Comparison of robust machine learning algorithms on outliers and imbalanced spam data

Dodo Zaenal Abidin[1], Jasmir Jasmir[2], Errisya Rasywir[3], Agus Siswanto[3]
[1]Department of Magister of Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[2]Department of Computer Engineering, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia
[3]Information System, Faculty of Computer Science, Universitas Dinamika Bangsa, Jambi, Indonesia

Article Info

Article history:

Received Nov 12, 2024
Revised Mar 22, 2025
Accepted Jul 2, 2025

Keywords:

Comparison
Imbalanced data
Machine learning
Outliers
Spam

ABSTRACT

Effective spam detection is essential for data security, user experience, and organizational trust. However, outliers and class imbalance can impact machine learning models for spam classification. Previous studies focused on feature selection and ensemble learning but have not explicitly examined their combined effects. This study evaluates the performance of random forest (RF), gradient boosting (GB), and extreme gradient boosting (XGBoost) under four experimental scenarios: (i) without synthetic minority over-sampling technique (SMOTE) and outliers, (ii) without SMOTE but with outliers, (iii) with SMOTE and without outliers, and (iv) with SMOTE and with outliers. Results show that XGBoost achieves the highest accuracy (96%), an area under the curve-receiver operating characteristic (AUC-ROC) of 0.9928, and the fastest computation time (0.6184 seconds) under the SMOTE and outlier-free scenario. Additionally, RF attained an AUC-ROC of 0.9920, while GB achieved 0.9876 but required more processing time. These findings emphasize the need to address class imbalance and outliers in spam detection models. This study contributes to developing more robust spam filtering techniques and provides a benchmark for future improvements. By systematically evaluating these factors, it lays a foundation for designing more effective spam detection frameworks adaptable to real-world imbalanced and noisy data conditions.

Corresponding Author:

Dodo Zaenal Abidin
Department of Magister of Information System, Faculty of Computer Science
Universitas Dinamika Bangsa Jambi
Jenderal Sudirman Street, Thehok, South Jambi, Jambi, Indonesia
Email: dodozaenalabidin@gmail.com

1. INTRODUCTION

The rapid growth of technology-based applications and the massive volume of digital data have highlighted the need for efficient data processing and analysis. Spam remains one of the most challenging issues for both researchers and practitioners due to the disruptions it causes for users and service providers [1], [2]. It negatively impacts user experience and poses significant security and trust concerns for organizations [3].

Machine learning algorithms have been widely applied in spam detection with promising results [4]. However, new challenges emerge, particularly when datasets contain outliers and class imbalances, which can significantly affect model performance [5]. Addressing these challenges requires further investigation into robust machine learning strategies, particularly in evaluating their effectiveness in handling outliers and class imbalance [6]. Several studies have explored different aspects of spam classification. Asdaghi and

Soleimani [7] investigated the impact of dimensionality reduction on spam classification, particularly in unbalanced datasets, and concluded that random forest (RF) achieved an accuracy of 94.86%. Similarly, Fayaz *et al*. [8] proposed an ensemble model using gradient boosting (GB) and extreme gradient boosting (XGBoost), which improved classification accuracy to 84.74% for gradient boosting machine (GBM) and 85.59% for XGBoost on unbalanced datasets.

Despite these advancements, limited research has directly benchmarked the performance of these algorithms in handling both outliers and class imbalance within spam detection datasets. While previous studies focused primarily on dimensionality reduction and ensemble learning, they did not explicitly examine how outlier handling and data imbalance impact classification performance in spam detection systems. Benchmarking studies are crucial to evaluating algorithm robustness under real-world noisy and imbalanced spam data.

Misclassification can result in losing relevant information and additional false positives (FPs) that are harmful to users and organizations [9], [10]. Hence the need for finding and studying algorithms capable of assuaging these concerns [11], [12]. The objective of this study is to fill the mentioned gap by comparing the performance of three robust machine learning algorithms, RF, GB, and XGBoost, across different spam detection datasets facing outliers and imbalance data issues.

By carrying out this type of analysis on the Spambase data set (4,601 samples and 57 relevant features) we hope to provide additional insights into how well each of these algorithms will manage data with such characteristics. The Spambase dataset from the UCI Machine Learning Repository is a good source for more insights on this problem [13]. Also, in this study, a new styling and approach to evaluating algorithms like accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) area under the curve (AUC) will be discussed. These metrics were chosen since each offers an alternative perspective on the quality of the algorithm regarding spam identification and reduction in misclassifications [14], [15]. In particular, ROC AUC shows in clear terms how well the model can distinguish between the spam and non-spam classes which is key to performance evaluation on imbalanced datasets [16], [17]. Through this contribution, we aim to lay a foundation for future research in spam detection and robust learning algorithms, ultimately enhancing the effectiveness of existing spam classification systems.

2. METHOD

This study evaluates the performance of machine learning models in handling outliers and class imbalance in spam detection. The research follows a structured process involving data collection, exploratory analysis, data preprocessing, model testing, and evaluation. The procedural steps are illustrated in Figure 1.

Figure 1 illustrates the experimental setup for spam detection simulation. The process begins with the spam dataset, followed by exploratory data analysis (EDA) to examine feature distribution, outliers, and class imbalance. In data preprocessing, data cleaning, feature engineering, and synthetic minority over-sampling technique (SMOTE) are applied to handle class imbalance and prepare the dataset for model training. During model testing, three machine learning algorithms RF, GBM, and XGBoost are evaluated under four experimental conditions, considering the presence or absence of SMOTE and outliers. Each algorithm is trained and tested to assess its performance on imbalanced and noisy data.

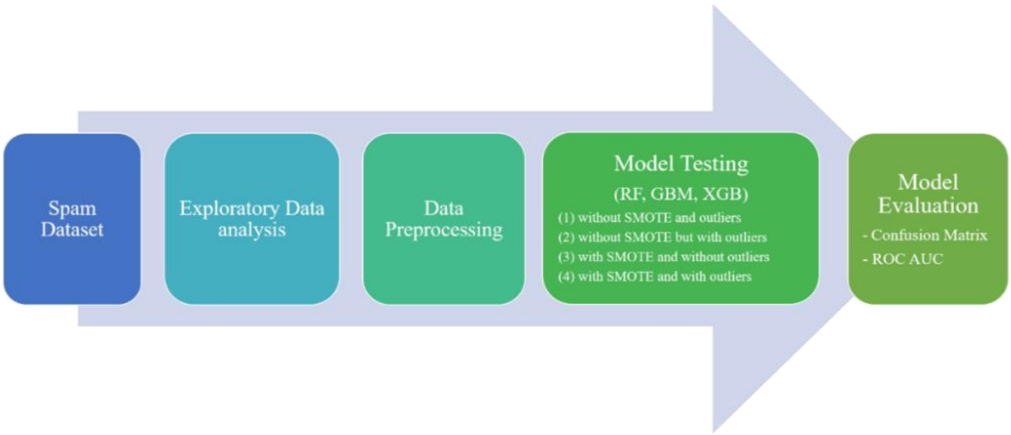
Spam Dataset
Exploratory Data analysis
Data Preprocessing
Model Testing
(RF, GBM, XGB)
(1) without SMOTE and outliers
(2) without SMOTE but with outliers
(3) with SMOTE and without outliers
(4) with SMOTE and with outliers
Model Evaluation
- Confusion Matrix
- ROC AUC

Figure 1. Stages of analysis and evaluation of spam dataset

Finally, model evaluation is conducted using confusion matrix, precision, recall, F1-score, and ROC-AUC to measure classification accuracy and robustness. This structured experimental setup highlights the impact of data preprocessing techniques on spam classification, ensuring the selection of the most effective model for real-world implementation.

2.1. Dataset

The Spambase dataset is available as part of the UCI Machine Learning Repository and can be used for spam detection, where it includes 4,601 emails with their classifications Frederickson *et al.* [18]. It has 57 attributes of real values (after normalizations) and one target feature representing whether a given e-mail is considered spam, we will consider as marginally positive examples only those for which its outputs are between to go from not being spam to the others. The attributes of this dataset are suited for tasks such as automatically filtering out spam emails from a user's email account [13], [19]. The data has already been cleaned and converted to numerical form, so there is no need for normalization here.

The datasets have mostly balanced class distribution, but further analysis is necessary to mitigate the potential impact of class imbalance on ML algorithms. These datasets will be split into training and testing sets, typically at a 70/30 ratio, ensuring sufficient data for model training while preserving an independent test set for unbiased evaluation. Data preprocessing techniques, including feature scaling and outlier removal, will be applied before training to enhance model performance. Additionally, cross-validation will be used to assess generalizability and prevent overfitting, ensuring that the trained models perform consistently across different data distributions.

2.2. Exploratory data analysis

Importance of the EDA stage: it is extremely important to perform an initial analysis before modeling as this helps a researcher understand different characteristics of the Spambase dataset [20], [21]. EDA helps in understanding the discrepancies between spam and non-spam emails as well as in finding outliers that might affect algorithm performance. In addition, EDA helps in choosing the right features and how to treat data (use of more robust algorithms for Outliers/unbalanced data).

The EDA results clearly stated that the no missing values in the Spambase dataset so we did not apply anything to this. Having said that, there are some outliers present in the features word_freq_george and word_freq_remove which might hurt model performance. This can be mitigated with techniques like trimming, or using stronger algorithms such as RF, GB, and XGBoost. Word_freq_your and char_freq_$ are important features in distinguishing spam as they exhibit the strongest positive correlation with the target (spam). Based on these correlations, feature selection can be performed to enhance model performance. Although the data consists of 60.6% non-spam and 39.4% spam, the class imbalance can be mitigated using SMOTE.

Summary: the dataset is now fully pre-processed. The implemented pipeline detects class imbalance and applies SMOTE to generate synthetic samples along the standard deviation, ensuring a balanced representation of spam and non-spam data. Additionally, highly correlated features with the target variable significantly influence model predictions, making feature selection a crucial step in improving classification accuracy and reducing noise.

2.3. Data preprocessing

Data preprocessing steps are performed to guarantee the quality and reliability of the analysis [22]. We first split the features and targets from X with the independent variables (features) and y with the dependent variable (targets), where 'Class' is the target column. Then, the SMOTE technique is applied to correct class imbalance by generating synthetic samples of the minority class, ensuring the model learns effectively from both spam and non-spam instances [23].

The resampled data is then divided into train and test sets using the train_test_split function, allocating 70% for training and 30% for testing. The stratified parameter is used to maintain the original class distribution across both sets, which is crucial for preventing biased learning and ensuring consistent model evaluation. Additionally, feature scaling and outlier removal are performed to enhance model performance by reducing the influence of extreme values and improving overall classification stability and generalization. These preprocessing steps play a critical role in improving classification accuracy and ensuring robust spam detection. By addressing outliers and class imbalance, this approach allows machine learning models to generalize better, enhance predictive reliability, and adapt effectively to real-world spam filtering challenges [5].

2.4. Model testing

This Spambase dataset is used to test algorithm models by examining different conditions of data imbalance and outliers so that the performance evaluation and comparison of algorithms like RF, GB, and

XGBoost can obtain accurate and reliable results, providing deeper insights into model robustness and classification effectiveness across varying data distributions.

This was tested in different scenarios. So, without SMOTE and outliers are used as a baseline for algorithm performance on the dataset that has not been altered. The second one, without SMOTE but with outliers, analyzes how outliers affect algorithm performance in the presence of imbalance. Third, with SMOTE and without outliers assesses performance improvement using SMOTE while also accounting for the presence of outliers. Finally, the with SMOTE and with outliers tests. The algorithm on SMOTE and outlier customized datasets as an overview of what the algorithm is good at when working too complex. Through these tests, we expect to determine the performances by which it functions best and what some of the best approaches may be in order to improve its accuracy and reliability as a spam classification model [24], [25]. Results will also help in the selection of the best algorithm for similar datasets later.

2.5. Model evaluation

Comparative performance assessment of algorithmic models like RF, GB, and XGBoost is expected to give insights into their ability in spam data classification by considering aspects such as class imbalance and outlier variables. Hence, both the confusion matrix and ROC AUC are important in this case as they provide a good understanding of the accuracy score [26]. In the confusion matrix, correct and incorrect predictions are represented as counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [26]. From these values calculation of accuracy, precision, recall, and F1-score can be done which gives an idea about the performance of the model in classifying spam or non-spam mail.

On the other hand, ROC AUC indicates how well a model distinguishes between positive (spam) and negative (non-spam) classes as thresholds change [27]. The AUC score measures how well predictions are ranked, rather than their absolute values (which is a problem with imbalanced data). As such, a higher AUC value indicates better classification performance. This study can use these two metrics to evaluate which algorithm performs best under real conditions providing a deeper understanding of the effectiveness of both. So, these outcomes will make a fundamental ground for utilizing machine-learning algorithms on spam datasets in past studies and in real-time cases which can efficiently work with spam identification techniques in respective scenarios.

2.6. Learning model

2.6.1. Random forest

RF is an ensemble learning algorithm that merges multiple decision trees, using majority voting to determine the outcome [28]. It is resilient to outliers due to its bagging approach, which creates random subsets of the data, thereby reducing the influence of outliers on the overall model [29]. When dealing with imbalanced data, RF can be optimized with techniques such as SMOTE, class weighting, or threshold tuning to better address the minority class. Additionally, feature importance analysis helps in selecting the most relevant attributes, improving model interpretability and reducing noise. The combination of these techniques enhances classification performance, particularly in spam detection, where distinguishing between legitimate and spam messages is crucial. In general, RF is a strong and versatile algorithm, effective at managing outliers and can be further refined to handle data imbalance more effectively.

2.6.2. Gradient boosting

GB is an ensemble method that constructs models in a sequential manner to rectify earlier mistakes [30]. It is not very resistant to outliers, as substantial errors from these outliers can influence the subsequent trees. This issue can be mitigated by employing a more robust loss function, like Huber loss, or by eliminating outliers during the data preprocessing phase. For data imbalance, GB can be optimized by adjusting class weights or using resampling techniques such as SMOTE [31], [32]. This step helps to handle minority classes to make the model more accurate. Overall, GB is robust, but requires customization to handle outliers and data imbalance.

Despite its sensitivity to noisy data, GB remains a powerful method for spam classification due to its ability to capture complex patterns. Hyperparameter tuning, including learning rate adjustment and tree depth regulation, plays a crucial role in improving performance. When combined with proper preprocessing and feature selection, GB can effectively enhance classification accuracy while mitigating the impact of outliers and class imbalance.

2.6.3. XGBoost

With optimizations like L1 and L2 regularization and tree pruning, XGBoost is a strong and effective ensemble learning algorithm [33]. By minimizing outlier influence through tree pruning and reducing the impact of large errors through regularization, these features increase their robustness to outliers [34]. XGBoost can be used in conjunction with resampling methods like SMOTE to address data imbalance by allowing class weight adjustments that give minority classes more weight. This makes it a useful algorithm for a variety of machine learning applications by increasing prediction accuracy on underrepresented classes [35].

Beyond handling outliers and class imbalance, XGBoost excels in spam detection due to its ability to efficiently process large datasets while maintaining high predictive performance. Its GB mechanism ensures that weak learners are sequentially improved, reducing misclassification rates. Additionally, tuning hyperparameters such as learning rate, max depth, and subsample ratio further refines its effectiveness, making it a competitive choice for spam classification tasks.

3. RESULTS AND DISCUSSION

This section presents the results and analysis of how machine learning algorithms perform when handling outliers and imbalanced spam data. The study focuses on three algorithms - RF, GB, and XGBoost known for their effectiveness in classification tasks under challenging conditions. The primary objective is to evaluate their robustness in real-world spam detection scenarios and assess their suitability for improving classification performance.

Spam detection is often hindered by the imbalanced distribution of spam and non-spam emails. To address this, this study applies four experimental scenarios: (i) without SMOTE and outliers, (ii) without SMOTE but with outliers, (iii) with SMOTE and without outliers, and (iv) with SMOTE and with outliers. These scenarios provide a comprehensive assessment of model behavior under different data conditions, ensuring a thorough evaluation of their strengths and weaknesses in spam classification.

Model evaluation is conducted using performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The confusion matrix provides insights into classification errors, while ROC-AUC measures the balance between TPs and FPs, aiding in threshold optimization. By analyzing these metrics, this study explores not only quantitative results but also the practical implications of model performance. This contributes to the development of more effective and robust spam detection systems, offering insights into their applicability in real-world settings.

3.1. Random forest model evaluation results

This section presents the evaluation results for the RF model, a key component of our study that compares machine learning algorithms robust to outliers and class imbalance in spam data. RF was selected due to its proven effectiveness in handling high-dimensional and class-imbalanced datasets. We performed a systematic evaluation across four test conditions: (1) without SMOTE and outliers, (2) without SMOTE but with outliers, (3) with SMOTE and without outliers, and (4) with SMOTE and with outliers. Each scenario offers insights into the model's adaptability and robustness in facing data challenges.

The results will be analyzed using appropriate metrics to assess the performance of the RF model in spam detection, supporting ongoing improvements in spam filtering systems. The evaluation includes accuracy, precision, recall, and AUC-ROC. A detailed summary of the findings is presented in Table 1.

Table 1 provides a performance comparison of four RF classifier scenarios, distinguished by the application of SMOTE and the presence of outliers. All scenarios maintained an accuracy of 0.96, demonstrating strong classification capability, stability, and robustness in handling imbalanced data while minimizing the impact of outliers on overall model performance and predictive reliability. Precision also remained at 0.96 across all scenarios, but scenario 3 (with SMOTE and without outliers) achieved the highest recall (0.97) and AUC-ROC (0.9920), indicating improved positive case detection. Conversely, scenario 4 recorded the lowest AUC-ROC (0.9910), reinforcing prior findings that outliers can slightly degrade model performance.

Table 1. Impact of SMOTE and Outliers on RF Classification Performance

Scenario	Accuracy	Precision	Recall	F1 score	AUC-ROC	Computation time
(1) without SMOTE and outliers	0.96	0.96	0.96	0.96	0.9919	0.7793 seconds
(2) without SMOTE but with outliers	0.96	0.96	0.96	0.96	0.9911	0.9116 seconds
(3) with SMOTE and without outliers	0.96	0.96	0.97	0.96	0.9920	0.8028 seconds
(4) with SMOTE and with outliers	0.96	0.96	0.96	0.96	0.9910	0.9018 seconds

To contextualize these results, Soleimani [7] reported 94.86% accuracy for RF in web spam detection using feature selection on the WEBSPAM-UK2007 dataset. While their study focused on feature selection, our research examines the impact of SMOTE and outlier handling in email spam classification. Although accuracy values are not directly comparable due to different datasets and preprocessing methods, both studies highlight RF's effectiveness in handling imbalanced data. Computation time varied, with Scenario 1 (without SMOTE and outliers) being the fastest at 0.7793 seconds, while scenario 2 (without SMOTE but with outliers) took the longest at 0.9116 seconds. Despite increasing computation time, SMOTE improved recall and AUC-ROC, justifying the trade-off.

In conclusion, applying SMOTE without outliers resulted in the highest recall and AUC-ROC scores, improving the model's ability to classify minority classes. This method is highly recommended for imbalanced datasets, and effective outlier handling can further enhance predictive accuracy. Ongoing monitoring of performance metrics is essential for sustaining model effectiveness. The performance evaluation of the RF model is shown in Figure 2.

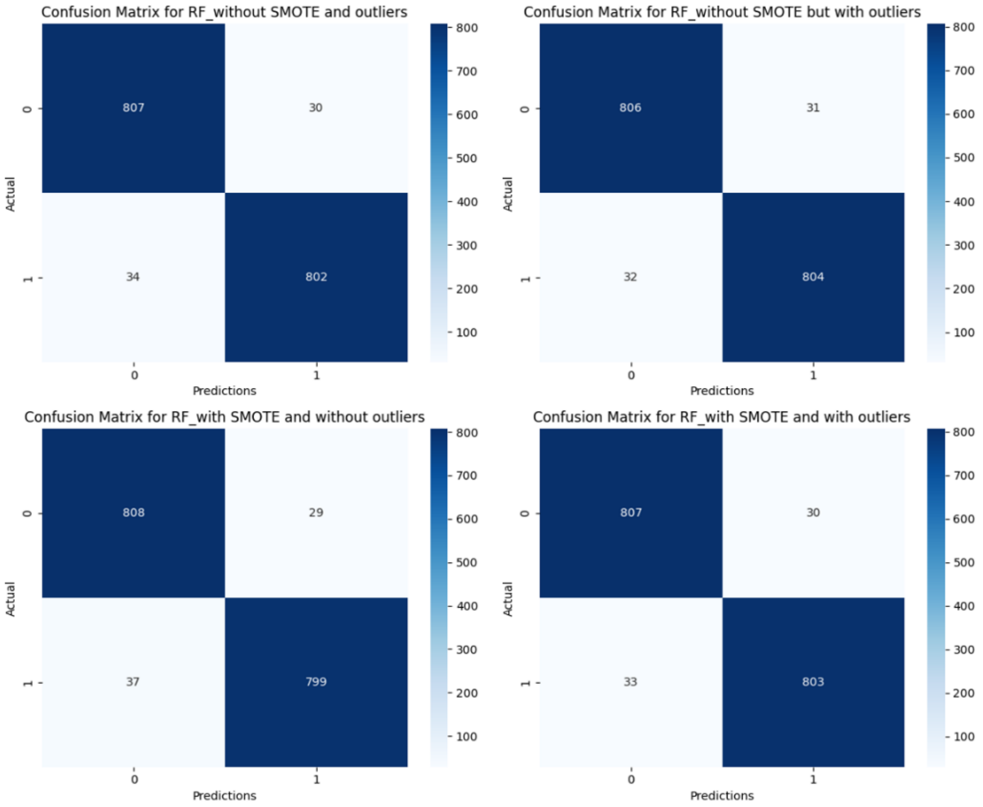

Figure 2. Comparison of confusion matrices for the RF model: impact of outliers and SMOTE

In the first scenario, the model showed strong performance with 807 TNs, 30 FPs, 34 FNs, and 802 TPs, indicating effective classification despite some misclassification in the positive class. The second scenario showed similar performance but with slightly higher FNs (32) and FPs (31), indicating that outliers can affect accuracy. With the application of SMOTE without outliers, the model captured more TNs (808) but misclassified more positives, resulting in 37 FNs. In the scenario with SMOTE and outliers, TNs remained high at 807, with 33 FNs, indicating better handling of positive class predictions.

Overall, the RF model classifies events effectively, although the introduction of SMOTE often increases FN results, and outliers affect predictions by increasing false classifications. These factors can influence overall model reliability and decision-making accuracy. To evaluate the performance of the model across different classes, the results are presented in Figure 3.

To measure the performance of machine learning algorithms, we use the ROC and AUC evaluation methods. In Figure 3, we can see a graph depicting the four ROC curves of various RF models. TP is measured on the vertical axis, representing the proportion of positive classes accurately identified by the model, while FP on the horizontal axis shows the proportion of negative classes misclassified as positive.

The results on all curves show an AUC value of 0.99, indicating outstanding performance in distinguishing between positive and negative classes. AUC values close to 1 suggest that the model predicts the class with high precision and robustness. Furthermore, a comparison between models shows similar performance, indicating that the addition of SMOTE and the presence of outliers do not significantly impact classification ability in this scenario. The consistently high AUC values across all conditions highlight the model's capability to generalize well across different data distributions. These results reinforce the effectiveness of the RF model, demonstrating its strong adaptability and reliability in handling spam classification challenges.

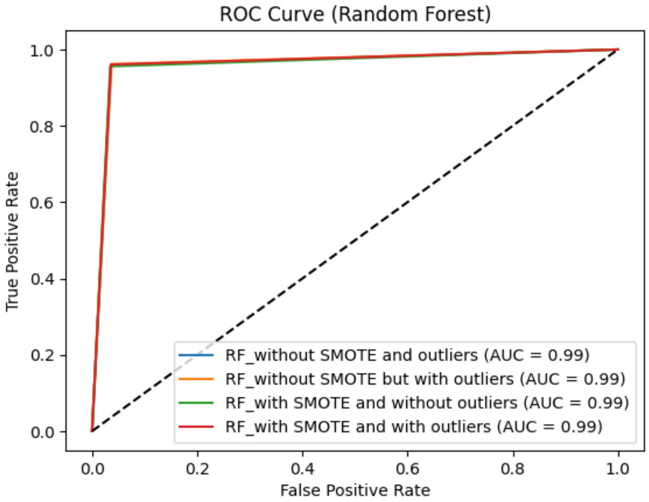

Figure 3. Performance evaluation of the RF model using ROC curve and AUC

3.2. Gradient boosting model evaluation results

This section provides an evaluation of the GB model, which serves as the central focus in comparing various machine learning algorithms capable of effectively managing outliers and integrating imbalanced class distributions in spam data. GB was selected for its exceptional capacity to process high-dimensional datasets and to address unbalanced class distributions through its iterative boosting approach. This technique systematically enhances the performance of weak learners in successive stages, making the model especially suited for complex classification tasks. As a result, GB consistently enhances prediction accuracy and precision, adapting well to the unique characteristics of challenging datasets. Its structured boosting process is particularly advantageous in scenarios where precise classification of minority classes is critical, further reinforcing its role as a powerful tool in spam detection and other classification challenges.

To evaluate the effects of SMOTE and outliers on the classification performance of the gradient-boosting model, we carried out a series of experiments. The results, detailed in Table 2, present metrics like AUC-ROC and computation time, offering insights into how the accuracy, precision, recall, and computational efficiency of the model are influenced by outliers and the use of SMOTE to balance the dataset. This analysis emphasizes the critical role of preprocessing methods and computational factors in enhancing the model’s predictive accuracy and efficiency.

Table 2 shows the comparison results of the GB model performance in various scenarios, presenting key metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and computation time. The GB model was evaluated across four scenarios, varying based on SMOTE application and outlier presence. All scenarios achieved similar accuracy levels ranging from 0.94 to 0.95, indicating consistent classification performance.

Precision and recall values remained stable between 0.94 and 0.95, showing that the model effectively predicted positive cases with minimal FPs and negatives. The F1 score ranged between 0.94 and 0.95, confirming a well-balanced model. These stable metrics suggest that GB maintains high classification reliability across different conditions.

Table 2. Impact of SMOTE and outliers on GB classification performance

Scenario	Accuracy	Precision	Recall	F1 score	AUC-ROC	Computation time
(1) without SMOTE and outliers	0.95	0.95	0.95	0.95	0.9870	2.6641 seconds
(2) without SMOTE but with outliers	0.94	0.94	0.94	0.94	0.9876	3.1801 seconds
(3) with SMOTE and without outliers	0.95	0.95	0.95	0.95	0.9874	2.2660 seconds
(4) with SMOTE and with outliers	0.95	0.95	0.95	0.95	0.9876	2.1856 seconds

Testing results in all scenarios yielded statistically similar AUC-ROC values between 0.9870 and 0.9876, demonstrating the model's strong ability to distinguish spam from non-spam emails. However, computation time varied, with the scenario without SMOTE and outliers taking the longest (3.1801 seconds), while the scenario with SMOTE and outliers resulted in the shortest (2.1856 seconds). SMOTE appears to impact computational efficiency, but further investigation is needed to confirm its role in reducing processing time.

Overall, the GB model demonstrates strong performance across all scenarios. The combination of SMOTE and outlier handling provides the best efficiency while maintaining high accuracy, precision, recall, F1 score, and AUC-ROC. This evaluation confirms the robustness and adaptability of GB in handling imbalanced datasets and noisy data. Future research could explore integrating feature selection with SMOTE to further optimize classification performance while maintaining computational efficiency. The performance evaluation of the GB model is presented in Figure 4.

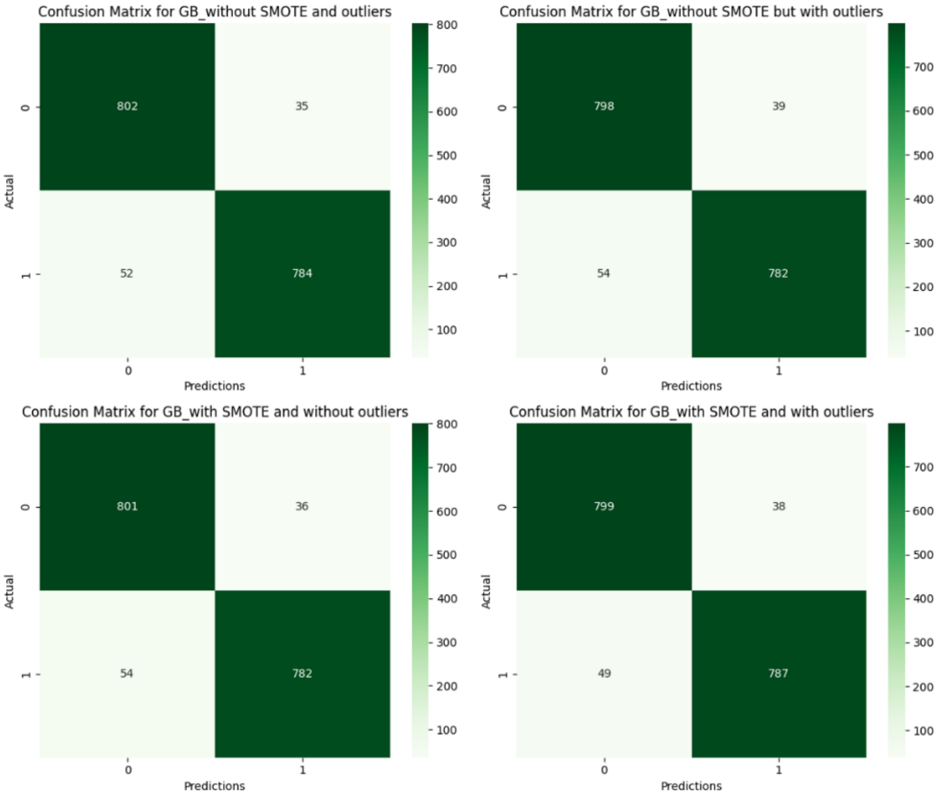

Figure 4. Comparison of confusion matrices for the GB model: impact of outliers and SMOTE

The confusion matrix in Figure 4 shows the performance of the GB algorithm across four scenarios. The test results in all scenarios show high accuracy values, with noticeable TN and TP values, such as 802 TN and 784 TP values in the first matrix, indicating strong predictive ability. When comparing the impact of SMOTE, the results are similar for GB with SMOTE and without outliers and GB without SMOTE and outliers, indicating a limited effect on classification performance. However, introducing outliers increased FPs and FNs, with 39 FP and 54 FN in the GB without SMOTE and outliers scenario highlighting the detrimental effect of outliers.

In summary, the GB algorithm maintained strong performance in scenarios with and without SMOTE, but caution is required in the presence of outliers. This analysis underscores the importance of evaluating factors such as SMOTE and outliers to understand their effect on model performance. To evaluate the performance of the model across different classes, the results are presented in Figure 5.

The ROC curve for the GB algorithm evaluates its classification performance across the four scenarios. All curves showed consistent performance, with an impressive AUC of 0.99 across all scenarios, regardless of the presence of SMOTE or outliers. This indicates that GB maintains high discriminative ability, effectively distinguishing between classes, while remaining robust against class imbalance and variations caused by outliers.

The high AUC indicates that the model excels at distinguishing between positive and negative classes. The GB model also successfully keeps the FN rate low, which is crucial in situations where missing positive instances is highly detrimental. Overall, the analysis underscores the strong performance of GB in handling class imbalance and outliers, validating its reliability for classification tasks within the dataset.

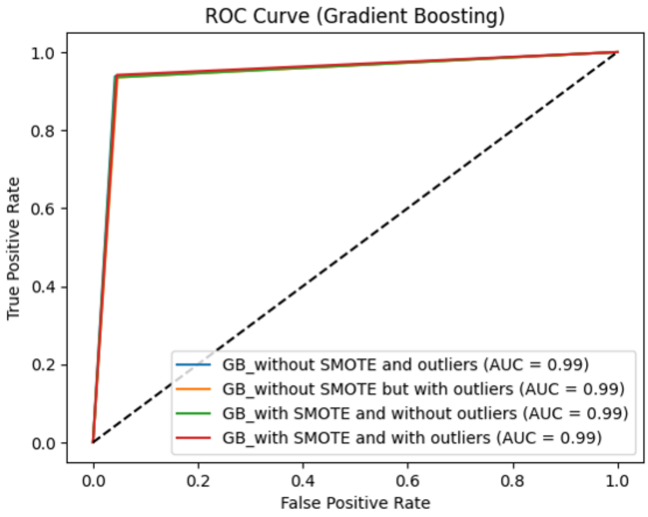

Figure 5. Performance evaluation of the GB model using ROC curve and AUC

3.3. XGBoost model evaluation results

This section details the evaluation results of the XGBoost model, focusing on its effectiveness in handling outliers and maintaining class continuity in spam data. XGBoost was selected for its robust capability to manage high-dimensional datasets via an efficient boosting algorithm and powerful regularization features that enhance accuracy and reduce overfitting. Its decision tree-based framework enables adaptive learning from past errors, making XGBoost particularly well-suited for tackling complex classification tasks.

To evaluate the impact of SMOTE and outliers, we conducted the experiments shown in Table 3, which include metrics such as AUC-ROC and computation time. The results highlight how outliers and the application of SMOTE for balancing the dataset affect accuracy, precision, recall, and computational efficiency. This underscores the importance of preprocessing methods in improving model performance.

XGBoost performance remained stable across four scenarios (with/without SMOTE and outliers), maintaining an accuracy, precision, recall, and F1 score of approximately 96%. This indicates that neither SMOTE nor outliers significantly impacted overall classification performance and model consistency. AUC-ROC values ranged from 0.9923 to 0.9928, with a slight increase when SMOTE was applied, confirming consistent class separation. The impact of SMOTE and outliers on computational efficiency was also analyzed. Scenario 3 (SMOTE without outliers) had the fastest execution time (0.6184 seconds), while scenario 4 (outliers included, no SMOTE) took the longest (2.2914 seconds).

Fayaz *et al.* [8] reported XGBoost achieving 85.59% accuracy in classifying spam product reviews using feature selection and ensemble learning. While their study focused on feature selection, our research applies SMOTE and outlier handling for email spam classification. As these studies use different datasets and preprocessing techniques, accuracy values are not directly comparable. However, both findings indicate XGBoost's robustness in handling imbalanced data across different domains.

These results emphasize the importance of selecting preprocessing techniques based on dataset characteristics. Future research should explore hybrid approaches integrating feature selection and class balancing for further optimization. The performance evaluation of the XGBoost model is presented in Figure 6.

Table 3. Impact of SMOTE and outliers on XGBoost classification performance

Scenario	Accuracy	Precision	Recall	F1 score	AUC-ROC	Computation time
(1) without SMOTE and outliers	0.96	0.96	0.96	0.96	0.9923	1.4206 seconds
(2) without SMOTE but with outliers	0.96	0.96	0.96	0.96	0.9925	2.2914 seconds
(3) with SMOTE and without outliers	0.96	0.96	0.96	0.96	0.9928	0.6184 seconds
(4) with SMOTE and with outliers	0.96	0.96	0.96	0.96	0.9928	0.7840 seconds

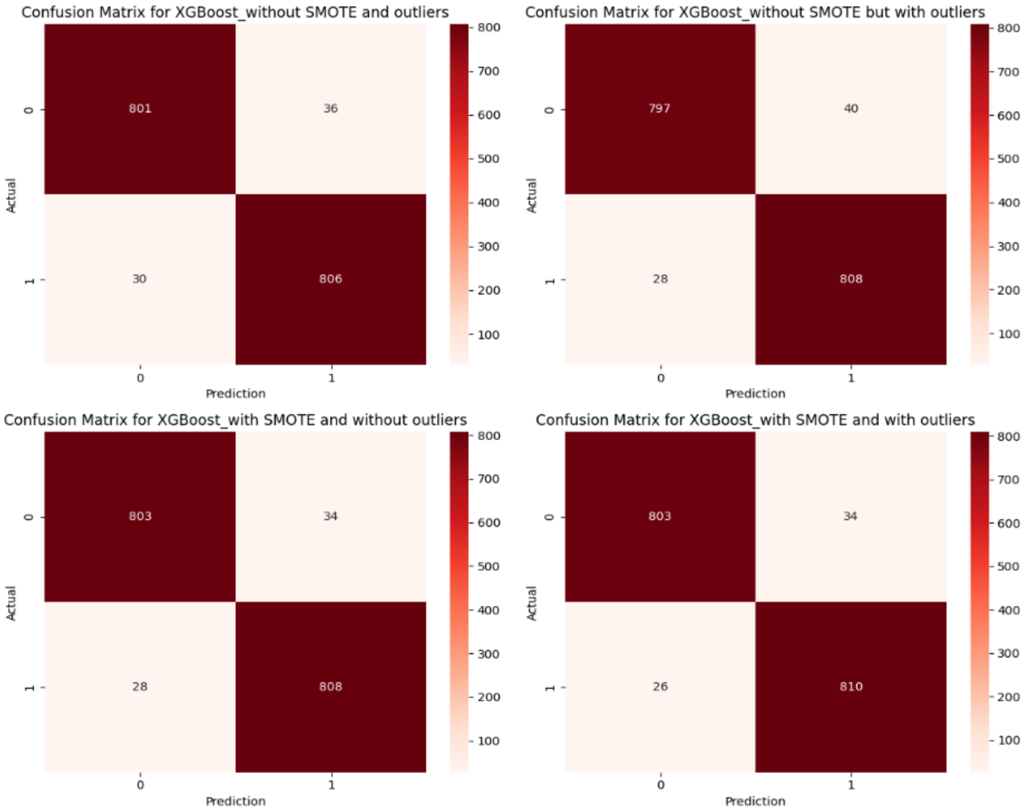

Figure 6. Comparison of confusion matrices for the XGBoost model: impact of outliers and SMOTE

Based on the combined confusion matrix graph for the XGBoost algorithm, this analysis illustrates the prediction results with and without the application of SMOTE and the presence of outliers. The findings highlight variations in misclassification rates, particularly in minority classes, demonstrating how SMOTE improves recall while outliers slightly affect overall model stability and precision. In the XGBoost model without SMOTE and outliers, the model recorded 806 TPs, 801 TNs, 36 FPs, and 30 FNs. These results reflect high accuracy in predicting both classes, indicating that the model effectively distinguishes between them under these conditions.

When evaluating the XGBoost model without SMOTE but with outliers, the performance showed 808 TPs and 797 TNs. However, the number of FPs increased to 40, while FNs slightly decreased to 28. This suggests that while the model's ability to correctly identify positive cases improved, the presence of outliers had a minor impact, slightly affecting the overall predictive performance.

After applying SMOTE to the XGBoost model without outliers, the count of TP stayed at 808, while TN rose to 803 and FP dropped to 34. This reflects a significant improvement in class balance, enhancing the model's ability to predict the negative class more accurately. Additionally, for the XGBoost model with both SMOTE and outliers, the TP increased slightly to 810, TN remained unchanged at 803, FP stayed at 34, and FN decreased to 26. This indicates that combining SMOTE with outliers improved the detection of positive classes while preserving overall model performance.

Overall, the assessment of XGBoost reveals consistent performance across different scenarios, with SMOTE proving essential for effectively balancing the classes. The impact of outliers seems to be minimal, indicating that XGBoost is resilient enough to manage noise in the data without a substantial decline in accuracy. To thoroughly evaluate the model's performance across various classes, the results are illustrated in Figure 7, which visualizes the distribution of prediction results and offers additional insights into the model's effectiveness.

The combined ROC curve analysis for the XGBoost algorithm shows that all models perform exceptionally well. Regardless of whether SMOTE is applied or outliers are present, every model configuration achieved an AUC of 0.99, which is very close to the optimal value of 1. This demonstrates the model's outstanding classification capability. There is no substantial difference among the results of each model, suggesting that neither the use of SMOTE nor the existence of outliers significantly impacted the models' predictive capabilities. XGBoost consistently demonstrates the ability to accurately predict both positive and negative classes.

Overall, XGBoost has demonstrated exceptional robustness, suggesting that class balancing with SMOTE and the presence of outliers have minimal impact on its classification performance. The model achieves a high TP rate while keeping FPs to a minimum, ensuring reliable spam detection and maintaining stability across varying data distributions and preprocessing conditions.

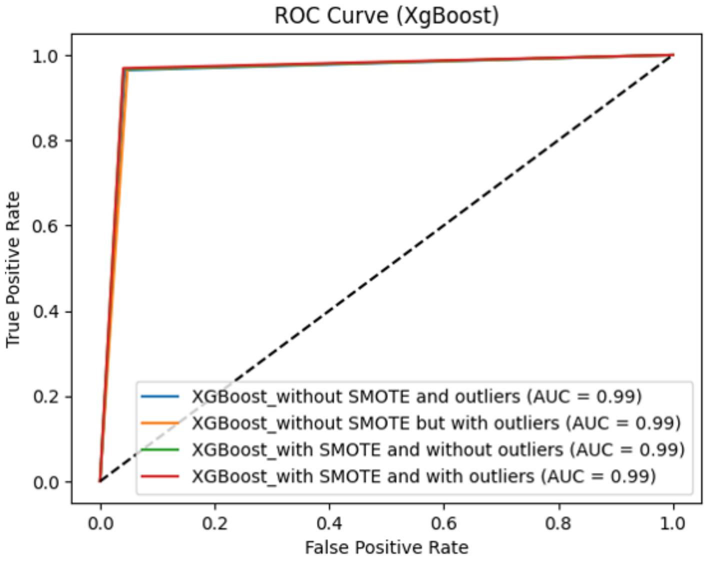

Figure 7. Performance evaluation of the XGBoost model using ROC curve and AUC

3.4. Best algorithm selection for imbalanced data with outliers

In this section we compare the best performance evaluation of three robust machine learning algorithms, the algorithms evaluated are RF, GB, and XGBoost. The algorithm models were tested on unbalanced and outlier spam data. Testing is done using four different scenarios. The objective of this stage is to find out the best-performing algorithm by comparing key metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and computation time. The comparison results are presented in Table 4. The data in Table 4 are the best algorithm results, tested in each scenario.

Based on the data in Table 4, the three algorithms RF, GB, and XGBoost were evaluated in their respective best scenarios, in applying SMOTE with or without outliers. The RF algorithm tested in the With SMOTE and without outlier scenarios achieved an accuracy of 0.96, with precision and F1 scores also at 0.96. Its recall value of 0.97 shows its effectiveness in identifying positive cases, making it ideal for situations where minimizing FNs is very important. Furthermore, it has a high AUC-ROC value of 0.9920, which illustrates its strong performance in discriminating between classes, while the computation time has a medium value of 0.8028 seconds.

Table 4. Best scenario performance for RF, GB, and XGBoost

Algorithm	Scenario	Accuracy	Precision	Recall	F1 score	AUC-ROC	Computation time
RF	with SMOTE and without outliers	0.96	0.96	0.97	0.96	0.9920	0.8028 seconds
GB	with SMOTE and with outliers	0.95	0.95	0.95	0.95	0.9876	2.1856 seconds
XGBoost	with SMOTE and without outliers	0.96	0.96	0.96	0.96	0.9928	0.6184 seconds

In contrast, GB evaluated in the with SMOTE and with outlier scenarios, produced an accuracy value of 0.95, with precision, recall, and F1 scores at 0.95. Although the GB model performed well, it still had a lower AUC-ROC score of 0.9876. as well as a longer computation time of 2.1856 seconds. These results show that although GB is a viable option for applications that require fast processing.

While the XGBoost algorithm has the best results in the with SMOTE and without outlier scenarios, from the test results it gets a value of 0.96 on the accuracy, precision, and recall metrics, and the F1 score value of 0.96. And AUC-ROC value of 0.9928. By looking at its superior ability to distinguish classes, and produce the fastest computation time at 0.6184 seconds. Based on these results, XGB is the best-performing model, making it very suitable for time-sensitive applications. To find out the performance comparison results of the algorithm, see Figure 8.

The graph in Figure 8 illustrates the performance comparison results of three machine learning algorithms namely RF, GB, and XGBoost. In the graph there are six performance metrics displayed, namely precision, recall, F1 score, AUC-ROC, and computation time shown as an orange line above the bar. The analysis shows that the RF and XGBoost algorithms have very good performance, with AUC-ROC values close to 1. While the GB algorithm is slightly lower, but still shows good performance with values around 0.95 for all metrics. Computation time is also a concern, where XGBoost has the lowest computation time (0.6184 seconds), followed by RF (0.8028 seconds), while GB takes the highest time (2.1856 seconds). This shows that although GB gives good results, it takes longer than the other two algorithms, which may be an important consideration in algorithm selection for time-efficient applications. In conclusion, both RF and XGBoost provide better performance compared to GB. If performance is the top priority, both algorithms are highly recommended. However, if time efficiency is also important, then XGBoost is a better choice than the other algorithms.

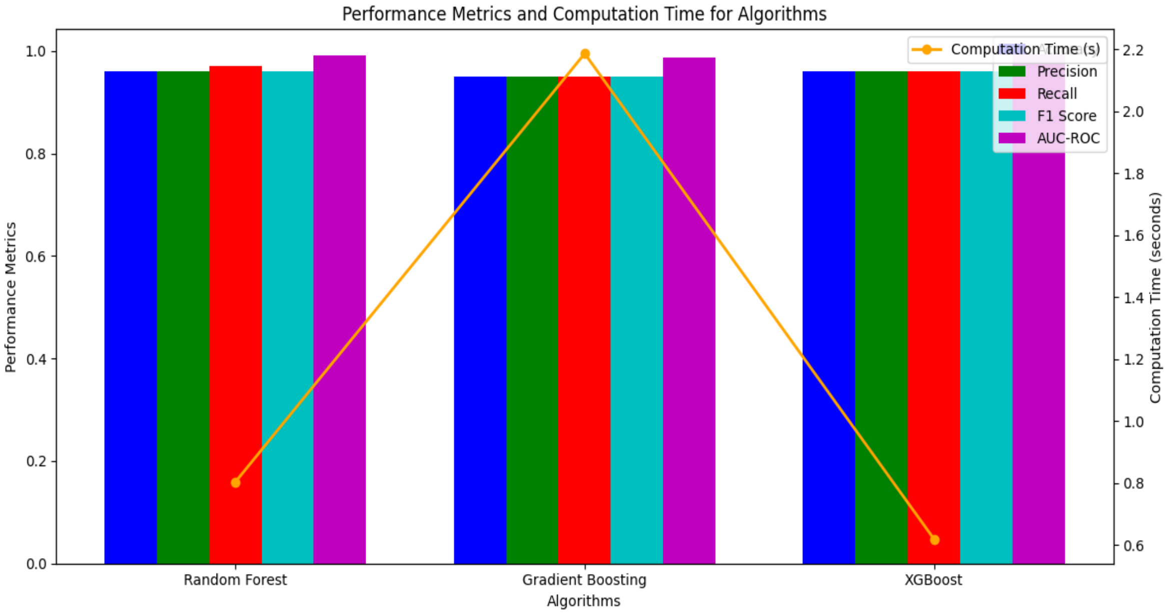

Figure 8. Performance comparison and computation time: RF, GB, XGBoost

4. CONCLUSION

This study evaluated the performance of RF, GB, and XGBoost in classifying imbalanced spam data with outliers under four experimental scenarios, analyzing key metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and computation time. The results indicate that XGBoost outperforms other models, achieving 96% accuracy, an AUC-ROC of 0.9928, and the fastest computation time (0.6184 seconds), making it suitable for real-time spam detection. RF also demonstrated strong performance in handling imbalanced data, achieving 96% accuracy and 0.97 recall but requiring a moderate computation time (0.8028 seconds), while GB, with 95% accuracy, showed effective classification but required longer computation (2.1856 seconds), making it less ideal for time-sensitive applications. These findings emphasize the importance of addressing class imbalance and outliers in spam classification, highlighting the impact of preprocessing techniques such as SMOTE and outlier handling on model performance. However, this study is limited to three machine learning models and specific datasets, which may not generalize to all spam detection cases. Future research should explore advanced ensemble methods, Bayesian optimization for hyperparameter tuning, and alternative outlier handling techniques, such as isolation forest or local outlier factor (LOF), while also expanding dataset diversity and evaluating deep learning approaches to further enhance spam detection accuracy and robustness.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support provided by the leadership of Dinamika Bangsa Foundation and the Rector of Universitas Dinamika Bangsa for funding this research. Their institutional commitment played a key role in facilitating the completion of this study. We also thank the Laboratory of Artificial Intelligence at Universitas Dinamika Bangsa for providing the computing facilities and technical assistance during the experimentation process.

FUNDING INFORMATION

This research was funded by the internal research grant scheme of the Dinamika Bangsa Foundation through Universitas Dinamika Bangsa, Jambi. The support covered computational resources, publication preparation, and research-related activities. No external funding agency or grant number is associated with this research.

AUTHOR CONTRIBUTIONS STATEMENT

All authors contributed to the conception and design of the study. They were collectively involved in the development of methodology, implementation of the experiments, analysis and interpretation of the results, as well as the writing and revision of the manuscript. Specific contributions of each author are detailed below according to the CRediT taxonomy.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dodo Zaenal Abidin	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Jasmir		✓				✓		✓	✓	✓	✓	✓		
Errisya Rasywir	✓		✓	✓			✓			✓	✓		✓	✓
Agus Siswanto					✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this research. No financial, personal, or professional relationships exist that could have appeared to influence the findings, methodology, or conclusions presented in this paper. The study was conducted independently under the internal research funding scheme of Universitas Dinamika Bangsa.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/94/spambase, DOI: 10.24432/C53G6X. The dataset is publicly accessible and requires no special permission for academic use.

REFERENCES

[1] L. He, X. Wang, H. Chen, and G. Xu, “Online spam review detection: a survey of literature,” *Human-Centric Intelligent Systems*, vol. 2, no. 1–2, pp. 14–30, May 2022, doi: 10.1007/s44230-022-00001-3.

[2] S. Rao, A. K. Verma, and T. Bhatia, “A review on social spam detection: challenges, open issues, and future directions,” *Expert Systems with Applications*, vol. 186, p. 115742, Dec. 2021, doi: 10.1016/j.eswa.2021.115742.

[3] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, “A review of spam email detection: analysis of spammer strategies and the dataset shift problem,” *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1145–1173, May 2023, doi: 10.1007/s10462-022-10195-4.

[4] Y. Kontsewaya, E. Antonov, and A. Artamonov, “Evaluating the effectiveness of machine learning methods for spam detection,” *Procedia Computer Science*, vol. 190, pp. 479–486, 2021, doi: 10.1016/j.procs.2021.06.056.

[5] V. W. de Vargas, J. A. S. Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, “Imbalanced data preprocessing techniques for machine learning: a systematic mapping study,” *Knowledge and Information Systems*, vol. 65, no. 1, pp. 31–57, Nov. 2023, doi: 10.1007/s10115-022-01772-8.

[6] S. Lusito, A. Pugnana, and R. Guidotti, “Solving imbalanced learning with outlier detection and features reduction,” *Machine Learning*, vol. 113, no. 8, pp. 5273–5330, Dec. 2024, doi: 10.1007/s10994-023-06448-0.

[7] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Systems*, vol. 166, pp. 198–206, Feb. 2019, doi: 10.1016/j.knosys.2018.12.026.

[8] M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, "Ensemble machine learning model for classification of spam product reviews," *Complexity*, vol. 2020, pp. 1–10, Dec. 2020, doi: 10.1155/2020/8857570.

[9] J. K. Afriyie *et al.*, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions," *Decision Analytics Journal*, vol. 6, p. 100163, Mar. 2023, doi: 10.1016/j.dajour.2023.100163.

[10] N. Pérez-Díaz, D. Ruano-Ordás, F. Fdez-Riverola, and J. R. Méndez, "Boosting accuracy of classical machine learning antispam classifiers in real scenarios by applying rough set theory," *Scientific Programming*, vol. 2016, pp. 1–10, 2016, doi: 10.1155/2016/5945192.

[11] O. Alghushairy *et al.*, "An efficient support vector machine algorithm based network outlier detection system," *IEEE Access*, vol. 12, pp. 24428–24441, 2024, doi: 10.1109/ACCESS.2024.3364400.

[12] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, Jul. 2020, doi: 10.1186/s40537-020-00327-4.

[13] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt, "Spambase-classifying email as spam or non-spam." UCI Machine Learning Repository, 1999, [Online]. Available: https://doi.org/10.24432/C53G6X.

[14] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[15] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance comparison of anti-spam technology using confusion matrix classification," *IOP Conference Series: Materials Science and Engineering*, vol. 879, no. 1, p. 12076, Jul. 2020, doi: 10.1088/1757-899X/879/1/012076.

[16] J. Muschelli, "ROC and AUC with a binary predictor: a potentially misleading metric," *Journal of Classification*, vol. 37, no. 3, pp. 696–708, Dec. 2020, doi: 10.1007/s00357-019-09345-1.

[17] Y. Guo, Z. Mustafaoglu, and D. Koundal, "Spam detection using bidirectional transformers and machine learning classifier algorithms," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 1, pp. 5–9, Apr. 2023, doi: 10.47852/bonviewJCCE2202192.

[18] C. Frederickson, M. Moore, G. Dawson, and R. Polikar, "Attack strength vs. detectability dilemma in adversarial machine learning," in 2018 International Joint Conference on Neural Networks (IJCNN), Jul. 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489495.

[19] K. Kotipalli and S. Suthaharan, "Modeling of class imbalance using an empirical approach with spambase dataset and random forest classification," in *RIIT 2014 - Proceedings of the 3rd Annual Conference on Research in Information Technology*, Oct. 2014, pp. 75–80, doi: 10.1145/2656434.2656442.

[20] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727–4735, Oct. 2019, doi: 10.35940/ijitee.L3591.1081219.

[21] T. Purwoningsih, H. B. Santoso, and Z. A. Hasibuan, "Online learners' behaviors detection using exploratory data analysis and machine learning approach," in *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, Oct. 2019, pp. 1–8, doi: 10.1109/ICIC47613.2019.8985918.

[22] B. Aruna Kumara and M. M. Kodabagi, "Efficient data preprocessing approach for imbalanced data in email classification system," in *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, Oct. 2020, pp. 338–341, doi: 10.1109/ICSTCEE49637.2020.9277221.

[23] A. Sharma, P. K. Singh, and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30655–30665, 2022, doi: 10.1109/ACCESS.2022.3158977.

[24] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, doi: 10.1016/j.jksuci.2021.01.014.

[25] L. Guangjun, S. Nazir, H. U. Khan, and A. U. Haq, "Spam detection approach for secure mobile message communication using machine learning algorithms," *Security and Communication Networks*, vol. 2020, pp. 1–6, Jul. 2020, doi: 10.1155/2020/8873639.

[26] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[27] A. J. Bowers and X. Zhou, "Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes," *Journal of Education for Students Placed at Risk*, vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

[28] Z. Ilhan Taskin, K. Yildirak, and C. H. Aladag, "An enhanced random forest approach using CoClust clustering: MIMIC-III and SMS spam collection application," *Journal of Big Data*, vol. 10, no. 1, Mar. 2023, doi: 10.1186/s40537-023-00720-9.

[29] L. Wang, G. Zeng, and B. Huang, "Naive Bayesian algorithm for spam classification based on random forest method," *Journal of Physics: Conference Series*, vol. 1486, no. 3, p. 32021, Apr. 2020, doi: 10.1088/1742-6596/1486/3/032021.

[30] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Aug. 2020, doi: 10.1007/s10462-020-09896-5.

[31] N. Cahyana, S. Khomsah, and A. S. Aribowo, "Improving imbalanced dataset classification using oversampling and gradient boosting," in *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, Oct. 2019, pp. 217–222, doi: 10.1109/ICSITech46713.2019.8987499.

[32] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, no. 1, Sep. 2020, doi: 10.1186/s40537-020-00349-y.

[33] J. Jasmir, D. Z. Abidin, F. Fachruddin, and W. Riyadi, "Experimental of information gain and AdaBoost feature for machine learning classifier in media social data," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 36, no. 2, pp. 1172–1181, Nov. 2024, doi: 10.11591/ijeecs.v36.i2.pp1172-1181.

[34] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.

[35] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognition Letters*, vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.

BIOGRAPHIES OF AUTHORS

Dodo Zaenal Abidin is senior lecture at Universitas Dinamika Bangsa Jambi, Indonesia. He received his Bachelor in information systems in 2001 from Universitas Nurdin Hamzah Jambi, Indonesia. and Master degree in information technology in 2008 from Universitas Putra Indonesia YPTK Padang, Indonesia. He receives a Doctor in informatics engineering at Universitas Sriwijaya Palembang, Indonesia in 2022. His research interest is data mining, machine learning, and deep learning for natural language processing and its application. He can be contacted at email: dodozaenalabidin@gmail.com.

Jasmir is senior lecture at Universitas Dinamika Bangsa Jambi, Indonesia. He received his Bachelor in computer engineering in 1995 and Master degree in information technology in 2006 from Universitas Putra Indonesia YPTK Padang, Indonesia. He receives a Doctor in informatics engineering at Universitas Sriwijaya Palembang, Indonesia in 2022. His research interest is data mining, machine learning, and deep learning for natural language processing and its application. He can be contacted at email: ijay_jasmir@yahoo.com.

Errissya Rasywir received the Bachelor degree (S.Kom) in computer science from the Sriwijaya University. She received the Master degree (M.T) in informatics Master STEI from the Institut Teknologi Bandung (ITB). In 2024, she started her Doctoral Program in Sriwijaya University in Biomedic Computer Science Filed. She is a Lecture of computer science in the informatics engineering, Dinamika Bangsa University (UNAMA). In addition, she is serving as Head of the research group (LPPM) on UNAMA. Her research interests are in data mining, artificial intelligent (AI), natural language processing (NLP), machine learning, and deep learning. She can be contacted at email: errissya.rasywir@gmail.com.

Agus Siswanto is senior lecture at Universitas Dinamika Bangsa Jambi, Indonesia. He received his Bachelor in 2008 and Master computer in 2011 from Universitas Bina Dharma Palembang, Indonesia. His research interest is machine learning and deep learning for natural language processing and its application. He can be contacted at email: wantodbz@gmail.com.