Comparative Analysis of Machine Learning and Deep Learning Models for Spam Email Detection Toward Sustainable and Secure Digital Systems

Phuc Hau Nguyen
Faculty of Information Technology, Electric Power University, 129823 Hanoi, Vietnam
phuchauptit@gmail.com

Keywords: Email Spam Detection, Machine Learning, Deep Learning, Naive Bayes, SVM, BERT.

Abstract: This paper proposes a comparative spam email detection framework that integrates traditional machine learning and deep learning models with a blockchain-supported auditability layer to enhance digital security, transparency, and trustworthiness. The study evaluates six representative classification approaches, including Naive Bayes, Support Vector Machine, Random Forest, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT). A unified experimental workflow is implemented consisting of data preprocessing, feature extraction, model training, and comparative evaluation using Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. In addition to classification performance analysis, the framework incorporates a lightweight blockchain-based verification mechanism that records classification metadata and cryptographic hashes to ensure integrity and traceability of spam filtering results. Experimental findings demonstrate that deep learning models, particularly BERT and LSTM, achieve superior contextual understanding and higher detection accuracy, while traditional machine learning methods provide lower computational complexity and faster execution suitable for lightweight environments. The proposed framework contributes a reproducible benchmarking methodology for intelligent spam detection and demonstrates how blockchain-supported auditability can improve transparency and reliability in AI-driven cybersecurity systems.

1 INTRODUCTION

Email communication remains one of the most widely used digital services in modern society, supporting personal interaction, business operations, education, and electronic commerce. However, the rapid growth of Internet services has also led to a significant increase in unsolicited and malicious email traffic. Spam emails frequently contain phishing links, fraudulent advertisements, malware attachments, and social engineering content that threaten both cybersecurity and user privacy. According to recent cybersecurity reports, spam messages continue to represent a substantial proportion of global email traffic, creating financial losses, reducing productivity, and increasing computational overhead for digital infrastructures [1].

Traditional spam filtering systems are primarily based on supervised machine learning techniques such as Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF). These approaches typically rely on manually engineered textual features including word frequency, n-grams, and TF-IDF representations to distinguish spam from legitimate emails. Although such methods demonstrate acceptable performance and low computational complexity, they often struggle to capture semantic relationships and contextual dependencies in complex multilingual messages [2], [3].

Recent advances in deep learning have significantly improved natural language processing and text classification tasks. Architectures such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT) enable automatic extraction of semantic features and contextual information from textual data [4] - [6]. In particular, transformer-based models have demonstrated superior capability in understanding contextual meaning and identifying sophisticated spam patterns, including obfuscated content and multilingual phishing messages.

Despite these advances, most existing studies

focus primarily on improving classification accuracy while paying limited attention to transparency, traceability, and integrity of classification results. Modern enterprise email systems increasingly require not only accurate spam detection but also reliable auditing mechanisms capable of verifying classification outcomes in distributed environments. In this context, blockchain technology provides a promising solution due to its immutability, decentralized verification, and resistance to tampering.

This study proposes a comparative spam detection framework with blockchain-supported auditability that combines traditional machine learning and deep learning approaches within a unified experimental environment. The proposed framework integrates a lightweight blockchain-based verification layer that records classification metadata and cryptographic hashes to improve transparency and trustworthiness of filtering results without interfering with the learning process.

The main contributions of this study are summarized as follows:

1) Development of a unified experimental framework for comparative evaluation of machine learning and deep learning models for spam email detection;
2) Integration of a blockchain-supported auditability mechanism for ensuring integrity and traceability of classification results;
3) Comparative analysis of six representative models, including NB, SVM, RF, LSTM, CNN, and BERT, under identical preprocessing and evaluation conditions;
4) Performance evaluation using standardized metrics including Accuracy, Precision, Recall, F1-score, and ROC-AUC;
5) Investigation of the balance between classification performance, computational efficiency, and system transparency.

The overall architecture of the proposed framework is presented in Figure 1, while Table 1 illustrates an example of the dataset structure used during the experimental evaluation. The findings of this study contribute to the development of secure, scalable, and transparent intelligent spam filtering systems suitable for modern digital communication infrastructures.

2 RELATED WORK

2.1 Studies Using Traditional Machine Learning Models

Traditional machine learning (ML) methods remain the foundation of many email spam detection systems due to their simplicity, interpretability, and low computational requirements. Among these approaches, Naive Bayes (NB) is one of the earliest and most widely used classifiers in text categorization tasks. It assumes conditional independence between features and estimates the probability of an email belonging to the spam class using Bayes' theorem [2]. Despite its efficiency and scalability, NB is limited in capturing semantic relationships between words, which reduces its effectiveness in complex linguistic contexts.

Support Vector Machine (SVM) has been widely adopted as a more robust alternative to NB. It aims to find an optimal separating hyperplane between spam and non-spam classes by maximizing the margin between data points [3]. SVM generally performs well in high-dimensional feature spaces and demonstrates strong generalization ability; however, its computational cost increases significantly with large datasets and complex feature representations.

Random Forest (RF), as an ensemble learning method, improves classification stability by combining the outputs of multiple decision trees. This approach reduces overfitting and enhances predictive performance, particularly in imbalanced datasets. Nevertheless, RF models often suffer from limited interpretability when compared to simpler linear models.

Overall, traditional ML methods remain practical for lightweight spam filtering systems; however, their reliance on manually engineered features (e.g., TF-IDF, n-grams) restricts their ability to capture deep semantic information in textual data.

The main characteristics of traditional machine learning models, including their advantages and limitations, are summarized in Table 2.

Table 1: Example of the dataset used in this study.

ID	Email Subject	Simplified Content	Label
001	"You won an iPhone 17!"	"Click the link to claim your prize"	Spam
002	"Department meeting on Monday"	"Dear faculty members…"	Ham
003	"0% Bank Loan Offer"	"Register now to enjoy the promotion"	Spam
004	"Semester report submission"	"Reminder: please submit before May 30"	Ham

Table 2: Advantages and disadvantages of traditional ML models.

Model	Advantages	Disadvantages
Naive Bayes	Fast training, low resource requirement, effective on large datasets	Independence assumption limits semantic representation
SVM	Clear class separation, robust to noise	High computational cost, limited scalability
Random Forest	Strong generalization, effective on imbalanced data	Limited interpretability, sensitive to tree count

Table 3: Comparison between deep learning and traditional ML models.

Model Category	Representative Models	Feature Extraction	Contextual Understanding	Average Accuracy	Computational Cost
Traditional ML	NB, SVM, RF	Manual (TF - IDF, n-gram)	Limited	85-93%	Low
Deep Learning	LSTM, CNN, BERT	Automatic (Embeddings)	High	94-98%	High

2.2 Studies Using Deep Learning Models in NLP

The emergence of deep learning (DL) has significantly advanced spam detection by enabling automatic feature learning from raw text data. Recurrent architectures such as Long Short-Term Memory (LSTM) networks were designed to address the limitations of standard RNNs by introducing memory gates that capture long-term dependencies in sequences [6]. LSTM models are particularly effective in modeling contextual relationships in email content, improving detection accuracy for complex spam patterns.

Convolutional Neural Networks (CNN), originally developed for image processing, have also been successfully adapted for text classification tasks. CNN-based models extract local semantic features through convolutional filters, allowing them to identify meaningful n-gram patterns within text sequences [5]. These models typically achieve strong performance on benchmark datasets while maintaining moderate computational efficiency.

More recently, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have set new performance standards in natural language processing tasks. BERT leverages bidirectional self-attention mechanisms to capture deep contextual relationships within text, enabling superior understanding of language semantics [4]. In spam detection tasks, BERT consistently outperforms traditional ML and earlier DL models, particularly in cases involving obfuscated or context-dependent spam content.

A comparative overview of traditional machine learning and deep learning models is presented in Table 3, highlighting differences in feature extraction, contextual understanding, and computational cost.

2.3 Comparative Analysis and Research Gap

Existing studies generally confirm that deep learning models outperform traditional machine learning approaches in terms of accuracy and contextual understanding. However, this improvement often comes at the cost of increased computational complexity and resource consumption. In contrast, traditional ML models remain attractive for real-time and resource-constrained environments due to their efficiency and interpretability.

Despite extensive research in both directions, most existing works evaluate models in isolation and focus primarily on classification performance metrics. There is still a lack of unified comparative frameworks that assess both traditional and deep learning approaches under identical experimental conditions, especially when considering system-level requirements such as transparency, auditability, and security.

Furthermore, very few studies address the issue of result integrity and traceability in spam detection systems. This motivates the integration of blockchain-based mechanisms, which can provide immutable logging and verification of classification outputs, thereby enhancing trust in automated filtering systems.

3 METHODOLOGY

3.1 Data Description

The study utilizes three datasets: Enron Spam, Ling-Spam, and a self-collected Vietnamese email corpus. After data cleaning and duplicate removal, the final dataset contains 10,500 emails, including 5,200 legitimate messages and 5,300 spam samples. The statistical characteristics of the datasets are summarized in Table 4.

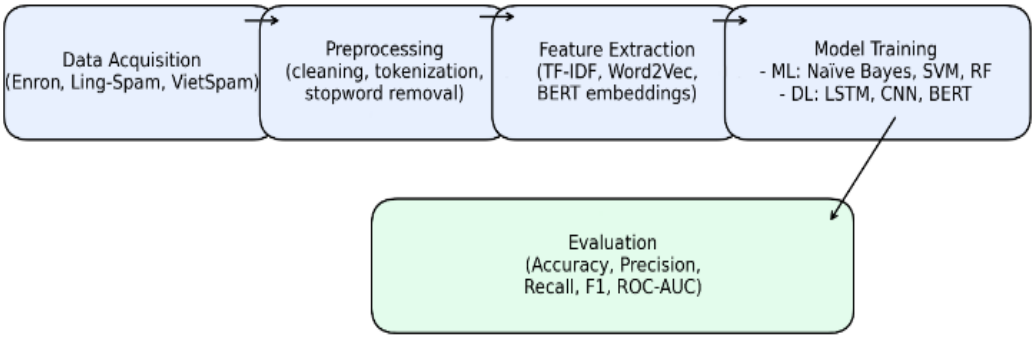

Figure 1: General workflow of the proposed system.

Table 4: Dataset statistics.

Dataset	Language	Emails	Spam (%)	Features (TF-IDF)
Enron	English	5,000	48,6	12,453
Ling-Spam	English	3,500	51,4	9,276
VietSpam	Vietnamese	2,000	53,0	10,101

The preprocessing stage includes several standard text-processing operations. First, all emails are normalized by converting text to lowercase and removing special characters, HTML tags, and URLs. Next, tokenization is performed using WordPiece for English texts and RDRSegmenter for Vietnamese data [7]. Stopword elimination is then applied using predefined language-specific stopword lists. Finally, textual information is transformed into numerical representations using TF-IDF vectorization for traditional machine learning approaches and semantic embeddings such as Word2Vec and BERT embeddings for deep learning models.

3.2 Experimental Models

The experimental framework includes six classification models representing both traditional machine learning and modern deep learning approaches:

1) Naive Bayes (NB) - a probabilistic classifier commonly applied in text categorization tasks;
2) Support Vector Machine (SVM) - a supervised learning method designed to determine optimal decision boundaries between classes;
3) Random Forest (RF) - an ensemble learning approach based on multiple decision trees and majority voting;
4) Long Short-Term Memory (LSTM) - a recurrent neural network architecture capable of capturing long-range contextual dependencies in textual sequences [6];
5) Convolutional Neural Network (CNN) - a deep learning model that extracts local semantic patterns from textual data through convolution operations;
6) Bidirectional Encoder Representations from Transformers (BERT) - a transformer-based language model fine-tuned for multilingual spam classification tasks [4].

The processing pipeline follows a sequential-parallel architecture designed to improve both classification accuracy and scalability. The workflow begins with raw email acquisition from benchmark datasets containing both spam and legitimate messages. During preprocessing, irrelevant textual noise such as special symbols, HTML tags, and stopwords is removed, while tokenization and normalization techniques are applied to standardize the data.

After preprocessing, the email content is converted into numerical feature representations using TF-IDF, Word2Vec, or BERT embeddings to capture semantic relationships within the text. The extracted features are then simultaneously processed by the six-machine learning and deep learning models [8]. Traditional models such as NB, SVM, and RF provide efficient baseline classification, whereas LSTM, CNN, and BERT enable deeper contextual and semantic analysis [4], [6].

In addition to the classification framework, a lightweight blockchain-based verification layer is conceptually integrated to improve data integrity and transparency. Each classified email is associated with a cryptographic hash value stored within a permissioned blockchain ledger. This mechanism ensures that classification outcomes remain tamper-

resistant and traceable in distributed environments.

Figure 2 presents the workflow of the proposed algorithm. The blockchain module functions independently from the training procedures and operates as a post-classification verification component. It records metadata such as prediction labels, timestamps, and content hashes to guarantee the integrity and auditability of classification results. A Practical Byzantine Fault Tolerance (PBFT)-inspired consensus mechanism is assumed to maintain low latency suitable for real-time spam filtering systems.

```
Algorithm 1: Blockchain-Integrated Spam Detection Pipeline
Input:
   Raw email dataset D = {e1, e2, ..., en}
Output:
   Predicted labels Y = {y1, y2, ..., yn} with blockchain-verified records
Initialize preprocessing modules and tokenizers
Initialize ML models: NB, SVM, RF
Initialize DL models: LSTM, CNN, BERT
Initialize blockchain ledger B and hash function H(·)
for each email ei ∈ D do
    // Preprocessing
    Clean text (remove HTML, special characters, URLs)
    Normalize text (lowercasing, tokenization)
    Remove stopwords
    // Feature Extraction
    if ML model then
        xi ← TF-IDF(ei)
    else
        xi ← Embedding(ei)  // Word2Vec or BERT embedding
    end if
    // Model Inference (parallel execution)
    y_NB   ← NB.predict(xi)
    y_SVM  ← SVM.predict(xi)
    y_RF   ← RF.predict(xi)
    y_LSTM ← LSTM.predict(xi)
    y_CNN  ← CNN.predict(xi)
    y_BERT ← BERT.predict(xi)
    // Model Selection / Aggregation
    yi ← SelectBest(y_NB, y_SVM, y_RF, y_LSTM, y_CNN, y_BERT)
    // Blockchain-based verification
    hi ← H(ei || yi || timestamp)
    Record transaction Ti = {hi, yi, timestamp}
    Append Ti to blockchain ledger B
end for
return Y = {y1, y2, ..., yn}
```

Figure 2: Workflow diagram of Algorithm 1.

Due to computational limitations, the current study focuses primarily on architectural integration and conceptual feasibility rather than large-scale deployment. Nevertheless, the proposed design establishes a foundation for secure, scalable, and transparent spam detection systems applicable to distributed enterprise infrastructures.

3.3 Blockchain-Based Auditability Layer

In addition to the machine learning and deep learning classification pipeline, this study integrates a lightweight blockchain-based auditability layer to enhance transparency, integrity, and traceability of spam detection results. Unlike traditional spam filtering systems that store only final labels, the proposed framework records essential metadata for each processed email in an immutable ledger.

Specifically, after the classification stage, each email instance is transformed into a cryptographic hash derived from its content. Together with the predicted label, timestamp, and model identifier, this hash is stored as a transaction in a permissioned blockchain network. This design ensures that any modification of email content or classification output can be easily detected, as even minor changes in input data produce a different hash value.

The blockchain module operates independently from the machine learning models and does not influence the training or inference process. Instead, it functions as a post-classification verification mechanism that guarantees auditability of results. A Practical Byzantine Fault Tolerance (PBFT)-inspired consensus mechanism is assumed to maintain consistency across distributed nodes while ensuring low latency suitable for real-time email filtering systems.

Due to computational and scalability constraints, the current implementation focuses on conceptual integration rather than full-scale deployment. However, the proposed architecture demonstrates how blockchain technology can be effectively combined with AI-based spam detection systems to provide verifiable and tamper-resistant classification logs. This approach is particularly relevant for enterprise environments where accountability, compliance, and data integrity are critical requirements.

3.4 Evaluation Metrics

The performance of all models is evaluated using five widely adopted classification metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC [9]. These metrics provide a comprehensive assessment of classification effectiveness, balancing predictive correctness, detection capability, and robustness against false positives and false negatives.

To ensure statistical reliability and minimize overfitting, the experiments are conducted using 10-fold cross-validation [10]. This validation strategy enables consistent benchmarking across both traditional machine learning and deep learning approaches, facilitating the identification of the most effective configuration for multilingual spam detection tasks.

4 RESULTS AND DISCUSSION

4.1 Experimental Results

The experimental evaluation compares the performance of six classification models, including three traditional machine learning methods (Naive Bayes, SVM, Random Forest) and three deep learning models (LSTM, CNN, BERT). All models were trained and tested under identical preprocessing and feature extraction conditions to ensure fair comparison. The results are summarized in Table 5 and visualized in Figure 3 and Figure 4.

The obtained results demonstrate a clear performance hierarchy between traditional and deep learning approaches. Among classical models, Random Forest achieves the highest accuracy (94.2%), followed by SVM (93.6%) and Naive Bayes (91.1%). These models show stable performance but are limited in capturing deep semantic relationships in email text.

Deep learning models significantly outperform traditional approaches. LSTM achieves 95.7% accuracy, CNN reaches 96.3%, while BERT obtains the highest performance with 97.8% accuracy and 0.993 ROC-AUC. These results confirm the superiority of transformer-based architectures in capturing contextual and semantic dependencies in textual data.

Figure 3 illustrates the comparative accuracy of all models, clearly showing the performance gap between classical and deep learning methods. Figure 4 further presents ROC curves, where deep learning models demonstrate stronger classification separability, especially in low false-positive regions.

4.2 Blockchain-Based Auditability Evaluation

To evaluate the feasibility of the proposed blockchain-supported auditability layer, a simulated analysis was conducted to estimate system overhead. The results indicate that the integration of blockchain introduces an additional processing latency of approximately 6-9% per email due to hash generation, transaction packaging, and consensus operations.

However, this overhead does not affect classification accuracy, as the blockchain module operates independently from the machine learning pipeline. Instead, it ensures that all classification outputs are immutably recorded, providing traceability and resistance to tampering. This feature is particularly important for enterprise and security-critical environments where auditability of automated decisions is required.

4.3 Discussion

The comparative analysis highlights a fundamental trade-off between performance and computational efficiency. Traditional machine learning models offer fast training and low resource consumption, making them suitable for lightweight or real-time email filtering systems [11]. However, their limited ability to capture semantic context restricts their effectiveness on complex or obfuscated spam messages.

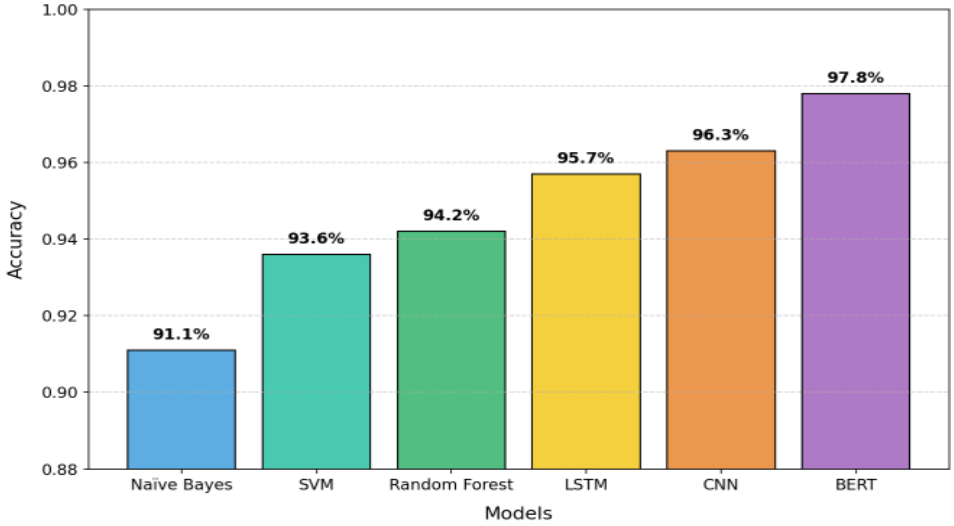

Figure 3: Accuracy comparison among models.

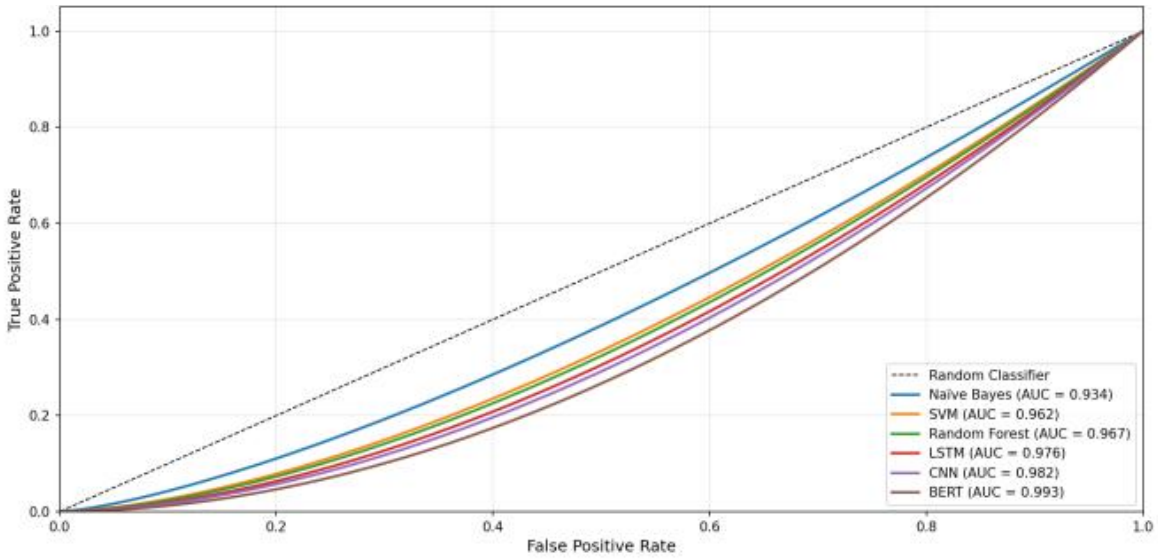

Figure 4: ROC curves for all models.

Table 5: Comparative performance of models.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Naive Bayes	0.911	0.893	0.873	0.883	0.934
SVM	0.936	0.921	0.907	0.914	0.962
Random Forest	0.942	0.935	0.918	0.926	0.967
LSTM	0.957	0.949	0.935	0.942	0.976
CNN	0.963	0.953	0.947	0.950	0.982
BERT	0.978	0.972	0.965	0.968	0.993

In contrast, deep learning models demonstrate superior performance due to their ability to learn contextual representations automatically. BERT, in particular, shows the highest robustness against advanced spam techniques such as word manipulation and multilingual phishing content.

The integration of blockchain adds a new dimension to spam detection systems by introducing transparency, integrity, and auditability. While it introduces additional computational overhead, the trade-off is justified in scenarios where trust and traceability are critical requirements.

Overall, the results suggest that future spam detection systems should not rely solely on accuracy-oriented optimization but should also consider system-level properties such as security, transparency, and verifiability. A hybrid architecture combining deep learning models with lightweight blockchain-based auditing represents a promising direction for next-generation intelligent cybersecurity systems.

6 CONCLUSIONS

This study presented a comprehensive comparative framework for spam email detection based on traditional machine learning and deep learning models, enhanced with a blockchain-supported auditability mechanism. The proposed system was designed to evaluate not only classification performance but also transparency, integrity, and traceability of prediction results in modern digital environments.

Experimental results demonstrated that deep learning models significantly outperform traditional machine learning approaches in terms of classification accuracy and semantic understanding. Among all tested models, BERT achieved the highest performance (97.8% accuracy and 0.993 ROC-AUC), followed closely by CNN and LSTM. Traditional models such as Naive Bayes, SVM, and Random Forest showed competitive efficiency and faster execution but were limited in capturing complex contextual relationships in email content.

In addition to performance evaluation, the study introduced a blockchain-based auditability layer that records classification metadata and cryptographic hashes in an immutable ledger. This mechanism ensures traceability and tamper resistance of spam detection outcomes without interfering with the learning process. Although it introduces a moderate computational overhead, the trade-off is acceptable in security-sensitive and enterprise-level applications where accountability is critical.

The main contribution of this work is the integration of machine learning, deep learning, and blockchain technologies into a unified spam detection framework that balances accuracy, efficiency, and trustworthiness. The findings highlight that future intelligent spam filtering systems should move beyond purely performance-driven optimization and incorporate transparency and verifiability as core design principles.

Future work will focus on extending the framework to large-scale real-world deployments, optimizing blockchain consensus mechanisms for lower latency, and exploring federated learning approaches to further enhance privacy and scalability in distributed email security systems.

REFERENCES

[1] Kaspersky Lab, “Spam and Phishing in Q4 2024,” Technical Report, 2024.

[2] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos and C. D. Spyropoulos, “An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering,” in Proceedings of ECML, 2000, pp. 9-17.

[3] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” in Proceedings of ECML, 1998, pp. 137-142.

[4] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of NAACL-HLT, 2019, pp. 4171-4186.

[5] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in Proceedings of EMNLP, 2014, pp. 1746-1751.

[6] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[7] H. Nguyen and A. Le, “RDRSegmenter: A Robust Vietnamese Word Segmentation Algorithm,” in Proceedings of PACLIC, 2016, pp. 265-272.

[8] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proceedings of ICLR, 2015.

[9] D. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation,” Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011, [Online]. Available: https://doi.org/10.48550/arXiv.2010.16061.

[10] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in Proceedings of IJCAI, 1995, pp. 1137-1145.

[11] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, U.K.: Packt, 2019.