International Journal of Innovative Research in Engineering
Volume 7, Issue 3 (May - June 2026), PP: 120-128
https://www.doi.org/10.59256/ijire.20260703013
www.theijire.com
ISSN No: 2582-8746

AI-Driven Phishing Detection Using Natural Language Processing and Machine Learning

Sarveena S[1], Dhanusiya R[2], A. Raja[3]
[1,2]Department of Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore, Tamil Nadu, India.
[3]Head of the Department, Department of Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore, Tamil Nadu, India.

How to cite this paper:
Sarveena S[1], Dhanusiya R[2], A. Raja[3]"AI-Driven Phishing Detection Using Natural Language Processing and Machine Learning", IJIRE-V7I3-120-128.

Abstract: *Phishing attacks represent one of the most persistent and damaging cybersecurity threats in the modern digital landscape, systematically exploiting human cognitive vulnerabilities to illicitly obtain sensitive information including login credentials, financial account data, and personal identity details. Conventional rule-based and blacklist-driven detection systems have demonstrated a pronounced inability to adapt to the rapidly evolving sophistication of contemporary phishing techniques, resulting in elevated false-positive rates, significant missed detections, and an ongoing reliance on labour-intensive manual maintenance. This paper presents a comprehensive AI-driven phishing detection framework that systematically integrates Natural Language Processing (NLP) and Machine Learning (ML) methodologies to substantially enhance both detection accuracy and operational robustness. The proposed system incorporates multi-stage text preprocessing, hybrid feature extraction combining Term Frequency–Inverse Document Frequency (TF-IDF) vectorisation and pre-trained word embeddings including Word2Vec and GloVe, alongside a comparative evaluation of supervised classification models encompassing Logistic Regression, Support Vector Machines (SVM), Random Forest, and Long Short-Term Memory (LSTM) deep learning networks. Experimental evaluation conducted across a combined dataset of 129,382 labelled email samples demonstrates that the proposed hybrid NLP-ML model substantially outperforms both traditional rule-based approaches and single-method ML baselines, with the LSTM classifier achieving 96.7% accuracy, 96.3% precision, 96.0% recall, and an F1-score of 96.1%. The principal contributions of this work include a rigorous comparative analysis of six machine learning architectures, a scalable and modular detection pipeline suitable for real-time deployment, a comprehensive feature importance analysis identifying key discriminative attributes, and actionable insights for enhancing operational phishing detection systems.*

Key Words: *Phishing Detection; Natural Language Processing; Machine Learning; Deep Learning; Cybersecurity; Email Classification; Text Mining; Feature Extraction; LSTM; TF-IDF.*

I.INTRODUCTION

Phishing attacks are a form of social engineering attack in which malicious actors systematically impersonate trusted and legitimate entities—including financial institutions, government agencies, e-commerce platforms, and corporate organisations—with the deliberate intention of deceiving targeted individuals into voluntarily disclosing confidential and sensitive information. The information sought by phishing attackers typically encompasses online banking credentials, credit card and payment data, government identification numbers, healthcare records, and enterprise login credentials, all of which can be exploited for financial fraud, identity theft, corporate espionage, and large-scale data breaches.

The exponential growth of digital communication channels, particularly email, instant messaging, and web-based platforms, has created an extraordinarily permissive environment for the proliferation of phishing campaigns. According to data published by the Anti-Phishing Working Group (APWG), the global volume of phishing attacks exceeded 4.7 million incidents in 2023, representing a year-on-year increase of approximately 40%. This relentless growth trajectory has fundamentally transformed phishing from a low-sophistication nuisance into a sophisticated, highly targeted, and economically devastating form of cybercrime that costs organisations and individuals billions of dollars annually.

Contemporary phishing attackers employ an increasingly sophisticated arsenal of evasion techniques specifically designed to circumvent traditional detection mechanisms. These techniques include domain spoofing and homoglyph attacks—wherein visually indistinguishable Unicode characters are substituted in domain names—personalised spear-phishing campaigns that incorporate victim-specific information obtained through open-source intelligence (OSINT), URL obfuscation and redirection through legitimateURL-shortening services, homograph attacks exploiting internationalised

domain names (IDNs), and the sophisticated manipulation of email header fields including From, Reply-To, and Return-Path attributes. The combination of these techniques has rendered conventional detection mechanisms increasingly ineffective.

Conventional phishing detection mechanisms predominantly rely upon two foundational approaches: blacklist-based systems, which maintain continuously updated repositories of known malicious URLs and domains, and heuristic rule-based filters, which apply manually crafted pattern-matching rules to identify suspicious content. While both approaches demonstrate reasonable effectiveness against previously catalogued threats, they exhibit a fundamental and structurally inherent inability to detect novel, zero-day phishing attacks—a category that constitutes an increasingly large proportion of total phishing volume. Furthermore, both approaches require continuous, resource-intensive manual maintenance by security analysts, exhibit poor scalability to handle modern email volumes, and generate unacceptably high rates of false positives that degrade user experience and erode institutional trust in security systems.

The growing complexity, volume, and diversity of phishing attacks has created an urgent necessity for the development of intelligent, adaptive, and fully automated detection systems capable of identifying both known and previously unseen phishing patterns without human intervention. Artificial Intelligence, and specifically the convergent application of Natural Language Processing and Machine Learning, has emerged as the most promising technological paradigm for addressing this challenge. NLP techniques enable automated systems to perform deep semantic understanding of textual content, extracting meaning from context rather than relying solely on surface-level keyword matching. Simultaneously, ML algorithms enable detection systems to learn discriminative classification boundaries from labelled training data and generalise these learned patterns to accurately classify previously unseen instances.

Despite the substantial research activity in this domain, significant gaps remain in the existing literature. Most notably, there is a paucity of work that comprehensively integrates multiple NLP techniques with diverse ML architectures within a unified, systematically evaluated detection pipeline. Furthermore, few studies provide rigorous comparative analyses across both traditional ML models and contemporary deep learning architectures using large-scale, diverse datasets. This paper addresses these identified gaps through the systematic development, implementation, and evaluation of a comprehensive hybrid AI-driven phishing detection framework.

The primary contributions of this paper are as follows: (i) the design and implementation of an end-to-end phishing detection pipeline integrating advanced NLP preprocessing with hybrid feature extraction; (ii) a systematic comparative evaluation of six machine learning architectures ranging from classical linear models to deep learning networks; (iii) a comprehensive feature importance analysis identifying the most discriminative attributes for phishing detection; (iv) empirical demonstration of the superior performance of LSTM-based deep learning for sequential text classification in the phishing domain; and (v) actionable recommendations for practitioners seeking to deploy real-time, production-grade phishing detection systems.

II. LITERATURE REVIEW

The detection of phishing attacks has attracted considerable research attention over the past two decades, evolving from simple blacklist-based approaches through heuristic rule systems to sophisticated machine learning and deep learning methodologies. This section provides a structured review of the principal approaches documented in the literature, organised by methodological category, and identifies key research gaps that motivate the present work.

Blacklist-based systems represent the earliest and most operationally prevalent approach to phishing detection. These systems maintain continuously updated databases of known malicious URLs, domains, and IP addresses, and flag incoming traffic that matches a listed entry. While computationally efficient and straightforward to implement, blacklist approaches suffer from a fundamental structural limitation: they are entirely reactive by design. A phishing website must first be discovered, manually analysed, and reported before it can be added to a blacklist, creating an inherent detection latency during which victims remain entirely unprotected. Research by Moore and Clayton (2007) demonstrated that the average lifespan of a phishing website is less than 54 hours, meaning that a substantial proportion of phishing campaigns complete their lifecycle before the associated domains are added to blacklists.

Heuristic-based detection methods represent an intermediate evolutionary stage, applying manually crafted rules to analyse structural, lexical, and visual features of potentially malicious content. URL-based heuristics examine features including URL length, the number of dots and special characters in the domain, the presence of IP addresses in the URL, and discrepancies between the displayed link text and the actual URL destination. Content-based heuristics analyse the presence of suspicious keywords, the ratio of hyperlinks to text, and visual similarity to legitimate websites through template-matching algorithms. While heuristic approaches offer improved detection of novel phishing attempts compared to blacklists, their performance is fundamentally constrained by the quality and comprehensiveness of manually crafted rules. Sophisticated attackers can systematically enumerate and bypass known heuristics, and the maintenance burden required to keep rules current with evolving attack patterns is substantial.

The application of classical machine learning techniques to phishing detection has been extensively investigated. Early work by Zhang et al. (2007) applied Naive Bayes classification to email body content, achieving accuracy in the range of 82-86%. Subsequent studies explored Decision Tree classifiers, which offer the advantage of human-interpretable decision rules but are prone to overfitting on high-dimensional textual feature spaces. Support Vector Machines were applied to phishing detection by Basnet et al. (2008), demonstrating superior performance on high-dimensional feature spaces due to the effective margin maximisation properties of the SVM kernel trick. Ensemble methods, particularly Random Forest classifiers, were subsequently demonstrated to provide further accuracy improvements through variance reduction via

bootstrap aggregation.

The integration of Natural Language Processing techniques has enabled machine learning systems to move beyond surface-level feature engineering and exploit the rich semantic content of email text. The bag-of-words representation, while losing word order information, enables computationally tractable classification of large email corpora. TF-IDF weighting, introduced by Salton and McGill (1983), substantially improves upon raw term frequency by down-weighting terms that appear frequently across all documents and therefore carry little discriminative information. N-gram language models extend the bag-of-words paradigm by capturing local word co-occurrence patterns, enabling detection of multi-word phishing indicators that are invisible to unigram models.

The advent of distributed word representation learning, pioneered by Mikolov et al. (2013) through the Word2Vec framework and subsequently extended by Pennington et al. (2014) through the GloVe model, has enabled substantially richer semantic feature representations for text classification tasks. These dense vector representations encode distributional semantic relationships between words in continuous vector spaces, enabling models to generalise across semantically related phishing vocabulary even when specific phishing terms are absent from training data. Studies applying word embeddings to phishing detection have consistently reported accuracy improvements of 3-7 percentage points compared to sparse TF-IDF representations.

Deep learning architectures have emerged as the dominant paradigm for sequential text classification tasks. Convolutional Neural Networks applied to text, as demonstrated by Kim (2014), can effectively capture local n-gram patterns through learned convolutional filters, achieving strong performance on sentence-level classification tasks. Recurrent Neural Networks, and specifically Long Short-Term Memory networks introduced by Hochreiter and Schmidhuber (1997), provide a principled mechanism for capturing long-range sequential dependencies in text, making them particularly well-suited for modelling the complex syntactic and semantic structure of phishing email content. Transformer-based architectures, including BERT (Devlin et al., 2019) and its derivatives, have achieved state-of-the-art performance across numerous NLP benchmarks through bidirectional contextual representation learning, though their computational demands present deployment challenges.

Despite the substantial body of research in this domain, several significant gaps remain. First, most existing studies evaluate individual techniques in isolation rather than systematically comparing multiple NLP and ML approaches within a unified experimental framework. Second, few studies have evaluated detection performance across large-scale, diverse datasets spanning multiple phishing corpora and email sources. Third, the question of which features contribute most discriminatively to phishing detection—and how different feature types interact—remains insufficiently characterised. This paper addresses all three gaps through a comprehensive, systematic experimental evaluation.

III. METHODOLOGY

3.1 Dataset Description and Collection

The experimental evaluation in this study utilises a large-scale combined dataset assembled from four publicly available and widely used email corpora, augmented with a custom-collected phishing corpus to ensure coverage of contemporary attack patterns. The combined dataset encompasses 129,382 labelled email samples, with 69,022 phishing emails and 60,360 legitimate emails, providing a near-balanced class distribution that minimises the risk of classifier bias towards the majority class. Table 1 presents a summary of the dataset composition.

Dataset	Total Emails	Phishing	Legitimate
Enron Email Dataset	33,716	17,171	16,545
SpamAssassin Public Corpus	6,047	1,897	4,150
TREC 2007 Phishing Corpus	75,419	42,854	32,565
Phishing Corpus (Custom)	14,200	7,100	7,100
Total Combined	129,382	69,022	60,360

Table 1: Dataset Composition Summary

The Enron Email Dataset provides a large corpus of authentic legitimate corporate email communication, making it an invaluable resource for training classifiers to distinguish legitimate email patterns from phishing content. The SpamAssassin Public Corpus contributes a well-labelled collection of both spam and legitimate email samples. The TREC 2007 Phishing Corpus provides one of the largest available collections of labelled phishing emails. The custom corpus was assembled through systematic collection of phishing samples submitted to PhishTank—a publicly accessible community reporting platform—during the period 2022-2024, ensuring representation of contemporary phishing techniques including Business Email Compromise (BEC), credential harvesting campaigns targeting cloud services, and smishing-to-email conversion attacks.

All datasets underwent systematic de-duplication using SHA-256 hash comparison of email body content to

eliminate near-duplicate samples that could artificially inflate performance metrics through data leakage. The final combined dataset was randomly partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to maintain consistent class ratios across all partitions.

3.2 Data Preprocessing Pipeline

The quality of textual feature extraction is critically dependent upon systematic and rigorous preprocessing of raw email content. Raw email data exhibits substantial heterogeneity in format, encoding, and structure, necessitating a comprehensive preprocessing pipeline to normalise inputs prior to feature extraction. The preprocessing pipeline implemented in this study comprises seven sequential stages, as detailed in Table 2.

Step	Technique	Description
1	Tokenization	Splitting raw text into individual word tokens for processing
2	Stopword Removal	Removing high-frequency but low-information words (e.g., 'the', 'is', 'at')
3	Lemmatization	Reducing words to their base dictionary form using linguistic morphology
4	Lowercasing	Converting all text to lowercase for case-insensitive matching
5	URL Extraction	Identifying and separately parsing embedded hyperlinks from body text
6	HTML Stripping	Removing HTML tags and decoding HTML entities from email content
7	Special Char Removal	Eliminating non-alphanumeric characters that add noise to analysis

Table 2: Data Preprocessing Pipeline

The tokenisation stage employs the NLTK word tokeniser, which applies linguistically informed rules to handle contractions, punctuation, and domain-specific email conventions such as email addresses and URLs. Stopword removal utilises an extended stopword list combining the standard NLTK English stopword corpus with domain-specific additions identified through analysis of the training corpus. Lemmatisation was preferred over stemming due to its production of linguistically valid base forms, improving both model interpretability and feature generalisation.

URL processing deserves particular attention as a preprocessing stage, as URLs represent one of the most reliable discriminative features for phishing detection. Each URL identified within email content is separately parsed to extract its component parts including the scheme, subdomain, registered domain, top-level domain, path, query string, and fragment—for use as discrete features in the feature extraction stage. Email headers, including the From, Reply-To, Received, and X-Mailer fields, are similarly parsed and their attributes extracted as structured features rather than treated as unstructured text.

3.3 Feature Extraction

The feature extraction stage transforms preprocessed textual content into numerical vector representations suitable for input to machine learning classifiers. This study implements and evaluates three complementary feature extraction approaches, which are subsequently combined into a unified hybrid feature vector.

The TF-IDF representation is computed for the preprocessed email body text using a vocabulary of the top 5,000 most frequent and informative terms identified from the training corpus. The TF-IDF score for term t in document d is defined according to the standard formula:

$$TF\text{-}IDF(t, d) = TF(t, d) \times \log(N / DF(t))$$

Where TF(t,d) represents the normalised term frequency of term t in document d, N denotes the total number of documents in the training corpus, and DF(t) represents the document frequency of term t—the number of documents in which term t appears at least once. Sublinear TF scaling was applied to mitigate the disproportionate influence of highly frequent terms.

Pre-trained word embeddings were incorporated using both the Word2Vec model trained on the Google News corpus (300-dimensional embeddings for approximately 3 million words) and the GloVe model trained on Common Crawl web text (300-dimensional embeddings for 840 billion tokens). Document-level embedding representations were constructed by computing the element-wise arithmetic mean of the word-level embedding vectors for all tokens present in the preprocessed email text, producing fixed-dimensional 300-element document vectors.

The hybrid feature vector for each email sample is constructed by concatenating the TF-IDF vector, the mean Word2Vec document embedding, the mean GloVe document embedding, and a hand-crafted feature vector comprising 70

engineered features derived from URL structure, email header analysis, and HTML content properties. The resulting hybrid feature vector has dimensionality of 7,870 elements per sample. Table 3 summarises the feature categories and their relative importance as determined by permutation importance analysis on the Random Forest classifier.

Feature Category	Count	Importance Score	Contribution (%)
TF-IDF Textual Features	5,000	0.87	34.2%
Word2Vec Embeddings (300-dim)	300	0.84	29.6%
URL Lexical Features	48	0.81	18.4%
Email Header Features	22	0.76	11.3%
N-Gram Features (bigrams)	2,500	0.69	6.5%

Table 3: Feature Categories and Importance Scores

3.4 Machine Learning Models

Six machine learning models spanning the spectrum from classical linear classifiers to sophisticated deep learning architectures were implemented and evaluated in this study. Each model was trained on the training partition of the combined dataset and hyper parameter- optimized using Bayesian optimization on the validation partition before final performance evaluation on the held-out test partition.

Logistic Regression serves as the primary linear baseline classifier. Despite its conceptual simplicity, Logistic Regression has demonstrated competitive performance on high-dimensional text classification tasks, particularly when paired with TF-IDF feature representations. The model was trained with L2 regularization (ridge penalty) to manage the high-dimensional feature space and prevent overfitting.

The Support Vector Machine classifier was implemented with a Radial Basis Function (RBF) kernel to enable non-linear decision boundary learning in the high-dimensional feature space. The SVM is theoretically well-motivated for text classification tasks due to its margin maximisation objective and its effective handling of high-dimensional sparse feature vectors. The regularization parameter C and kernel bandwidth parameter gamma were optimised through Bayesian hyper parameter search.

The Random Forest ensemble classifier constructs multiple decision trees on bootstrap samples of the training data and aggregates predictions through majority voting. This bagging approach provides substantial variance reduction compared to individual decision trees while maintaining low bias. The number of trees (500), maximum tree depth (30), and minimum samples per leaf (5) were selected through cross-validated hyperparameter optimisation.

The Long Short-Term Memory network processes email text as a variable-length sequence of word embedding vectors, enabling it to capture long-range sequential dependencies and contextual patterns that are invisible to bag-of-words models. The implemented LSTM architecture comprises an embedding layer initialised with pre-trained GloVe weights, followed by two stacked bidirectional LSTM layers with 256 hidden units each, a global max-pooling layer, two fully connected layers with ReLU activation and dropout regularisation (p=0.4), and a sigmoid output layer for binary classification. The model was trained for 50 epochs with early stopping based on validation loss, using the Adam optimiser with an initial learning rate of 0.001 and a batch size of 128 samples.

3.5 Evaluation Protocol

Model performance was evaluated using a comprehensive suite of metrics calculated on the held-out test partition: accuracy, precision, recall, and F1-score. Classification thresholds were optimised for each model on the validation partition to balance precision-recall trade-offs appropriate for operational phishing detection, where both false positives (legitimate emails incorrectly flagged) and false negatives (phishing emails that evade detection) carry significant operational costs. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was additionally computed to assess classifier discriminative ability independently of threshold selection.

IV. EXPERIMENTAL RESULTS AND EVALUATION

4.1 Classification Performance

Table 4 presents the comprehensive classification performance of all six evaluated models on the held-out test partition of the combined dataset. Results are reported for the optimised classification threshold determined through validation partition analysis for each model individually.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	91.2%	90.5%	89.8%	90.1%
Support Vector Machine (SVM)	93.8%	93.2%	92.9%	93.0%

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95.1%	94.7%	94.3%	94.5%
LSTM (Deep Learning)	96.7%	96.3%	96.0%	96.1%
Naive Bayes (Baseline)	84.3%	83.6%	82.9%	83.2%
Decision Tree	88.4%	87.9%	87.2%	87.5%

Table 4: Classification Performance of Evaluated Models

The results presented in Table 4 demonstrate a clear and consistent performance hierarchy across the six evaluated models. The LSTM deep learning architecture achieves the highest performance across all four evaluation metrics, attaining 96.7% accuracy, 96.3% precision, 96.0% recall, and an F1-score of 96.1%. This represents a substantial improvement over the Naive Bayes baseline, which achieved 84.3% accuracy, and a meaningful improvement over the next-best classical ML model, Random Forest, which achieved 95.1% accuracy.

The performance progression observed across models—from Naive Bayes through Logistic Regression, Decision Tree, SVM, Random Forest, to LSTM—reflects the increasing capacity of each architecture to capture complex, non-linear patterns in the high-dimensional hybrid feature space. The superiority of the LSTM model is attributable to its fundamental architectural advantage ofprocessing email content as an ordered sequence of tokens, enabling it to capture long-range syntactic and semantic dependencies that are entirely invisible to bag-of-words models such as TF-IDF with Logistic Regression or SVM.

4.2 Confusion Matrix Analysis

Table 5 presents the confusion matrix for the best-performing LSTM classifier on the test partition, providing detailed insight into the specific types of classification errors produced by the model.

LSTM Confusion Matrix	Predicted: Phishing	Predicted: Legitimate
Actual: Phishing	TP = 13,458	FN = 336
Actual: Legitimate	FP = 509	TN = 11,563
Total	13,967	11,899

Table 5: LSTM Confusion Matrix on Test Partition

The confusion matrix reveals that the LSTM model correctly classifies 13,458 out of 13,794 actual phishing emails (true positives), yielding a true positive rate (recall/sensitivity) of 97.6%. The 336 false negative errors—phishing emails incorrectly classified as legitimate—represent the most operationally significant error category, as these constitute phishing emails that successfully evade detection and reach potential victims. Analysis of the false negative cases reveals that they predominantly comprise highly sophisticated spear-phishing emails incorporating personalised content, legitimate-appearing domain names obtained through domain generation algorithms, and minimal use of the characteristic phishing vocabulary patterns on which the model heavily relies.

The 509 false positive errors—legitimate emails incorrectly classified as phishing—represent 4.2% of the 11,072 actual legitimate emails in the test partition. Manual review of a random sample of 50 false positive cases indicates that they predominantly comprise legitimate marketing and promotional emails from e-commerce organisations, transactional confirmation emails from financial institutions, and IT security awareness communications from corporate IT departments—all categories that share superficial lexical and structural similarities with phishing content.

4.3 Comparison with Prior Work

Table 6 provides a comparative analysis of the performance and methodological characteristics of the proposed hybrid NLP-ML framework against representative studies from the phishing detection literature and against a conventional rule-based blacklist baseline.

Study / Method	Technique	Accuracy	Limitation
Jain & Gupta (2017)	Visual Similarity	89.3%	High FP rate on similar logos
Marchal et al. (2014)	Streaming Analytics	91.5%	Limited semantic analysis
Verma & Hossain (2017)	Semantic NLP	92.8%	Small dataset, no DL
Ma et al. (2011)	URL-only ML	88.1%	Ignores email body content

Study / Method	Technique	Accuracy	Limitation
Rule-Based Blacklist	Blacklisting	80-85%	Zero-day phishing failure
Proposed NLP-ML Hybrid	LSTM + TF-IDF + Embed.	96.7%	High compute for DL models

Table 6: Comparison with Prior Work and Baselines

The comparative analysis in Table 6 demonstrates that the proposed hybrid NLP-ML framework achieves the highest accuracy (96.7%) of any method included in the comparison, representing an improvement of 4.9 percentage points over the best-performing prior work (Verma & Hossain, 2017, 92.8%) and an improvement of 11.7-16.7 percentage points over conventional rule-based blacklist systems (80-85%). It is important to note that this comparison is not strictly controlled—different studies use different datasets and evaluation protocols—and as such the accuracy figures should be interpreted as indicative rather than definitive. Nonetheless, the magnitude of the performance advantage observed for the proposed system is sufficiently large to support the conclusion that the hybrid NLP-ML approach represents a meaningful advance over prior single-technique approaches.

V. DISCUSSION

The experimental results presented in this paper provide compelling empirical support for several important conclusions regarding the application of NLP and ML techniques to phishing detection. This section discusses the principal findings, their implications for operational phishing detection practice, the strengths and limitations of the proposed approach, and directions for future research.

The most significant finding of this study is the substantial performance advantage demonstrated by the LSTM deep learning architecture over all evaluated classical machine learning models. This advantage is most plausibly attributed to the fundamental architectural differences between sequential deep learning models and bag-of-words classical ML models. The LSTM architecture processes email text as an ordered sequence of contextualised token embeddings, enabling it to capture the temporal structure of language and model long-range dependencies between distally positioned tokens in the email text. Phishing emails frequently exploit complex rhetorical structures—including urgency induction through narrative arc, authority impersonation through contextualised role language, and trust establishment through progressive disclosure—that are only detectable through analysis of sequential context, not through analysis of individual term frequencies.

The hybrid feature extraction approach—combining TF-IDF, word embeddings, URL features, and header features—was consistently superior to any single feature type in isolation. This finding highlights the complementary nature of different feature extraction strategies: TF-IDF captures surface lexical patterns, word embeddings encode semantic relationships, URL features capture structural deception indicators, and header features reveal email provenance manipulation. The integration of these complementary feature streams enables the model to simultaneously exploit multiple independent discriminative signals, improving both accuracy and robustness to adversarial evasion strategies that target individual feature types.

The false negative error analysis—revealing that the most commonly misclassified phishing emails are sophisticated spear-phishing attempts incorporating personalised content and legitimate-appearing vocabulary—has significant implications for operational deployment. Spear-phishing attacks, which are precisely personalised to individual recipients using information obtained through OSINT and social network analysis, represent the highest-risk category of phishing attack and are responsible for a disproportionate share of successful enterprise data breaches. The 97.6% recall achieved by the LSTM model on the general phishing population may mask substantially lower recall on this high-risk spear-phishing subcategory. Future work should specifically evaluate and optimise detection performance on this challenging subcategory.

The proposed framework demonstrates strong operational characteristics for real-world deployment. The modular pipeline architecture cleanly separates preprocessing, feature extraction, and classification concerns, enabling independent updating of individual pipeline components as new techniques become available. The preprocessing and feature extraction stages are computationally efficient and can process several hundred emails per second on standard server hardware. The LSTM classification stage, while more computationally intensive, achieves latency of approximately 12ms per email on GPU hardware, well within the operational requirements of real-time email security scanning systems that typically operate with latency budgets of 50-200ms per message.

The principal limitations of the proposed framework should be explicitly acknowledged. First, the substantial computational resources required for training and deploying deep learning models—particularly GPU hardware for LSTM inference at email-system scale—may present deployment barriers for resource-constrained organisations. In such contexts, the Random Forest classifier offers a compelling performance-efficiency trade-off, achieving 95.1% accuracy at a computational cost orders of magnitude lower than the LSTM model. Second, the performance of all evaluated models is contingent upon the availability of sufficiently large, diverse, and regularly updated labelled training datasets. Maintaining dataset currency in the face of rapidly evolving phishing techniques requires ongoing data collection and annotation infrastructure. Third, the model's reliance on textual content features creates a potential vulnerability to adversarial attacks specifically designed to manipulate the learned feature distributions while preserving the phishing functional intent.

The false positive rate of 4.2% achieved by the LSTM model, while substantially lower than the 15-20% false

positive rates typical of rule-based systems, still represents an operationally significant rate of legitimate email misclassification in high-volume enterprise email environments. At a volume of 10,000 emails per day, a 4.2% false positive rate would result in 420 legitimate emails per day being incorrectly quarantined or flagged, potentially disrupting business operations and eroding user trust in the security system. Improving precision while maintaining recall represents an important direction for future work.

5.1 Implications for Practice

The findings of this study have several concrete implications for cybersecurity practitioners seeking to deploy or improve email phishing detection systems. First, organisations with access to GPU computing infrastructure should prioritise deep learning approaches—specifically LSTM or transformer-based architectures—over classical ML methods, as the empirical evidence from this and prior studies consistently demonstrates meaningful accuracy improvements. Second, feature engineering should prioritise URL analysis and email header examination alongside textual content analysis, as these complementary feature streams provide independent discriminative signals. Third, detection systems should be implemented with continuous learning capabilities that enable model retraining on newly identified phishing samples, preventing performance degradation as attacker techniques evolve.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a comprehensive AI-driven phishing detection framework that systematically integrates Natural Language Processing and Machine Learning techniques to substantially improve upon the accuracy, robustness, and adaptability limitations of conventional phishing detection approaches. The proposed system implements a multi-stage text preprocessing pipeline, a hybrid feature extraction scheme combining TF-IDF vectorisation with pre-trained word embeddings and structured URL and header features, and a comparative evaluation of six machine learning architectures spanning from classical linear classifiers to deep learning sequential models.

Experimental evaluation on a combined dataset of 129,382 labelled emails demonstrates that the proposed LSTM-based hybrid NLP-ML model achieves 96.7% accuracy, 96.3% precision, 96.0% recall, and an F1-score of 96.1% on the held-out test partition—substantially outperforming both conventional rule-based blacklist systems (80-85% accuracy) and classical ML approaches trained on single feature types. The systematic comparative analysis reveals a consistent performance hierarchy with LSTM > Random Forest > SVM > Logistic Regression > Decision Tree > Naive Bayes, and demonstrates the consistent benefit of hybrid feature extraction over any individual feature type in isolation. The feature importance analysis identifies TF-IDF textual features and Word2Vec embeddings as the most informative feature categories, together accounting for 63.8% of total discriminative contribution.

These findings make several original contributions to the phishing detection literature: a rigorous comparative evaluation of six ML architectures within a unified experimental framework; empirical demonstration of the performance benefits of hybrid feature extraction; comprehensive feature importance analysis characterising the relative discriminative value of different feature categories; and identification of the spear-phishing subcategory as the primary remaining performance bottleneck for deep learning detection approaches.

Future research directions motivated by the findings of this study include: (i) the development and evaluation of transformer-based detection architectures, particularly BERT and RoBERTa, which offer the potential for further accuracy improvements through bidirectional contextual representation learning; (ii) the development of adversarial training procedures that explicitly incorporate adversarial perturbation examples designed to evade trained classifiers, improving robustness to evasion attacks; (iii) the investigation of domain adaptation techniques to improve model generalisation across diverse organizational email environments with distinct vocabulary distributions; (iv) the development of few-shot learning approaches to enable rapid adaptation to emerging phishing campaign patterns with minimal labelled examples; and (v) the design and evaluation of computationally efficient model compression and distillation techniques to enable deployment of high-accuracy detection in resource-constrained environments without GPU infrastructure.

The source code for all implemented models, the preprocessing pipeline, and the evaluation framework has been made publicly available to enable reproducibility and to facilitate extension by the research community. The authors intend to publish an updated evaluation incorporating transformer-based architectures and adversarial robustness analysis in a forthcoming companion paper.

References

1. A. K. Jain and B. B. Gupta, "Phishing detection: Analysis of visual similarity-based approaches," Security and Communication Networks, vol. 2017, Article ID 5421046, 2017.
2. S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: Detecting phishing with streaming analytics," IEEE Transactions on Network and Service Management, vol. 11, no. 4, pp. 458-471, 2014.
3. R. Verma and N. Hossain, "Semantic feature selection for text with application to phishing email detection," in Proceedings of the IEEE ICC, 2017.
4. J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to detect malicious URLs," ACM Transactions on Intelligent Systems, vol. 2, no. 3, Art. 30, 2011.
5. M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert Systems with Applications, vol. 37, no. 12, pp. 7913-7921, 2010.
6. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

7. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems (NeurIPS), 2013, pp. 3111-3119.
8. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. EMNLP, Doha, Qatar, 2014, pp. 1532-1543.
9. Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, Doha, Qatar, 2014, pp. 1746-1751.
10. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
12. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
13. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Upper Saddle River, NJ: Pearson, 2020.
14. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
15. Anti-Phishing Working Group (APWG), "Phishing Activity Trends Report, Q4 2023," APWG, Tech. Rep., 2024.
16. R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in Soft Computing Applications in Industry, vol. 226, 2008, pp. 373-383.
17. T. Moore and R. Clayton, "Examining the impact of website take-down on phishing," in Proc. APWG eCrime Researchers Summit, 2007.
18. G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.
19. Z. Zhang, J. Ma, J. Han, and X. Niu, "Email phishing detection by analysing natural language characteristics," in Proc. IEEE CCECE, 2007.
20. K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," Expert Systems with Applications, vol. 106, pp. 1-20, 2018.