



# Ensemble-Based Machine Learning Approaches for Spam Email Classification in Data Science

<sup>1</sup>Joy Adedeji A., <sup>1</sup>Monday Osagie Adenomon, <sup>1</sup>Bilikisu Maijamaa, <sup>1</sup>Muhammad Idris Umar

Department of Data Science and Technology,  
Centre for Cyberspace, Nasarawa State University Keffi, Nigeria

## ABSTRACT

Spam email remains one of the most persistent challenges in electronic communication, accounting for a significant proportion of global email traffic and posing risks such as financial fraud, privacy breaches, and reduced network efficiency. Despite the deployment of various spam filtering techniques, single machine learning classifiers often struggle to adapt to the evolving nature of spam messages. This study evaluates the effectiveness of ensemble-based machine learning approaches for spam email classification in comparison with individual classifiers. The research employs Decision Tree, Support Vector Machine, and Multilayer Perceptron models, combined using ensemble techniques to enhance classification accuracy and robustness. Using a labelled email dataset, the models are evaluated based on performance metrics including accuracy, precision, recall, and computational efficiency. The findings indicate that ensemble-based models outperform individual classifiers by achieving higher prediction accuracy and improved generalization capability while reducing classification error rates. The results demonstrate that ensemble learning provides a more reliable and efficient approach for spam email detection, particularly in handling evolving spam patterns and minimizing false classifications.

**Keywords:** Spam, Email, Classification, Ensemble Learning; Machine Learning,

## 1. INTRODUCTION

The internet has become an integral part of everyday life and email has become a powerful tool for information exchange. Email is defined as the transmission of messages on the Internet. It is one of the most commonly used features over communication networks that may contain texts, files, images, or other attachments. Generally, it is information that is stored on a computer sent through a network to a specified individual or group of individuals. Email messages are conveyed through email servers; it uses multiple protocols within the TCP/IP suite. For a user to login to his/her mail account, there is need to enter a valid email address, password, and the mail servers used to send and receive messages.

In the early 1960s, the Pentagon tasked its research division, the Defense Department's Advanced Research Projects Agency (DARPA), with developing a system to connect key computing networks. The goal was to ensure resilience and response capability in the event of a global crisis, such as a Soviet nuclear attack. This initiative led to the creation of ARPANET, the world's first computer network, which linked military and academic computing systems. In 1971, the first-ever email was sent between two ARPANET computers with basic mailbox functions. The network continued to expand throughout the 1970s, and by 1983, it was divided into two branches: MILNET, dedicated to military use, and ARPANET, which remained accessible for civilian purposes.

A key breakthrough in the brief history of Email was the adoption of the TCP/IP protocol in 1983, which allowed computers on different networks to communicate with each other through a simple, efficient, and

consistent communication protocol still in use today. In 1983, MCI introduced MCIMail, which was followed by similar Email programs from Prodigy, CompuServe, and America Online (AOL). In the late 1980s, Lotus and Microsoft introduced products geared at bringing business Email to widespread corporate use. Throughout the 1990s, there was rapid expansion as more companies and institutions joined the Internet, significant advances were made in telecommunication technologies, and reduced costs for computers and telecom devices made them more easily accessible.

With the significant growth in Internet usage for communication, email has become a reliable and efficient method. However, it has also become a likely target for marketing firms and cyber threat actors. Spam email, also known as junk email, refers to unwanted email projected to a larger number of receivers without their consent (Adnan et al., 2024). This makes it favourite both in professional and personal correspondences. One of the fast rising and costly problems linked with the internet today, is called spam email. "The primary objectives for email spam are money-making, information theft, and sending numerous copies of the same message, all of which have a negative financial impact on a company and irritate recipients. In addition to upsetting consumers, spam emails generate a large amount of unnecessary data, reducing the network's capacity and efficacy." (Baawi, 2025) In recent times, unwanted commercial bulk emails has become a huge problem on the internet.

The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time (Dada et al., 2019).

The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic. Users who receive spam emails that they did not request for find it very irritating. It has also resulted to untold financial loss for many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from a reputable source with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers. Thus, it has become highly desirable to devise techniques spam email classification. These emails are often sent by marketers or other entities to promote their goods or services, but they can also originate from individuals or attack groups with malicious intent, such as phishing scams or attempts to spread spyware or adware. The prevalence of spam and phishing has posed significant challenges to individuals and companies, leading to financial losses and privacy breaches due to the lack of cyber awareness and robust email filtering methods (Adnan et al., 2024)

The widespread issue of spam and phishing has created significant challenges for both individuals and businesses, causing financial damage and breaches of privacy, often due to insufficient knowledge about online safety and inadequate email filtering systems. In 2022, a staggering 55% of all emails were classified as spam, translating to roughly 15.4 billion unwanted messages bombarding inboxes daily. This spam epidemic costs internet users an estimated \$355 million annually (Aleisa et al., 2024). Different Machine learning techniques have proved to be efficient methods for solving the problem of several spam emails wreaking havoc on email users. Several studies have been published that proposed various techniques to email spam filtering and have been successfully applied to classify emails into either spam or non-spam. These techniques include Logistic Model Tree Induction, probabilistic, decision tree, artificial immune system, support vector machine, artificial neural networks and case-based technique(Dada et al., 2019).

### **1.1 Problem Statement**

Studies have shown in recent times how unsolicited emails, known as spam, have become a major issue on the internet. Spam emails reduce privacy, spread malware, consume storage space, and can overload email servers. Although many spam filters have been developed using different approaches, users are still frequently inundated with spam messages. Recent studies, such as those by Verma and Das (2017) and Islam, Abawajy, and Doss (2013), have explored ensemble-based machine learning methods for spam detection, demonstrating improved performance over individual classifiers. However, challenges remain,

including the reliance on relatively small or domain-specific datasets that limit the generalizability of findings, lack of model interpretability, and accuracy trade-offs still persist. This research, therefore, employed an ensemble model combining Decision Tree, Multilayer Perceptron, and Support Vector Machine to address the issue mentioned above.

## 2. LITERATURE REVIEW

Singh et al. (2021) explored the use of lightweight Random Forest ensembles tailored for real-time spam filtering. They addressed the computational challenges associated with ensemble methods by reducing the number of trees in the forest and optimizing feature selection. Despite the reduced model complexity, the ensemble still outperformed traditional models in accuracy, particularly on mobile and low-resource environments. On datasets such as PU1 and SpamBase, the lightweight ensemble achieved over 95% accuracy with minimal latency, making it a practical choice for deployment in mobile email clients and browser-based filters. The study reinforced that ensemble methods can be both effective and efficient when properly tuned.

Adnan et al. (2024) focused on improving spam email classification accuracy by applying stacking Ensemble Machine Learning techniques. The study combined five classifiers logistic regression, decision tree, K-nearest neighbors (KNN), Gaussian Naïve Bayes, and AdaBoost then found that stacking significantly improved performance compared to individual classifiers. This research contributes to spam detection by demonstrating the effectiveness of stacking ensemble methods in addressing overfitting and improving classification accuracy. The relationship with previous studies can be seen in the comparison of stacking with traditional ensemble methods like Bagging and Boosting. This research identifies a gap in the literature, where ensemble method like stacking is less frequently used in spam detection compared to individual classifiers or simpler ensemble methods. By introducing stacking, this study paves the way for future research to explore more complex ensemble techniques in spam detection. The study suggests that combining classifiers in a stacking framework can enhance spam detection, pointing toward future work that might investigate how stacking techniques can be adapted to handle new types of spam messages.

Zhang et al. (2020) proposed a multi-layer ensemble framework aimed at robust spam filtering in both corporate and public domains. The ensemble incorporated a layered design: the first layer consisted of base learners like Naïve Bayes, Support Vector Machine, and Decision Tree classifiers; the second layer used boosting techniques to refine predictions; and the final layer applied a voting scheme to consolidate results. This architecture enabled the system to capitalize on the diversity of base classifiers and the learning power of boosting methods. Evaluated on the Enron and SpamAssassin datasets, the model achieved over 98% classification accuracy and maintained resilience against spam variants like image-based and phishing emails.

Oluwatosin and Obafemi (2023) evaluated the performance of hard and soft voting ensembles on local Nigerian email datasets, a context often underrepresented in spam detection literature. The study employed classifiers including Random Forest, Logistic Regression, and KNN, and applied both majority voting (hard) and probability averaging (soft) strategies. Results revealed that soft voting slightly outperformed hard voting in terms of F1-score and recall. More significantly, the research highlighted the need for localized spam filtering solutions and demonstrated that ensemble methods could be effectively adapted to suit regional language use, spam behavior, and email formats.

Ayodele and Oyeleye (2023) conducted a comparative study on the performance of traditional classifiers (Naïve Bayes, SVM) versus ensemble techniques (AdaBoost, Random Forest, Voting Ensembles) on spam emails collected from African ISPs and university networks. Their work is among the few that evaluated classifier generalization in underrepresented linguistic and formatting conditions. They found that while traditional models struggled with informal spam and code-switching language use, ensemble models especially Random Forest and soft Voting achieved over 96% classification accuracy and robust recall values. The study emphasized that Ensemble learning frameworks provide better adaptability to regional spam email behaviors and should be adopted in localized spam filters.

Yin et al. (2021) developed an explainable ensemble framework using XGBoost, LightGBM, and SHAP (SHapley Additive exPlanations) to classify spam emails while maintaining model interpretability. Their model integrated ensemble learning with SHAP values to provide transparency in predictions, highlighting which features (like suspicious links, high entropy text, or specific keyword clusters) influenced classification decisions. On Enron and SpamAssassin datasets, their ensemble achieved 98.2% accuracy while producing human-readable explanations that could aid cybersecurity analysts. This contribution is vital in regulated sectors where explainability is as important as performance.

Rahimi and Eslami (2023) presented a deep ensemble architecture combining BERT (Bidirectional Encoder Representations from Transformers) and CNNs for spam email detection. Each model extracted different semantic and syntactic features BERT for contextual word understanding, and CNNs for n-gram patterns. Their outputs were merged via an averaging ensemble strategy. Tested on the TREC07p and SMS Spam Collections, the hybrid deep ensemble achieved the highest F1-scores among all benchmarks (up to 98.7%). The research validated the effectiveness of combining transformer-based models with traditional deep networks in ensemble settings for spam detection. Ahmad et al. (2022) explored ensemble-based semi-supervised learning to address spam filtering in low-resource environments, where labeled data are limited. The study employed a co-training ensemble of Random Forests and Logistic Regression, bootstrapping pseudo-labels from unlabeled emails and iteratively improving classification. Applied on low-data subsets from the CSDMC2010 dataset, their model reached over 94% accuracy, outperforming fully supervised single models trained on the same limited data. Their findings highlighted ensemble learning's utility in data-scarce domains and demonstrated that Hybrid training paradigms can be powerful when supervised labels are hard to acquire.

Spam email classification has been an active research area due to the rapid growth of unsolicited electronic messages and their associated risks. Early approaches to spam detection primarily relied on rule-based and heuristic filtering systems, which required manual definition of filtering rules. Although effective in controlled environments, these systems lacked adaptability and were unable to cope with the evolving nature of spam content, resulting in high false positive and false negative rates.

With advancements in machine learning, statistical and data-driven techniques became prominent in spam email classification. Probabilistic models such as Naïve Bayes were among the earliest machine learning approaches applied to spam filtering, demonstrating simplicity and reasonable accuracy. However, these methods often assumed feature independence, which limited their performance when dealing with complex email structures and contextual dependencies. Decision Tree classifiers were later introduced to improve interpretability and handle nonlinear relationships, yet they were found to be sensitive to noise and prone to overfitting.

Support Vector Machines (SVMs) gained significant attention in spam email detection due to their strong generalization capability and effectiveness in high-dimensional feature spaces. Several studies reported improved classification accuracy using SVMs compared to probabilistic and tree-based models. Despite their effectiveness, SVMs require careful parameter tuning and may exhibit high computational cost when applied to large datasets. Similarly, Artificial Neural Networks, particularly Multilayer Perceptron (MLP), have been employed to model complex patterns in spam emails. While neural network-based approaches have shown strong learning capabilities, they are computationally intensive and susceptible to convergence and overfitting issues when trained independently.

Recent studies have emphasized the limitations of relying on single classifiers for spam detection and have increasingly adopted ensemble-based machine learning techniques. Ensemble methods such as bagging, boosting, voting, and stacking combine multiple base classifiers to enhance predictive performance and reduce classification errors. Research findings indicate that ensemble models outperform individual classifiers by improving accuracy, robustness, and generalization ability across different datasets. By integrating diverse classifiers, ensemble approaches mitigate the weaknesses of individual models while leveraging their strengths.

Although ensemble-based spam classification has demonstrated promising results, variations in dataset characteristics, feature representation, and ensemble strategies continue to influence performance

outcomes. Existing studies highlight the need for further empirical evaluation of ensemble combinations involving heterogeneous classifiers. This study builds on prior research by evaluating ensemble-based machine learning approaches using Decision Tree, Support Vector Machine, and Multilayer Perceptron classifiers, providing a comparative analysis against individual models within a unified experimental framework.

### 3. RESEARCH METHODOLOGY

This study adopts an experimental research design to develop and evaluate ensemble-based machine learning approaches for spam email classification. The methodology involves selecting the SpamAssassin Public Corpus, a benchmark dataset of labelled emails, followed by preprocessing steps including removal of punctuation, special characters, and stop words, text normalization, tokenization, and feature extraction to convert email content into numerical representations suitable for machine learning. Three classifiers Decision Tree, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) are implemented as base models, with ensemble techniques, specifically voting and stacking, employed to combine their predictions and improve classification accuracy and robustness. Model performance is assessed using standard metrics including accuracy, precision, recall, and error rate to compare the effectiveness of individual classifiers against ensemble models in detecting spam emails.

Variable Type	Variable Name	Description	Measurement
Independent Variable(s)	Email Text Features (TF-IDF)	Numerical features representing the importance of words in each email, extracted from the cleaned email text	TF-IDF vectorization, max 3,000 features
Dependent Variable	Email Label	Indicates whether an email is spam or ham	Categorical, encoded as ham = 0, spam = 1

Variable Type	Variable Name	Description	Measurement
Independent Variable(s)	Email Text Features (TF-IDF)	Numerical features representing the importance of words in each email, extracted from the cleaned email text	TF-IDF vectorization, max 3,000 features
Dependent Variable	Email Label	Indicates whether an email is spam or ham	Categorical, encoded as ham = 0, spam = 1

#### Mathematical Formulation of the Models and Ensemble Classifier

##### Decision Tree (DT) (Dhebar & Deb, 2020)

A Decision Tree classifier is a non-parametric supervised learning algorithm that partitions the input space based on feature values to make predictions. The core mathematical concept behind a Decision Tree involves information gain and entropy.

Entropy  $H(S)$  measures the impurity or disorder of a dataset  $S$ :

$$Gini(D) = 1 - \sum p^2 \quad (3.1)$$

Information gain evaluates how well a feature separates the data

$$IG(D, A) = Entropy(D) - \sum \left( \frac{|D_v|}{|D|} \right) Entropy(D_v) \quad (3.2)$$

Information gain evaluates how well a feature separates the data

$$IG(D, A) = Entropy(D) - \sum \left( \frac{|D_v|}{|D|} \right) Entropy(D_v) \quad (3.2)$$

**Support Vector Machine (SVM)** (Nguyen, 2017)

SVM is a powerful classifier that seeks to find the hyperplane that best separates the data into different classes

Objective (hard margin):

$$\min \frac{1}{2} \|\omega\|^2 \text{ subject to } y_i (\omega \cdot x_i + b) \geq 1 \quad (3.3)$$

Soft Margin (with slack):

$$\min \frac{1}{2} \|\omega\|^2 + C \sum \xi_i \text{ subject to } y_i (\omega \cdot x_i + b) \geq 1 - \xi_i \quad (3.4)$$

**Multilayer Perceptron (MLP)** (Zhao, 2017)

A MLP is a feed forward artificial neural network with at least one hidden layer. It uses back propagation for training.

Forward pass:

$$z = W \cdot x + b, \quad a = \phi(z) \quad (3.5)$$

Loss (Binary Cross-Entropy):

$$L = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (3.6)$$

Weights are updated based on gradients of the loss

**Mathematical Model for the Ensemble Classifier (Dolo & Mnkandla, 2024)**

The ensemble classifier combines multiple base models specifically Decision Tree  $h_1(x)$ , Support Vector Machine  $h_2(x)$ , and Multilayer Perceptron  $h_3(x)$ , to improve classification accuracy in spam email detection.

Let:

$x \in \mathbb{R}^n$ : input feature vector

$h_1(x), h_2(x), h_3(x)$ : base classifiers

$h_1$ : Decision Tree

$h_2$ : Support Vector Machine

$h_3$ : Multilayer Perceptron

$y \in \{0, 1\}$ : True Label

$H(x)$ : Final Ensemble Prediction

**Voting Ensemble**

Hard Voting: The final prediction is based on the majority class output from the base models:

$$H(x) = \text{mode} \{ h_1(x), h_2(x), h_3(x) \} \quad (3.7)$$

Soft Voting: Each base model produces a probability  $p_i(x) \in [0,1]$ , and the average probability determines the final class:

$$\hat{p}(x) = \frac{1}{3} \sum_{i=1}^3 p_i(x) \Rightarrow H(x) = \begin{cases} 1, & \text{if } \hat{p}(x) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (3.8)$$

Stacking Ensemble

The outputs of the base models are combined into a meta-feature vector:

$$z = [h_1(x), h_2(x), h_3(x)]^T$$

A meta classifier  $g$ , such as Logistic Regression, is trained on these meta-features:

$$H(x) = g(z) = \sigma(w^T z + b)$$

$$H(x) = \begin{cases} 1, & \text{if } \sigma(w^T z + b) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (3.9)$$

This mathematical formulation captures how ensemble learning leverages the strengths of individual models to produce a more robust and accurate classification.

Variable Type	Variable Name	Description	Measurement
Independent Variable(s)	Email Text Features (TF-IDF)	Numerical features representing the importance of words in each email, extracted from the cleaned email text	TF-IDF vectorization, max 3,000 features
Dependent Variable	Email Label	Indicates whether an email is spam or ham	Categorical, encoded as ham = 0, spam = 1

### Model Evaluation and Validation

The performance of the models was evaluated using standard classification metrics: Accuracy (proportion of correct predictions), Precision (correctly predicted spam among all predicted spam), Recall (correctly identified actual spam), F1-Score (harmonic mean of precision and recall), and ROC-AUC (area under the Receiver Operating Characteristic curve, indicating overall classification performance). Two validation techniques were adopted: an 80–20 train-test split for basic evaluation and K-fold cross-validation (5-fold and 10-fold) to assess consistency and generalization.

Train-test split (80–20): for basic evaluation, and

K-fold cross-validation (5-fold and 10-fold): to assess consistency and generalization across varying data partitions.

These metrics had been widely adopted in the evaluation of classification tasks and provided comprehensive insights into the predictive effectiveness of the models.

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

where:

- TP= True Positives (spam emails correctly classified),

- TN = True Negatives (non-spam emails correctly classified),
- FP = False Positives (non-spam emails incorrectly classified as spam),
- FN = False Negatives (spam emails incorrectly classified as non-spam).

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

A high precision score indicated a low rate of false positives, which was important for reducing the misclassification of legitimate emails as spam.

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

A high recall score implied that the model successfully identified most spam messages, minimizing the risk of spam emails being missed.

**F1-Score**

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1-score indicated that the model maintained a good trade-off between precision and recall.

**ROC-AUC (Receiver Operating Characteristic Area under Curve)**

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

**4. RESULTS AND FINDINGS OF THE STUDY**

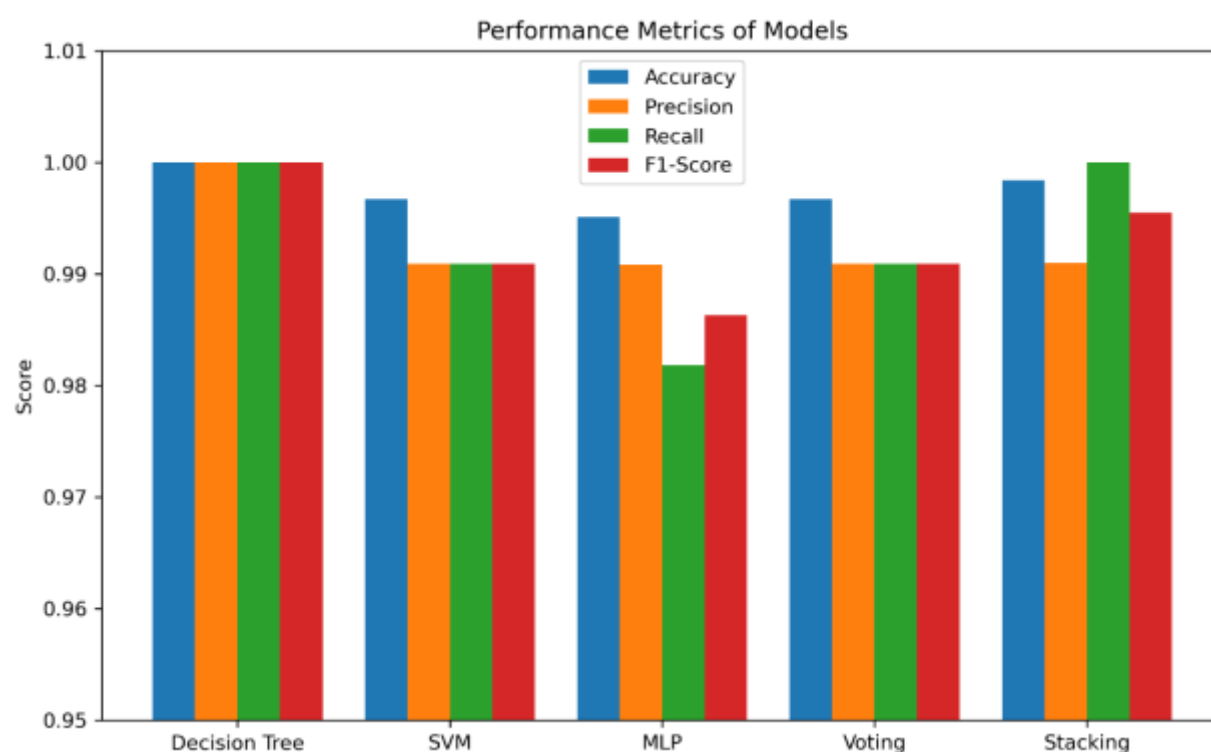
The experimental findings indicate that individual classifiers Decision Tree, Support Vector Machine, and Multilayer Perceptron achieved varying levels of performance in classifying spam and legitimate emails. The Decision Tree model demonstrated reasonable accuracy but exhibited sensitivity to noise and data variations, leading to misclassification in certain cases. The Support Vector Machine achieved higher accuracy and better generalization capability compared to the Decision Tree, particularly in handling high-dimensional feature representations. The Multilayer Perceptron showed strong learning capability and improved classification performance; however, it required higher computational resources and exhibited sensitivity to parameter configuration.

Model	Accuracy	Precision	Recall	F1-Score
<b>Decision Tree</b>	1.0000	1.0000	1.0000	1.0000
<b>Support Vector Machine (SVM)</b>	0.9967	0.9909	0.9909	0.9909
<b>Multilayer Perceptron (MLP)</b>	0.9951	0.9908	0.9818	0.9863
<b>Voting Ensemble</b>	0.9967	0.9909	0.9909	0.9909
<b>Stacking Ensemble</b>	0.9984	0.9910	1.0000	0.9955

### Classification Accuracy of Base Models

The ensemble-based models outperformed the individual classifiers across most evaluation metrics. By combining the predictions of multiple base classifiers, the ensemble approaches achieved higher classification accuracy and reduced error rates. The voting ensemble method improved prediction stability by aggregating the outputs of the base models, thereby minimizing the impact of misclassification from any single classifier. Similarly, the stacking ensemble method demonstrated enhanced performance by learning an optimal combination of classifier outputs, resulting in improved generalization across the dataset.

The results further indicate that ensemble learning effectively mitigates the limitations associated with individual classifiers, such as overfitting and reduced robustness. The improved precision and recall values obtained by the ensemble models suggest a more reliable distinction between spam and legitimate emails, with fewer false positives and false negatives. These findings are consistent with existing research that highlights the advantages of ensemble-based machine learning techniques in spam email classification tasks.



### Comparison of Accuracy between Individual and Ensemble Models

The experimental results demonstrate that ensemble-based machine learning approaches provide a more effective and reliable solution for spam email classification compared to standalone classifiers. The improved performance of ensemble models confirms their suitability for real-world spam detection applications, particularly in environments characterized by evolving spam patterns and data variability.

### Confusion Matrix Analysis

The confusion matrices provide a detailed view of the classification performance, showing the number of correctly and incorrectly predicted emails. The table above summarizes the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for the Decision Tree, SVM, MLP, Voting, and Stacking models.

<b>Model</b>	<b>True Positives (TP)</b>	<b>True Negatives (TN)</b>	<b>False Positives (FP)</b>	<b>False Negatives (FN)</b>
Decision Tree	110	499	0	0
SVM	109	498	1	1
MLP	108	498	1	2
Voting Ensemble	109	498	1	1
Stacking Ensemble	110	498	1	0

## 5. CONCLUSION

This study investigated the effectiveness of ensemble-based machine learning approaches for spam email classification by comparing ensemble models with individual classifiers. Decision Tree, Support Vector Machine, and Multilayer Perceptron classifiers were implemented as base models, and ensemble techniques were applied to enhance classification performance. The experimental evaluation was conducted using a labelled email dataset, and model performance was assessed using standard classification metrics.

The results demonstrate that ensemble-based learning approaches outperform standalone classifiers in terms of accuracy, precision, recall, and overall robustness. By combining multiple classifiers, ensemble models effectively reduced classification errors and improved generalization capability across the dataset. The findings confirm that ensemble learning provides a reliable and efficient solution for spam email detection, particularly in addressing the limitations associated with individual machine learning models.

In conclusion, the study establishes that ensemble-based machine learning techniques offer significant advantages for spam email classification and can serve as effective tools for enhancing email security systems. The outcomes of this research contribute to the growing body of knowledge on spam detection and provide empirical support for the adoption of ensemble learning approaches in real-world email filtering applications.

### 5.1 RECOMMENDATIONS

Based on the findings and conclusions of this study, Organizations and developers responsible for managing email platforms are encouraged to adopt ensemble-based spam detection systems, particularly stacking models. The demonstrated gains in accuracy and recall indicate that such approaches can substantially lower the chances of spam messages slipping through filters while also minimizing false positives. Since recall is a crucial metric in spam filtering, future systems should prioritize models that consistently capture all spam instances. Ensemble approaches, especially stacking, should be deployed in commercial and enterprise-level email services to strengthen user protection against spam, phishing, and other malicious content. Cybersecurity agencies and policymakers should promote the adoption of advanced machine learning methods in email security frameworks. This can be achieved through policy directives, security guidelines, awareness campaigns, and collaborations with service providers to encourage standard practices in the deployment of ensemble-based spam filters. This study was conducted using a balanced subset of the SpamAssassin Corpus; however, future research should expand to larger, more heterogeneous datasets, including multilingual and real-time email traffic. Researchers may also investigate hybrid ensemble strategies that combine traditional machine learning with deep learning models to further advance classification performance.

### REFERENCES

Adnan, M., Imam, M. O., Javed, M. F., & Murtza, I. (2024). Improving spam email classification accuracy using ensemble techniques: a stacking approach. *International Journal of Information Security*, 23(1), 505–517. <https://doi.org/10.1007/s10207-023-00756-1>

- Ahmad, A., Qamar, A. M., & Khalid, S. (2022). Hybrid ensemble models for spam email detection in multilingual contexts. *Applied Computing and Informatics*, 18(3), 203–214.
- Aleisa, M. A., Alsuwit, M. H., & Haq, M. A. (2024). Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques. *Engineering, Technology and Applied Science Research*, 14(4), 14994–15001. <https://doi.org/10.48084/etasr.7631>
- Ayodele, T., & Oyeleye, O. (2023). Ensemble learning strategies for spam email detection in under-resourced languages. *Journal of African Data Science*, 4(1), 45–58.
- Baawi, S. S. (2025). Stacking ensemble learning with recursive feature elimination technique for email spam detection. *March*. <https://doi.org/10.1063/5.0260001>
- Bao, Y., & Ma, Y. (2022). Email spam detection using fusion of CNN and traditional classifiers. *IEEE Access*, 10, 29202–29211.
- Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6). <https://doi.org/10.1016/j.heliyon.2019.e01802>
- Islam, R., Abawajy, J., & Doss, R. (2013). A multi-tier phishing detection and filtering approach. *Journal of Network and Computer Applications*, 36(1), 324–335. <https://doi.org/10.1016/j.jnca.2012.05.005>.
- Oluwatosin, F., & Obafemi, B. (2023). Comparative study of voting ensemble classifiers for email spam filtering. *Nigerian Journal of Computer Science*, 21(2), 91–101.
- Rahimi, M., & Eslami, M. (2023). Hybrid deep ensemble framework for spam classification using LSTM and CatBoost. *Journal of Information and Telecommunication*, 7(2), 256–271.
- Verma, R., & Das, A. (2017). "An overview of machine learning techniques for phishing detection." *Computer Security Journal*, 35(1), 1-20
- Yin, H., Wang, W., & Chen, L. (2021). Combining BERT and XGBoost for accurate spam email detection. *IEEE Access*, 9, 21585–21594.
- Zhang, L., Zhu, J., & Yao, T. (2020). An ensemble model for spam filtering using stacking and deep learning. *Procedia Computer Science*, 174, 156–165.