Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 6, No. 3, pp. 192-196, 2025

Comparative Analysis of Machine Learning Models for Email Spam Detection

Mugi Lestari[1*], Yasir Salih[2], Alim Jaizul[3]

[1]*Communication in Research and Publications, Bandung, Indonesia*
[2]*Department of Mathematics, Faculty of Education, Red Sea University, SUDAN*
[3]*Research Collaboration Community, Bandung, Indonesia*

**Corresponding author email:* mu2lestari@gmail.com

Abstract

The development of information technology has driven a significant increase in the use of email as a primary communication tool across various sectors. Spam emails have become a serious issue that can disrupt productivity, threaten data security, and compromise user privacy. Conventional rule-based spam filtering systems are no longer considered effective in countering increasingly sophisticated and adaptive spam attack patterns. A more dynamic and accurate approach is required through the utilization of Machine Learning. This study aims to analyze and compare the performance of several Machine Learning algorithms in detecting spam emails, namely Extra Trees Classifier, Random Forest, Support Vector Machine (SVM) with an RBF kernel, and CatBoost. The methodology involves data acquisition from the SMS Spam Collection Dataset, data preprocessing through text cleaning and feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method, followed by model training and evaluation using Accuracy, F1 Score, and ROC AUC metrics. The results show that the Extra Trees Classifier achieved the best performance, with an Accuracy of 97.29%, an F1 Score of 0.8814, and a ROC AUC of 0.9868. Tree-based ensemble models, particularly Extra Trees and Random Forest, demonstrated superior capability in maintaining a balance between precision and recall. The SVM (RBF) recorded the highest AUC value but presented a trade-off in the form of a higher number of False Negatives. The findings of this research serve as a reference for the development of more adaptive and effective Machine Learning–based spam detection systems.

Keywords: Comparison, Machine learning models, Spam email detection, Classification.

1. Introduction

The development of digital communication has triggered an explosion in the volume of emails every day. Amidst legitimate communication, spam emails have become increasingly sophisticated, employing various manipulative techniques, ranging from phishing and financial scams to the spread of malware designed to be challenging to detect with the naked eye (Alkhalil et al., 2021). Spam attacks not only disrupt productivity but also pose a serious threat to data security and user privacy worldwide.

Conventional rule-based spam filters are increasingly inadequate in dealing with dynamic and evolving spam attack patterns. New adaptive patterns often go undetected by static rule-based systems, necessitating more innovative and more flexible detection solutions (Blanzieri and Bryl, 2008). Machine Learning (ML) is a promising approach due to its ability to learn from data and automatically recognize new patterns, offering hope for the creation of more accurate spam filter systems that are resistant to increasingly complex attacks (Singh et al., 2025).

Machine Learning (ML)-based spam filter systems are a necessity in today's digital age due to their ability to learn from data and automatically recognize new patterns. Unlike conventional methods that rely on static rules, ML models can adapt to the ever-evolving variety and complexity of spam attacks. The accuracy and robustness of ML models in detecting spam play a crucial role in maintaining the security of digital communication, ensuring that legitimate messages are delivered appropriately while minimizing the risk of harmful emails (Ozkan-Okay et al., 2024).

Although various Machine Learning algorithms have been applied in spam detection, selecting the most effective model remains a challenge because the performance of each algorithm is highly dependent on the characteristics of the data and the feature representation method used. Most previous studies have focused on specific models without conducting comprehensive comparative analyses, particularly of more modern ensemble and boosting algorithms. This study aims to conduct a comparative analysis of several Machine Learning models in detecting spam emails.

2. Methods

2.1. Data Collection

The data used in this study was obtained from the Kaggle platform under the title "SMS Spam Collection Dataset." This dataset contains a collection of text messages (SMS) that have been manually labeled into two categories, namely ham (non-spam messages) and spam (spam messages). The dataset consists of two primary columns, namely Category, which indicates the label (ham or spam), and Message, which contains the message content.

2.2. Data Preprocessing

Data preprocessing was carried out to convert raw text data into a numerical form that Machine Learning algorithms could process. The first step was to remove duplicate data to ensure that no message was recorded more than once in the dataset (Ruskanda, 2019). Text cleaning is performed, which includes converting all characters to lowercase, removing numbers, punctuation marks, and other non-alphabetic characters. Words included in the English stopwords list are also removed because they are considered to have no significant contribution to the meaning of the message context. After the text was cleaned, it was transformed into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. At this stage, only the top 3,000 features were selected based on word frequency to reduce the data dimension while retaining important information from the messages.

2.3. Modeling

This study uses four Machine Learning algorithms to detect spam emails, namely Extra Trees Classifier, Support Vector Machine (SVM) with RBF kernel, Random Forest Classifier, and CatBoost Classifier.

a) Extra Trees
Extra Trees is a decision tree-based ensemble algorithm that builds multiple trees with random split point selection. The final prediction is generated through a majority voting process of all trees (Almakhya et al., 2022).

b) Support Vector Machine (SVM) with RBF Kernel
SVM works by finding the optimal hyperplane that separates two data classes with maximum margin. The use of the Radial Basis Function (RBF) kernel enables SVM to project data into a higher-dimensional space, allowing it to handle non-linear classification cases (Guido et al., 2024).

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \quad (1)$$

where γ is the kernel parameter that controls the range of influence of a data point.

c) Random Forest
Random Forest is an ensemble method that builds multiple decision trees from randomly selected subsets of data. This model combines the prediction results from all trees using majority voting (Sun et al., 2024).

$$\hat{y} = mode(h_1(x), (h_2(x), \ldots, (h_n(x) \quad (2)$$

d) CatBoost Classifier
CatBoost is a decision tree-based boosting algorithm that adopts gradient boosting techniques with special handling of categorical features and prediction shift bias reduction. This model is built iteratively by focusing training on the prediction errors of the previous model (Qian et al., 2023).

$$\widehat{C_j} = \frac{1}{|S_j|} \sum_{x_i \in S_j} y_i \quad (3)$$

2.4. Evaluation

This study employs traditional metrics, namely accuracy and F1 score, as primary measures.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (4)$$

$$F1 - Score = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

where TP (true positive) is the actual positive label and is predicted to be positive, while TN (true negative) is the exact negative label and is predicted to be negative, FP (false positive) occurs when negative data is classified as positive, and FN (false negative) when positive data is classified as negative.

3. Results and Discussion

3.1. Model performance evaluation

The results of model performance testing against four Machine Learning algorithms are shown in Table 1.

Table 1. Model performance evaluation

No	Model	Accuracy	F1 Score	ROC AUC
1	Extra Trees	0.9729	0.8814	0.9868
2	Random Forest	0.9709	0.8684	0.9896
3	SVM (RBF)	0.9700	0.8646	0.9921
4	CatBoost	0.9671	0.8496	0.9750

Testing of four Machine Learning algorithms showed that the Extra Trees Classifier model performed best with an Accuracy of 97.29%, an F1 Score of 0.8814, and a ROC AUC of 0.9868. This model excelled in maintaining a balance between spam and ham predictions with a low error rate. The Random Forest Classifier model ranked second with an accuracy of 97.09%, an F1 score of 0.8684, and an ROC AUC of 0.9896. Although its AUC value is slightly higher than that of Extra Trees, its F1 score is lower, indicating that this model is somewhat less optimal in terms of precision and recall balance.

The Support Vector Machine (SVM) model with an RBF kernel exhibits competitive performance, achieving an Accuracy of 96.99%, an F1 Score of 0.8646, and the highest ROC AUC among all models at 0.9921. The high AUC value indicates the SVM's ability to distinguish between spam and ham classes with precision at various thresholds. However, its F1 Score is slightly below that of Random Forest and Extra Trees. The CatBoost Classifier ranks fourth with an Accuracy of 96.71%, an F1 Score of 0.8496, and an ROC AUC of 0.9750. Although its performance is slightly lower than that of other models, CatBoost still demonstrates stable performance with an accuracy above 96% and an AUC close to 0.98, confirming its ability to handle text classification despite its more complex boosting approach.

3.2. Confusion Matrix Analysis

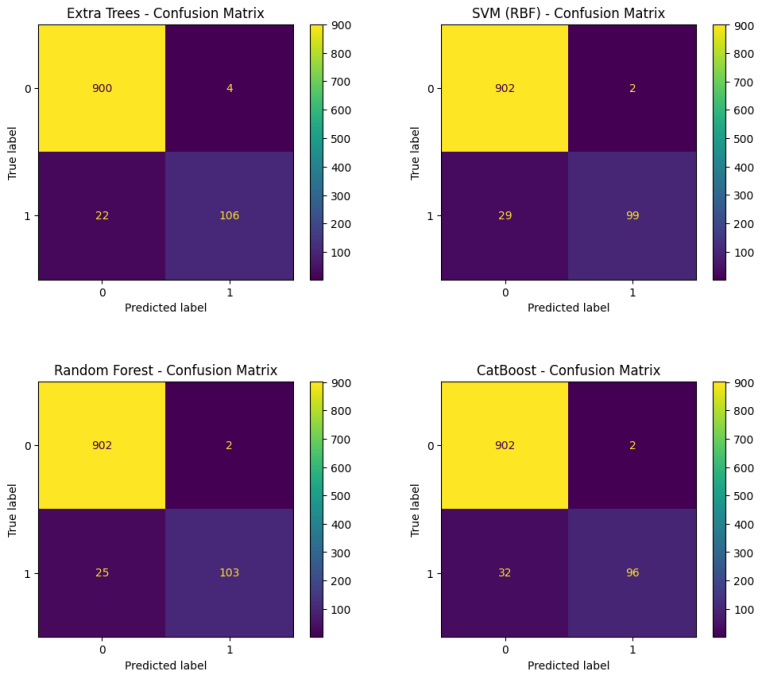

Figure 1: Comparison of confusion matrices

The Confusion Matrix is used to evaluate the performance of model predictions in more detail, particularly in terms of the distribution of correct predictions (True Positives and True Negatives) and classification errors (False Positives and False Negatives). The results of the Confusion Matrix for the four models tested can be seen in Figure 1 above.

This model successfully classified 900 ham messages correctly (True Negative) and 106 spam messages correctly (True Positive). However, 22 spam messages were incorrectly classified as ham (False Negative), and 4 ham messages were incorrectly classified as spam (False Positive). Nevertheless, the error rate of this model is relatively low with good prediction balance. SVM (RBF) demonstrated a strong capability in correctly classifying 902 ham messages; however, 29 spam messages were not detected (False Negatives). Only two ham messages were incorrectly classified as spam

(False Positives). Although the False Negative rate is higher than that of Extra Trees, this model still has good discriminative power, as evidenced by the highest ROC AUC value.

Random Forest correctly classified 902 ham messages and 103 spam messages as spam (True Positive). However, this model still produced 25 False Negatives and 2 False Positives. In general, Random Forest's performance is similar to Extra Trees, although it is slightly inferior in terms of spam prediction balance. CatBoost correctly classified 902 ham messages and 96 spam messages as spam (True Positive). However, this model recorded the highest number of False Negatives, with 32 spam messages incorrectly classified as ham, as well as 2 False Positives. This indicates that while CatBoost is relatively stable in its predictions, its sensitivity to spam messages is slightly lower compared to other models.

3.3. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) Curve is used to evaluate a model's ability to distinguish between spam and ham classes at various probability thresholds. The closer the curve is to the upper left corner of the graph, the better the model's performance in maximizing the True Positive Rate (TPR) while minimizing the False Positive Rate (FPR). An Area Under the Curve (AUC) value close to 1 indicates that the model has excellent classification capabilities.

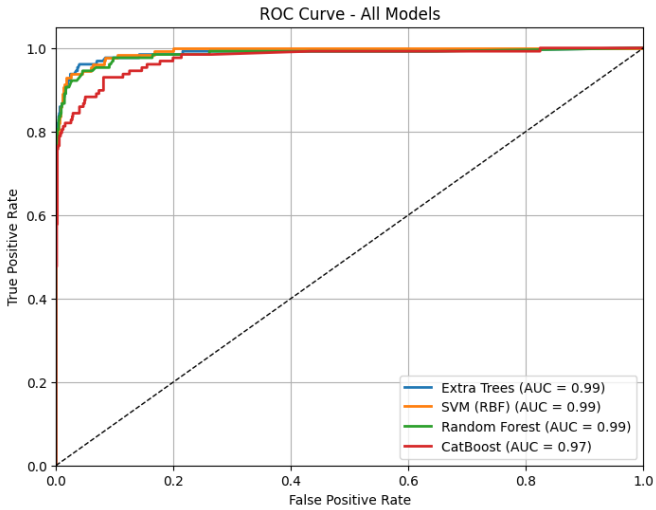

Figure 2: Comparison of ROC Curve graphs

In Figure 2, all four models tested demonstrated excellent ROC curve performance, with AUC values exceeding 0.97. The Extra Trees, Random Forest, and SVM (RBF) models all achieved an AUC of 0.99, indicating that all three have a high level of accuracy in separating spam and ham classes at various thresholds. The curves of these three models are almost touching the upper y-axis, reflecting their excellent performance in detecting spam. The CatBoost model has an AUC of 0.97, which is slightly lower than that of the other three models. The CatBoost ROC curve appears somewhat lower at the beginning (at a very small FPR), indicating that this model is more prone to false negatives at low thresholds, as also reflected in the previous Confusion Matrix results. The ROC Curve shows that Extra Trees, Random Forest, and SVM (RBF) have highly competitive detection capabilities and are balanced in minimizing classification errors. Although CatBoost exhibits stable performance, it lags slightly behind the others in terms of sensitivity (True Positive Rate).

4. Conclusion

Based on the evaluation results, the Extra Trees model demonstrated the best performance, with an Accuracy value of 97.29%, an F1 Score of 0.8814, and an ROC AUC of 0.9868. This was followed by Random Forest and SVM (RBF), which also recorded competitive evaluation results. Although SVM recorded the highest AUC value of 0.9921, this model had a higher number of false negatives compared to Extra Trees, resulting in a slightly lower balance of precision and recall. CatBoost showed stable performance but was relatively lower than the other three models in terms of F1 Score and sensitivity to spam. These results indicate that decision tree-based ensemble models, particularly Extra Trees and Random Forest, have an advantage in handling text classification for accurate and balanced spam detection. On the other hand, SVM remains a strong choice in scenarios where precise classification margins are critically needed.

References

Al Mahkya, D., Notodiputro, K. A., & Sartono, B. (2022). Extra trees method for stock price forecasting with rolling origin accuracy evaluation. Media Statistika, 15(1), 36-47.

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1), 63-92.

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An overview on the advancements of support vector machine models in healthcare applications: a review. Information, 15(4), 235.

Ozkan-Okay, M., Akin, E., Aslan, Ö., Kosunalp, S., Iliev, T., Stoyanov, I., & Beloev, I. (2024). A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. IEEe Access, 12, 12229-12256.

Qian, L., Chen, Z., Huang, Y., & Stanford, R. J. (2023). RETRACTED: Employing categorical boosting (CatBoost) and meta-heuristic algorithms for predicting the urban gas consumption.

Ruskanda, F. Z. (2019). Study on the effect of preprocessing methods for spam email detection. Indonesian Journal on Computing (Indo-JC), 4(1), 109-118.

Singh, G., Acharya, H. B., & Kwon, M. (2025). Programmable Data Planes for Network Security. arXiv preprint arXiv:2507.22165.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications, 237, 121549.