FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 1, 2026, 490-499

Development of an Ensemble Model for Email Filtering and Classification

Bolaji A. OMODUNBI[1*], Hammed A. OLASUNKANMI[2]

[1, 2] Department of Computer Engineering, Federal University Oye-Ekiti, Ekiti State, Nigeria

[1*]bolaji.omodunbi@fuoye.edu.ng, [2]olahammed103@gmail.com

Abstract

This research presents an advanced ensemble-driven framework for email classification that integrates conventional machine learning algorithms with deep transfer learning methods. A well-structured dataset comprising spam, phishing, and legitimate email samples was assembled and processed using both TF-IDF representations and contextual embeddings derived from BERT. The proposed approach combines Naïve Bayes, Support Vector Machine, and BERT models through a weighted voting strategy, with weights fine-tuned via cross-validation. Experimental evaluation based on performance metrics such as accuracy, precision, recall, and F1-score indicates notable effectiveness, with the model achieving 98% accuracy, 97% precision, 98% recall, and 98% F1-score. Furthermore, the system demonstrated a high capability in identifying phishing emails while minimizing false negative rates. These results highlight the advantage of integrating traditional machine learning techniques with deep learning models to achieve a more reliable and efficient email classification system.

Keywords: *Ensemble Learning, Email Classification, Spam Detection, Machine Learning, BERT.*

1.0 Introduction

Electronic mail has become a fundamental channel for communication across personal, educational, financial, and organizational domains in today's digital environment. However, its increasing adoption has also created opportunities for cyber threats, especially in the form of spam and phishing attacks. Unsolicited messages not only clutter users' inboxes and hinder efficiency but also place unnecessary strain on network bandwidth and storage infrastructure. More importantly, phishing emails are deliberately crafted to manipulate recipients into disclosing confidential information, which can lead to identity compromise, financial damage, and serious security violations. As attackers continue to adopt more advanced and deceptive strategies, conventional rule-based systems and single-model filtering approaches are no longer sufficient for reliably identifying harmful emails.

To address these issues, various machine learning techniques have been applied to email classification tasks. Algorithms such as Naïve Bayes and Support Vector Machine (SVM) are widely recognized for their effectiveness. Naïve Bayes offers simplicity and efficiency, particularly when dealing with high-dimensional textual features, while SVM is known for its strong capability to handle complex data distributions and maintain good generalization performance. More recently, developments in Natural Language Processing (NLP) have introduced deep learning models such as Bidirectional Encoder Representations from Transformers (BERT), which significantly improve the understanding of contextual information in text. Unlike traditional methods that rely heavily on handcrafted features, BERT leverages pre-trained language representations to capture deeper semantic relationships, thereby enhancing the detection of more subtle and sophisticated phishing attempts.

Although these techniques have demonstrated considerable success individually, relying on a single model often leads to inconsistent performance when faced with diverse and evolving email threats. This challenge has led to increased interest in ensemble learning strategies, which combine multiple models to achieve better predictive performance. By integrating different classifiers, ensemble methods can exploit the unique strengths of each approach, resulting in improved accuracy, reduced error rates, and greater system stability.

In response to these challenges, this study proposes a hybrid ensemble framework that combines traditional machine learning models (Naïve Bayes and SVM) with a deep learning model (BERT) through a weighted voting scheme optimized using cross-validation. In contrast to approaches that depend solely on either classical algorithms or computationally intensive deep learning models, the proposed method achieves a balance between efficiency and performance. Furthermore, the study incorporates a hybrid dataset consisting of locally sourced institutional emails alongside publicly available datasets, thereby enhancing diversity and ensuring practical relevance. The major contributions of this work include: (i) the development of a scalable ensemble architecture that integrates both statistical and contextual text representations; (ii) the application of an optimized weighted voting mechanism to improve classification outcomes; and (iii) an extensive evaluation demonstrating enhanced accuracy and reduced misclassification in detecting spam and phishing emails.

Accordingly, this research focuses on designing and implementing an ensemble-based email filtering and classification system that unifies Naïve Bayes, Support Vector Machine, and BERT within a single framework. By merging conventional machine learning approaches with advanced deep transfer learning techniques, the proposed

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

system aims to deliver higher classification accuracy and more reliable phishing detection. Ultimately, this work contributes to improving the security of email communication by offering a robust and scalable solution capable of adapting to the continuously evolving landscape of email-based threats.

2.0 Related Works

The detection of spam and phishing emails has received extensive research focus in recent years, largely due to the growing complexity and frequency of email-based cyber threats. Initial research efforts concentrated on traditional machine learning algorithms, whereas more recent studies have shifted toward ensemble, hybrid, and deep learning techniques to improve classification accuracy and robustness.

Najah *et al.* (2024) introduced an ensemble framework based on a stacking approach that integrates Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN), with Logistic Regression serving as the meta-learner. Their model, evaluated using 10-fold cross-validation, achieved an accuracy of 95.8%, surpassing the performance of individual classifiers. Despite this improvement, the approach increased computational demands and did not incorporate deep learning models, thereby limiting its capability to capture more intricate patterns in email data.

Similarly, Alsudani *et al.* (2024) proposed a hybrid model that combines Crow Search Optimization with Feedforward Neural Networks and Long Short-Term Memory (LSTM) networks for spam detection. This method demonstrated strong performance and effectively reduced false positive rates by modeling sequential relationships in email content. However, the inclusion of complex optimization strategies may hinder scalability, particularly in real-time deployment scenarios.

Anirudh *et al.* (2024) developed a phishing detection system using a stacking ensemble of SVM and XGBoost as base classifiers, with Logistic Regression as the meta-classifier. By applying TF-IDF feature extraction to a dataset of 18,650 emails, the model achieved improved detection accuracy. Nevertheless, the reliance on traditional feature engineering and the associated computational overhead pose limitations in terms of scalability and efficiency.

In another study, Temidayo and David (2023) implemented an optimized ensemble model combining Random Forest and XGBoost with extensive hyperparameter tuning. Their approach enhanced classification performance and mitigated overfitting issues. However, the use of exhaustive grid search increased computational cost, and the evaluation on a single dataset limited the generalizability of the findings.

Douzi *et al.* (2023) examined the evolution of spam detection techniques from conventional Bag-of-Words and Bayesian models to modern embedding-based approaches such as Word2Vec and Paragraph Vector–Distributed Memory (PV-DM). Their analysis highlighted that integrating TF-IDF with embedding techniques improves contextual representation of text. Nonetheless, they emphasized the need for continuous model adaptation to effectively respond to evolving spam strategies.

Nallabariki *et al.* (2021) conducted a review of classical machine learning algorithms, including Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines, using benchmark datasets such as SpamAssassin and Enron. Their findings underscored the importance of feature selection and dimensionality reduction in improving model performance. However, the study lacked comprehensive experimental validation and did not address real-world deployment challenges.

Azeez and Ologe (2022) explored the use of ensemble learning for phishing detection by combining classifiers such as KNN, Decision Tree, Random Forest, Naïve Bayes, and SVM. While their ensemble approach improved classification accuracy, the use of limited datasets and reliance on predefined features may restrict its ability to adapt to emerging and dynamic threats.

Earlier work by Sharma and Bhardwaj (2018) reviewed a range of machine learning and hybrid techniques for spam detection, emphasizing the benefits of combining algorithms such as Naïve Bayes with cryptographic methods like Secure Hash functions. Although hybrid approaches enhanced detection rates, challenges including dataset imbalance and computational inefficiency remained significant concerns.

Overall, existing studies demonstrate that ensemble and hybrid models generally outperform single classifiers in detecting spam and phishing emails. However, many of these approaches either depend heavily on traditional machine learning methods, involve high computational complexity, or lack effective integration with deep learning techniques. Furthermore, the reliance on limited or outdated datasets reduces their ability to handle modern and evolving threats. These limitations reveal a critical research gap, emphasizing the need for scalable, context-aware ensemble frameworks that effectively combine classical and deep learning methods to deliver robust and adaptive email classification systems.

3.0 Methodology

In this study, the Multinomial Naïve Bayes algorithm was adopted due to its suitability for handling text classification tasks characterized by discrete term frequencies. This approach performs particularly well when

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

applied to feature representations derived from TF-IDF, as it effectively models word occurrence patterns within documents.

The Support Vector Machine (SVM) classifier was implemented with a Radial Basis Function (RBF) kernel to enable the modeling of non-linear relationships within high-dimensional data spaces. This configuration enhances the model's ability to differentiate between intricate patterns present in email content, thereby improving classification performance.

For the deep learning component, a pre-trained transformer model, BERT (bert-base-uncased), was utilized to capture contextual and semantic relationships in textual data through bidirectional encoding. The model was further fine-tuned on the prepared dataset using a learning rate of 2×10^{-5}, a batch size of 16, and a training duration of three epochs. Optimization during training was carried out using the Adam optimizer to facilitate stable and efficient convergence.

To evaluate the effectiveness of the proposed models, the dataset was partitioned into training and testing subsets using an 80:20 ratio. Additionally, cross-validation techniques were employed during the training phase to improve model reliability and ensure better generalization across unseen data.

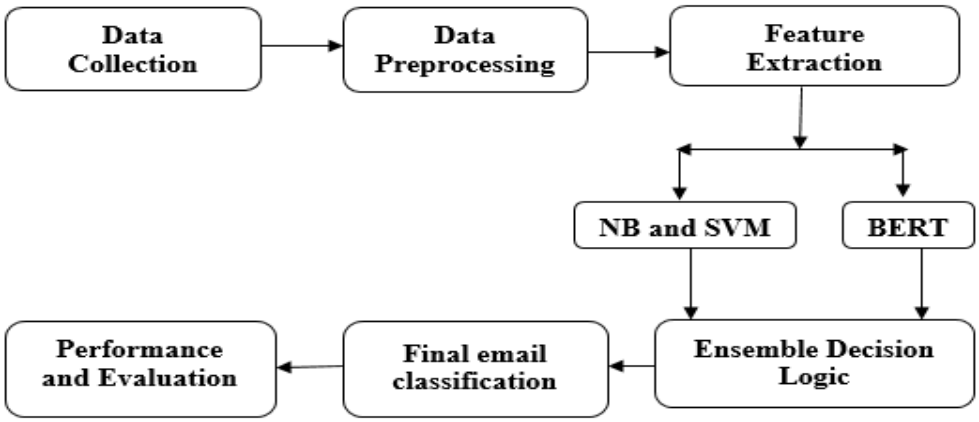

Figure 1 Block Diagram of the Deep Learning Models

3.1 Dataset Description

This study uses a hybrid dataset collected from Osun State Polytechnic, Iree and Federal Polytechnic, Offa, combined with a public dataset from Kaggle. The dataset includes email attributes such as sender, subject, body, and spam/legitimate labels. After preprocessing and merging, the final dataset contains 2,500 emails, consisting of 1,250 spam and 1,250 legitimate messages.

3.2 Data Preprocessing

Prior to model development, a series of preprocessing steps were applied to prepare the email dataset for effective classification. The dataset, which comprised both locally sourced emails and samples obtained from a Kaggle repository, was first harmonized to ensure consistency before integration.

For the traditional machine learning models, namely Naïve Bayes and Support Vector Machine (SVM), standard text cleaning procedures were carried out. These included converting all text to lowercase, eliminating punctuation and common stop words, and segmenting the text into tokens. The cleaned data was subsequently converted into numerical form using the TF-IDF technique to facilitate model training.

In contrast, the preprocessing requirements for the BERT model were minimal. This is because the built-in tokenizer of BERT inherently manages text segmentation, the inclusion of special tokens, and sequence padding, thereby streamlining the preparation process for deep learning-based classification.

3.3 Model Selection

A range of machine learning and deep learning techniques were assessed for their effectiveness in email filtering and classification, with consideration given to performance accuracy, computational efficiency, and compatibility with textual data. Following a comparative evaluation, three models Naïve Bayes (NB), Support Vector Machine (SVM), and Bidirectional Encoder Representations from Transformers (BERT) were identified as the most suitable and subsequently integrated into the proposed ensemble framework.

3.3.1 Naïve Bayes (NB)

Naïve Bayes is a probabilistic classification algorithm commonly applied in text analysis tasks because of its computational efficiency and ability to handle high-dimensional feature spaces. It operates based on Bayes' theorem, estimating the likelihood that a given email belongs to a particular category by analyzing the distribution

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

and occurrence of words within the text. Formula for calculating Naïve Bayes is represented by equation 1. The posterior probability of a class given an input feature set is expressed as:

$$P(C \mid X) = \frac{\mathrm{P(X \mid C)P(C)}}{P(X)} \quad (1)$$

Where: C represents the class label (spam or legitimate), X denotes the feature vector of the email, P(C I X) is the posterior probability of class C given features X, P(X I C) is likelihood of features given class C, P(C) is the prior probability of class C, and P(X) is the probability of the feature vector.

3.3.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm that separates data into classes by identifying an optimal hyperplane in a high-dimensional space. It is effective for text classification and helps identify subtle patterns in email content. Formula for calculating Support Vector Machine is represented by equation 2, 3, 4 and 5. SVM Optimization is defined as:

$$min\frac{1}{2}\left||w|\right|^2 + C\sum_{i=1}^{n}\xi i \quad (2)$$

SVM Constraints is expressed as:

$$y_i(w \cdot x_i + b) \geq 1 - \xi i, \ \xi i \geq 0 \quad (3)$$

SVM Decision Function (Linear) is expressed as:

$$f(x) = w^T x + b \quad (4)$$

SVM Kernel Form is expressed as:

$$f(x) = \sum_{i=1}^{n} \alpha \mathrm{iyiK}(x_i, x) + b \quad (5)$$

Where: w is the weight vector, x is the input feature vector, b is the bias term, α_i are the Lagrange multipliers, y_i are the class labels, and K (x_i, x) is the kernel function (e.g., RBF kernel).

3.3.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a deep learning architecture built on the transformer framework, designed to produce context-aware text representations through bidirectional processing of input sequences. By leveraging self-attention mechanisms, it effectively models semantic relationships within email content, enabling a deeper understanding of contextual meaning. Formula for calculating Bidirectional Encoder Representations from Transformers is represented by equation 6. The contextual representation of a token is given as:

$$h_i \in \mathbb{R}^d \quad (6)$$

Where: h_i is the contextual embedding of token i and d is the dimensionality of the embedding space.

3.4 Ensemble Decision Logic

The proposed ensemble framework integrates the predictions of Naïve Bayes (NB), Support Vector Machine (SVM), and BERT using a weighted voting mechanism. In this approach, each classifier contributes to the final decision according to a weight derived from its performance on validation data, such that models with higher accuracy exert greater influence on the outcome. The final classification whether an email is spam or legitimate is determined by aggregating the weighted outputs and selecting the class with the highest combined score. This strategy enhances both the accuracy and the stability of the overall system. Formula for calculating the Ensemble Model is represented by equation (7) and is expressed as:

$$\hat{y} = \arg\max_{c \epsilon C} \sum_{m=1}^{M} w_m \ . \ 1\{h_m(x) = C\} \quad (7)$$

Where: ŷ is the final predicted class, C is the set of possible classes (spam or legitimate), M is the total number of models, w_m is the weight assigned to model mmm, h_m (x) is the prediction of model mmm, and 1(·) is the indicator function (equals 1 if the condition is true, otherwise 0).

3.5 Model Evaluation

The effectiveness of the proposed ensemble model, along with its individual components Naïve Bayes, Support Vector Machine (SVM), and BERT was assessed using widely accepted classification performance metrics. These evaluation measures offer a clear indication of how well each model differentiates between spam and legitimate email messages. The key evaluation metrics used in this study are discussed below:

i. **Accuracy Score:** Accuracy measures the proportion of correctly predicted emails (both spam and legitimate) out of the total number of emails processed. It offers a general measure of overall model performance. This metric is represented in equation 8.

$$\text{Accuracy Score} = \frac{\mathrm{TP + TN}}{\mathrm{TP+TN+FP+FN}} \times 100 \quad (8)$$

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

Where: TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

ii. **Precision (Positive Predictive Value):** Precision quantifies the ratio of emails classified as spam that were actually spam. A high precision score shows that the model is active at avoiding the misclassification of legitimate emails as spam. Formula for calculating precision is represented by equation 9.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (9)$$

iii. **Recall (Sensitivity or True Positive Rate):** Recall evaluates how effectively the model is able to detect all actual spam emails within the dataset. A high recall value indicates that the classifier successfully identifies the majority of spam messages, minimizing the number of spam emails that go undetected. This is represented by equation 10.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (10)$$

iv. **F1-Score:** The F1-score is the harmonic mean of precision and recall. It provides a balanced evaluation, particularly useful when the dataset is imbalanced between spam and legitimate emails. F1 score is represented in equation 11.

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (11)$$

3.6 Model Implementation

The Naïve Bayes, Support Vector Machine (SVM), and BERT models were deployed within a Streamlit-based framework to build an interactive web application for email spam classification. This interface enables users to submit email content and obtain real-time prediction results. Incoming text is subjected to the same preprocessing pipeline used during model training to ensure consistency. For the Naïve Bayes and SVM classifiers, TF-IDF is applied to transform textual data into numerical feature vectors, whereas the BERT model utilizes its native tokenizer to generate contextual embeddings. The pre-trained models are saved and reloaded using Joblib, after which each model independently performs classification on the processed input.

3.7 Data Distribution Analysis

A complete study of the labeled email dataset was conducted to understand feature characteristics and class balance, which are critical for model performance and reliability. Emails were categorized as spam or legitimate, and key features were examined, including email length, uppercase word count, punctuation density, keyword frequency, and number of URLs.

The analysis revealed distinct patterns between spam and legitimate emails. Spam emails generally exhibited shorter lengths, higher capitalization, denser punctuation, and contained more promotional keywords and URLs, reflecting common strategies used by spammers to attract attention or evade simple filtering rules. In contrast, legitimate emails tended to have longer content, more natural punctuation patterns, and fewer promotional elements.

The dataset demonstrated a moderately balanced class distribution, with spam and legitimate emails occurring in roughly comparable proportions. This balance reduces potential bias in model exercise and supports the effectiveness of the selected features for unique between the two classes. Understanding these distributions provides insight into why certain classifiers, such as BERT, which captures contextual relationships, can outperform traditional models that rely primarily on surface-level statistical features.

3.8 Application of the Developed Email Classification System

The developed email classification system provides a user-friendly interface that allows users to input or paste email text for analysis. Once submitted, the system preprocesses the text and classifies it using an ensemble model combining Naïve Bayes, Support Vector Machine (SVM), and BERT. Based on the prediction, the system determines whether the email is spam or not spam. Spam messages are blocked from reaching the user's inbox, while legitimate messages are allowed through. This implementation proves the success of group knowledge for email classification through a simple and responsive web interface, as illustrated in figure 2: The user Interface of the Implemented Model.

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

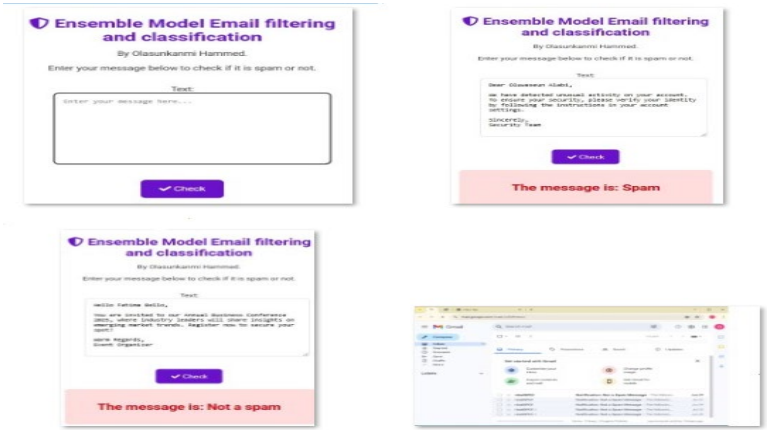

Figure 2: The User interface of the implemented model

4.0 Results and Discussion

The investigational outcomes determine that the proposed group classic completes greater recital equaled to separate classifiers. Specifically, the model attained 98% accuracy, 97% precision, 98% recall, and 98% F1-score. The improved performance is attributed to the integration of complementary learning techniques, where Naïve Bayes captures statistical patterns, SVM provides optimal decision boundaries, and BERT extracts deep contextual features. The weighted voting mechanism further enhances classification reliability by prioritizing high-performing models.

4.1 Individual Model Evaluation

Each base classifier was evaluated independently using Accuracy, Precision, Recall and F1-score to evaluate its success in email classification.

Table 1: Performance evaluation of Individual Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	94.00	93.00	94.00	93.00
SVM	96.00	95.00	96.00	96.00
BERT	97.00	97.00	96.00	96.00

The evaluation of individual classifiers revealed notable differences in performance. The Naïve Bayes (NB) model achieved reasonable accuracy but occasionally misclassified legitimate emails as spam, indicating a tendency toward false positives. This limitation is consistent with the dataset analysis, where certain legitimate emails contained features commonly associated with spam, such as high punctuation density or capitalization.

The Support Vector Machine (SVM) demonstrated improved stability among precision and recall compared to NB, reflecting its ability to better separate the feature space and handle borderline cases. SVM's hyperplane-based classification allowed for more consistent discrimination between spam and legitimate emails, reducing misclassification rates.

The BERT model achieved the highest overall performance across all evaluation metrics. Its superior accuracy, precision, recall, and F1-score can be attributed to its ability to capture contextual and sequential relationships in email text, which traditional classifiers cannot exploit. By understanding semantic patterns and the context of disputes inside emails, BERT effectively distinguished nuanced spam from legitimate messages, addressing limitations observed in NB and SVM.

Overall, these results highlight the progressive improvement from classical statistical models to deep learning approaches, emphasizing the value of contextual embeddings for robust email classification.

4.2 Ensemble Model Evaluation

To enhance classification robustness, a voting-based ensemble strategy was implemented by uniting the forecasts of Naïve Bayes (NB), Support Vector Machine (SVM), and BERT. The ensemble aggregates the crops of the separate models using a weighted voting mechanism, where the last period label for each email is determined using a weighted voting mechanism, where each model contributes based on its assigned weight.

The ensemble approach leverages the complementary strengths of the base classifiers: NB effectively captures statistical text features, SVM provides robust separation of feature space, and BERT contributes contextual understanding of email content. By integrating these replicas, the ensemble mitigates the faintness of separate

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

classifiers, such as NB's tendency for false positives or SVM's sensitivity to feature overlap, resulting in more balanced and reliable predictions.

The ensemble model achieved 98% accuracy, 97% precision, 98% recall, and 98% F1-score, outperforming all individual classifiers. Cross-validation confirmed the constancy of these results, with minimal variance across folds, indicating that the typical simplifies fine to hidden files. These findings demonstrate that weighted voting not only improves overall performance but also reduces misclassifications by reconciling conflicting predictions from the base classifiers.

Overall, the ensemble evaluation highlights the effectiveness of combining classical machine learning and deep learning models to create a context-aware, robust, and high-performing email spam detection system.

Table 2: Performance Metrics for the Ensemble Model

Model	Accuracy	Precision	Recall	F1-Score
Ensemble (Vote)	98%	97%	98%	98%

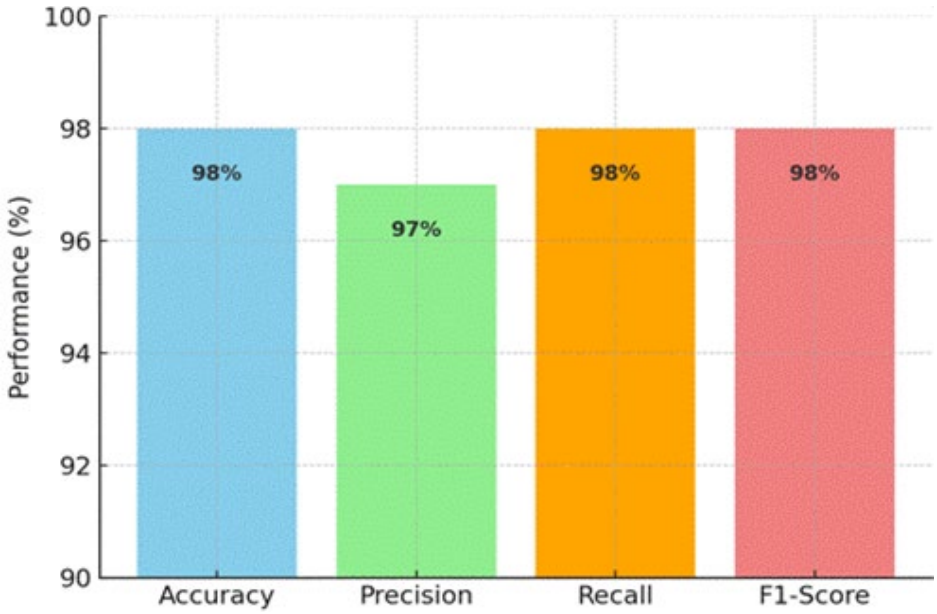

Figure 3: performance metrics of the Ensemble Model

The proposed ensemble framework attained an overall accuracy of 98% and an F1-score of 98%, marginally outperforming the strongest individual model, BERT. In addition, it recorded a precision of 97% and a recall of 98%, reflecting a balanced capability in correctly classifying both malicious and legitimate email messages. These outcomes indicate that the weighted voting strategy successfully leverages the complementary strengths of Naïve Bayes, Support Vector Machine, and BERT, while simultaneously reducing the impact of individual model weaknesses and minimizing classification errors.

Through the combination of multiple model predictions, the ensemble produced more stable and reliable decisions, particularly in ambiguous or borderline cases where individual classifiers may produce conflicting outputs. Furthermore, the application of cross-validation during evaluation confirmed that the model's performance remains consistent across different data partitions, thereby reducing concerns of overfitting or random performance gains. Overall, these findings emphasize the effectiveness of integrating traditional machine learning techniques with deep learning models in developing a more adaptive and context-aware email spam detection system.

4.3 Confusion Matrix Analysis

To further analyze model performance, the misperception medium of the ensemble model was examined, as shown in Table 3.

Table 3: Confusion Matrix for Ensemble

	Predicted Spam	Predicted Legitimate
Actual Spam	1,210	40
Actual Legitimate	30	1,220

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

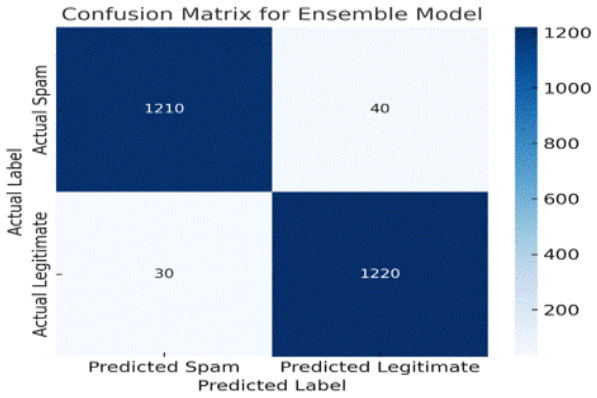

Figure 4: Confusion Matrix for the Ensemble Model

The group perfect attained a balanced trade-off between precision and recall, outperforming individual classifiers by reducing classification errors. It misclassified 40 spam emails as legitimate (false negatives) and 30 legitimate emails as spam (false positives), indicating improved accuracy in email classification.

4.4 Model Comparison and Summary

The overall performance of the developed models is summarized in Table 4. The comparison includes Naïve Bayes, Support Vector Machine (SVM), BERT and the proposed ensemble model evaluated using Accuracy, Precision, Recall and F1-score.

Table 4: Overall Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	94%	93%	94%	93%
SVM	96%	95%	96%	96%
BERT	97%	97%	96%	96%
Ensemble (Vote)	98%	97%	98%	98%

The experimental results demonstrate a clear performance progression from conventional machine learning techniques to deep learning, and ultimately to the ensemble model. Among the individual classifiers, Naïve Bayes recorded the lowest performance, while Support Vector Machine (SVM) exhibited a noticeable improvement in classification accuracy. The BERT model further enhanced performance owing to its ability to capture deep contextual and semantic relationships within textual data, resulting in more accurate predictions. The ensemble model achieved the best overall performance, attaining 98% accuracy and F1-score.

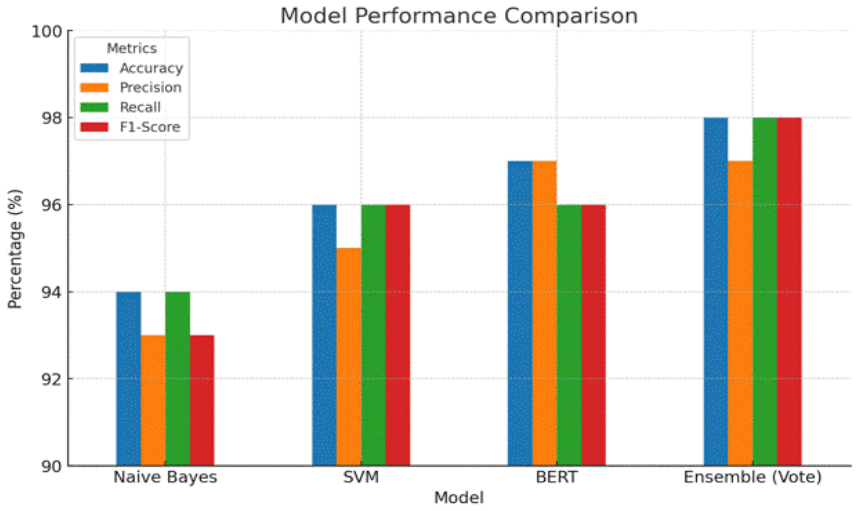

Figure 5: Overall Model Performance Comparison

The ensemble model consistently achieved the highest scores across all evaluation metrics, demonstrating superior effectiveness in distinguishing between spam and legitimate emails. This improvement is attributed to the

combination of different classification strategies, which enhances robustness and reduces individual model limitations. These results highlight the effectiveness of ensemble learning for email spam detection tasks.

4.5 Model Comparison with Existing Studies

A comparative analysis with prior studies further demonstrates the superiority of the proposed method. For instance, Najah *et al.* (2024) reported an accuracy of 95.8% using a stacking-based ensemble approach, while Anirudh *et al.* (2024) and Temidayo and David (2023) also obtained strong performance using hybrid machine learning frameworks. In comparison, the model proposed in this study achieved an accuracy of 98%, indicating a clear improvement over existing approaches.

This performance gain can be attributed to the combination of deep contextual feature extraction using BERT with traditional machine learning classifiers, alongside the use of a weighted voting mechanism that enhances the stability and reliability of final predictions.

Overall, the findings indicate that the proposed ensemble framework provides a more effective and scalable solution for addressing contemporary challenges in email classification tasks.

5.0 Conclusion

This study developed a high-performance ensemble-based email classification system by integrating Naïve Bayes, Support Vector Machine (SVM), and BERT within a unified framework. The proposed model achieved superior classification results, demonstrating its effectiveness in detecting spam and phishing emails with high accuracy and reliability. The combination of classical machine learning and deep learning techniques enabled the system to capture both statistical and contextual features, significantly improving classification performance. The weighted voting ensemble strategy further enhanced robustness by reducing misclassification errors. Overall, the study provides a scalable and efficient solution for secure email filtering, contributing to advancements in intelligent cybersecurity systems.

6.0 Recommendation

Future research may estimate the future ensemble model using greater and added varied datasets to improve generalizability across different domains. Further studies may also explore advanced transformer architectures such as RoBERTa or DistilBERT to enhance contextual feature learning and classification performance. Additionally, multilingual spam detection and real-time deployment within operational email systems could be investigated to assess scalability and practical applicability. Finally, future work may examine the model's robustness against adversarial spam and evolving phishing techniques.

References

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, 2007, pp. 60–69.

[2] M. A. Ahmed, M. A. Qureshi, and M. Khan, "A comparative study of machine learning and deep learning models for spam detection," *Journal of Computer Science*, vol. 18, no. 2, pp. 123–136, 2022.

[3] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, pp. 1–6, 2014.

[4] M. Al-Ajeli, F. Al-Anazi, and A. Al-Mutairi, "Comparative analysis of machine learning algorithms for spam filtering," *Journal of Computer Networks and Communications*, vol. 2020, pp. 1–9, 2020.

[5] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in *Proc. 11th ACM Symposium on Document Engineering*, 2011, pp. 259–262.

[6] A. Alsudani, L. Alzubaidi, and J. Zhang, "Optimizing spam detection using hybrid models combining CSO with FFNN and LSTM," *Expert Systems with Applications*, vol. 234, p. 119552, 2024.

[7] R. Anirudh, N. Kumar, and M. Shubham, "A stacking ensemble model for phishing email classification," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 1, pp. 204–210, 2024.

[8] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 160–167.

[9] T. O. Ayodele, K. S. Adewole, and E. O. Omidiora, "Email summarization and classification based on user activity," *Journal of Computer Science and Its Applications*, vol. 14, no. 1, pp. 100–112, 2007.

[10] N. Bacanin, S. Vukovic, and M. Tuba, "SCA-ML: A hybrid spam detection system based on machine learning and swarm intelligence," *Applied Soft Computing*, vol. 121, p. 108772, 2022.

[11] M. T. Banday and T. R. Jan, "A profile-based architecture for detecting email spam using artificial neural networks," *International Journal of Computer Applications*, vol. 6, no. 6, pp. 9–14, 2009.

[12] M. Beaman and H. Isah, "Feature extraction methods for phishing detection: A systematic review," *Computer Science Review*, vol. 45, p. 100497, 2022.

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026

[13] T. Bhuiyan, M. M. A. H. Rony, and R. M. Rahman, "An improved email spam classification technique using support vector machine," *International Journal of Computer Applications*, vol. 178, no. 22, pp. 25–30, 2019.
[14] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[15] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, pp. 63–92, 2008.
[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[17] X. Carreras and L. Márquez, "Boosting trees for anti-spam email filtering," in *Proc. International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2001, pp. 58–64.
[18] D. Champa, J. Thomas, and R. Ghosh, "Phishing email detection using deep contextualized embeddings," *IEEE Access*, vol. 12, pp. 56022–56033, 2024.
[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[20] G. V. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2007.
[21] G. V. Cormack, "Email spam filtering: A systematic review," *Information Retrieval*, vol. 11, no. 3, pp. 203–250, 2008.
[22] G. V. Cormack and T. R. Lynam, "Spam corpus creation for TREC," in *Proc. Conference on Email and Anti-Spam (CEAS)*, 2005.
[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.
[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Received: 25/03/2026; Revised: 13/04/2026; Accepted: 17/04/2026; Published: 25/04/2026