2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM 2025)
27-28 June 2025, Gazipur-1707, Bangladesh

Hybrid Email Filtering Using TF-IDF and BERT Enhanced Spam Detector

Sumit Kanti Sarker[1], Mohammad Shoaib Rahman[2], Abdullah Al Prince[3], Abdul Al Mahmud Riaz[4], Md Shakhauat Hossan Sumon[5], Md. Tauhidur Rahman Sakib[6], Sifat Sohanoor[7], and Md Abu TAlha[*8]
[1]*Department of Computer Science & Engineering (CSE), Faculty of Science & Technology (FST)*
American International University Bangladesh, Dhaka 1229, Bangladesh
[2]*Department of Computer Science & Engineering (CSE), Faculty of Modern Science*
Leading University (LU), Sylhet-3112, Bangladesh
[3]*School of Business Analytics, Collins College of Business*
The University of Tulsa, Tulsa, OK 74104, United States
[4]*Department of Computer Science, Faculty of Computing, Engineering and Science*
University of South Wales, South Wales, CF37 1dl, UK
[5]*Department of Electrical and Computer Engineering, School of Engineering & Physical Sciences*
North South University (NSU), Dhaka-1229, Bangladesh
[6]*Department of Computer Science and Engineering, School of Data and Sciences*
BRAC University, Dhaka-1212, Bangladesh
[7]*Department of Computer Science and Engineering (CSE), School of Engineering & Physical Sciences*
North South University (NSU), Dhaka-1229, Bangladesh
[8]*Department of Electrical and Electronics Engineering (EEE), Faculty of Engineering (FE)*
American International University Bangladesh, Dhaka-1229, Bangladesh

Email- sumitsarker141@gmail.com, shoaibse86@gmail.com, abp4247@utulsa.edu, almahmudr1@gmail.com, sumon.eee.cse@gmail.com, tauhidurrahmansakib@gmail.com, sifatsohanoor7@gmail.com, abutalha8324@gmail.com

***Abstract*—Email spam detection is a crucial cybersecurity challenge requiring accurate and efficient filtering methods. This study introduces the TF-IDF and BERT Enhanced Spam Detector (TBESD), a hybrid deep learning model that integrates statistical and contextual text analysis for robust spam classification. TF-IDF captures essential n-gram patterns, while BERT embeddings extract deep semantic features. These feature vectors are combined and processed through a multilayer perceptron (MLP) with dropout regularization and ReLU activation for optimal classification. The methodology includes text preprocessing (normalization, stopword removal, tokenization), feature extraction (TF-IDF, BERT embeddings), and model training using the Adam optimizer with binary cross-entropy loss. Evaluated on a publicly available dataset, TBESD achieved an impressive accuracy of 98.88%, precision of 98.44%, recall of 99.44%, F1-score of 98.94%, and an AUC score of 0.9990. This study demonstrates the superiority of hybrid models over conventional methods, highlighting the role of advanced NLP techniques in cybersecurity.**

Keywords—Spam Detection, TF-IDF, BERT, Hybrid Model, TBESD, TBESD, AUC Score

I. INTRODUCTION

In today's digital landscape, email remains one of the most widely used communication tools across personal, professional, and governmental domains. However, the increasing volume of spam emails poses significant challenges, leading to reduced productivity, security threats, and privacy breaches. Unsolicited and malicious emails often serve as vectors for phishing attacks, malware distribution, and identity theft, making efficient spam detection a critical cybersecurity concern.

Al-Shanableh et al. proposed a novel ensemble meta-learning technique for detecting email spam. This method integrates multiple machine-learning algorithms, including Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN), using a stacking approach [1]. The stacking process consolidates the predictions into a final Logistic Regression classifier. This ensemble approach achieved an improved detection accuracy of 95.8%, demonstrating that combining less robust models can result in a more effective spam detection system. Khattab et al. compared Naive Bayes, KNN, and SVM algorithms for spam detection. Naive Bayes demonstrated superior performance, achieving an accuracy of 97%, sensitivity of 93%, and specificity of 87% [2]. However, the researchers did not investigate more sophisticated approaches such as deep learning. Janez-Martino et al. proposed a topic-based approach for classifying spam emails using agglomerative hierarchical clustering [3]. Their study examined 16 different pipelines combining various text representation methods with classification algorithms. Although this method demonstrated innovative aspects in topic-based classification and cross-language assessment, it faces limitations in handling larger datasets and addressing complex spam characteristics. Liu et al. utilized the DistilBERT deep-learning algorithm for spam detection [4]. This streamlined version of BERT demonstrated impressive results with 93% classification accuracy. The approach, using pretrained models, identified intricate patterns, surpassed conventional methods, and minimized computational demands. Ayo et al. proposed a hybrid correlation-based deep learning model for email spam classification, integrating fuzzy inference to reduce misclassification and categorize spam severity [5]. Using rule-based feature selection, the model achieved an F1-score of 96.5% for spam and 96.4% for non-spam, along with better accuracy (94.0017%), low error rate (5.9983%), and fast processing time (0.5s). Bouke et al. introduced a transparent

2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM) | 979-8-3315-5542-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/NCIM65934.2025.11160311

979-8-3315-5542-9/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.

approach using a Random Forest classifier [6]. This method incorporated Hybrid Random Sampling to balance the data and employed the Gini Index for feature selection, resulting in an accuracy of 94.8%. Although this technique demonstrated enhanced transparency and performance, it might encounter difficulties with extremely large datasets.

Mwavali et al. propose a supervised machine learning model tailored to address phishing attacks in Kenyan financial institutions, demonstrating how AI can enhance email security within a specific context [7]. This research exemplifies the practical deployment of machine learning in real-world security challenges. Alzahrani et al. introduce an explainable AI framework that employs ensemble learning techniques to boost the accuracy of spam detection from text, focusing on both performance and interpretability [8]. Their approach not only improves detection rates but also ensures transparency in the decision-making processes of AI models. Bhardwaj et al. investigate the integration of sentiment analysis with machine learning for detecting fake social media news and spam, emphasizing the role of sentiment scoring in identifying deceptive content [9]. They suggest that similar techniques could prove effective for email spam detection. However, as spam tactics continue to evolve, there is an increasing need for more advanced methods to combat these challenges [10][11].

Traditional spam detection methods, such as rule-based filtering and classical machine learning models like Naïve Bayes and Support Vector Machines (SVM), have demonstrated effectiveness but struggle with evolving spam patterns and contextual ambiguities. Recent advancements in deep learning and Natural Language Processing (NLP) have enabled more sophisticated classification techniques. Particularly, transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) offer deep contextual understanding, improving classification accuracy.

This research proposes a hybrid spam detection model—TF-IDF and BERT Enhanced Spam Detector (TBESD)—that leverages statistical feature extraction through Term Frequency-Inverse Document Frequency (TF-IDF) and deep semantic learning via BERT embeddings. By combining these complementary approaches, TBESD effectively captures both word importance and contextual meaning, enhancing classification performance. A multilayer perceptron (MLP) with dropout regularization and ReLU activation is employed for robust classification.

The key contributions of this study include:

- Hybrid Model Design – Integrating TF-IDF and BERT embeddings to improve spam detection accuracy.
- Comprehensive Preprocessing Pipeline – Implementing text normalization, stopword removal, tokenization, and feature extraction.
- Performance Evaluation – Comparing TBESD with conventional machine learning and deep learning models using accuracy, precision, recall, F1-score, and AUC metrics.

The structure of this paper's remaining sections is as follows: Section II delineates the methodology employed in this research, encompassing data preprocessing, feature extraction methods, model architecture, and training protocols. Section III presents the experimental results, highlighting the model's performance across various evaluation metrics, including accuracy, precision, recall, and the Confusion Matrix. Section IV concludes the paper with an analysis of the findings and suggestions for future research directions.

II. METHODOLOGY

This study's methodology focuses exclusively on detecting email spam using hybrid TF-IDF and BERT-based deep learning models. The process encompasses data acquisition, data preprocessing, feature extraction, model architecture, and model evaluation.

A. Dataset Overview

The study's dataset, which was obtained from Kaggle, comprised 83,446 records of email combinations, including both legitimate and spam emails. The dataset consisted of two columns: "text" and "label." The "text" column contained the email content, while the "label" column utilized binary classification, where 1 indicated unwanted emails (Spam) and 0 indicated legitimate emails (Ham). The dataset sample is shown in Table I.

TABLE I. DATASET OF SPAM AND HAM EMAIL

Label	Text
1	Viiiiiiiagraa\nonly for the ones that want t.....
0	got ice thought look az original message iec.....
*83,446 rows * 2 columns for True datasets*	
1	Yo our wom an ne eds an escapenumber in ch.....
0	Author jra date escapennumber eaca.....

B. Data Preprocessing

To ensure that the data are prepared for feature extraction and model training, data preprocessing is essential to achieve optimal accuracy. This process transforms unstructured email content into a structured format that is suitable for analysis and classification. The following sections provide a comprehensive elucidation of the processing techniques employed for text normalization, text cleaning, stopword removal, and tokenization.

Text cleaning: During the initial data processing stage, various cleaning techniques are implemented to enhance the quality of input data for effective spam identification. Initially, all Uniform Resource Locators (URLs), including web links and internet addresses commencing with "http," "www," or "https," are extracted to mitigate their influence on the classification algorithm. Subsequently, user mentions (e.g., "@username") and hashtags (e.g., "#hashtag") are eliminated, as they provide negligible value for spam detection. Furthermore, all non-alphabetic characters and numerals are replaced with spaces, ensuring that only word-based content remains. This process facilitates concentration on the most relevant textual information for classification, thereby improving model efficiency and interpretability.

Normalization and Stopword removal: To standardize text processing, all electronic mail messages were initially converted to lowercase messages, thereby eliminating variations caused by capitalization. This standardization procedure facilitates the consistency of word representations. Subsequently, each electronic mail message was subjected to tokenization, which is the process of segmentation into

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.

individual lexical units. During this phase, common words that do not contribute significant semantic value, referred to as stopwords (e.g., "the," "is," "and"), were removed utilizing a predefined lexicon. By concentrating on semantically meaningful words, this step enhances the efficacy of feature extraction and classification, ultimately improving the performance of spam detection systems.

Tokenization: Tokenization is the process of segmenting text into discrete units, typically words or subwords, to facilitate analysis and processing. In email spam detection, each electronic message is parsed into tokens, enabling the extraction of salient linguistic patterns. This step standardizes textual representation and prepares it for subsequent preprocessing tasks such as stopword removal and feature extraction. Effective tokenization enhances the efficacy of text classification models by ensuring that each lexical unit is treated as a distinct entity, thereby contributing to an improved spam detection performance.

C. Feature Extraction

Subsequent to text preprocessing, feature extraction is conducted to transform the processed text into numerical representations suitable for classification.

TF-IDF Feature Extraction: The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique is utilized to transform textual data into numerical representations appropriate for machine learning algorithms. This methodology assesses the significance of words based on their frequency within a document while mitigating the impact of commonly occurring terms across the corpus. To enhance feature representation, both unigram and bigram features are extracted, capturing individual words and adjacent word pairs that contribute to textual semantics. The vocabulary is constrained to the 5,000 most salient words, ensuring that the most informative terms are retained while reducing dimensionality and computational complexity. This approach augments the capacity of the classification model to differentiate between spam and legitimate electronic mail communications effectively.

BERT Feature Extraction: BERT (Bidirectional Encoder Representations from Transformers) is employed to generate contextualized word embeddings for enhanced text representation. The BERT tokenizer encodes the text into token embeddings, ensuring consistency in input processing. The pre-trained BERT model extracts deep semantic features from the text, with the CLS token representation from the final hidden state serving as a comprehensive summary of each input sequence. This feature extraction process enhances the model's capacity to capture complex linguistic patterns, thereby improving spam classification accuracy.

D. Model Architecture

The proposed hybrid model integrates statistical and deep learning-based features for enhanced spam detection. The architecture comprises three primary components: TF-IDF-based statistical feature extraction, BERT-based semantic embeddings, and a Multi-Layer Perceptron (MLP) classifier for classification. The integration of these techniques augments both lexical and contextual comprehension of email content.

Input Representations: The model employs two distinct feature representations to enhance the efficacy of spam classification. The TF-IDF feature representation is a 5,000-dimensional vector obtained through term frequency-inverse document frequency (TF-IDF) transformation, which captures salient statistical n-gram patterns in the text. This representation emphasizes the significance of words based on their frequency while mitigating the impact of commonly occurring terms.

Conversely, the BERT embedding representation is a 768-dimensional vector derived from a pre-trained BERT-based uncased model. The CLS token representation from the final hidden state of BERT is utilized to encode the semantic context of the entire email, capturing the intricate linguistic relationships that conventional statistical methods may fail to discern. By integrating these two complementary representations—one predicated on word frequency distributions and the other on contextualized word embeddings—the model achieves a more comprehensive analysis of email content, resulting in enhanced spam detection performance.

TABLE II. PROPOSED HYBRID MODEL LAYOUT PROCEDURE

Layer Name	Type	Output Shape	Description
TF-IDF Input	Input Layer	(5000,)	Converts text into numerical vectors using n-gram features.
BERT Embedding Input	Input Layer	(768,)	Extracts contextual word embeddings from pre-trained BERT
Concatenation	Merge Layer	(5768,)	Merges TF-IDF and BERT features into a single representation
Dense Layer 1	Fully Connected	(256,)	Learns high-level spam patterns (ReLU activation)
Dropout Layer 1	Regularization	(256,)	Prevents overfitting (dropout rate = 0.5)
Dense Layer 2	Fully Connected	(128,)	Further refines spam classification (ReLU activation)
Dropout Layer 2	Regularization	(128,)	Prevents overfitting (dropout rate = 0.5)
Output Layer	Fully Connected	(1,)	Sigmoid activation for binary classification (spam/ham)

Feature Fusion Layer: To effectively integrate both statistical and contextual information, the TF-IDF and BERT feature representations are concatenated into a single 5,768-dimensional feature vector. This integration combines the strengths of word frequency-based statistical patterns from TF-IDF with deep contextual embeddings from BERT, enabling the model to learn from both sources concurrently. By utilizing these complementary representations, the model attains a more comprehensive understanding of the text, enhancing its capacity to differentiate between spam and legitimate emails, thereby improving classification performance.

Hidden Layers: The concatenated feature vector, representing both statistical and contextual text representations, is processed through a Multi-Layer Perceptron (MLP) to facilitate effective classification. The

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.

MLP comprises two fully connected layers designed to extract high-level patterns and enhance spam detection accuracy.

The initial dense layer consists of 256 neurons with ReLU (Rectified Linear Unit) activation, enabling non-linearity and facilitating the network's ability to learn complex spam-related features. To mitigate overfitting and improve model generalization, a dropout layer with a 50% dropout rate is implemented, randomly deactivating half of the neurons during training. This regularization technique ensures robustness by reducing dependency on specific neurons.

Subsequently, a second dense layer with 128 neurons and ReLU activation further refines the learned feature representations, capturing intricate relationships within the textual data. An additional dropout layer with a 50% dropout rate is introduced to mitigate overfitting risks, promoting improved generalization on unseen data. This structured approach to feature transformation and classification enhances the model's capacity to differentiate between spam and legitimate emails effectively, leveraging both statistical and deep contextual representations for robust performance.

Output Layer: The output layer of the model comprises a solitary neuron with a sigmoid activation function, which converts the acquired feature representations into a probability score ranging from 0 to 1. This probability denotes the likelihood of an electronic mail message being classified as spam. A threshold value of 0.5 is employed for decision-making, wherein electronic mail messages with a probability exceeding 0.5 are classified as spam (1), while those below this threshold are categorized as ham (0). The sigmoid function ensures a continuous probability distribution, rendering the model appropriate for binary classification tasks by providing interpretable confidence scores.

E. Advance Analytical Analysis

Advanced Analytical Analysis outlines a systematic process for email spam detection, integrating TF-IDF and a BERT-based deep learning model. This study employs a systematic approach to email spam detection, incorporating feature extraction, hybrid modeling, and performance assessment for high accuracy. The process commences with comprehensive text preprocessing, where email data undergoes cleaning and refinement. This entails normalizing text, converting to lowercase, and eliminating non-informative elements such as URLs, special characters, mentions, and hashtags. Common stopwords are removed, and tokenization and lemmatization are applied to create a structured word representation. The study utilizes a hybrid feature extraction method, combining statistical and contextual text representations. TF-IDF transforms text into a high-dimensional vector space, capturing crucial n-gram patterns and emphasizing spam-relevant words. BERT-based embeddings encode deep contextual word relationships, facilitating the understanding of linguistic nuances. The fusion of TF-IDF and BERT embeddings results in a 5768-dimensional hybrid feature vector for spam classification.

A Multi-Layer Perceptron (MLP) neural network processes these hybrid features. The network architecture comprises fully connected layers to learn complex spam patterns. The input layer merges TF-IDF (5000 dimensions) and BERT embeddings (768 dimensions). Hidden layers contain 256 and 128 neurons, activated by ReLU. Dropout regularization (50%) is applied to each hidden layer to mitigate overfitting. The output layer has a single neuron with a sigmoid activation function, generating a spam probability score.

To ensure optimal training and performance, the model was compiled utilizing carefully selected optimization techniques. The Binary Cross-Entropy loss function was employed, as it is well-suited for binary classification tasks by effectively handling probability-based outputs. To optimize the weight updates and enhance the convergence speed, the Adam optimizer was used with a learning rate of 0.0001. The dataset was split into training and testing subsets, with 80% allocated for training and 20% allocated for testing, ensuring an objective evaluation of the model's generalization capability. To mitigate overfitting and augment the training efficiency, batch processing with a batch size of 32 was implemented in conjunction with early stopping methodologies.

Model performance was evaluated using multiple metrics, including Accuracy, Precision, Recall, and the F1-score, providing a comprehensive assessment of classification effectiveness. Additionally, the Confusion Matrix was employed to visualize the distribution of correct and incorrect classifications, offering insights into false positives and false negatives. Furthermore, the Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC-ROC) score were utilized to measure the ability of the model to distinguish between spam and legitimate emails across different threshold values. A higher AUC-ROC score indicates stronger classification capability, enhancing the interpretability of the model's performance.

III. Result and Discussion

The efficacy of the proposed hybrid email spam detection system, which integrates TF-IDF and a BERT-based deep learning model, was evaluated utilizing multiple performance metrics. These metrics encompassed accuracy, precision, recall, F1-score, confusion matrix analysis, and the ROC curve. This section presents the results of the comprehensive performance evaluation, providing a detailed examination of the model's effectiveness in identifying spam emails.

A. Model Performance Metrics

Fig. 1 illustrates a comparative analysis between the model's training accuracy (represented by a blue line) and validation accuracy (depicted in green) across 40 epochs during the training process. Initially, both the training and validation accuracies exhibited a rapid increase, indicating effective learning from the dataset. As training progressed, the curves plateaued, and the training accuracy marginally exceeded the validation accuracy. Despite minor fluctuations, the validation accuracy remained close to the training accuracy, which was 98.88%, suggesting the model's capacity to generalize effectively to novel data without significant overfitting. The high accuracy levels attained by both training and validation were 99.11% and 98.88%, respectively, demonstrating that the model successfully identified salient features for spam detection, resulting in robust performance. This graphical representation underscores the efficacy of the proposed model and its ability to maintain consistent performance across diverse datasets.

Fig. 2 illustrates the efficacy of the model through four critical evaluation metrics: accuracy, precision, recall, and F1-score. These measurements collectively demonstrate the effectiveness of the proposed spam detection system. With an

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.

accuracy of 98.88%, the model exhibits high overall reliability, while a precision of 98.44% indicates its capacity to accurately identify spam emails without misclassifying legitimate ones. The recall, which denotes the model's ability to detect all spam emails, reaches 99.44%. Furthermore, the F1-score of 98.94% ensures an optimal balance between precision and recall. The utilization of a bar chart format facilitates visual comparison and interpretation of these metrics.

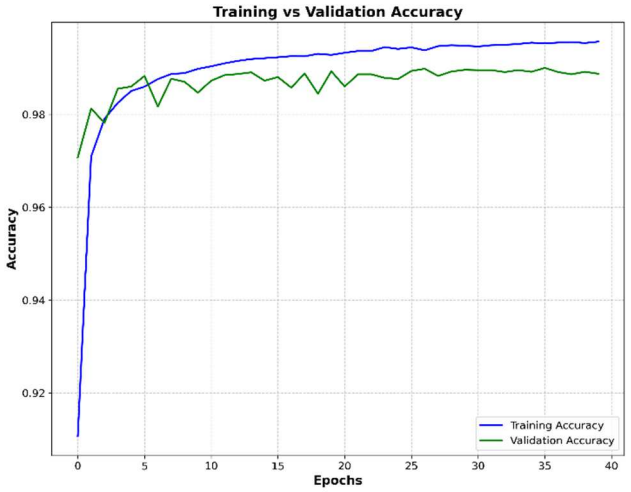


Fig. 1. Training and Validation Accuracy.

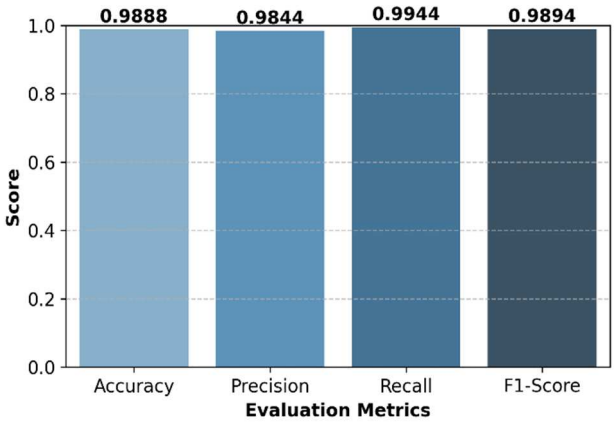


Fig. 2. Evaluation Metrics for Spam Detection.

B. *Confusion Matrix Analysis*

The confusion matrix analysis in Fig. 3 provides a comprehensive evaluation of the email spam classification model's performance by illustrating the relationship between actual and predicted classifications. The matrix demonstrates that the model correctly identified 7,800 legitimate (Ham) emails as Ham (True Negatives) and 8,703 spam emails as Spam (True Positives). However, some errors occurred, with 138 Ham emails incorrectly classified as Spam (False Positives) and 49 Spam emails erroneously categorized as Ham (False Negatives). The substantial number of True Positives and True Negatives underscores the model's proficiency in accurately distinguishing between spam and legitimate emails. Furthermore, the relatively low counts of False Positives and False Negatives indicate the model's effectiveness in minimizing misclassifications, contributing to its robust overall performance. This evaluation emphasizes the model's reliability and accuracy, rendering it a viable option for practical email spam detection implementations.

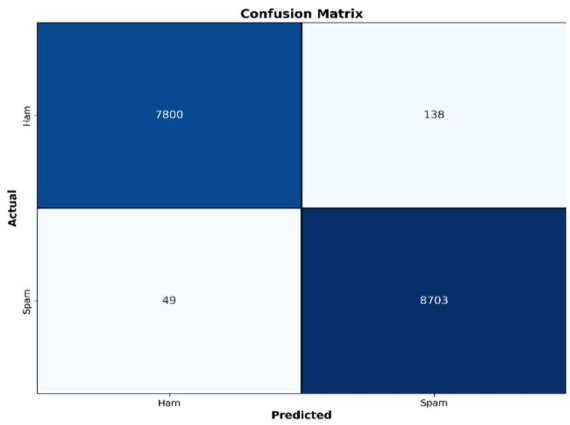


Fig. 3. Confusion Matrix for Spam Detection.

C. *Operating Characteristics (ROC) Curve*

Fig. 4 presents a Receiver Operating Characteristic (ROC) curve, which provides a comprehensive evaluation of the model's capacity to distinguish between spam and non-spam emails.

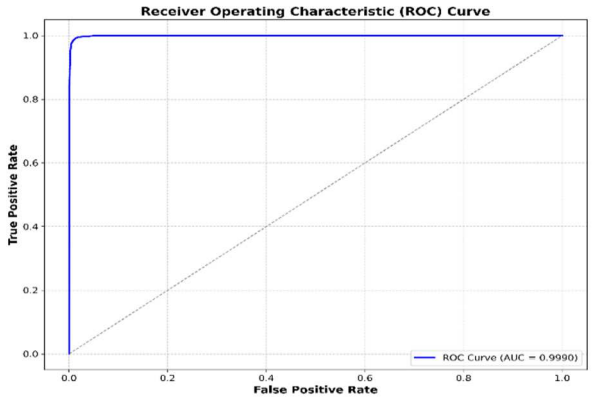


Fig. 4. Receiver Operating Characteristics (ROC) Curve.

This curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across various classification thresholds. The steep ascent of the curve near the origin, in conjunction with an area under the curve (AUC) of 0.9990, indicates the exceptional performance of the model, approaching optimal classification. The substantial AUC value demonstrates the model's proficiency in maximizing True Positives, while minimizing False Positives across different thresholds. This outstanding performance underscores the robustness and reliability of the model in classification, rendering it highly suitable for practical spam detection applications.

D. *Feature Importance Analysis*

Analyzing the importance of features improves the interpretation of the logistic regression model. Extract and visualize model coefficients that represent feature weights. The higher the absolute coefficient, the greater the impact on spam classification. The top 20 features determined by their weights were visualized using a word cloud to highlight their contribution, which is shown in Table III and Fig. 5. In addition, locally explained models are used to generate cases of case-level explanations. Lime determines the most influential term for each prediction, which ensures transparency of the model decisions. These explanatory AI techniques improve the interpretation of the model and help test its consistency with domain knowledge.

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.

TABLE III. TOP FEATURES AND WEIGHTS

Top Features	Weights	Top Features	Weights
houston	-19.8883	subject	-28.8919
listinfo	-19.9621	debian	-27.1926
ethz	-20.2558	thanks	-30.1963
php	-20.7167	samba	-30.4836
enron	-21.1128	wrote	-34.7055
nytimes	-21.5818	vince	-23.7749
speakup	-22.8678	louise	-22.8678

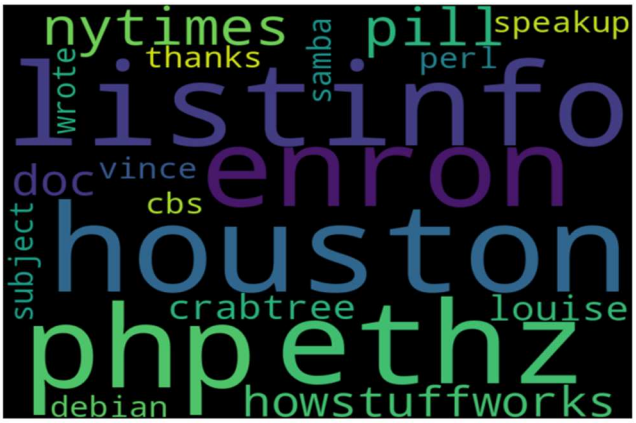


Fig. 5. Top Features Contributing to Spam Classification.

E. Comparative Analysis

The newly developed TF-IDF and BERT enhanced Spam Detector (TBESD), an innovative hybrid approach, demonstrated superior performance with a 98.88% accuracy rate, significantly outperforming various existing machine learning and deep learning techniques in identifying email spam. Table IV illustrates that the TBESD hybrid model's accuracy surpasses that of DistilBERT (93.00%) as reported by Liu et al. [4], and the Hybrid Correlation-based Deep Learning model with a Fuzzy Inference System (HCDL-FIS) documented by Ayo et al. [5], which achieved 94.0017%. Furthermore, it outperformed the Random Forest classifier (94.80%) mentioned by Bouke et al. [6] and the Ensemble Model (95.80%) presented by Al-scannable et al. [1]. While TF-IDF-based logistic regression proves efficient and expedient, rendering it suitable for impromptu email filtering in smaller datasets, it encounters difficulties with complex patterns. The introduced TF-IDF & BERT Enhanced Spam Detector (TBESD) hybrid model excels at recognizing both statistical and semantic text patterns, rendering it highly effective for classifying email spam.

TABLE IV. TOP ACCURACY COMPARISON OF THE PROPOSED TBESD HYBRID MODEL WITH OTHER EXISTING METHODS

References	Model	Accuracy
Liu et al. [4]	DistilBERT	93.00%
Ayo et al. [5]	HCDL-FIS	94.0017%
Bouke et al. [6]	Random Forest	94.80%
Al-shanableh et al. [1]	Ensemble Model	95.80%
Proposed TBESD hybrid model		98.88%

IV. CONCLUSION

This research introduced the TF-IDF, and BERT Enhanced Spam Detector (TBESD), a hybrid deep learning model that integrates statistical and contextual text analysis for accurate email spam classification. By combining TF-IDF for capturing n-gram patterns with BERT embeddings for deep semantic understanding, the proposed model effectively distinguishes between legitimate and spam emails. The multilayer perceptron (MLP) classifier, enhanced with dropout regularization and ReLU activation, further optimizes classification performance. Experimental results on a publicly available spam dataset demonstrated superior performance, achieving 98.88% accuracy, 98.44% precision, 99.44% recall, 98.94% F1-score, and an AUC score of 0.9990. These findings confirm that hybrid models significantly outperform conventional spam detection approaches, improving both accuracy and robustness.

Despite its high efficiency, the model can be further optimized for real-time implementation and adversarial spam detection. Future research will explore transformer-based architectures like RoBERTa and GPT-based models to enhance classification accuracy while reducing computational costs and ensuring scalability for large-scale email filtering systems.

REFERENCES

[1] N. Al-shanableh, M. Alzyoud, and E. Nashnush, "Enhancing Email Spam Detection Through Ensemble Machine Learning: A Comprehensive Evaluation Of Model Integration And Performance," *Communications of the IIMA,* vol. 22, no. 1, Sep. 2024.

[2] S. A. Khattab, R. N. Alkaied, and M. K. S. Ma'aitah, "Comparison of Three Machine Learning Models for the Detection of Emails Spam," *Research Square,* Jan. 23, 2023.

[3] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach," *Appl. Soft Comput.,* vol. 139, no. 110226, p. 110226, 2023.

[4] T. Liu, S. Li, Y. Dong, Y. Mo, and S. He, "Spam Detection and Classification Based on DistilBERT Deep Learning Algorithm," *Applied Science and Engineering Journal for Advanced Research,* vol. 3, no. 3, Art. no. 3, May 2024.

[5] F. E. Ayo, L. A. Ogundele, S. Olakunle, J. B. Awotunde, and F. A. Kasali, "A hybrid correlation-based deep learning model for email spam classification using fuzzy inference system," *Decision Analytics Journal,* vol. 10, p. 100390, Mar. 2024.

[6] M. A. Bouke, O. I. Alramli, and A. Abdullah, "XAIRF-WFP: a novel XAI-based random forest classifier for advanced email spam detection," *Int. J. Inf. Secur.,* vol. 24, no. 1, p. 5, Oct. 2024.

[7] A. Mwavali, "Combating phishing in Kenya: A supervised learning model for enhanced email security in Kenyan financial institutions," *International Journal of Technology and Systems,* vol. 9, no. 4, pp. 23–36, 2024.

[8] A. Alzahrani, "Explainable AI-based framework for efficient detection of spam from text using an enhanced ensemble technique," *Eng. Technol. Appl. Sci. Res.,* vol. 14, no. 4, pp. 15596–15601, 2024.

[9] A. Bhardwaj, S. Bharany, and S. Kim, "Fake social media news and distorted campaign detection framework using sentiment analysis & machine learning," *Heliyon,* vol. 10, no. 16, p. e36049, 2024.

[10] N. Banjare and P. Gani, "A review on hybrid feature selection based deep learning model for enhanced email spam detection," *International Journal For Multidisciplinary Research,* vol. 7, no. 2, 2025.

[11] S. K. Sarker et al., "Email spam detection using logistic regression and explainable AI," in 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2025, pp. 1–6.

Authorized licensed use limited to: University of Malaya. Downloaded on September 18,2025 at 18:33:32 UTC from IEEE Xplore. Restrictions apply.