Spam Detection Using an Advanced Hybrid Model

Tahany Kmail, Marah Hawa, and Ahmad Hasasneh *

Department of Natural, Engineering, and Technology Sciences, Faculty of Graduate Studies,
Arab American University (AAUP), Ramallah, Palestine
Email: t.kmial@student.aaup.edu (T.K.); m.hawa1@student.aaup.edu (M.H.); Ahmad.Hasasneh@aaup.edu (A.H.)
*Corresponding author

***Abstract*—Emails are now extensively used across diverse domains, including business and education. However, the growing prevalence of spam poses a persistent challenge for users, leading to wasted time, resource consumption, and compromised data privacy. As spam volumes continue to rise, traditional detection techniques such as blacklists and content-based filters are proving increasingly insufficient against the evolving sophistication of fraudulent tactics. To address this issue, this study introduces a novel hybrid model that integrates a Transformer, a Multilayer Perceptron (MLP), and Bidirectional Encoder Representations from Transformers (BERT). This model is distinguished by its ability to capture the rich contextual nuances of communication, enhance classification accuracy, and detect complex patterns within lengthy texts. Its robustness and capacity to generalize to new threats were validated using a large and diverse dataset. The results indicate that the proposed model effectively balances precision and adaptability, outperforming previous approaches that often relied on limited datasets or exhibited poor generalization. Beyond serving as a classification tool, the model functions as an integrated system capable of continuous updates, making it a practical solution for improving the security of modern email systems. It achieves a high accuracy rate of 94%. In addition to underscoring the value of hybrid and advanced models in combating spam, this study provides a solid foundation for future research aimed at increasing effectiveness, improving adaptability, and minimizing the adverse impacts of spam on users and organizations.**

***Keywords*—spam, hybrid model, Multilayer Perceptron (MLP), transformers, classification**

I. Introduction

Electronic messages have become an integral part of daily life, greatly enhancing communication between individuals and institutions. Since the advent of email in the mid-1990s [1], it has revolutionized various sectors, including business, healthcare, education, and industry. However, this growing dependence on email has also given rise to new cybersecurity threats.

The most prominent threats include spam attacks, malware, and various forms of electronic exploitation. Studies show that spam constitutes over 50% [2] of global email traffic, highlighting the considerable challenges users face in managing such messages. Although traditional spam detection methods—such as blacklists and keyword-based filtering—have been widely used, their effectiveness has declined due to the increasingly sophisticated techniques employed by spammers to bypass these defenses. Consequently, there has been growing interest in Machine Learning (ML) and Deep Learning (DL) approaches, which are capable of recognizing complex patterns and adapting to evolving spam strategies. Over the past two decades, researchers have proposed a range of techniques for distinguishing spam from ham messages, including Real-Time Blackhole Lists [3], blacklists [4], and content-based filters [5]. Nonetheless, research continues to focus on developing more accurate and efficient solutions. In recent years, there has been a notable shift toward Artificial Intelligence (AI)-based approaches for spam detection. The successful application of ML techniques in this domain is well established, and the advancement of DL methods has further strengthened detection capabilities.

These studies demonstrate that ML and DL provide effective frameworks for addressing spam detection; however, they still face challenges such as managing false positives and false negatives [6–8]. In this study, a hybrid model is proposed that combines Transformer, Multilayer Perceptron (MLP), and Bidirectional Encoder Representations from Transformers (BERT) techniques to improve the accuracy of spam detection. The model leverages deep text analysis and captures contextual patterns, enabling it to better adapt to evolving email threats. By evaluating the model on a large and diverse dataset, the study aims to strike an optimal balance between detection accuracy and performance efficiency, making it well-suited for real-world email systems. This research contributes to the enhancement of electronic security systems by developing more accurate and effective methods for spam detection, while also offering a deeper understanding of ML and DL techniques in this domain.

This section provides an overview of the growing volume of spam, particularly in email systems. The structure of the paper is organized as follows: Section II reviews related work; Section III outlines the proposed

Manuscript received April 10, 2025; revised May 6, 2025; accepted June 11, 2025; published September 19, 2025.

methodology; Section IV presents the results; Section V discusses the findings; and Section VI concludes the study.

II. Literature Review

Beginning in 2015, researchers began to recognize the limitations of traditional systems in handling the growing volume of spam and the increasingly sophisticated tactics employed by attackers, such as content manipulation and obfuscation. Consequently, research shifted toward ML and DL techniques that emphasize content analysis and feature extraction [9].

A. Machine Learning Methods

In the field of spam detection, ML techniques have been widely applied to achieve high accuracy in message classification. ML systems depend on training data to extract patterns and build models capable of generalizing and predicting new messages. According to the literature, methods such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forests (RF), Logistic Regression (LR), Decision Tree (DT), and k-Nearest Neighbors (k-NN) have been employed with notable success, achieving high accuracy rates across numerous studies [10]. For example, Cota and Zinca [11] demonstrated that applying RF to various datasets yielded strong performance in spam message classification. Despite these achievements, several challenges persist, including false positives, instability in feature extraction, and high computational complexity. The dynamic and evolving nature of spam techniques further necessitates the development of more adaptive and flexible approaches. This underscores the importance of exploring hybrid or ensemble models that combine multiple algorithms to improve accuracy and minimize false positives [12]. Additionally, adversarial attacks—where spammers use misleading words or insert ham terms to deceive classifiers present a significant obstacle. Such manipulative content can degrade the quality and performance of ML models over time [13]. Based on prior studies summarized in Table I, it is evident that researchers place considerable value on ML techniques due to their proven effectiveness in detecting spam messages.

B. Deep Learning Techniques

Deep learning is one of the most advanced techniques used for spam detection. It employs Deep Neural Networks (DNNs) with MLPs to extract features from data and analyze complex patterns. One of the key advantages of DL is its ability to automatically learn important features without the need for manual feature engineering. This adaptability allows it to effectively handle new messages and evolving threats. Kamilaris and Prenafeta-Boldú [14] highlighted the broader impact of deep learning across various domains, particularly in agriculture, showcasing its strength in handling image-based classification tasks. Deng and Yu [15] classified DL techniques into three main categories: supervised, unsupervised, and hybrid, and reviewed their various applications in text processing and information retrieval [14]. Networks such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and DNNs are used in spam classification, where they help distinguish between ham and spam messages through deep text analysis. Baccouche *et al*. [16] proposed an LSTM model for spam analysis on social media platforms, while AbdulNabi and Yaseen [17] improved the BERT model by comparing it with traditional ML methods such as NB and k-NN, which led to enhanced classification efficiency. In addition, Abdullahi and Kaya [18] proposed a deep learning-based method using DNN for detecting email and SMS spam. Their study compared the performance of the DNN model with traditional machine learning classifiers such as SVM, Naïve Bayes, Decision Tree, Random Forest, and Logistic Regression, demonstrating the superiority of the deep learning approach. Rafat *et al*. [19] also discussed the role of text preprocessing in improving classification accuracy using the SpamAssassin dataset, concluding that DL outperforms traditional ML techniques. Despite the significant progress in spam detection using DL, challenges remain, including the need to enhance model interpretability, test models on extensive real-world datasets, and improve their adaptability to increasingly sophisticated spam messages.

C. Hybrid Techniques

To overcome the limitations of individual models, recent research has focused on combining ML and DL techniques or integrating DL with fuzzy analysis methods. Gazal and Juneja [20] introduced a hybrid model that merges pre-filtering with a fuzzy composite evaluator, demonstrating superior performance compared to many standalone techniques. Alauthman [21] proposed a hybrid botnet spam detection model that integrates Random Forest with LSTM. Similarly, Srinivasarao and Sharaff [22] introduced a Fuzzy-Based Recurrent Neural Network combined with Harris Hawk Optimization (FRNN-HHO) for the post-classification of spam and ham messages. Their proposed architecture was evaluated using three distinct datasets: SMS, Email, and SpamAssassin. The method achieved high accuracy on both the SMS and Email datasets.

A common issue in many models is bias toward the majority class (ham). Since the number of ham messages is significantly higher, the model may become less effective in real-world environments where data is more balanced. Although the model might appear accurate, the disproportionately small representation of the minority class can limit its suitability for certain practical applications [23, 24]. Existing approaches to spam detection still exhibit knowledge gaps, as most prior studies rely on relatively small datasets [25], often sourced from email or SMS platforms. For example, some studies used limited datasets, such as 5171 emails in Ref. [26] and 1360 spam versus 4360 regular messages in Ref. [27]. The limited size of these datasets may hinder the generalizability of the proposed models' outcomes. This study addresses this gap by utilizing a large, integrated dataset containing 148,746 samples, enhancing data representativeness and improving the generalizability of the results across diverse real-world scenarios. Additionally, the paper adopts a hybrid approach

combining BERT and MLP, an approach that has been rarely explored in previous studies, particularly in the context of large-scale datasets.

Unlike traditional studies that rely on classical machine learning algorithms such as SVM, NB, and K-NN, we developed a more advanced framework based on hybrid techniques that combine DL and traditional learning. The key features of this research are: 1) Designing a hybrid model that combines BERT and MLP: While most previous studies used BERT or MLP separately, we combined these two techniques to extract the best features from each. BERT enables the extraction of in-depth and accurate features from texts, while MLP allows for highly efficient classification. 2) Improved model generalization: By significantly increasing the size of the dataset used, the model is no longer constrained by small, context-limited datasets. It is now capable of handling diverse real-world environments, enhancing its accuracy and effectiveness across various scenarios. 3) Extensive comparison with traditional models: Not only did we develop a new model, but we also compared it with several traditional models such as SVM, NB, and K-NN, demonstrating the superior performance of our approach, especially when dealing with large and imbalanced datasets. 4) Reducing ham bias: A common problem with spam classification models is their tendency to classify ham messages more accurately due to the uneven distribution of data. In our study, we optimized data balancing and model tuning strategies to deliver fairer and more efficient performance in real-world settings.

Overall, this study contributes to expanding the scope of scientific research in the field of spam detection by 1) providing a more advanced framework that combines DL and traditional learning techniques to achieve higher performance. 2) conducting an in-depth comparative analysis that demonstrates the superiority of the proposed model over traditional methods. 3) Utilizing a larger and more representative dataset that reflects real-world environments, thereby making the results more generalizable. This contribution aims not only to improve the accuracy of spam detection models but also to make a fundamental impact on detection methodologies by offering solutions that are more adaptable to modern challenges. It advances the frontiers of scientific research in the domains of text processing and spam analysis more effectively.

There are notable gaps in current spam detection approaches, primarily due to the reliance on relatively small datasets, often sourced from email or SMS platforms. For instance, some studies utilized limited datasets, such as 5171 emails in Ref. [26] and 1360 spam versus 4360 regular messages in Ref. [27]. Such limited dataset sizes can hinder the generalizability of the proposed models' outcomes.

TABLE I. MACHINE LEARNING METHOD FOR EMAIL SPAM

References	Dataset	Method	Accuracy (%)
[23]	Lings Spam	k-Nearest Neighbors (K-NN), Support Vector Machine (SVM)	98
[24]	Lings Spam	K-NN	98.06
[25]	Spam base	K-NN, Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB) and SVM	98.09
[26]	From Kaggle (contains 5171 email)	SVM, Random Forests (RF), and NB	98.41
[27]	From Kaggle (contains 4360 ham samples, and 1368 spam samples)	K-NN, SVM, DT, LR, RF and NB	99
[28]	Turkish Emails, CSDMC 2010, and Enron	SVM, LR, and NB	98.91
[29]	CSDMC 2010	K-NN, SVM, DT, RF, NB, and AdaBoost	95.97

In this study, we address this limitation by employing a large, integrated dataset containing 148,746 samples, which enhances data representativeness and improves the generalizability of the results across diverse real-world scenarios. Furthermore, the study introduces a hybrid approach that combines BERT and MLP—an approach rarely explored in prior research, particularly in the context of large-scale datasets. The primary innovation lies in integrating the BERT text processing model with an MLP to improve classification accuracy, achieving up to 94%, thereby outperforming many traditional models cited in the literature. The study also addresses the bias commonly found in traditional models toward the majority class (ham) by balancing the dataset and refining evaluation strategies, resulting in a model that is more effective in real-world environments with varied data distributions.

Overall, this study contributes to expanding the scope of scientific research in the field of spam detection by:

1) Providing a larger and more diverse dataset, which enhances the model's generalizability to real-world scenarios. 2) Employing advanced hybrid techniques (BERT and MLP), which are rare in this field, to improve classification accuracy. 3) Developing a more efficient framework for large-scale spam processing, pushing the boundaries of research in the application of DL techniques for text analysis.

III. MATERIALS AND METHODS

This section illustrates the proposed methodology, as shown in Fig. 1, to build a hybrid model for classifying spam and ham emails. The process begins with data collection and labeling, where a large body of data related to spam and ham in English is collected. The data is then preprocessed by formatting and filtering the data to ensure its quality and alignment for analysis. This is followed by the trait extraction phase, where advanced techniques are used to extract features and features related to spam, ham from the data. Next, the models are created using Bert and MLP techniques, and these models are very strongly analyzed to detect spam emails. In the final stage, performance is evaluated using specific metrics such as accuracy and F1-score to ensure the effectiveness in detecting spam on social media platforms.

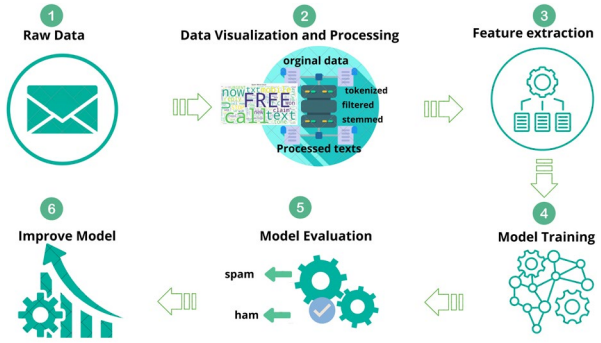


Fig. 1. Proposed methodology.

A. Dataset Description

The dataset used in this study was collected from the Kaggle platform and comprises two large, high-quality corpora. The first dataset, titled "Spam and ham Emails" [30], contains 9990 labeled samples divided into two binary classes: spam and ham. The second dataset [31] includes 138,756 email samples, also categorized into two binary classes: spam (1) and ham (0). These datasets were chosen for their relevance, size, and diversity, making them suitable for training and evaluating deep learning models for spam detection. The final merged dataset consists of labeled samples where "1" denotes spam (i.e., emails containing offensive or irrelevant promotional content), and "0" denotes ham emails. Before training, several preprocessing steps were applied to standardize the inputs: text normalization, lowercasing, removal of punctuation and HTML tags, elimination of duplicates, and label unification across the two datasets. This preprocessing ensured clean and consistent inputs compatible with the BERT tokenizer. Although the datasets are diverse and substantial, potential limitations such as class imbalance and domain specificity (e.g., email structure and language style) were acknowledged. These factors were addressed during training by applying appropriate techniques, such as balanced batch sampling, to mitigate bias. The data used is illustrated in Fig. 2.

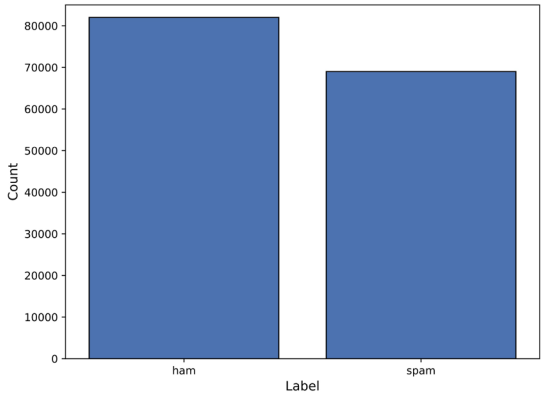
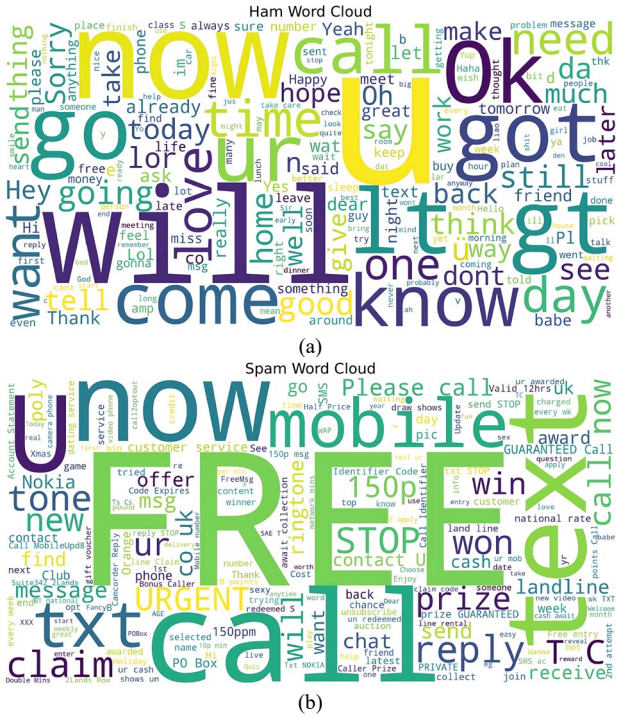


Fig. 2. Distribution of the message.

B. Data Visualization and Pre-processing

For spam and ham emails to know the most frequent words, this is demonstrated by Fig. 3, showing the word sizes, where big words are used a lot.



(a)

(b)

Fig. 3. Word cloud (a) positive; (b) negative.

Pre-processing is a crucial step in ML and DL techniques since it prepares the dataset for model training by cleaning and preparing it. In this research, the study has applied the Natural Language Processing (NLP) technology to address several email-related problems. It was used as depicted in Fig. 4.

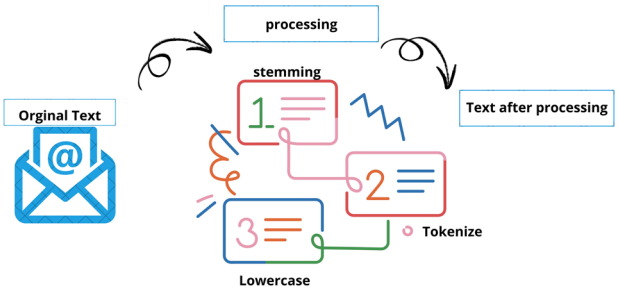


Fig. 4. Data preprocessing.

C. Embedding Models

1) Bidirectional Encoder Representations from Transformers (BERT)

BERT is a deep language model designed to extract numerical representations of text that capture the full meaning and context of words. In this study, email texts were processed using BERT to generate embeddings for each email. These high-quality numerical representations were then used as inputs for a classification model. BERT enhances the understanding of subtle meanings in emails, even when they contain similar or ambiguous words. Consequently, BERT improved the accuracy of classifying emails into "Spam" and "ham" [32].

2) Multi-layer perceptron

One of the fundamental artificial neural networks used in various classification tasks is the MLP. Multilayer

Perceptron consists of an input layer that receives data, hidden layers that process patterns and extract features, and an output layer that produces the final classification. In this study, numerical representations (embeddings) extracted from BERT were used as input to the MLP, which classifies emails as "Spam" or "ham". MLP is characterized by its ability to learn complex patterns in textual data, which contributed to achieving strong classification results [33].

3) Hybrid model (BERT + MLP)

The proposed hybrid model combines BERT to extract deep semantic features from texts and an MLP to classify data based on the extracted representations. Mathematically, each text input x, x is transformed into a numerical representation $E(x)$, $E(x)$ in Eq. (1) [32] using the BERT model, where the representation is calculated as follows in Eq. (1):

$$E(x) = BERT(X) \in \mathbb{R}^d \quad (1)$$

where d represents the dimensionality of the vector representation generated by the last layer of BERT. This representation is then passed to a MLP neural network, which consists of multiple hidden layers with a nonlinear activation function, such as ReLU, as shown in Eq. (2) [33]:

$$hi = ReLU(Wi.hi - 1 + bi) \quad (2)$$

where Wi and bi in Eq. (2) represent the weights and biases for each layer, which are adjusted during the training process using the backpropagation algorithm. In the final layer, as shown in Eq. (3), a SoftMax function is used to generate the probability of classifying the sample as either "Spam" or "ham", as described in Eq. (3) [33] below:

$$P\left(\frac{Y}{X}\right) = softmax(w_{out}.h_{final} + b_{out}) \quad (3)$$

This structure enhances the model's ability to understand linguistic context with greater accuracy compared to traditional algorithms such as SVM and NB, thereby contributing to improved classification performance when dealing with large and diverse datasets.

Additionally, Fig. 5 below visually illustrates the technological infrastructure of the proposed model. It begins with a textual input, which is passed to BERT to be converted into a digital representation. These features are then forwarded to the MLP network, where three layers—input, hidden, and output—are used to determine the most appropriate class. The figure also demonstrates how the training and classification processes are carried out, enhancing the interpretability and reproducibility of the proposed model. So, the integration of deep contextual analysis provided by BERT with classification performed by MLP enables robust text processing. The flow from raw text input to AI-powered classification accuracy is clearly outlined. This allows users and researchers to understand the key stages of the processing pipeline, facilitating the development of a future-proof model. Additionally, it supports the adoption of the model in traditional environments that demand high performance and transparent interpretation of results.

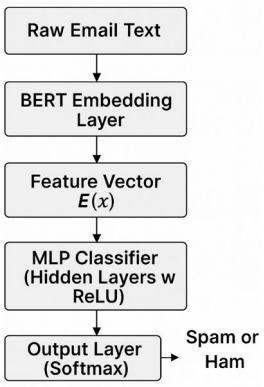


Fig. 5. Data flow for the proposed model.

Fig. 5 illustrates the data flow in the proposed model. Text inputs are processed by BERT to extract semantic features, which are then passed to a multi-layer MLP network for final classification.

D. Model Generation and Evaluation

In this study, Python was used to develop a model for data classification. The model was evaluated using several performance metrics, including accuracy, precision, recall, and F1-score [34]. These metrics were calculated using the equations presented and based on the values shown in Table II.

TABLE II. THE FORMULA OF EVALUATION MEASURES

Evaluation Measure	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1	$\frac{2 \times (precision \times recall)}{precision + recall}$

TP denotes the True Positives, TN the True Negatives, FP the False Positives, and FN the False Negatives.

E. Model Improvement

In this study, a hybrid model combining BERT and MLP is developed to classify emails as either "Spam" or "ham". Initially, the email texts are processed using a BERT tokenizer, which converts them into input IDs and attention masks. These tokenized representations are then passed into the model. BERT, a pre-trained deep language model, analyzes the text and generates high-quality numerical embeddings that capture the semantic meaning of the content. Specifically, the representation of the Classification token (CLS), which encapsulates the overall meaning of the input, is extracted. This embedding is then fed into a MLP, composed of several densely connected layers with ReLU activation functions and dropout layers

to help mitigate overfitting. The final output layer generates a single probability score indicating the likelihood that the email is classified as "Spam". This integration of BERT and MLP enables the model to harness BERT's deep contextual understanding of language alongside the MLP's capability to learn and generalize categorical patterns effectively. Table III presents the hyperparameters used in the hybrid model.

TABLE III. HYPERPARAMETERS OF THE HYBRID MODEL

Hyper parameter	value
Dropout Rate	0.3
Activation Function	Sigmoid
Optimizer	Adam
Loss Function	binary_crossentropy

IV. RESULTS

Experiments were conducted to analyze the performance of models used to classify texts using ML and DL algorithms, while modifying parameters to improve accuracy. These experiments aimed to compare the effectiveness of different models and to select the most appropriate one based on the available data. A hybrid model was developed using BERT and MLP technologies to classify emails into "spam" and "ham" categories, using a large dataset of approximately 149,000 samples. The model demonstrated excellent classification performance, achieving an accuracy of up to 94%. In this section, performance results will be presented in detail based on several metrics.

A. Key Performance Metrics

Several metrics were used to evaluate the model's performance comprehensively. The results are as follows in Table IV.

TABLE IV. THE RESULT OF MODEL PERFORMANCE

Measures	Value
Accuracy	0→94% 1→94%
Precision	0→94% 1→95%
Recall	0→95% 1→94%
F1_score	0→95% 1→94%

In Table IV, the values "0" and "1" denote the class labels used in the classification task. Specifically, "0" represents ham emails, while "1" corresponds to spam emails. This binary representation simplifies the evaluation of model performance. The results illustrate the model's effectiveness in distinguishing between the "Spam" and "ham" classes while maintaining a strong balance between precision and recall. The hybrid model (BERT & MLP) demonstrated excellent performance in email classification, achieving an accuracy of 94% on a dataset comprising 148,746 samples. Compared to traditional models, it provides detailed metrics for each category, including precision, recall, and F1-score. The model achieved a precision of 94% for the Spam category, meaning that 94% of the emails classified as spam were correctly identified. Similarly, recall was 94%, indicating that the model successfully detected 94% of all actual spam emails. The F1-score, which harmonizes precision and recall, was also 94%, reflecting the model's consistent ability to detect spam while minimizing false positives. These findings highlight the robustness of the model in processing large-scale textual data and underscore its strong generalization capability, maintaining high accuracy across diverse input scenarios.

B. Confusion Matrix

The confusion matrix in Fig. 6 illustrates the model's classification performance across different classes. It shows the counts of correct and incorrect predictions for each class, offering a clear view of the types of errors the model may make. This visual tool is crucial for analyzing classification errors and their distribution. In this study, the confusion matrix revealed that the model correctly classified 1007 ham messages while misclassifying 51 ham messages as spam. Conversely, the model accurately identified 881 spam messages but failed to recognize and misclassified 59 spam messages as ham. These results indicate balanced performance, with a low error rate and strong capability to detect spam messages—the primary objective of the classification system.

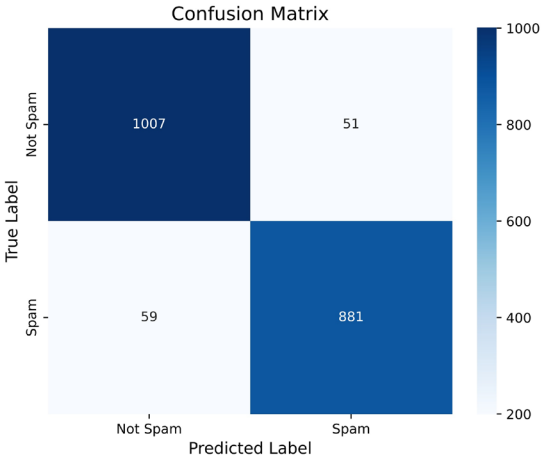


Fig. 6. Confusion matrix.

C. ROC Curve

The Receiver Operating Characteristic (ROC) curve is a key tool for evaluating the performance of classification systems, as it illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across all possible decision thresholds. In this study, the proposed model achieved an Area Under the Curve (AUC) of 99%, demonstrating a strong ability to distinguish between spam and ham messages. Notably, the AUC represents the model's overall discriminative performance, independent of any specific classification threshold. In contrast, metrics such as precision, recall, and accuracy are calculated at a fixed threshold, typically set at 0.5. Therefore, the slight difference between the high AUC (99%) and the operational metrics (precision and recall at 94%) is expected and highlights the effect of

threshold selection on real-world performance. Fig. 7 presents the ROC curve.

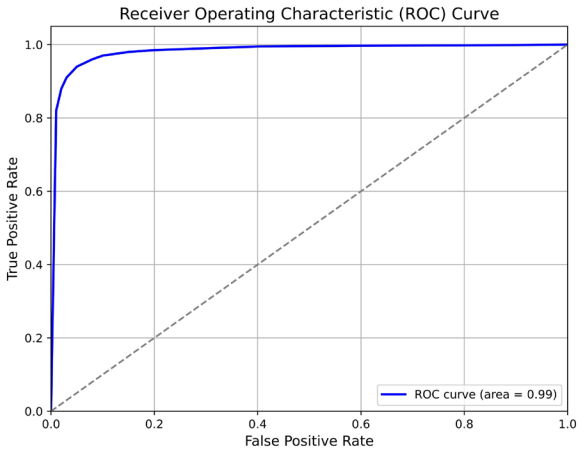


Fig. 7. ROC curve for the hybrid model.

D. Analysis of Model Performance Comparison with Previous Studies in Spam Classification

When comparing the model with previous studies, it is evident that the use of BERT + MLP resulted in improved accuracy compared to the best-performing traditional models. The hybrid model also demonstrated superior performance when handling larger and more balanced datasets, as well as in managing imbalanced datasets. In our experiment, we utilized the Enron dataset [35], as shown in Table V, which contained imbalanced data, thereby increasing the complexity of the spam classification task.

TABLE V. A COMPARISON BETWEEN THE PROPOSED SYSTEM AND OTHERS IN CURRENT STUDIES

Algorithm	ROC-AUC (%)	Accuracy (%)	F1-Score AVG (%)	Recall Spam (%)	Recall ham (%)
SVM	Not mentioned	95.97	95	95.33	95
Glove + MLP	Not mentioned	96	96	94	94
Our proposed	99	95	96	96	96

Our proposed algorithm (MLP + BERT) demonstrated robust performance and outperformed the (GloVe + MLP) model when evaluated on the same Enron dataset. Although the traditional model achieved an accuracy of 96%, this result is influenced by data imbalance, as the algorithm tends to classify most messages as "important" due to the dominance of that class within the dataset. While conventional models such as SVM have shown solid performance, the hybrid model surpassed them across all major evaluation metrics, particularly in terms of F1-Score and recall for both the spam and ham classes.

To emphasize the effectiveness of the proposed model, a direct comparison was made with the highest-performing traditional algorithms reported in previous literature, as shown in Table VI. Ref. [27] reports the best accuracy of 99%, achieved using a combination of traditional algorithms such as K-NN, SVM, and Random Forest on a modified dataset from the Kaggle platform. In contrast, the current model, based on transformer technologies (BERT and MLP), achieved an accuracy of 94%, but its real advantage lies in achieving an ROC-AUC score of 99%—a qualitative achievement not previously recorded in any other study. This metric reflects the model's exceptional ability to distinguish between spam and ham messages, regardless of data balance.

TABLE VI. A COMPARISON BETWEEN THE PROPOSED MODEL AND THE STUDY IN REF. [27]

Approach	Dataset	Accuracy (%)	ROC-AUC (%)	Notes
[27]	Kaggle	99	Not mentioned	High accuracy on modified data, customized dataset, and ensemble methods.
Proposed approach	Enron	94	99	Advanced transformer model with high performance, real dataset.

Although some studies have reported relatively higher accuracy, the proposed model has the advantage of generalizability when tested on real-world data such as the Enron dataset [35], making it more robust and suitable for practical applications than traditional models, which may be overly tailored to specific environments.

In turn, the proposed algorithm takes advantage of the deep BERT representation, which captures the full context of each word in a sentence, enhancing the ability to distinguish between plain text and spam even in complex situations. Importantly, the algorithm addressed the issue of bias toward the larger category (ham) by balancing the model during training, resulting in a well-balanced precision and recall between the two categories. Although the model's accuracy was relatively close (95%), the proposed algorithm excelled in more realistic evaluation metrics, such as ROC-AUC (99%), demonstrating its high generalizability in real-world environments, particularly in information security applications and email filtering systems.

E. Statistical Validation of Model Performance

To evaluate the statistical significance of the performance improvement achieved by the proposed hybrid model, a paired t-test was conducted. The proposed model combines a Transformer-based architecture, and its performance was compared against a baseline model implemented using a simple RNN architecture without any enhancements. Both models were trained on the same dataset under identical conditions, and the classification accuracy was recorded over 27 training epochs. The results of the statistical test are as follows:

T-statistic = 22.17;

p-value = 2.07×10^{-18}.

These values indicate a substantial and statistically significant difference in performance, confirming that the improvement is not due to random variation. The extremely low p-value ($p < 0.05$) supports the robustness of the proposed model and validates its superiority over the

baseline architecture. Table VII shows the statistical validation for the proposed model and the baseline model.

TABLE VII. A COMPARISON BETWEEN THE PROPOSED MODEL AND THE BASELINE MODEL

Model	Average Accuracy	Notes
RNN Baseline	86%	Simple RNN Without enhancement
Our proposed	94%	Transformers-based Hybrid model
p-value	2.07e−18	Statistically significant

V. DISCUSSION

Our experiments demonstrated that the proposed model, trained on a large dataset of 148,746 emails, achieved a classification accuracy of 94% in distinguishing important messages from spam. This result strongly highlights the model's effectiveness in handling diverse, large-scale data, enhancing its potential for practical applications in improving email systems.

In this paper, several advanced spam classification models using DL techniques were studied, and the results demonstrated high performance in accuracy and other metrics. However, when analyzing the limitations of previous studies, several points can be identified to improve future models. For example, although the use of a BERT model has improved classification efficiency on the Spam Base dataset [36], it requires long input sequences, which increases computational complexity and training time. This issue can be addressed by applying dimensionality reduction techniques or feature selection strategies to reduce the input size while maintaining quality. On the other hand, CNN-based models have shown good performance in identifying random content on Twitter [36], but the complexity of the network structure may hinder deployment speed in real-world environments. This problem can be mitigated by applying model compression techniques or relying on lighter networks. For the LSTM model used with the WEBSPAM-2007 dataset [37], the need to continuously improve the algorithm to handle large amounts of data remains a challenge. Improved gradient algorithms or hybrid models that reduce computational load can be explored. The study that used Turkish data [34] achieved ideal accuracy, but the small sample size may limit the model's generalizability on a wider scale. Therefore, it is advisable to collect more diverse data to enhance the model's stability. Finally, in our study, we used a hybrid system that combined BERT and MLP, where the two models worked together to classify spam with very high accuracy. After training on a large and balanced dataset of 148,746 samples, the model achieved an accuracy of 94%. This large amount of data helped improve the model's ability to generalize and detect diverse patterns in messages, demonstrating the effectiveness of the hybrid solution in addressing the challenges mentioned above. Furthermore, recent studies have shown a clear diversity in algorithmic approaches used to detect spam in different types of data, whether SMS or emails. For example, the LSTM model was used in Ref. [38] to improve the accuracy of SMS classification [39]. Although these studies achieved high accuracy rates ranging from 95.3% to 97.6%, some limitations should be considered. One such limitation is that the models rely on limited datasets, which may reduce their ability to generalize when dealing with messages that are more complex or structurally different from those used in training. In addition, reducing the number of hidden layers in Ref. [39], while it helps reduce training time, may sometimes affect the depth of the representation extracted from the texts and make the model more sensitive to linguistic noise. Similarly, the study in Ref. [40] focused on improving the hyperparameter tuning process of LSTM models using diverse datasets such as Spam Base and Ling-Spam, reaching an accuracy of 98.9%. However, these approaches rely on intensive parameter tuning, which may require significant human expertise, increase training time, and make it difficult to generalize the model on a large scale unless automated hyperparameter optimization methods are employed. The study in Ref. [41] employed word embedding in a CNN model to achieve 97.1% accuracy on a dataset of self-generated emails. Although CNNs demonstrated efficiency in extracting spatial features of texts, using self-generated data may not accurately reflect the true diversity of actual spam messages, which may lead to a lack of generalizability to different environments or new types of attacks. Table VIII shows a comparison between the proposed algorithms and previous studies.

TABLE VIII. COMPARATIVE EVALUATION WITH PRIOR STUDIES

Approach	Dataset	Accuracy (%)	Model used	Dataset(size)
[23]	Ling's spam	98	KNN, SVM	Small data
[24]	Ling's spam	98.06	KNN	Small data
[25]	Spam Base	98.09	KNN, DT, SVM, NB, LR	Small data
[26]	Kaggle	98.41	SVM, RF, NB	5171 Email
[27]	Kaggle	99	KNN, DT, SVM, NB, LR, RF	5728 Email
[28]	Turkish Email Enron, CSDMC 2010	98.91	SVM, NB, LR	Small data
[29]	CSDMC 2010	95.97	KNN, DT, SVM, NB, RF, AdaBoost	Small data
Our Proposed Approach	Integrated data set	94	BERT + MLP	148,746 Email

In general, it can be concluded that previous studies have made significant progress in classifying spam messages; however, several common challenges remain inadequately addressed. Most notable are the size and diversity of the datasets used, the nature of the models requiring precise parameter tuning, and the potential decline in model performance when confronted with new or sophisticated types of spam messages. Therefore, incorporating broader DL strategies, employing modern natural language processing methods, and creating more diverse datasets represent promising directions to overcome these limitations and enhance the accuracy and reliability of models in the long term. Our model demonstrates superior generalization ability due to its

training on a large dataset, compared to most studies that relied on smaller datasets.

Although some studies have reported higher accuracy rates (up to 99%), these results may be misleading because they often use small and imbalanced datasets, which can limit the model's effectiveness when applied to new, unseen data. Our study addresses the data imbalance problem more effectively by utilizing BERT and MLP, whereas previous studies mainly focused on traditional algorithms.

Our model demonstrates superior generalization capability, primarily due to its training on a large and diverse dataset, unlike many prior studies that relied on smaller, often imbalanced datasets. While some studies have reported higher accuracy scores (e.g., 99%), such results may be misleading, as they are typically achieved on limited and skewed data, which compromises the model's ability to perform effectively on unseen data. In contrast, our study addresses the data imbalance issue by integrating BERT and MLP, whereas earlier research predominantly focused on traditional ML algorithms. Furthermore, the results of the paired t-test confirm the statistical significance of the observed performance improvements, reinforcing the effectiveness of incorporating Transformer-based architectures, particularly in enhancing model generalization and classification accuracy. These findings are consistent with and extend previous research that highlights the advantages of attention-based models in capturing complex patterns. Fig. 8 presents a comparison of accuracy between our model and those in prior studies.

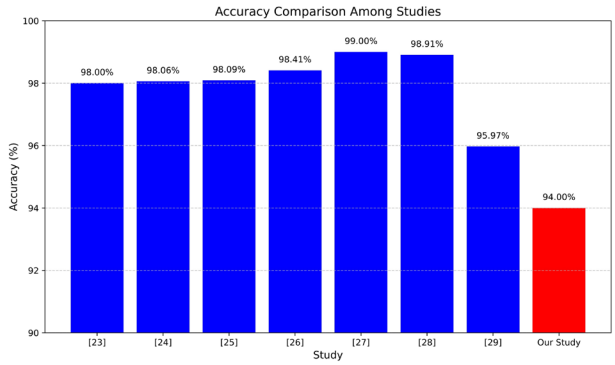


Fig. 8. Accuracy comparison between our study and the previous study.

Fig. 8 presents a comparative analysis of model accuracy across various studies, including our proposed framework. While some previous studies, such as Ref. [27], achieved slightly higher accuracy (99%), it is important to note that these models were trained on significantly smaller datasets. For example, the dataset used in Ref. [26] contained only 5171 emails, whereas our study leveraged a much larger dataset of 148,746 samples, ensuring a more representative a generalizable model.

Moreover, traditional ML algorithms such as K-NN, SVM, and NB were predominantly employed in prior studies. Although effective for smaller datasets, their performance may degrade when applied to large-scale, real-world data. In contrast, our approach incorporates a hybrid DL model combining BERT and MLP, a strategy not extensively explored in large-scale spam classification. Despite achieving a slightly lower accuracy (94%), our model offers several advantages:

(1) Better generalization due to a larger and more diverse dataset.
(2) Hybrid architecture that leverages Dl techniques rather than relying solely on traditional ML models.
(3) Improved real-world applicability, as models trained on small datasets often fail to perform well on unseen data.

Additionally, data imbalance remains a critical factor in spam classification. Some prior studies used highly balanced datasets, which simplifies the classification task. However, our dataset presents a more realistic distribution, providing a better benchmark for evaluating real-world spam detection systems. Future work could explore advanced techniques such as data augmentation and cost-sensitive learning to further improve performance.

VI. Conclusions

In this study, we developed a hybrid system (BERT & MLP) for classifying spam and ham emails, focusing on overcoming challenges related to data scarcity and imbalance by using a large, balanced dataset. This approach achieved a high accuracy of 94%, outperforming several previous studies that faced data limitations or relied on less advanced models. This superiority reflects the hybrid system's effectiveness in capturing complex textual patterns and its adaptability to varying message characteristics. These results highlight the importance of adopting hybrid ML and DL approaches and pave the way for further research aimed at developing more accurate and reliable systems for spam message classification, addressing the ongoing challenges in this field. However, the absence of comparisons with commercial spam filtering systems (such as Gmail or Outlook) represents a limitation of this study, as it may affect the assessment of the model's real-world performance. Nonetheless, the results highlight the importance of adopting hybrid approaches in ML and DL, paving the way for further research toward developing more accurate and reliable email spam classification systems in light of the ongoing challenges in this domain. It is worth noting that the proposed hybrid model distinguishes itself from current industry standards by combining the advanced contextual understanding of BERT with the high adaptability of MLP. This architecture enables the model to effectively learn from dynamic data and adapt to emerging patterns. In contrast to many commercial systems that rely on closed architectures or static rule-based mechanisms, our model offers a transparent and flexible alternative that can be deployed across a variety of applications, positioning it as an academically grounded innovation with strong practical potential.

Conflict of Interest

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

TK led the research work, including the design of the methodology, implementation of experiments, and writing the initial draft of the manuscript; MH contributed to data collection, preprocessing, and result analysis; AH supervised the research, provided critical feedback, and contributed to the review and refinement of the final manuscript; all authors had approved the final version.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. Ahmad Hasasneh for his valuable support and guidance throughout this research. The authors also acknowledge the Arab American University for providing the necessary resources and academic environment to complete this work.

REFERENCES

[1] K. Deshpande, J. Girkar, and R. Mangrulkar, "Security enhancement and analysis of images using a novel Sudoku-based encryption algorithm," *Journal of Information and Telecommunication*, vol. 7, no. 3, pp. 270–303, 2023. doi: 10.1080/24751839.2023.2183802

[2] E. G. Dada, J. S. Bassi, H. Chiroma *et al.*, "Machine learning for email spam filtering: Review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, e01802, 2019. doi: 10.1016/j.heliyon.2019.e01802

[3] S. Dhanaraj and V. Karthikeyani, "A study on e-mail image spam filtering techniques," in *Proc. of the 2013 International Conf. on Pattern Recognition, Informatics and Mobile Engineering*, Salem, 2013, pp. 49–55. doi: 10.1109/ICPRIME.2013.6496446

[4] A. Bhowmick and S. M. Hazarika, "Machine learning for e-mail spam filtering: Review, techniques and trends," arXiv preprint, arXiv:1606.01042, 2016.

[5] C. Laorden, X. Ugarte-Pedrero, I. Santos *et al.*, "Study on the effectiveness of anomaly detection for spam filtering," *Inf Sci.*, vol. 277, pp. 421–444, 2014. doi: 10.1016/J.INS.2014.02.114

[6] S. Zavrak and S. Yilmaz, "Email spam detection using hierarchical attention hybrid deep learning method," *Expert Syst Appl.*, vol. 233, 120977, 2023. doi: 10.1016/J.ESWA.2023.120977

[7] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," *Future Generation Computer Systems*, vol. 102, pp. 524–533, 2020. doi: 10.1016/J.FUTURE.2019.09.001

[8] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning," *Computer Networks*, vol. 206, 108826, 2022. doi: 10.1016/J.COMNET.2022.108826

[9] A. A. Abdo, K. Alhajri, A. Alyami *et al*., "AI-based spam detection techniques for online social networks: Challenges and opportunities," *Journal of Internet Services and Information Security*, pp. 78–103, 2023. doi: 10.58346/JISIS.2023.I3.006

[10] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A systematic review of deep learning techniques for phishing email detection," *Electronics*, vol. 13, no. 19, 3823, 2024. https://doi.org/10.3390/electronics13193823

[11] R. P. Cota and D. Zinca, "Comparative results of spam email detection using machine learning algorithms," in *Proc. 2022 14th International Conf. on Communications (COMM)*, 2022. doi: 10.1109/COMM54429.2022.9817305

[12] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002. doi: 10.1145/505282.505283

[13] C. Bansal and B. Sidhu, "Machine learning based hybrid approach for email spam detection," in *Proc. 2021 9th International Conf. on Reliability, Infocom Technologies and Optimization*, 2021. doi: 10.1109/ICRITO51393.2021.9596149

[14] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput Electron Agric*, vol. 147, pp. 70–90, 2018. doi: 10.1016/J.COMPAG.2018.02.016

[15] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014. doi: 10.1561/2000000039

[16] A. Baccouche, S. Ahmed, D. Sierra-Sosa *et al*., "Malicious text identification: Deep learning from public comments and emails," *Information 2020*, vol. 11, no. 6, 312, 2020. doi: 10.3390/INFO11060312

[17] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," *Procedia Comput Sci*, vol. 184, pp. 853–858, 2021. doi: 10.1016/J.PROCS.2021.03.107

[18] A. A. Abdullahi and M. Kaya, "A deep learning based method to detect email and SMS spams," in *Proc. 2021 International Conf. on Decision Aid Sciences and Application (DASA)*, 2021, pp. 430–435. doi: 10.1109/DASA53625.2021.9681921

[19] K. F. Rafat, Q. Xin, A. R. Javed *et al*., "Evading obscure communication from spam emails," *Mathematical Biosciences and Engineering*, vol. 19, no. 2, pp. 1926–1943, 2022. doi: 10.3934/MBE.2022091

[20] Gazal and K. Juneja, "Two-phase fuzzy feature-filter based hybrid model for spam classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 10339–10355, 2022. doi: 10.1016/J.JKSUCI.2022.10.025

[21] M. Alauthman, "Botnet spam e-mail detection using deep recurrent neural network," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, 2020. doi: 10.30534/ijeter/2020/83852020

[22] U. Srinivasarao and A. Sharaff, "SMS sentiment classification using an evolutionary optimization based fuzzy recurrent neural network," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42207–42238, 2023. doi: 10.1007/s11042-023-15206-2

[23] L. N. Vejendla, B. Bysani, A. Mundru *et al.*, "Score-based support vector machine for spam mail detection," in *Proc. 2023 7th International Conf. on Trends in Electronics and Informatics (ICOEI)*, 2023, pp. 915–920. doi: 10.1109/ICOEI56765.2023.10125718

[24] T. Georgieva-Trifonova, "Research on filtering feature selection methods for e-mail spam detection by applying K-NN classifier," in *Proc. 2022 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2022. doi: 10.1109/HORA55278.2022.9799999

[25] P. Thakur, K. Joshi, P. Thakral *et al.*, "Detection of email spam using machine learning algorithms: A comparative study," in *Proc. 2022 8th International Conf. on Signal Processing and Communication (ICSC)*, 2022, pp. 349–352. doi: 10.1109/ICSC56524.2022.10009149

[26] V. Sunjaya, S. Senjaya, J. Utama *et al.*, "Content-based spam classifying algorithms in email," in *Proc. 2022 3rd International Conf. on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future (AiDAS),* 2022, pp. 305–310. doi: 10.1109/AIDAS56890.2022.9918759

[27] Y. Kontsewaya, E. Antonov, and A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," *Procedia Comput Sci*, vol. 190, pp. 479–486, 2021. doi: 10.1016/J.PROCS.2021.06.056

[28] B. K. Dedeturk and B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Appl Soft Comput.*, vol. 91, 106229, 2020. doi: 10.1016/J.ASOC.2020.106229

[29] N. Saidani, K. Adi, and M. S. Allili, "A semantic-based classification approach for an enhanced spam detection," *Comput Secur.*, vol. 94, 101716, 2020. doi: 10.1016/J.COSE.2020.101716

[30] Bagavathypriya. Spam ham dataset. *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/bagavathypriya/spam-ham-dataset

[31] Z. Mustafa. Spam and ham classification balanced dataset. *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/zubairmustafa/spam-and-ham-classification-balanced-dataset

[32] J. Devlin, M. W. Chang, K. Lee *et al*., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018.

[33] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos Environ*, vol. 32, no. 14–15, pp. 2627–2636, 1998. doi: 10.1016/S1352-2310(97)00447-0

[34] E. H. Tusher, M. A. Ismail, M. A. Rahman *et al*., "Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3467996

[35] J. Al Rashid. Enron spam email detection PyTorch. *Kaggle*. [Online]. Available: https://www.kaggle.com/code/japeralrashid/enron-spam-email-detection-pytorch

[36] Z. Alom, B. Carminati, and E. Ferrari, "A deep learning model for Twitter spam detection," *Online Soc Netw Media*, vol. 18, 100079, 2020. doi: 10.1016/J.OSNEM.2020.100079

[37] A. Makkar and N. Kumar, "An efficient deep learning-based scheme for web spam detection in IoT environment," *Future Generation Computer Systems*, vol. 108, pp. 467–487, 2020. doi: 10.1016/J.FUTURE.2020.03.004

[38] H. Yang, Q. Liu, S. Zhou *et al*., "A spam filtering method based on multi-modal fusion," *Applied Sciences*, vol. 9, no. 6, 1152, 2019. doi: 10.3390/APP9061152

[39] S. Gadde, A. Lakshmanarao, and S. Satyanarayana, "SMS spam detection using machine learning and deep learning techniques," in *Proc. 2021 7th International Conf. on Advanced Computing and Communication Systems (ICACCS),* 2021, pp. 358–362. doi: 10.1109/ICACCS51430.2021.9441783

[40] F. Wei and T. Nguyen, "A lightweight deep neural model for SMS spam detection," in *Proc. 2020 International Symposium on Networks, Computers and Communications (ISNCC)*, 2020. doi: 10.1109/ISNCC49221.2020.9297350

[41] V. S. Vinitha, D. K. Renuka, and L. A. Kumar, "Long short-term memory networks for email spam classification," in *Proc. the 2023 International Conf. on Intelligent Systems for Communication, IoT and Security (ICISCoIS),* 2023, pp. 176–180. doi: 10.1109/ICISCOIS56541.2023.10100445