their approach demonstrated that decentralized learning can achieve competitive performance while preserving user privacy, making it suitable for organizations with strict data-sharing constraints [6]. With the emergence of generative AI, researchers have begun investigating AI-generated phishing emails. Tooher and Lallie (2025) proposed a two-stage deep learning framework that combines DistilBERT with persuasion-principle analysis based on Cialdini's psychological principles. Instead of relying solely on textual features, their framework analyzes behavioral persuasion cues to detect sophisticated AI-generated phishing emails, achieving high detection accuracy while improving model interpretability [7].

Recent deep learning architectures have also incorporated multiple information sources. Rashed and Ozcan (2026) proposed a dual-layer architecture that combines LSTM and transformer models to analyze both email content and malicious attachments. Their framework demonstrated that integrating attachment-level analysis significantly improves phishing and spam detection compared with single-layer content-based models [8]. Several recent conference papers have focused on practical AI applications in email processing. Hemannth et al. (2024) proposed a phishing detection model combining Horse Herd Optimization with Random Forest classification, showing that bio-inspired optimization can improve feature selection and classifier performance [9]. Industrial solutions have also begun incorporating artificial intelligence into operational security platforms. Palo Alto Networks' Cortex XSOAR phishing email classifier applies machine learning and natural language processing to automate phishing investigation and response, illustrating the practical adoption of AI-based phishing detection in Security Operations Centers (SOCs). Similarly, industry reports emphasize that modern phishing detection increasingly depends on contextual language understanding rather than solely relying on static email metadata [10].

Overall, phishing detection has evolved from feature-based methods to deep learning and LLM-based approaches. Feature-based methods are simple and interpretable, while deep learning models provide better contextual understanding but require more data. The proposed HPDnet combines both strengths by using a BiLSTM branch to learn contextual features and a parallel dense branch to process handcrafted heuristic features, improving detection accuracy even with limited labeled data.

3. RESEARCH METHODOLOGY

HPDnet is a two-branch neural network designed for binary classification of emails as phishing or legitimate. The model processes the input email in two parallel branches. The first branch cleans and tokenizes the email text before learning contextual information using a BiLSTM network. The second branch extracts handcrafted heuristic features, such as structural and lexical phishing indicators. The outputs of both branches are then concatenated and passed through a final classification layer, which predicts the probability that the email is phishing.

3.1 Text Preprocessing

The input email text, including the subject and sender information (when available), is first converted to lowercase. URLs and email addresses are replaced with special tokens (such as <url> and <email>) instead of being removed, because their presence can help identify phishing emails. Unnecessary special characters are removed, and extra spaces are cleaned. The processed text is then converted into word tokens, and each email is padded or truncated to a fixed length of 250 tokens so that all emails have the same input size for the neural network.

3.2 Handcrafted Feature Extraction

Along with tokenization, the model also extracts seven simple features directly from the original email text: URL count, exclamation marks, urgency phrases, uppercase ratio, generic greeting presence, normalized email length, and link-like token count. These features capture common phishing patterns, are scaled using a standard scaler, and are processed through a separate dense branch. Unlike PILFER's feature set, they rely only on the email content and do not require external lookups, making the model lightweight and usable offline.

3.3 Model Architecture

The text branch starts with an embedding layer (128 dimensions) that converts each word into a dense vector while ignoring padding tokens through masking. The embedded sequence is then processed by two stacked BiLSTM layers (64 and 32 units per direction) to learn contextual information and generate a fixed-length representation of the email. A dense layer (64 units, ReLU) with dropout further refines this representation.

The feature branch processes the seven scaled heuristic features using a dense layer (32 units, ReLU) with dropout to learn a compact feature representation. The outputs of both branches are concatenated, passed through another dense layer (32 units, ReLU) with dropout to learn combined patterns, and finally fed to a sigmoid output layer, which predicts the probability of the email being a phishing message.

3.4 Training Procedure

The model is trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy as the loss function. Early stopping (patience = 3) monitors the validation loss and restores the best model weights to prevent overfitting. The model with the lowest validation loss is saved. During training, accuracy, AUC, precision, and recall are monitored on both the training and validation datasets.

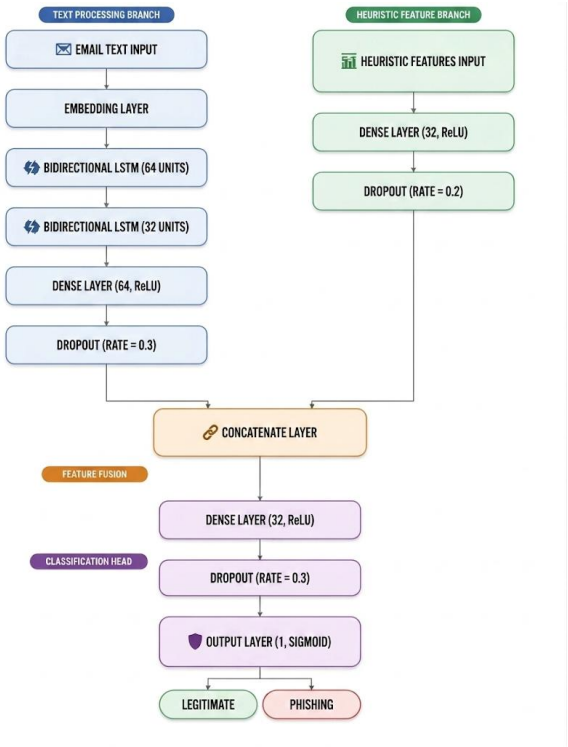

Fig 1: HPDnet - Hybrid Phishing Detection Network

4. EXPERIMENTAL SETUP & RESULTS

The proposed model was evaluated on a synthetic dataset containing 800 email samples (400 phishing and 400 legitimate). The emails were created using realistic templates representing common scenarios such as account verification, payment failures, shipment updates, meeting reminders, and receipts. The dataset was divided into 70% training, 15% validation, and 15% testing using stratified sampling.

Since the dataset is template-based, the reported performance is expected to be higher than that on real-world emails. Future work will validate the model using publicly available datasets such as the Nazario phishing corpus and the Enron email dataset.

The model was implemented in Python using TensorFlow/Keras for deep learning, scikit-learn for data preprocessing and feature scaling, and pandas for data handling. The tokenizer vocabulary was limited to 10,000 words, and email sequences were padded or truncated to 250 tokens. The model was trained for up to 15 epochs with a batch size of 32, using early stopping to prevent overfitting.

Model performance was evaluated on the test dataset using standard classification metrics: Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). These metrics provide a comprehensive assessment of phishing email detection performance.

On the held-out test set of 120 emails (60 phishing and 60 legitimate), the proposed hybrid BiLSTM model achieved an accuracy of 97.5% and an ROC-AUC of 0.992. For phishing emails, the model obtained a precision of 0.98, recall of 0.97, and F1-score of 0.98. For legitimate emails, it achieved a precision of 0.97, recall of 0.98, and F1-score of 0.98. The training and validation losses decreased consistently, and early stopping helped prevent overfitting.

These results indicate that the proposed model is effective for phishing email classification on the synthetic dataset. However, since the dataset was generated from predefined templates, it does not represent the full diversity of real-world emails. Therefore, the results should be considered as a proof of concept that validates the proposed preprocessing pipeline, hybrid architecture, and training process. Future work will evaluate the model using real-world phishing datasets to assess its practical performance and generalization ability.

Table 1: Classification Results

Class	Precision	Recall	F1-Score	Support
Legitimate	0.97	0.98	0.98	60
Phishing	0.98	0.97	0.98	60
Overall Accuracy				97.5%
ROC-AUC				0.992

5. CONCLUSION & FUTURE WORK

This paper presented a hybrid BiLSTM-based phishing email detection model that combines contextual text learning with lightweight heuristic features. By integrating sequential text representations with phishing-specific indicators, the model achieves accurate and efficient email classification. Experimental results on a synthetic dataset demonstrate the effectiveness of the proposed approach while maintaining low computational complexity, making it suitable for practical phishing detection. Future work will evaluate the proposed model on real-world phishing datasets such as the Nazario and Enron corpora to assess its generalization capability. The model can be further enhanced by incorporating email headers, attachment analysis, attention mechanisms, and transformer-based embeddings to improve detection of sophisticated and AI-generated phishing emails.

6. REFERENCES

[1] I. Fette, N. Sadeh, and A. Tomasic, “Learning to detect phishing emails,” in Proc. 16th Int. Conf. World Wide Web (WWW '07), Banff, AB, Canada, 2007, pp. 649–656.

[2] Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014.

[3] Halgas, L., Agrafiotis, I., & Nurse, J. R. C. (2019). Catching the Phish: Detecting Phishing Attacks using Recurrent Neural Networks (RNNs). In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 317-336). Springer.

[4] Bagui, S., & Nandi, N. (2021). Machine learning and deep learning for phishing email classification using one-hot encoding. Journal of Computer Science, 17.

[5] Salloum, Said & Gaber, Tarek & Vadera, Sunil & Shaalan, Khaled. (2022). A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques. IEEE Access. 10. 65703-65727. 10.1109/ACCESS.2022.3183083.

[6] Chen, P., & Kuo, L. (2021). Federated Phish Bowl: LSTM-based decentralized phishing email detection. arXiv.

[7] Tooher, P., & Lallie, H. (2025). A two-stage deep learning framework for AI-driven phishing email detection based on persuasion principles. Computers, 14(12), 523.

[8] Rashed, S., & Ozcan, C. (2026). A Novel Dual-Layer Deep Learning Architecture for Phishing and Spam Email Detection. Electronics, 15(3), 630. https://doi.org/10.3390/electronics15030630

[9] Hemannth, P. & Chinta, Mukesh & Satya, S. & Devasena, P. (2024). A Phishing Email Detection Model Based on Horse Herd Optimization and Random Forest Algorithms. 1-6. 10.1109/APCIT62007.2024.10673699.

[10] Palo Alto Networks. Phishing email classifier in Cortex XSOAR. https://www.paloaltonetworks.com/blog/security-operations/phishing-email-classifier-in-cortex-xsoar/