IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

Hybrid Deep Learning and Transformer-Based Approach for Accurate Email Spam Classification

Dhanashri Shukla[1], Suyash Shrivastava[2]

Student, Dept of Computer Science and Engineering, SR Institute of Management and Technology, Lucknow[1]

Head of Department, Computer Science and Engineering, SR Institute of Management and Technology, Lucknow[2]

Abstract: Due to the rapid increase of unsolicited and malevolent messages, the problem of spam detection in email has become a serious issue. In this paper, we propose a comprehensive multi-model framework for spam classification integrating traditional machine learning, deep learning and transformer-based approaches. The system uses feature-based models such as TF-IDF with Logistic Regression and XgBoost, sequence based model like LSTM,CNN,BiLSTM with Attention, transformer-based DistilBERT models Experiments on a labeled email dataset demonstrate that CNN model is most performant with accuracy 98% and F1-score 0.96, BiLSTM with Attention and TF-IDF with XGBoost about 97 percent accurate. Transformer based models also provide competitive results with approx 96% accuracy. The outcome of the models underscores that hybrid and attention-based architectures are critical in improving classification performance, resisting attacks, and adjusting to changing patterns of spam.

Keywords: Email Spam Detection, Machine Learning, Deep Learning, CNN, BiLSTM with Attention, XGBoost, BERT, Text Classification, Natural Language Processing, Hybrid Models

I. INTRODUCTION

Email has continued to be one of the most popular communication platforms in personal and professional settings. Nevertheless, its active development has resulted in an increase in spam mail significantly, which is usually malicious in nature, including phishing links, advertisements, and other malware. Such undesirable messages not only decrease the productivity of users but also cause grave [1] cybersecurity risks. Conventional rule-based filtering systems based on predefined patterns and matching of keywords are no longer effective against the current spam techniques that keep on changing and advancing.

The main problem with email spam detection is how to successfully differentiate between legitimate (ham) and spam emails and deal with such problems as high data dimensionality, class [2] imbalance and dynamic spam patterns. Naive Bayes, Support Vector Machines and Logistic Regression are conventional machine learning models that deliver moderate performance but do not tend to identify intricate contextual [3] and semantic connections in textual data. Equally, standalone deep learning models such as LSTM can be highly susceptible to long-term dependencies and have massive computational needs.

This work fills the gap by proposing a holistic multi-model framework consisting of classical machine learning, deep learning, [4] and transformer-based models to maximize classification performance. The proposed method will implement a combination of TF-IDF based models with advanced architecture like CNN, BiLSTM with Attention and DistilBERT to utilize hybrid features that consist of both statistical features as well as contextual semantic representation. This hybrid approach helps in making the spam detection systems more resilient and dynamic over ever-changing threats. This paper makes the following main contributions: (1) we propose a unified multi-model architecture integrating machine learning, deep learning, and transformer-based methods; (2) we implement advanced preprocessing techniques and feature extraction methods to optimize data representation; (3) we provide a complete evaluation of performance indicators such as accuracy, precision, recall, F1-score, ROC-AUC and PR-AUC; (4) for hybrid models and attention-based ones to demonstrate their effectiveness by performing more comparative analysis with CNN achieving state-of-the-art 98% classification accuracy.

The rest of this paper is organized as follows. Section II covers the literature review of spam detection techniques. Section III presents the proposed methodology, which consists of system architecture, algorithm and hyperparameter tuning.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154100

Section IV deals with implementation and experimental results, including performance comparisons. Section V presents the conclusions of this paper and summarizes future research directions.

II. LITERATURE REVIEW

Email ranks among the most important communication tools in our everyday life; however, this also resulted in a huge surge of spam emails, and to the cybersecurity risks like phishing, malware propagation, and data stealing. As a significant portion of email traffic worldwide is found to be spam, rule-based approaches for the detection of spam emails are expected to be futile against sophisticated and adaptive attacks for spam [4], [18]. A solution that has started to gain attention in the research community` is using machine learning based techniques and artificial intelligence for spam detection and getting better robustness in systems.

Traditional studies in spam detection applied basic machine learning methods to get started, including Naivete Bayes, Support Vector Machines (SVM), Decision Trees, Logistic Regression and Random Forest. These techniques analyze the properties of a text inside the emails and their statistical values to determine if an email is spam or not. SVM has been found to work well for classification use cases and other models like Random Forest have also demonstrated better generalizability with lesser overfitting [7], [11]. These models no matter how effective, are susceptible to changes in spam patterns, noise and distortions of data; and so need more sophisticated techniques [10].

To solve the above problems, improved and hybridized machine learning methods have been suggested. Integration of Logistic Regression with Decision Trees is an example where the complexity and noise in data are reduced before classification, thus improving system performance [10]. Similarly, methods that integrate SMOTE with Genetic Programming have shown to properly handle imbalanced data and achieve better scores in terms of recall and precision [5]. Feature selection techniques have been developed, especially for its application in educational institutions, to enhance the performance and efficiency of the system [8]. Approaches that emphasize these aspects have been shown to improve spam classification systems and underline the importance of data preprocessing and feature engineering techniques.

Due to the increasing popularity of deep learning methods, scholars have resorted to the application of neural networks to detect spam. Long Short-Term Memory (LSTM) networks and other architectures capable of learning the context of messages and the order in which they are received in email to comprehend complex patterns better can be trained [3], [14]. Moreover, models based on transformers like BERT and DistilBERT have shown the state of the art performance with some of them even reaching more than 99 percent accuracy as they are able to learn the correlation between semantic and contextual properties of the text [6], [13]. They are much better than traditional methods especially when dealing with large and complicated data.

There are also recent studies conducted around hybrid deep models that integrate various feature extraction methods for enhanced detection accuracy [1]. For instance, a system that combines convolutional neural networks (CNN) for text recognition and discrete Fourier transform (DFT) techniques for frequency-based pattern recognition to detect lexical and obfuscated-based features of the spam. This framework also uses dynamic feature integration and dynamic thresholding with XGBoost for improved detection accuracy and reduced false negatives of obfuscated spam messages [1]. These advancements reflect the shift towards multidimensional analysis and adaptive feature selection in spam filtering.

One of the huge advancements in this space is with privacy-preserving & decentralized learning techniques. This implies that many LSTM models may be trained using multiple devices via federated learning, which is guaranteed not to share sensitive user information between training devices so as to improve both the security and performance of the platforms [2]. Additionally, adaptive learning frameworks and blockchain-based verification pathways have been identified as potential solutions for enhancing trust and resilience against both spoofing attacks and adversarial attacks [9].

Also, various machine learning methods such as tuned K-Nearest Neighbors (KNN) with TF-IDF and PCA have been tried to enhance feature representation while increasing classification accuracy [17]. Likewise, state of the art ensemble methods like XGBoost exceed other classifiers with sufficiently competitive precision and recall scores over benchmark datasets [15]. These approaches are based on combining a compact representation of the data with a powerful classifier. This is the evolution of email spam detection from classical models to current AI-based, advanced models that can render mail in the face of complexity and be dynamic. While traditional machine learning methods served as the basis, recent works applying deep learning, hybrid models and federated systems have greatly improved identification accuracy and agnosticism. However, new methods of spam detection, data privacy issues, and computational approaches have created continuous opportunities that need to evolve within this area [16].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

III. METHODOLOGY

3.1 Proposed System Architecture

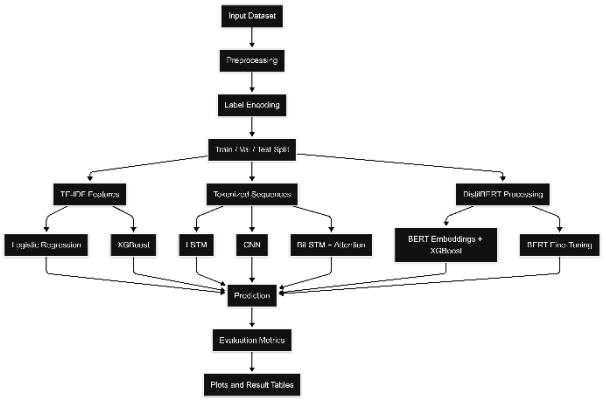

Figure 1: Architecture of the Multi-Model Email Spam Classification System

Here is the proposed spam detection approach illustrated in figure 1. It starts with the input of a dataset, which is then preprocessed and label encoded. It is then divided into training, validation and test sets. A series of parallel pipelines are then implemented, such as TF-IDF and machine learning, deep learning sequence models, and DistilBERT. These predictions are merged and then evaluated with performance measures and plots and tables of results.

3.2 Algorithm

Algorithm 1: Multi-Model Email Spam Classification Framework

Input

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where

- x_i: email text
- $y_i \in \{0,1\}$: class label (0 = ham, 1 = spam)
- N: total number of samples

Output

Predicted labels $\hat{y}_i$, probability scores p_i, and evaluation metrics:

$$\{\text{Accuracy}, \text{Precision}, \text{Recall}, F1, \text{ROC-AUC}, \text{PR-AUC}\}$$

Step 1: Text Preprocessing

Each input email is transformed as:

$$\tilde{x}_i = f_{\text{clean}}(x_i)$$

where

- $f_{\text{clean}}(\cdot)$: preprocessing function
- removes noise, URLs, emails, special characters
- converts text to lowercase

Step 2: Label Encoding

$$y_i = \begin{cases} 0, & \text{if ham} \\ 1, & \text{if spam} \end{cases}$$

Step 3: Dataset Splitting

$$\mathcal{D} \rightarrow \{\mathcal{D}_{tr}, \mathcal{D}_{val}, \mathcal{D}_{te}\}$$

where

- $\mathcal{D}_{tr}$: training set
- $\mathcal{D}_{val}$: validation set

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

- $\mathcal{D}_{te}$: testing set

Step 4: TF-IDF Feature Extraction

$$\mathbf{v}_i = \text{TF-IDF}(\tilde{x}_i)$$

where

- $\mathbf{v}_i \in \mathbb{R}^d$: feature vector
- d: vocabulary size

Step 5: Logistic Regression Model

$$p_i^{(LR)} = \sigma(\mathbf{w}^T \mathbf{v}_i + b)$$

where

- $\sigma(z) = \frac{1}{1+e^{-z}}$: sigmoid function
- $\mathbf{w}$: weight vector
- b: bias term

Step 6: XGBoost Model

$$p_i^{(XGB)} = \sum_{k=1}^{K} f_k(\mathbf{v}_i)$$

where

- f_k: decision tree function
- K: number of trees

Step 7: Tokenization and Padding

$$\mathbf{s}_i = \text{Pad}(\text{Tokenize}(\tilde{x}_i), L)$$

where

- $\mathbf{s}_i$: sequence vector
- L: maximum sequence length

Step 8: Embedding Layer

$$\mathbf{e}_i = \text{Embedding}(\mathbf{s}_i)$$

where

- $\mathbf{e}_i \in \mathbb{R}^{L \times d_e}$
- d_e: embedding dimension

Step 9: LSTM Model

$$\mathbf{h}_i = \text{LSTM}(\mathbf{e}_i)$$

$$p_i^{(LSTM)} = \sigma(\mathbf{W}_h \mathbf{h}_i + b_h)$$

where

- $\mathbf{h}_i$: hidden state vector

Step 10: CNN Model

$$\mathbf{c}_i = \text{Conv1D}(\mathbf{e}_i)$$

$$\mathbf{z}_i = \max(\mathbf{c}_i)$$

$$p_i^{(CNN)} = \sigma(\mathbf{W}_z \mathbf{z}_i + b_z)$$

where

- $\max(\cdot)$: global max pooling

Step 11: BiLSTM with Attention

$$\mathbf{H}_i = \text{BiLSTM}(\mathbf{e}_i)$$

Attention weights:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

where

$$e_t = \tanh\left(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a\right)$$

Context vector:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \, \mathbf{h}_t$$

Prediction:

$$p_i^{(Att)} = \sigma(\mathbf{W}_c \mathbf{c} + b_c)$$

Step 12: BERT Embedding Extraction

$$\mathbf{z}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \mathbf{h}_{it}$$

where

- $\mathbf{h}_{it}$: token embedding
- T_i: number of tokens

Step 13: BERT + XGBoost

$$p_i^{(BERT+XGB)} = g(\mathbf{z}_i)$$

where

- $g(\cdot)$: XGBoost classifier

Step 14: BERT Fine-Tuning

$$\mathbf{o}_i = \text{BERT}(\tilde{x}_i)$$

$$p_i^{(BERT)} = \text{Softmax}(\mathbf{o}_i)$$

Step 15: Final Prediction

$$\hat{y}_i = \begin{cases} 1, & p_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Step 16: Evaluation Metrics

Accuracy: Precision: Recall: F1-score:

Final Algorithm Steps

1. Load dataset $\mathcal{D}$
2. Preprocess text using $f_{\text{clean}}(\cdot)$
3. Encode labels into binary values
4. Split datasets into training, validation, and testing sets
5. Extract TF-IDF features $\mathbf{v}_i$
6. Train Logistic Regression and XGBoost models
7. Tokenize and pad sequences $\mathbf{s}_i$
8. Train LSTM, CNN, and BiLSTM + Attention models
9. Extract BERT embeddings $\mathbf{z}_i$
10. Train XGBoost using BERT embeddings
11. Fine-tune BERT classifier
12. Computer probability p_i for each model
13. Generate predictions $\hat{y}_i$
14. Evaluate using classification metrics
15. Output results and visualizations

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

3.3 Comparative Analysis Based on Hyperparameter Tuning

Table 1: Hyperparameter Tuning Comparison

Model	Key Parameters	Values
TF-IDF	d, n, df_{min}, df_{max}	40k, (1–2), 2, 0.95
Logistic Reg.	T, w_c	2000, balanced
XGBoost	K, d_{tree}, η, s, c	200, 6, 0.08, 0.9, 0.8
Tokenizer	V, L	20k, 150
Embedding	d_e	128
LSTM	h, p_d, p_r, B, E	96, 0.2, 0.2, 64, 20
CNN	f, k, p_d	128, 5, 0.3
BiLSTM+Att	h, α_t, p_d	64, learned, 0.3
BERT+XGB	L_b, B_b, K	64, 8, 120
BERT FT	η_b, E_b, B_b	2e-5, 20, 2

IV. IMPLEMENTATION AND RESULT ANALYSIS

4.1 Dataset

This dataset is a set of labelled email messages classified in ham (legitimate) and spam classes, thus it provides a realistic benchmark for spam detection tasks. Every record has the email text with its numerical representation of label, ham given as 0 while spam is given as 1 For instance, some lines before "enron methanol meter…" might be normal communication emails such as meeting details or files sharing and spam content (spammy information that proposes promotional or misguiding info ("photoshop windows office cheap..."). It has 3,672 ham emails and 1,482 spam emails so we have a class imbalance which will be compensated by defining stratified data splits also takin into consideration the class weighting during training. You should consider preprocessing steps for cleaning the, text normalization and tokenization to get better feature representation. The dataset is split up into training, validation and test sets to prevent overfitting allowing it to be a good benchmark for performance evaluation of multi-class machine learning algorithms / deep learning/ transformer based models.

4.2 Result analysis

Figure 2: Word Cloud Visualization for Ham and Spam Emails

Figure 2 : A word cloud of the most recurring words in ham and spam emails. Ham emails tend to use more conversational and information-based words whereas spam emails have a greater frequency of promotional- or persuasive-type words indicating different word usage patterns which can be exploited for classification.

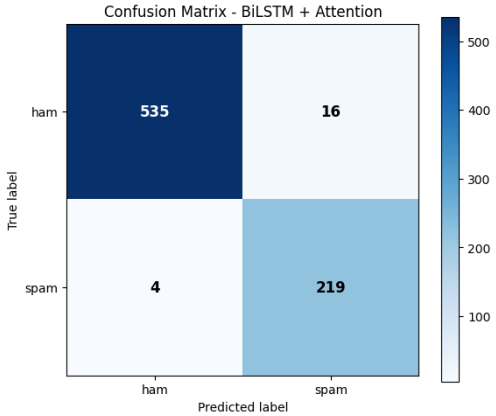

Figure 3: Confusion Matrix of BiLSTM with Attention Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

The results for classification of BiLSTM with Attention model are shown in this figure 3. Thus, the extremely high true positive and true negative values with much smaller false positives and false negatives confirm a model capturing adequate contextual information of emails.

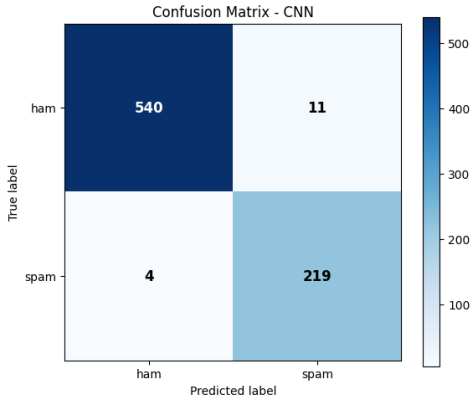

Figure 4: Confusion Matrix of CNN Model

This is detail of figure 4 Confusion Matrix of CNN model. Implying the CNN learns local features in order to correctly classify spam with high accuracy, precision and recall, but low true positives and false negatives.

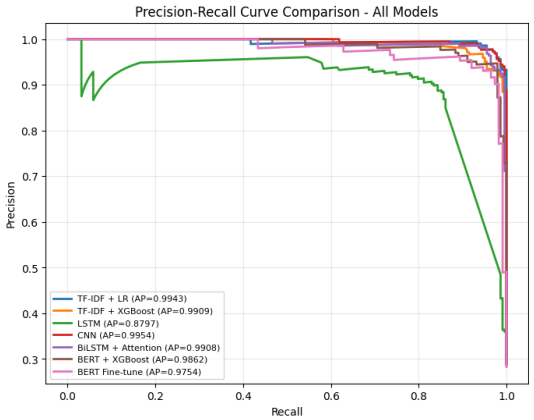

Figure 5: Precision–Recall Curve Comparison of All Models

Figure 5 shows precision-recall curves for all models. All models yield high precision and recall values, indicating that the transformer and ensemble approaches are more balanced to be used when class imbalance is present.

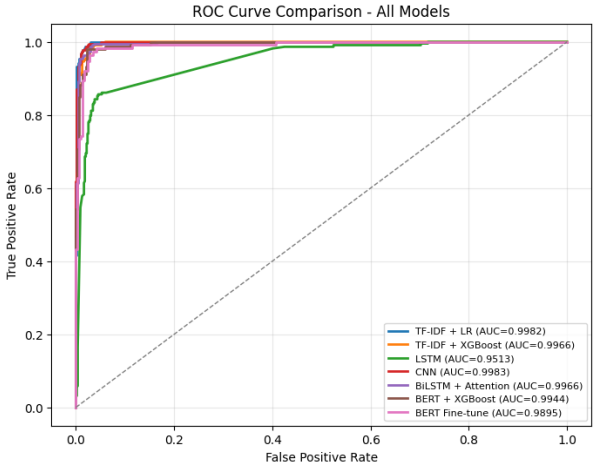

Figure 6: ROC Curve Comparison of All Models

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154100

Figure 6 Receiver Operating Characteristic (ROC) curves of the models. The curves appear near the top-left corner denoting strong classification performance, and several models achieve nearly perfect AUC scores.

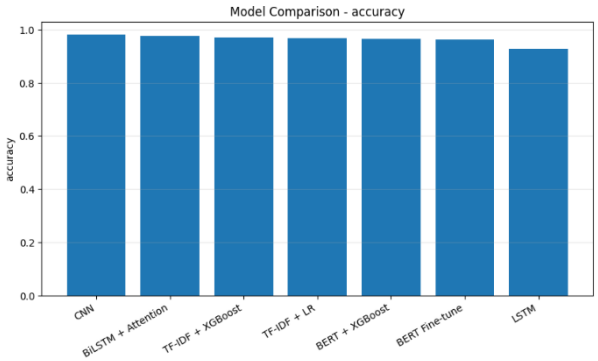

Figure 7: Accuracy Comparison of All Models

Figure 7 compares the accuracy of the model. Most of the models have high accuracies, while CNN, BiLSTM with Attention and hybrid based approaches seem to give better results which shows the effectiveness of deep learning and ensemble methods.

4.3 Result comparison

Table 2: Performance Comparison of Spam Classification Models

Rank	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
1	CNN	0.98	0.95	0.98	0.96	0.99	0.99
2	BiLSTM + Attention	0.97	0.93	0.98	0.95	0.99	0.99
3	TF-IDF + XGBoost	0.97	0.91	0.99	0.95	0.99	0.99
4	TF-IDF + LR	0.96	0.89	1.00	0.94	0.99	0.99
5	BERT + XGBoost	0.96	0.96	0.91	0.93	0.99	0.98
6	BERT Fine-tune	0.96	0.95	0.91	0.93	0.98	0.97
7	LSTM	0.92	0.88	0.85	0.8	0.95	0.87

The following table 2 shows a comparison of performance for all tested models in spam classification. CNN model has the best accuracy as well as F1-score, while BiLSTM with Attention and TF-IDIDF-XGboost following closely. As consistent with the previous results, Transformer-primarily based totally fashions additionally perform quite properly even as the LSTM version offers extensively lower numbers indicating that hybrid and interest-based totally architectures are effective.

4.4 Tested result

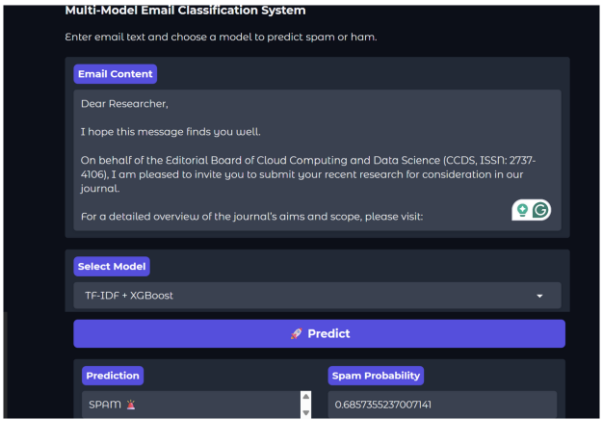

Figure 8: Multi-Model Email Spam Classification Interface

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154100

A multimodal email classifier is depicted in Figure 8 to identify spam messages. The users will enter the email contents, choose a machine learning model (TF-IDF + XGBoost algorithm) for predictions. We could see the output display: spam or not spam and the chance of it being was spam, this gives us an insight on how accurately our model is identifying if content of email is suspicious.

V. CONCLUSION

The study introduces a strong and holistic multi-model framework of email spam classification, combining the traditional machine learning, deep learning, and transformer-based methods into one architecture. The experimental assessment proves that deep learning models, especially CNN, performed the best with the accuracy rate of 98, F1-score of 0.96, and ROC-AUC of 0.99, which reflects a better ability to detect local textual patterns. BiLSTM with Attention model took it a step further to enhance the contextual understanding with 97 percent accuracy and high recall (0.98), and TF-IDF with XGBoost also did the same with 97 percent accuracy and high generalization. Transformer models, such as BERT with XGBoost and fine-tuned BERT, scored a reliable 96 percent accuracy, and therefore, it is possible to note their efficiency in semantic feature extraction despite the limitations of computational capabilities. The standalone LSTM model, however, demonstrated relatively lower performance (92% accuracy), which highlights the benefit of hybrid and attention-based architecture. In general, the findings can be supported by the idea that integration of several feature extraction and learning methods can greatly improve the classification accuracy, robustness, and suitability to sophisticated spam patterns. Future research can be aimed at applying the model to real-time settings, multilingual data, to enhance computational efficiency with lightweight transformer versions, and to find ways to use federated learning to achieve privacy-preserving spam detection systems.

REFERENCES

[1]. A. A. Abbood and A. A. Abdullah, “Email Spam Detection: A Novel Hybrid Approach Using Machine and Deep Learning Techniques,” *International Journal of Intelligent Engineering & Systems*, vol. 18, no. 7, 2025.
[2]. V. Sowmitha, R. Deebika, K. Vinoth, S. Suryaprakash, K. Kalaivendhan, and V. Karthi, “Federated Learning in Email Server Authentication for Spam Mail Detection Using LSTM Algorithm,” in *Proc. Int. Conf. Smart & Sustainable Technology (INCSST)*, Chikodi, India, 2025, pp. 1–6.
[3]. A. A. E. Damanik, H. H. Nuha, N. D. W. Cahyani, Setyorini, and M. A. B. Ismail, “Email Spam Detection using Long Short-Term Memory (LSTM) Network Method,” in *Proc. Int. Conf. Decision Aid Sciences and Applications (DASA)*, Manama, Bahrain, 2024, pp. 1–6.
[4]. M. Raza, N. D. Jayasinghe, and M. M. A. Muslam, “A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms,” in *Proc. Int. Conf. Information Networking (ICOIN)*, Jeju Island, South Korea, 2021, pp. 327–332.
[5]. M. Habib, H. Faris, M. A. Hassonah, J. Alqatawna, A. F. Sheta, and A. M. Al-Zoubi, “Automatic Email Spam Detection using Genetic Programming with SMOTE,” in *Proc. 5th HCT Information Technology Trends (ITT)*, Dubai, UAE, 2018, pp. 185–190.
[6]. K. Debnath and N. Kar, “Email Spam Detection using Deep Learning Approach,” in *Proc. Int. Conf. Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, 2022, pp. 37–41.
[7]. V. Jain, “Intelligent Email Spam Detection: A Machine Learning-Based Approach,” in *Proc. 5th Int. Conf. Trends in Material Science and Inventive Materials (ICTMIM)*, Kanyakumari, India, 2025, pp. 1574–1579.
[8]. A. Zaid, J. Alqatawna, and A. Huneiti, “A Proposed Model for Malicious Spam Detection in Email Systems of Educational Institutes,” in *Proc. Cybersecurity and Cyberforensics Conf. (CCC)*, Amman, Jordan, 2016, pp. 60–64.
[9]. V. R., D. Kamalin, S. Vanaja, C. H. Ramesh, G. Karuna, and G. Sabarinathan, “Enhanced Naïve Bayes Model for Intelligent and Secure Email Spam Detection Using Hybrid Feature Engineering and Blockchain-Based Verification,” in *Proc. Int. Conf. Metaverse and Current Trends in Computing (ICMCTC)*, Subang Jaya, Malaysia, 2025, pp. 1–5.
[10]. A. Wijaya and A. Bisri, “Hybrid decision tree and logistic regression classifier for email spam detection,” in *Proc. 8th Int. Conf. Information Technology and Electrical Engineering (ICITEE)*, Yogyakarta, Indonesia, 2016, pp. 1–4.
[11]. A. Yogaraj, T. R. D. Kumar, S. S. Babu, D. HemanthRaj, A. HerrickLinore, and P. Jagath, “Email Spam Detection using Machine Learning,” in *Proc. 8th Int. Conf. Circuit, Power & Computing Technologies (ICCPCT)*, Kollam, India, 2025, pp. 904–909.
[12]. N. RamojiRao, S. Anusuya, Harshavardhini, and C. B. Sembuli, “Analysis of Email Spam Detection Using Naive Bayes and Support Vector Machine Classification Algorithms,” in *Proc. IEEE 9th Int. Conf. Engineering Technologies and Applied Sciences (ICETAS)*, Bahrain, 2024, pp. 1–5.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154100

[13]. V. I. Del Rosario, B. D. P. Fernandez, and D. A. Padilla, "Email Spam Classification using DistilBERT," in *Proc. IEEE 15th Int. Conf. Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, Coron, Philippines, 2023, pp. 1–6.
[14]. G. S. Chauhan, R. Nahta, and N. Garg, "Email-based Spam Detection using Long Short-Term Memory," in *Proc. 4th OPJU Int. Technology Conf. (OTCON)*, Raigarh, India, 2025, pp. 1–5.
[15]. S. R. Kalluri, D. R. Billa, A. R. Chinthala, and M. M. Rana, "EmailSpamML: Optimized ML Algorithms for Spam Email Detection, Performance Insights, and Future Research Directions," in *Proc. 27th Int. Conf. Advanced Communications Technology (ICACT)*, Pyeong Chang, South Korea, 2025, pp. 59–62.
[16]. N. A. Al-Dmour, S. Kousar, A. Khan, A. Ihsan, T. Abbas, and A. Q. Saeed, "Enhancing Email Spam Detection Using Advanced AI Techniques," in *Proc. Int. Conf. Decision Aid Sciences and Applications (DASA)*, Manama, Bahrain, 2024, pp. 1–6.
[17]. S. Saklani, K. Thapa, and D. Singh, "Spam Email Detection using K-Nearest Neighbors: An Enhanced Approach," in *Proc. Int. Conf. Intelligent Computing and Control Systems (ICICCS)*, Erode, India, 2025, pp. 1156–1160.
[18]. A. Mohamed and S. K. Assayed, "Spam Email Detection Using Artificial Intelligence and Machine Learning Techniques," in *Proc. IEEE Smart World Congress (SWC)*, Calgary, Canada, 2025, pp. 213–216.

