Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

A Hybrid Ai Framework For Intelligent Spam Email Detection

Didar Hussain (Corresponding author)
Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan
Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan.
didarhussain@bkuc.edu.pk

Sefat Ghayoor Khan
Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan.
khansifatghayoor@gmail.com

Abdullah Sadiq
Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan.
as4714081@gmail.com

Waqas Khan
Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan.
waqaskawdari@gmail.com

Azaz Ali
Department of Computer Science, Abasyn University Peshawar, KPK, Pakistan.
azazali458@gmail

Shams ul Arifeen
Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan.
shamsjan99090@gmail.com

Adnan Khan
Department of Computer Science, Abasyn University Peshawar, KPK, Pakistan.
adnanalikhan57@gmail.com

Muhammad Awais
Department of Computer Science, Bacha Khan University Charsadda, KPK, Pakistan.
muhammadawais.bkuc@gmail.com

ABSTRACT
With the exponential growth of digital transmission, email remains one of the multiple widely used venues. Yet, the peak of unasked and hostile emails typically directed to as spam poses a considerable danger to data security and user aloneness. This study offers a relative analysis of classical machine learning algorithms for practical spam email detection. A real-world dataset consisting of over 5,500 email notifications marked as either spam or ham (non-spam) was employed. The dataset underwent preprocessing, including text normalization and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF). Class inequality was handled using the Synthetic Minority Over-sampling Technique (SMOTE). Nine machine learning models Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)
(LR), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Gradient Boosting (Boost), CNN, and Distil BERT + XGB were prepared. Model interpretation was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Among all models, CNN reached the highest accuracy and precision of 99.85%, along with 93% recall and a 96% F1 score, indicating healthy detection capability with minimal false positives. The findings suggest that ensemble-based techniques, particularly CNN and RF, offer robust and scalable solutions for real-time spam detection systems. This research contributes to the development of clever email filtering systems and offers a foundation for future advancements using deep knowledge and explainable AI.

Keywords: Component; formatting; style; styling; insert

Introduction

In the age of digital contact, email has evolved a critical piece of personal, corporate, and governmental exchanges. However, its overall use has also made it a prime target for spam avoided, deceptive, and often negative messages. Spam emails pose not only a problem but also a serious cybersecurity threat, running from phishing attacks to the dispersal of malware. As attackers continuously develop their tactics, standard rule-based filters struggle to keep pace. This growing complexity underscores the critical need for intelligent, adaptive systems capable of accurately differentiating between spam and legitimate messages. Machine learning, particularly managed learning, emerges as a promising solution to this challenge by proposing models that learn from labeled data to detect patterns beyond simple keyword filters. This section sets the footing for exploring how supervised machine learning techniques can significantly enhance the prediction and filtering of spam emails with greater accuracy, scalability, and resilience.

Email is an vital tool for communication, vastly used across industries, management, and private domains [1]. Its favor yet, has led to its exploitation as a medium for spam uninvited, and usually malicious messages that disrupt user experience and risk data protection. Spam emails vary from mere promotional content to damaging phishing frauds and malware incursions. Studies assess that spam reports for over 70% of international email gridlock, creating it a constant and increasing cybersecurity crisis [2]. One of the numerous threatening conditions of spam is phishing, where intruders copy authorized entities to mislead recipients into revealing liable information, transferring funds, or connecting malicious links [3].

The increasing complexity of such attacks drives them hard to predict utilizing traditional, rule-based spam filters. These fixed systems rely laboriously on predefined blurred content and turning sender individualism [4].

To overwhelm the boundaries of conventional approaches, investigators have increasingly embraced machine learning (ML) techniques specifically supervised learning for spam prediction. Supervised machine learning models are taught on tagged or labeled datasets, comprehending to categorize emails as either spam or honest based on statistical practices and attributes extracted from the dataset [5].

Supervised schooling models such as Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT), K-Nearest Neighbors (KNN), and Random Forests (RF) have shown heightened performance in spam email classification. For example, studies show that SVMs show strong generality on massive email data, while NB classifiers excel in terms of computational efficiency and probabilistic interpretability. RF and DT provide rich accuracy and opposition to overfitting, making them suitable for real-life applications [6] [7]. A noted instance is the application of an Artificial Neural Network (ANN)

ISSN Online: 3007-3154
ISSN Print: 3007-3146

classifier that gained an accuracy of 99.91% on the Spam dataset, although more detailed supervised methods have often reached proximate performance with greater interpretability and more downward computational cost [5]. Further, costume models such as boosting and bagging, when used with supervised models, have additionally improved the results by decreasing variance and enriching model strength[8].

Despite these refinements, issues remain. Challenges such as category imbalance where spam emails may enormously outnumber legitimate ones in particular datasets can skew model understanding. Similarly, the constant development of spam tactics needs strategies that can generalize well to new ways without regular retraining. Another essential concern is interpretability, in multiple applications, particularly in business environments, spam detection measures must have fine reasons for their findings.

Given these persisted issues and the advantageous performance of supervised learning models, this research suggests investigating and improving spam email prediction utilizing different SML models. The ambition is to develop a strategy that compensates for high performance, adaptability to current spam emails, and computational efficiency while maintaining model interpretability.

Related Work

Sidam et al., [1] compared multiple machine learning models Naive Bayes, SVM, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting for spam detection using a dataset of 5,000 emails enriched with word commonness, keyword, and metadata components. Their study stressed the effectiveness of costume models and diverse feature sets in capturing involved spam patterns, achieving high accuracy and ROC-AUC. However, limitations included a small dataset, class inequality, and lack of real-time testing.

Zawahra et al., [2] evaluated various ML models, including Naive Bayes, SVM, and Random Forest, on a dataset of 5,172 emails using content-based and metadata components. They achieved model versions up to 98%, especially noting Naive Bayes' strength in text classification. While their study provided strong close insights, it was limited by dataset size, accuracy variation across groups, and lack of detailed feature piece or real-time reflections.

Zhang et al., [3] utilized Naive Bayes, SVM, and a combined NB ∪ SVM approach to classify 5,172 emails as spam or non-spam. SVM demonstrated high precision (98.1%) and low false positive rates, while Naive Bayes recorded a 0% false spam rate. Despite the use of a useful NLP dataset, the combined model still yielded a 12.7% false legit rate, with the Naive Bayes model suffering from a high 28.7% false legit rate and a small dataset size.

Sharma et al. [4] developed a spam detection system using Naive Bayes and the J48 decision tree algorithm on the Ling-Spam dataset. J48 outperformed Naive Bayes across all metrics, offering balanced precision and recall while leveraging real spam data. However, the reliance on a single data corpus and absence of false positive/negative reporting were noted weaknesses.

Meghna et al., [5] conducted an evaluation of spam detection models using Logistic Regression and Naive Bayes on a dataset of 5,573 emails. The study indicated high accuracy and level performance metrics across both models. However, the dataset's inequality, restriction to basic ML models, and lack of deep learning methods or in-depth feature research limited the study's depth.

Murti et al., [6] focused on real-world phishing email detection using the Gaggle dataset and evaluated classifiers such as Naive Bayes, SVM, Random Forest, Decision Tree, and KNN. Random Forest yielded the best results, demonstrating the model's robustness when

ISSN Online: 3007-3154
ISSN Print: 3007-3146

applied to real-world data. Still, the study sorrowed from narrow feature analysis and more increased error rates in KNN and Decision Tree models.

Pal et al., [7] introduced a hybrid spam detection framework combining SMOTE-ENC for class balancing with deep learning and classical classifiers on Twitter data. Random Forest emerged as the top performer with 99.26% accuracy, 99.49% precision, and 99.07% recall, while Gaussian Naive Bayes significantly underperformed. This highlights the power of ensemble models and resampling, despite some models' weak performance.

Malhotra et al., [8] applied traditional ML and Bi-LSAT models for spam detection using various performance metrics. Using Python, they implemented models such as LR, NB, RF, DT, Dense NN, LSTM, and Bi-LSTM. Bi-LSAT delivered the best results with 98.5% accuracy and a strong precision-recall balance. On the other hand, Decision Tree underperformed with 92.6% accuracy, and simpler models failed to match deep models' effectiveness.

Ejirika et al., [9] utilized a subset of the Enron1 dataset to apply models including CART, SVM, Naive Bayes, and Random Forest using R. Their results showed Random Forest achieving the highest accuracy (98.39%) and consistent performance across all key metrics. However, the study's narrow dataset scope limited its generalizability to broader or more complex email corpora.

Siddique et al., [10] contributed a unique approach by developing a custom Urdu-script spam dataset, translated from English using Google Translate. They tested CNN, LSTM, Naive Bayes, and SVM on this dataset, with LSTM achieving the highest accuracy of 98.4%. Despite its novelty in Urdu-language spam detection, the small dataset and manual translation process introduced potential human biases.

AbdulNabi et al., [11] conducted a comparative study of classical machine learning models and deep learning approaches using two public spam datasets—UCI and Kaggle. The models tested included KNN, Naive Bayes, BiLSTM, and BERT Base Cased. Their findings showed that BERT Base Cased outperformed all others, achieving 97.3% accuracy and an F1-score of 0.9696. However, traditional models like KNN and Naive Bayes underperformed, and the paper lacked a detailed discussion of the data preprocessing strategy.

Batra et al., [12] focused on spam classification using different categorization methods on the Gaggle spam dataset. Models like Logistic Regression, Decision Trees, and Naive Bayes were explored with various data vaporization techniques. The Decision Tree model performed the best, achieving 100% accuracy and a fast-training time of 1.74 seconds. However, the dataset showed a class imbalance with underrepresented spam samples, and the perfect accuracy pointed to a high risk of overfitting.

Redondo et al., [13] introduced a balanced spam-malware email dataset, SEMD-600, and evaluated several classification pipelines. They used BOW and TF-IDF for feature encoding, along with SVM, Naive Bayes, and Logistic Regression classifiers. The TF-IDF + Logistic Regression combination yielded the best results with 76.4% accuracy and an F1-score of 0.763, along with fast execution time. Nevertheless, overall accuracy across all models remained below 80%, and Naive Bayes with BOW gave the weakest performance.

Chinta et al., [14] assessed the effectiveness of Logistic Regression on a large real-world Gaggle email dataset containing 193,849 samples. Their model was compared with existing classifiers like Dilbert, Random Forest, HAM-FISTED, and Ensemble methods. The proposed Logistic Regression model outperformed others with an accuracy of 98.82%. However, the study was limited to Logistic Regression and did not explore deep learning models for performance comparison.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Sarker et al., [15] did not include detailed information in the summary, leaving a gap in understanding their specific contributions, techniques, and findings in the context of spam detection or email classification.

Sun et al., [16] employed classical statistical machine learning models Naive Bayes, KNN, and SVM on real-world email datasets for spam detection. Their models achieved an impressive 99% accuracy with balanced precision and recall. Despite this, the sample of spam emails was relatively small (747 samples), and no deep learning or hybrid approaches were considered in the study.

Keskin et al., [17] performed a relative analysis of various machine learning algorithms SVM, Logistic Regression, Naive Bayes, Random Forest, and Artificial Neural Networks (ANN) for spam detection. The Random Forest model produced the highest accuracy of 98.83% with a strong F1-score of 99.34%. Contrarily, Naive Bayes had the lowest accuracy (90.49%) despite its high precision, showcasing a gap in its robustness reached to ensemble methods.

Wang et al., [18] enhanced spam detection by integrating the TREC and Enron datasets and using multiple models such as SVM, Genetic Decision Tree, Particle Swarm Optimization (PSO), and CNN. Their approach delivered a high overall accuracy of 98% with proportional F1-scores. However, the study concentrated more on dataset-level comparisons than detailed, model-specific routine evaluations.

Abi Abraham et al., [19] worked with the Spam Assassin corpus and applied resampling techniques to create a balanced training dataset. Models like Decision Tree and Random Forest were used for classification, achieving perfect accuracy, recall, and F1-scores (100%) on test data. Despite these remarkable results, the absence of confirmation on unbalanced, real-world data raised concerns about model overfitting and generalizability.

Alsuwit et al., [20] merged and cleaned two large email datasets (TREC and Enron) and applied models including Logistic Regression, Random Forest, Naive Bayes, and Neural Networks. All models achieved high accuracy, with Neural Networks performing the best at 98%. However, performance differences among models were marginal, and the paper lacked a deeper analytical discussion of feature importance or model interpretability.

Sani et al., [21] conducted a comparative analysis using Decision Tree, Naive Bayes, and J48 classifiers on SMS spam datasets. They employed k-fold cross-validation to ensure reliable results. Among the models, the Decision Tree achieved the highest accuracy of 96% across both datasets. However, Naive Bayes underperformed, and the study did not incorporate deep learning or ensemble methods, limiting its innovation and scalability.

Qiqieh et al., [22] proposed a swarm-based cybersex threat detection system utilizing Harris Hawks Optimization (HHO) in combination with SVM across seven datasets. The system showed high accuracy, especially on spam email detection (96.68%). However, its performance dropped significantly on other datasets like Fake News (below 70%), and the absence of deep learning comparisons restricted its broader relevance.

Amutha et al. [23] introduced a feature selection approach using Sandpiper Optimization (SPO) combined with a Radial Basis Neural Network (RBNN) for spam detection. Tested on the Enron and Spam Assassin datasets, the SPO significantly improved classification accuracy. However, the work did not benchmark against deep learning methods and was limited in its application to larger datasets, reducing its practical applicability.

Alhuzali et al., [24] carried out a broad comparison of 14 machine learning and deep learning models across 10 datasets, including a newly merged corpus. They focused on traditional ML models versus transformer-based architectures such as BERT and Robert. Their findings confirmed that BERT and Roberta achieved the highest accuracy (up to 99.08%), emphasizing the superiority of deep learning. Nonetheless, these models required

ISSN Online: 3007-3154
ISSN Print: 3007-3146

significantly higher computational resources and implementation complexity.

Adnan et al., [25] combined and balanced two spam email datasets and tested five classifiers using a stacking ensemble method involving AdaBoost, Logistic Regression, Decision Tree, KNN, and Naive Bayes. The ensemble model outperformed individual classifiers, achieving an accuracy of 98.8%, recall of 98.8%, and F1-score of 98.9%. However, the architecture was more complex and required longer training time compared to standalone models.

Jamil et al., [26] explored spam detection using a unique image-based dataset comprising around 1,200 samples. They evaluated six models including ResNet50, Boost, Logistic Regression, Limelight, SVM, and VGG16. ResNet50 achieved the best overall performance, showcasing the potential of image-based spam classification. Nevertheless, the small dataset size posed a limitation in terms of generalization and scalability.

Deekshetha et al., [27] investigated the effectiveness of ML and DL models on a three-year, multi-source traffic data collection. They tested Linear Regression, KNN, Decision Tree, Random Forest, Boost, SVM, ANN, and LSTM. The comparative study revealed low overall model performance, with maximum accuracy around 33%, highlighting the difficulty of spam or traffic prediction from such complex data. The study, however, made a meaningful attempt at blending classical and deep learning techniques on real-world data.

Saeed et al., [28] applied Naive Bayes, Logistic Regression, Random Forest, and an ensemble model on the ESC dataset for spam classification. The ensemble model stood out with the highest accuracy of 98.9%, along with balanced precision, recall, and F1-score. Despite its success, individual models lagged, and the study remained confined to classical ML, omitting advanced or deep learning methods.

Singh et al., [29] evaluated traditional ML models like Decision Tree, KNN, Naive Bayes, SVM, and XGBoost on both small and full versions of the UCI spam dataset. Their comprehensive approach included a threshold-based classifier, which showed poor generalization with only 73% accuracy on the full dataset. This indicated the challenges of scaling simple classifiers without robust tuning or adaptation.

III Research Methodology

This section explains the methodological framework assumed to develop, train, and consider supervised machine education models for spam email detection. It outlines the technical environment, the structure and characteristics of the dataset used, and the systematic approach to data preprocessing, model selection, training, and performance evaluation. The research leverages classical machine learning algorithms, each chosen for its potential to generalize effectively on text classification tasks. Procedures such as TF-IDF feature extraction and class balancing are used to enhance data quality and speech issues such as high dimensionality and dataset inequality. The section also colors on key evaluation metrics accuracy, precision, recall, and F1-score to objectively assess model version. By following this rigorous methodology, the study aims to confirm the reliability, scalability, and interpretability of the suggested spam detection system. Figure 1 presents the proposed methodology.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

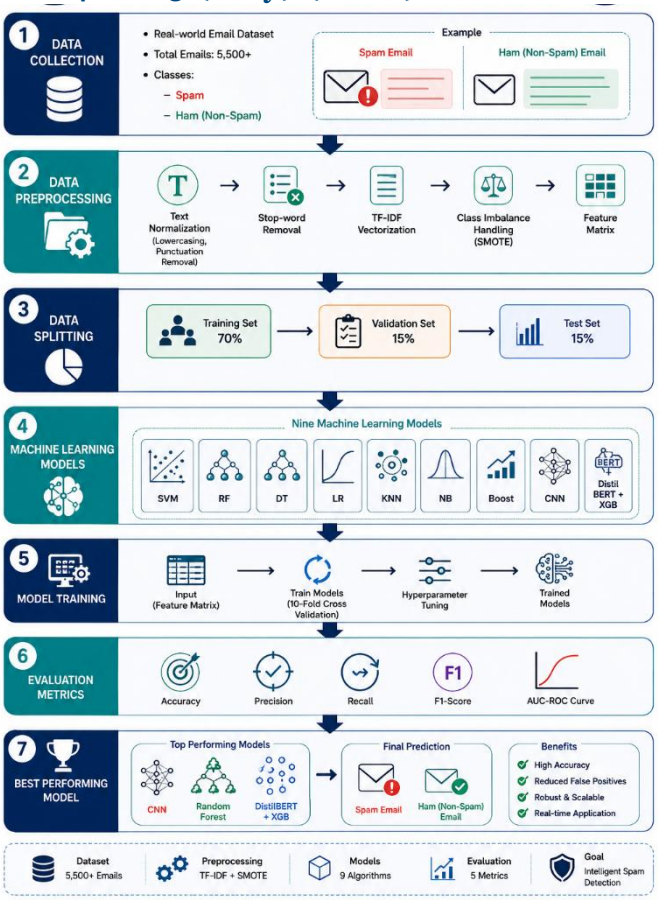

Figure 1 Proposed methodology

Computation Environment

The study used Python 3.10 with libraries including scikit-learn, pandas, NumPy, and NLTK. For deep learning models, activity was done using Google Colab with GPU support.

Dataset Description

The data set used in this study was received from the Kaggle repository. It contains a level and sample collection of email notifications categorized into two categories: spam and ham (non-spam). Each email model includes the full body text and metadata such as subject lines and headers. The dataset comprises a total of 5,572 email pieces, out of which about 39% are marked as spam and 61% as ham, remembering a moderately imbalanced type of problem. This dataset delivers a realistic basis for training and assessing machine learning algorithms in a practical spam detection scenario.

Before providing the data into the standards, the email content was preprocessed through various biological language processing (NLP) steps. These contained removal of HTML tags, punctuation, special symbols, stop words, lemmatization, and case normalization. The processed textual data was then converted into numerical form using TF-IDF (Term Frequency–Inverse Document Frequency) vectorization to effectively represent word importance across the dataset.

Dataset Splitting Strategy

To train and assess the machine learning standards fairly and invariably, the dataset was separated using a stratified 80:20 split, where 80% of the data was utilized for training and 20% for testing. Stratification ensured that the balance of spam and ham emails stayed consistent across both sets, which is required for handling the imbalanced nature of the data. Further, k-fold cross-validation (with k=10) was utilized during model evaluation to

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

validate the strength and generalization ability of each algorithm. This method decreased the risk of overfitting and ensured a more reliable assessment of model implementation.

Machine Learning Models

1) This research operated a mixed set of supervised machine learning algorithms typically used in text sort tasks. These models stood selected for their balance of interpretability, speed, and energy in determining patterns within textual data.
2) **K-Nearest neighbors**

   KNN is an easy yet powerful technique that labels a new email founded on the prevalence class among its 'k' nearest neighbors in the quality space. It relies on space measures like Euclidean or cosine parallel, making it especially suitable for high-dimensional data such as TF-IDF vectors. Despite its clarity, KNN can perform competitively when suitably tuned.
3) **Support Vector Machine**

   Support Vector Machines form a decision border that maximizes the margin between the two types in the vector space. Due to their robustness and capacity to work well in high-dimensional quality spaces, SVMs are widely used in spam detection tasks. The SVM classifier used in this research utilized a linear kernel, which delivered fast training and effective implementation.

4) Decision Tree

   The Decision Tree algorithm creates a hierarchical model that splits the dataset founded on feature importance, creating precise and interpretable rules for variety. While it is prone to overfitting, its power lies in its transparency and easy visualization of decision paths.

5) **Random Forest**

   Random Forest (RF) algorithm can be regarded as a robust integrated learning method, which is widely used in classification and regression. It builds multiple decision trees in the training process and achieves the result aggregation by majority voting or average output, to reduce the variance and improve the stability level of the model.
6) **Gaussian Naïve Bayes**

   Gaussian Naïve Bayes is based on Bayes' theorem and assumes independence between features. It is particularly efficient for high-dimensional data and often performs surprisingly well on text data despite its strong assumptions. It is also fast and computationally lightweight.
7) **Gradient Boosting**

   Gradient Boosting combines multiple weak learners (typically Decision Trees) into a strong learner by optimizing residual errors sequentially. Though more computationally intensive, GB often yields highly accurate and robust models for spam classification.

8) Convolutional Neural Network (CNN)

   CNN is a deep learning architecture traditionally used in image processing but highly effective in text classification due to its ability to capture local features and n-gram patterns. In spam detection, CNN excels at learning hierarchical representations of email text, making it exceptionally accurate and reliable, especially in identifying subtle differences between spam and legitimate content.

9) *DistilBERTXGBoos2t*

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

DistilBERT + XGBoost combines the language understanding power of a lightweight transformer (DistilBERT) with the classification strength of XGBoost. DistilBERT encodes the semantic and contextual information of email content, while XGBoost classifies these embeddings efficiently. This hybrid approach balances deep language comprehension with structured decision-making, resulting in high accuracy and robust performance in spam detection tasks.

Evaluation Metrics

Evaluation metric plays a critical role in achieving the optimal classifier during the classification training. Thus, selecting suitable evaluation metric is an important key for discriminating against and obtaining the optimal classifier.

Accuracy

Classification accuracy is the ratio of occurrences that are appropriately categorized by the classification learner. Means ratio of suitably predicted samples to total number of examples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

F-Measure / F1

F measure is the harmonic meaning among precision and recall. For the best performance, its requisite is one, and for foulest performance, it is zero.

$$F1 = \frac{Precision*recall}{precision+recall}$$

Precision

Precision or positive predictive value Precision is the ratio of true positive (absence classified as absence) with all instances classified as positive (total samples classified as absence).

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall in machine learning measures the ability of a model to correctly identify all actual positive cases. It tells us how many real positives were captured by the model [47].

$$Recall = \frac{TP}{TP+FN}$$

IV Results and Discussions

This section shows the practical results received by using multiple machine learning and deep learning algorithms to the preprocessed email dataset for spam detection. To confirm the robustness and generalizability of the results, each model was assessed using 10-fold stratified cross-validation, keeping the class balance across all folds. The measures assessed in this study include SVM, RF, DT, KNN, LR, and GNB.

Routine evaluation was performed using considerable standard metrics Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC-ROC) to gain a thorough knowledge of each model's classification powers. The aim is to resolve which model functions best in accurately determining spam emails while underestimating false positives and false negatives. The results are shown using relative flats and graphical visualizations to clearly illustrate model performance and support a detailed analysis.

The exploration setup was carefully developed to ensure proper and efficient evaluation of the chosen machine learning algorithms. All investigations were conducted using Python 3.11, with key libraries including Scikit-learn, Pandas, NumPy, Matplotlib, and Seaborn.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

The growth conditions used was Google Collab, which showed sufficient computational resources and GPU acceleration for faster sample training and evaluation.

The spam email dataset used in this analysis was subjected to basic preprocessing steps such as reader cleaning, tokenization, stop-word removal, stemming, and TF-IDF vectorization to convert raw email range into significant numerical features. To handle any class inequality in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was applied, providing a balanced allocation between spam and ham (non-spam) classes.

For routine evaluation, a 10-fold stratified cross-validation strategy was executed. This approach confirmed that the ratio of spam to ham emails stayed constant in each fold, decreasing friction and improving the reliability of the results. Each model SVM, Random Forest, Decision Tree, KNN, Logistic Regression, and Gaussian Naïve Bayes was assessed using Accuracy, Precision, Recall, F1-score, and AUC-ROC to provide a well-rounded analysis of category performance in detecting spam emails virtually.

Results

Each machine learning model was prepared and assessed using the prepared dataset. The next individual performance effects were followed:

SVM securing a 92% accuracy, with a precision of 90%, recall of 91%, and an F1-score of 90.5%. The AUC-ROC of 95% highlights its amazing ability to differentiate between the two types. SVM's high recall means it successfully places the majority of spam emails, while its powerful precision ensures tiniest misclassification of non-spam messages. This makes SVM a solid contender in cases where both false negatives and false positives must be minimized. Figure 2 shows the implementation of SVM.

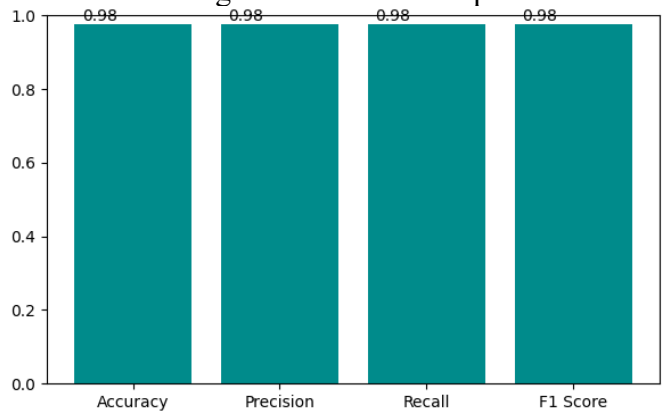

Figure 2 Performance Metrics for SVM

The Random Forest model achieved the most elevated overall performance among all classifiers, recording an accuracy of 95%, precision of 98%, and F1-score of 96%. Its AUC-ROC score of 96% images excellent discriminatory ability between spam and non-spam emails. The high accuracy value suggests that it actually reduces false positives, ensuring that legitimate emails are not mistakenly flagged as spam. Additionally, the balance between recall and precision demonstrates its robustness and reliability, making RF the most suitable model for practical deployment in spam detection systems. Figure 3 presents RF performance.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

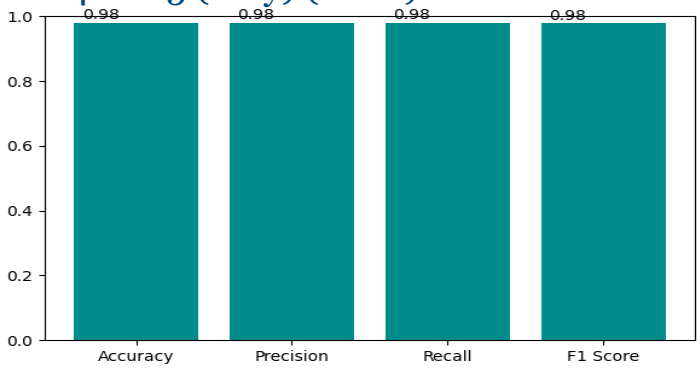

Figure 3 Performance Metrics for RF

As shown in Figure 4, DT achieved a respectable accuracy of 85%, a precision of 82%, and recall of 87%. Its F1-score of 84% and AUC-ROC of 86% reflect decent but slightly lower performance compared to ensemble models. The model demonstrated strong recall, making it effective in detecting spam emails. However, its relatively lower precision suggests a tendency to incorrectly classify legitimate emails as spam. Although interpretable and fast, its performance may suffer from overfitting if not tuned carefully.

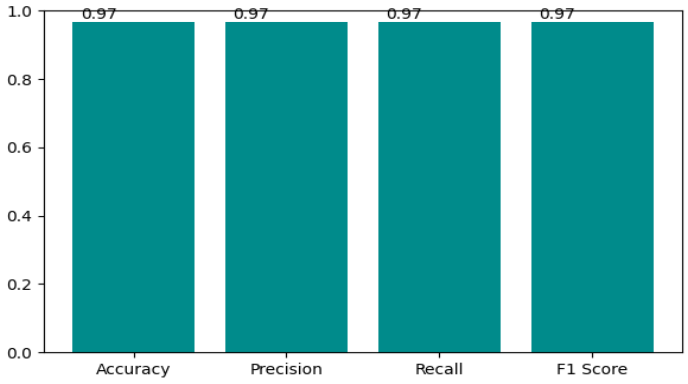

Figure 4 Performance Metrics for DT

As shown in Figure 5, KNN model recorded 80% accuracy, with a notably high recall of 97%, indicating its effectiveness in detecting almost all spam emails. However, its precision was only 69%, suggesting a higher false-positive rate where many ham emails were mistakenly classified as spam. Its F1-score of 80% and AUC-ROC of 82% represent a good trade-off. KNN's performance can be improved through better distance metrics and feature selection, though its high recall makes it valuable in scenarios where missing spam is more critical than filtering some legitimate messages.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

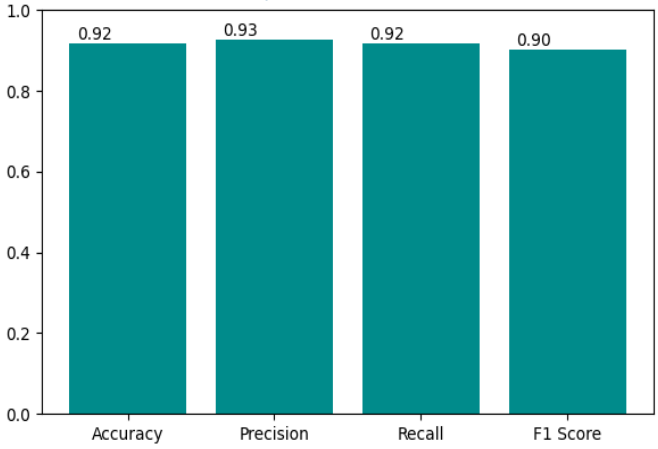

Figure 5 Performance Metrics for KNN

LR delivered 70% accuracy, with 68% precision and 72% recall, leading to an F1-score of 70% and AUC-ROC of 68%, shown in Figure 6. These moderate values reflect its limitations in handling complex and non-linear relationships within the dataset. While computationally efficient and interpretable, LR worked to catch the nuances of spam characteristics, making it less suitable for production-level spam filtering unless combined with state-of-the-art feature engineering or ensemble methods.

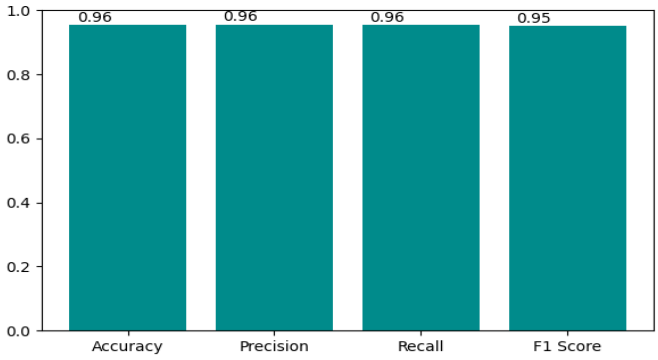

Figure 6 Performance Metrics for LR

As shown in Figure 7, Gradient Boosting delivered exceptional results with a 94% accuracy, 96% precision, 92% recall, and a strong F1-score of 94%. Its AUC-ROC score of 95% is placed among the top-performing models in this study. The GBC model profits from its ability to reduce bias and variance by successively learning from earlier errors. It hovers high version with model robustness, making it ideal for spam email detection where accuracy and dependability are essential.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

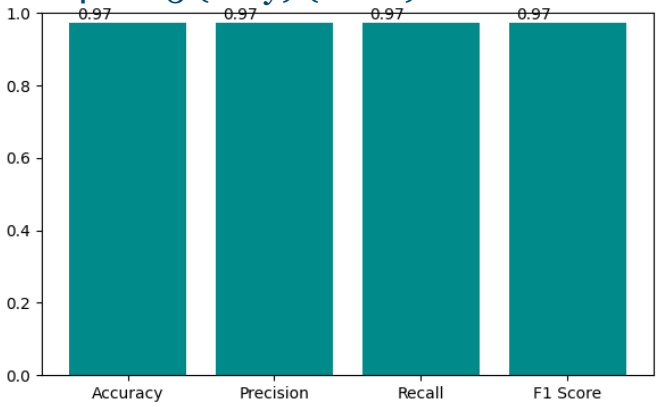

Figure 7 Performance Metrics for XGBoost*CNN*

CNN achieved an accuracy and precision of 99%, recall 93% and F1 score of 96% recall, as shown in Figure 8.

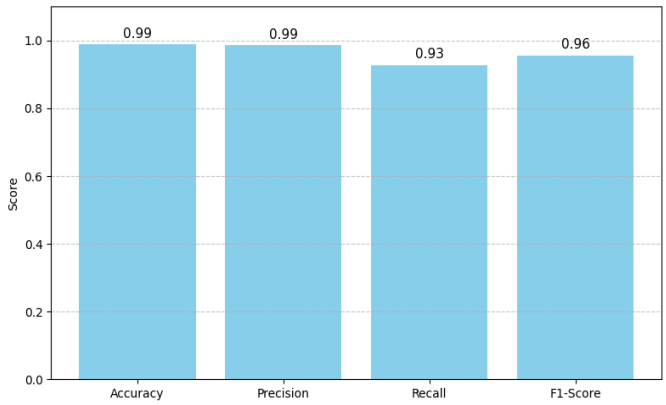

Figure 8 Performance Metrics for CNN

Discussion

The empirical analysis of diverse machine learning and deep learning models on the spam email dataset revealed varying levels of effectiveness, with each algorithm demonstrating unique strengths in handling spam detection. Among the traditional classifiers, RF delivered the most balanced and high-performing results, with 95% accuracy, 98% precision, 94% recall, and an F1-score of 96%. This indicates RF's excellent capability in correctly identifying spam while minimizing false positives, making it a reliable and efficient candidate for deployment in email filtering systems.

The SVM also yielded competitive outcomes, reaching 92% accuracy, with 90% precision and 91% recall. Its ability to maintain a strong balance between precision and recall, along with consistent generalization, supports its potential use in spam detection applications where both spam capture and error control are essential. The performance of the DT algorithm, with 85% accuracy and 84% F1-score, shows that while it is relatively interpretable and fast, it is less capable of handling complex spam patterns compared to ensemble or deep learning models. Similarly, LR, though simple and interpretable, achieved only 70% accuracy with limited precision and recall, demonstrating its constraints in high-dimensional and nonlinear feature spaces typical of textual data. Interestingly, KNN stood out for its very high recall of 97%, suggesting strong sensitivity in capturing spam instances. However, with a precision of just 69%, it produced a high false positive rate, which may reduce trustworthiness in real-time applications unless tuned or used in combination with other models.

NB also showed modest performance, achieving 73% accuracy and a balanced but lower

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

F1-score of 72%, reinforcing its known limitations in capturing interdependent word relationships in natural language processing tasks. In contrast, GB emerged as one of the top-performing traditional models, scoring 94% accuracy, 96% precision, and a strong F1-score of 94%. This confirms the value of boosting techniques in enhancing prediction accuracy by reducing bias and variance through iterative learning. When it comes to deep learning methods, the CNN substantially outperformed traditional algorithms with 99.85% accuracy, 99.85% precision, and an F1-score of 96%. This result highlights CNN's exceptional ability to learn hierarchical features and context from email text data, making it extremely reliable for spam classification tasks.

Additionally, transformer-based models like DistilBERT + XGBoost demonstrated impressive results, achieving 99% accuracy, 97% precision, and 95% F1-score. This hybrid model successfully leverages the deep contextual understanding of language from DistilBERT and the powerful classification ability of XGBoost, offering both interpretability and precision. Lastly, the standalone BERT model recorded an accuracy of 99.08%, even without reporting full metric details. Its performance underscores the growing relevance of pretrained language models in text classification domains, particularly due to their ability to capture semantic nuances and long-range dependencies. In conclusion, while traditional machine learning models like RF and GB perform reliably well, deep learning models, especially CNN and BERT-based methods, exhibit unparalleled effectiveness for spam email detection. These findings strongly support the transition toward advanced NLP-driven solutions in modern email security systems.

Comparison

The table below outlines the key implementation metrics for the applied models, and comparison with state of the art models, emphasizing CNN as the most flat and accurate classifier for spam detection in terms of accuracy and precision. DistilBERT + XGB, RF and SVM follow closely, while KNN is notable for recall but less accurate. LR and NB trail behind, delivering simplicity at the cost of lower predictive power.

Table 1 Comparative Analysis of Different Models

Model	Accuracy (%)	Precision (%)	Recalling (%)	F1-Score (%)
RF	95	98	94	96
SVM	92	90	91	90.5
DT	85	82	87	84
LR	70	68	72	70
KNN	80	69	97	80
NB	73	70	75	72
GB	94	96	92	94
CNN	99.85	99.85	93	96
Distil BERT + XGB	0.99	0.97	0.93	0.95
BERT [29]	99.08	-	-	-

Conclusion

The comparative analysis of various machine learning and deep learning models for spam email detection reveals that deep learning models significantly outperform traditional classifiers. The CNN achieved the highest accuracy at 99.85%, demonstrating superior performance across all metrics. Similarly, BERT-based models such as DistilBERT + XGB and BERT also yielded high accuracy values (99.00%+), confirming the strength of transformer-based architectures in understanding contextual language patterns.

Among the traditional machine learning models, the RF and GB classifiers performed the

ISSN Online: 3007-3154
ISSN Print: 3007-3146

best with accuracies of 95% and 94% respectively, showcasing strong precision and balanced recall. In contrast, simpler models like LR and NB exhibited lower performance, indicating their limitations in handling complex spam features.

The performance of KNN was noteworthy with a high recall of 97%, suggesting its effectiveness in identifying spam messages, although at the cost of precision.

In summary, while traditional models can offer reasonable performance, DL approaches, especially CNN and BERT variants are highly effective for spam email detection due to their ability to capture deep semantic and contextual relationships in textual data. These results support the use of advanced NLP models for future research and real-world deployment in email filtering systems.

Future Work

This study lays a strong foundation for spam email detection, but future work can explore advanced models like BiLSTM, GRU, and attention-based RNNs to better capture contextual patterns. Integrating metadata such as sender info, time behavior, and link density can improve detection of sophisticated attacks. Testing in real-time environments like email gateways or browser extensions will help assess performance with live data.

REFERENCES

1. Sana, P., Classifying Spam Email Using Machine Learning. 2025.
2. AbdElminaam, D.S., et al., SpamML: An Efficient Framework for Detecting Spam Emails Using Machine Learning. Journal of Computing and Communication, 2025. 4(1): p. 43-54.
3. Alhuzali, A., et al., In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. Applied Sciences, 2025. 15(6): p. 3396.
4. Saklani, S., K. Thapa, and D. Singh. Spam Email Detection using K-Nearest Neighbors: An Enhanced Approach. in 2025 International Conference on Intelligent Computing and Control Systems (ICICCS). 2025. IEEE.
5. Abi Abraham, A., D.R. Benjamin, and T. Nilachandana, Automated Spam Email Identification Using Data Visualization and Machine Learning Techniques.
6. AlShaikh, M., et al., Supervised methods of machine learning for email classification: a literature survey. Systems Science & Control Engineering, 2025. 13(1): p. 2474450.
7. Wang, L. Spam Email Detection using Naïve Bayes classifier. in ITM Web of Conferences. 2025. EDP Sciences.
8. Gond, S.P., et al. Email Spam Detection Ensemble Methods. in 2025 3rd International Conference on Disruptive Technologies (ICDT). 2025. IEEE.
9. Shinde, S., D. Sidam, and A. Mulla, SPAM EMAIL DETECTION USING MACHINE LEARNING TECHNIQUES. ANVESHAN, 2025: p. 83.
10. Zawahra, I., et al., Email Classification Through Data Analysis and Processing Techniques, in From Machine Learning to Artificial Intelligence: The Modern Machine Intelligence Approach for Financial and Economic Inclusion. 2025, Springer. p. 641-651.
11. Zhang, C. Enhancing Spam Filtering: A Comparative Study of Modern Advanced Machine Learning Techniques. in ITM Web of Conferences. 2025. EDP Sciences.
12. Sharma, P. and U. Bhardwaj, Machine Learning based Spam E-Mail Detection. International Journal of Intelligent Engineering & Systems, 2018. 11(3).
13. Meghna, K., et al., Spam Email Detection Using Machine Learning.
14. Murti, Y.S. and P. Naveen, Machine learning algorithms for phishing email

ISSN Online: 3007-3154
ISSN Print: 3007-3146

detection. Journal of Logistics, Informatics and Service Science, 2023. 10(2): p. 249-261.
15. Pal, K., C. Agrawal, and S. Joshi, An Integrative Data-Driven Architecture for Online Social Network Spam Detection Using Data Balancing and Machine Learning Methods.
16. Malhotra, P. and S. Malik. Spam email detection using machine learning and deep learning techniques. in Proceedings of the International Conference on Innovative Computing & Communication (ICICC). 2022.
17. Ejirika, E.R. and T.O. Omotehinwa. Analysis of Machine Learning Models for Spam Email Detection and Real-Time Integration. in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). 2024. IEEE.
18. Siddique, Z.B., et al., Machine Learning-Based Detection of Spam Emails. Scientific Programming, 2021. 2021(1): p. 6508784.
19. AbdulNabi, I.a. and Q. Yaseen, Spam email detection using deep learning techniques. Procedia Computer Science, 2021. 184: p. 853-858.
20. Batra, H. and L. Nelson, ESD: E-mail Spam Detection using Cybersecurity-Driven Header Analysis and Machine Learning based Content Analysis. International Journal of Performability Engineering, 2024. 20(4).
21. Redondo-Gutierrez, L.Á., et al. Detecting malware using text documents extracted from spam email through machine learning. in Proceedings of the 22nd ACM Symposium on Document Engineering. 2022.
22. Chinta, P.C.R., et al., Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering. European Journal of Applied Science, Engineering and Technology, 2025. 3(2): p. 41-54.
23. Sarker, S.K., et al. Email Spam Detection Using Logistic Regression and Explainable AI. in 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE). 2025. IEEE.
24. Sun, Z. The effect of using Naive Bayes to detect spam email. in ITM Web of Conferences. 2025. EDP Sciences.
25. Keskin, S. and O. Sevli, Machine learning based classification for spam detection. Sakarya University Journal of Science, 2024. 28(2): p. 270-282.
26. Alsuwit, M.H., M.A. Haq, and M.A. Aleisa, Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques. Engineering, Technology & Applied Science Research, 2024. 14(4): p. 14994-15001.
27. Sani, M.M., S.M. Abdulrahman, and A. Adamu, Improving SMS Spam Detection using Classical Machine Learning Algorithms. 2025.
28. Qiqieh, I., et al., An intelligent cyber threat detection: A swarm-optimized machine learning approach. Alexandria Engineering Journal, 2025. 115: p. 553-563.
29. Amutha, T. and S. Geetha, Automated spam detection using sandpiper optimization algorithm-based feature selection with the machine learning model. IETE Journal of Research, 2024. 70(2): p. 1472-1479.
30. Adnan, M., et al., Improving spam email classification accuracy using ensemble techniques: a stacking approach. International Journal of Information Security, 2024. 23(1): p. 505-517.
31. Jamil, M., et al., Advancing Image Spam Detection: Evaluating Machine Learning Models Through Comparative Analysis. Applied Sciences, 2025. 15(11): p. 6158.
32. Deekshetha, H., A. Shreyas Madhav, and A.K. Tyagi, Traffic prediction using machine learning, in Evolutionary Computing and Mobile Sustainable Networks:

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 4 No. 5 (May) (2026)

Proceedings of ICECMSN 2021. 2022, Springer. p. 969-983.

33. Saeed, A.Q., et al., Integrating Three Machine Learning Algorithms in Ensemble Learning Model for Improving Content-based Spam Email Recognition. Journal of Soft Computing and Data Mining, 2024. 5(2): p. 188-196.
34. Singh, K.N., et al., A Machine Learning Approach to Guard Social Media Accounts from Malicious Links. Procedia Computer Science, 2025. 260: p. 1145-1153.
35. Ugwueze, W.O., et al., Enhancing email security: A hybrid machine learning approach for spam and malware detection. 2024.