

# Enhancing the Accuracy of Airline Review Classification Using SMOTE and Grid Search with Cross Validation for Hyperparameter Tuning

Ekka Pujo Ariesanto Akhmad<sup>1,2\*</sup>, Kusworo Adi<sup>3</sup>, Aris Puji Widodo<sup>4</sup>

<sup>1</sup> Doctoral Program of Information System, School of Postgraduated Studies, Diponegoro University, Semarang 50241, Indonesia

<sup>2</sup> Seafaring Vocational Faculty, Universitas Hang Tuah, Surabaya 60111, Indonesia

<sup>3</sup> Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia

<sup>4</sup> Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia

\*Corresponding Author Email: [eka.pujo@hangtuah.ac.id](mailto:eka.pujo@hangtuah.ac.id)

## ARTICLE INFO

## ABSTRACT

Received: 26 Dec 2024

Revised: 12 Feb 2025

Accepted: 22 Feb 2025

This research aims to improve the classification accuracy of online airline reviews using SMOTE (Synthetic Minority Over-sampling Technique) and Grid Search with Cross Validation (CV) for hyperparameter tuning. The dataset includes 23,170 airline reviews processed through preprocessing, feature extraction, and data sharing stages. Ensemble models such as, LightGBM, XGBoost, GBM, and Random Forest were applied. SMOTE handles data imbalance, while Grid Search improves model performance. Results show that the optimized LightGBM achieved the highest accuracy of 99.10%, surpassing other models in precision, recall, and F1-score. This research provides important insights for airlines in understanding customer satisfaction and improving data-driven services.

**Keywords:** airline reviews, cross validation, grid search, lightgbm, SMOTE

## INTRODUCTION

The airline industry is one sector that relies heavily on customer feedback to improve services and flight experiences. In today's digital age, customer reviews have become an invaluable source of information for airlines to understand their customers' needs, preferences and satisfaction [1]. Customer reviews not only provide insights into the quality of services and facilities but also reflect the general perception of the airline brand [2].

However, processing and analyzing large volumes of customer reviews is a challenge for airlines [3]. This task is not only time- and resource-consuming but also requires a deep understanding of natural language and sentiment analysis [4]. In this context, the use of machine learning techniques to automatically classify customer reviews is becoming increasingly important [5]. This research aims to develop an efficient approach to classifying airline customer reviews using the LightGBM algorithm [6]. The main objective is to identify the positive, negative, or neutral sentiment of the reviews [7]. As such, this research not only covers the technical aspects of developing a classification model, but also contributes to further understanding of customer perceptions and preferences towards airlines[8], [9].

The primary goals of this research include:

1. Using customer experience metrics to understand customer loyalty and how satisfied customers are with their services and customer experience.
2. Create a classification model for airline customer reviews using ensemble learning boosting and random forest that is both accurate and efficient.
3. To increase the accuracy of airline review recommendation classification by modifying hyperparameters with the Grid Search CV and SMOTE approach.

4. Comparing the classification results with ensemble learning boosting and random forest classification approaches to evaluate the superiority of the LightGBM model.

This study consists of five sections. Section 1 is the introduction that provides the background, research problems, research objectives, research contributions, and research structure. Section 2 discusses the literature review related to review classification and the LightGBM algorithm. Section 3 explains the methodology used in this research. Section 4 presents the experimental results and analysis. Finally, Section 5 offers conclusions, practical implications, and suggestions for future research.

### **Reviews and Recommendations from Customers**

One way to quantify client satisfaction is through ratings. Enhancing the consumer experience and performance metrics both depend on it. Businesses use ratings as an indicator of the chance that buyers will use their products. One of the most trustworthy indicators of a company's potential for expansion is its Net Promoter Score (NPS), where the difference between promoters and critics is the Net Promoter Score [10]. The idea that NPS is a good measure of how well promotional efforts are working is backed by numerous other researchers [11]. By providing a satisfactory purchasing experience, one can sustain customers' interest in a product and encourage repeat purchases [12]. Numerous scholars are interested in comprehending consumer purchasing behavior due to the significance of client happiness. The quantity and quality of a product influence a consumer's decision to purchase it. Customers will be more inclined to buy the product if it has more positive evaluations [13].

### **Classification of Customer Reviews**

Classifying customer reviews involves organizing them according to particular labels or categories, such as sentiment (positive, negative, or neutral). In the context of airlines, customer review classification allows airlines to understand customers' opinions on service, comfort, safety, and other factors that affect the flight experience [14], [15], [16]. Customer review classification methods can involve machine learning techniques such as text classification and sentiment analysis [17], [18].

### **Ensemble Learning Methods**

Ensemble learning methods, such as stacking, boosting, and bagging, combine multiple models to enhance predictive accuracy, robustness, and generalizability. Stacking involves training a meta-model on the predictions of base models, boosting sequentially adjusts model weights to focus on errors, and bagging reduces variance by averaging predictions from multiple models trained on different subsets of data. In a variety of applications, these methods have proven to perform better than single models, including medical imaging, fraud detection, and autonomous decision-making, while also presenting challenges like high computational demands and potential overfitting [19].

Boosting, in particular, is an effective machine-learning approach that combines many models to improve generalization and prediction accuracy. Boosting works by sequentially training models, where each new model attempts to correct the errors of its predecessor. This method has been effectively utilized in a variety of disciplines, demonstrating its adaptability and efficacy.

While boosting is a powerful tool, it is essential to consider its computational cost and complexity, especially in large-scale applications. Additionally, the choice of base learners and the tuning of hyperparameters are critical to achieving optimal performance. Despite these challenges, the adaptability and effectiveness of boosting make it a valuable technique in diverse fields. This research uses three ensemble boosting methods consisting of Gradient Boosting Machine, LightGBM, XGBoost, and Random Forest.

## **LITERATURE REVIEW**

Using a variety of machine learning techniques, several prior studies have been carried out in the subject of customer review classification. While some research has suggested deep learning-based techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), others have employed more conventional techniques like Naive Bayes, Support Vector Machines (SVM), and Decision Trees. However, there is still a gap in the literature regarding the use of LightGBM for airline review classification [20].

Some commonly used traditional classification methods for classifying customer reviews include Naive Bayes, SVM, and Decision Trees. Naive Bayes is a simple probabilistic classification method based on Bayes' theorem. SVM is a classification method that constructs decision boundaries to separate different classes in the feature space. Decision Trees are a classification method that uses a series of rule-based decisions to classify instances based on their features. Although these methods have been used extensively in customer review classification, they may have limitations in coping with large and complex datasets [21].

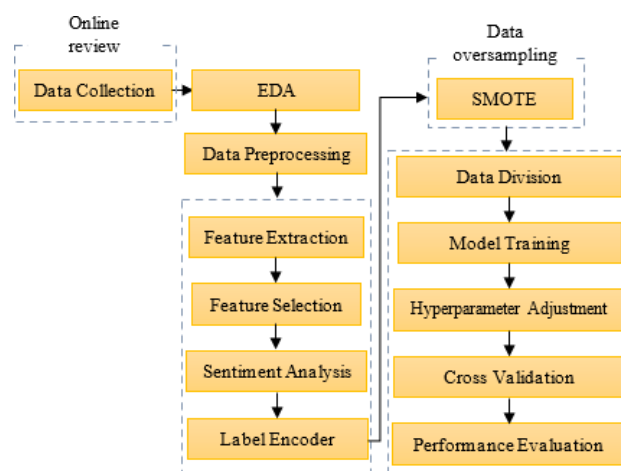
Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), two deep learning-based techniques, have demonstrated strong performance in the classification of complicated text, including customer reviews. CNNs are used to extract spatial features from text, while RNNs are suitable for modeling data sequences, such as the order of words in a review. However, training deep learning models requires large computational resources and large datasets, which may not always be available in the context of airline review classification [22], [23], [24], [25], [26], [27], [28], [29].

Optimizing a LightGBM model using grid search involves systematically exploring a predefined set of hyperparameter values to identify the best configuration for model performance. Traditional grid search can suffer from inefficiencies due to the low effective dimensionality problem, where only a subset of hyperparameters significantly impacts performance. Modified grid search methods, such as those ensuring equidistant projections or incorporating randomness, can improve efficiency and performance, sometimes outperforming classical grid search and even random search in certain contexts [30]. LightGBM, a gradient boosting framework, benefits from hyperparameter tuning to enhance its performance. In the context of automated theorem proving, grid search has been used to optimize decision tree models within LightGBM, favoring it over other frameworks like XGBoost due to its speed and stability on large datasets [31]. Additionally, LightGBM's robustness can be further enhanced by integrating topological data analysis, which improves classification accuracy in noisy environments. This strategy, while not directly connected to grid search, illustrates the need of complete feature engineering combined with hyperparameter tuning [32].

Studies show that using SMOTE (Synthetic Minority Over-sampling Technique) can increase the performance accuracy of machine learning models. For instance, a Random Forest model achieved an accuracy of 97.56% with SMOTE, compared to 92.21% without it [33]. Various algorithms, including KNN, Decision Trees, and Random Forest, were evaluated, demonstrating that SMOTE effectively mitigates bias in data, leading to better classification metrics such as precision and recall.

## METHOD

The methodology proposed in this study consists of several main stages, as depicted in Fig. 1. This research uses SMOTE to handle data imbalance and grid search with cross-validation for hyperparameter optimization. The airline review dataset is processed through preprocessing, feature extraction, and data-sharing stages. Ensemble models such as LightGBM, XGBoost, GBM, and Random Forest are used to evaluate performance.



**Fig. 1** The process flow diagram

This dataset contains 23,170 airline reviews from April 2005 to July 2023 for 417 popular airlines, collected using multiple-choice and free-text questions. Table 1 describes the dataset used in this research. The online reviews dataset consists of both textual and categorical data.

The Net Promoter Score (NPS) is a popular metric in the airline business for assessing customer loyalty and satisfaction. NPS is calculated using the following formula.

$$\text{NPS} = \% \text{ Promoters} - \% \text{ Detractors} \quad (1)$$

The scale ranges from 0 (unlikely) to 10 (very likely). Customers are divided into three categories based on their responses: Customers who give a score of 9 or 10 are classified as promoters. These clients are usually loyal and passionate. Customers who give a score of 7 or 8 are considered passive. These consumers are happy with the company's services but not enthusiastic enough to recommend it. Detractors rate their responses on a scale of 0 to 6. They are dissatisfied consumers who are unlikely to buy from the company again and may even discourage others from doing so.

The collected review data will then go through a series of preprocessing steps to clean and normalize the text. These steps include punctuation removal, tokenization, removal of conjunctions and stop words, and stemming or lemmatization to convert words into their base form. This preprocessing is necessary to prepare the text data for feature extraction and training using the ensemble learning algorithm.

Feature extraction is the process of generating new features from existing data. This research focuses on feature extraction, namely rewriting column names, deleting irrelevant features from the dataset, and handling missing values. A few Not a Number (NaN) values can be seen in the dataset preview. The customer review field frequently begins with a distinct character and a default language, followed by content identical to the route field. This redundant information can be removed from the customer review field, as well as any rows with NaN values, because they add no value. To account for the NaN values, we replace the medians in those columns.

**Table 1.** Dataset description for online airline reviews

| Feature         | Description                                                                                                 |
|-----------------|-------------------------------------------------------------------------------------------------------------|
| Airline         | the traveler's preferred airline name.                                                                      |
| Overall         | how satisfied was the passenger with the overall level of service during the flight, on a scale of 1 to 10? |
| Author          | the traveler's name who is writing reviews.                                                                 |
| Review_date     | the month and year of submission of the review.                                                             |
| Customer_review | customers' textual reviews of the airline.                                                                  |
| Aircraft        | types of airplane.                                                                                          |
| Traveller_type  | passenger types consist of solo/couple/family leisure, or business.                                         |
| Cabin           | a portion of an aircraft that passengers used (economy, premium economy, first class, business class).      |
| Route           | the route found in the review text.                                                                         |
| Date_flown      | the flight occurred in the month and year.                                                                  |
| Seat_comfort    | seat comfort is rated from 1 to 5.                                                                          |
| Cabin_service   | cabin staff members are rated from 1 to 5.                                                                  |
| Food_beve       | food and beverage quality is rated on a scale of 1 to 5.                                                    |
| Entertainment   | how satisfied was the traveler with the entertainment offered during the flight, on a scale of 1 to 5?      |
| Ground_service  | ground service is rated on a scale of 1 to 5.                                                               |
| Value_for_money | how satisfied was the traveler with the amount paid for the flight, on a scale of 1 to 5?                   |
| Recommended     | value mapping no to false and yes to true.                                                                  |

Feature extraction is a fundamental technique in machine learning and data analysis that aims to transform raw data into a set of features that may efficiently be used for modeling.

This process is essential for improving the performance and accuracy of machine learning models by reducing dimensionality and enhancing data quality. After preprocessing, important features will be extracted from the review text using natural language processing techniques.

VADER (Valence Aware Dictionary and Sentiment Reasoner) was used to analyze customer reviews from the Airline Quality website, classifying them into positive, negative, or neutral sentiments [34]. VADER is based on a lexicon evaluated by multiple human judges through a structured and methodical process [35]. Each word in the lexicon is assigned a sentiment valence depending on its polarity and intensities. Polarity in the text reflects whether its sentiment is positive or negative. One of the initial phases in feature engineering is to narrow down the features that are most relevant and necessary for analysis. This reduces the data's dimensionality and prevents overfitting. This research uses statistical metrics such as correlation to select the most influential features. A label encoder is used in this research to convert categorical and string data into numeric data that can be easily understood by the model.

SMOTE is used to balance the dataset and ensure that minority classes are properly represented. SMOTE's success is measured using metrics including accuracy, precision, recall, and F1-score, which combined imply improved model performance.

The processed review data will be divided into training set and testing set. The training and testing sets of data comprise the full set; 30% of the total data is used for testing, and the remaining 70% is used for training. It is important to ensure that this split reflects the true distribution of reviews to avoid bias in the evaluation of the model's performance.

A model is developed, and three boosting algorithms and Random Forest are considered for the training and evaluation of the model. The algorithms that are used are Gradient Boosting, Extreme Gradient Boosting (XGBoost), LightGBM, and Random Forest.

Gradient Boosting Machine Algorithm. It has been demonstrated that using sentiment representation rather than raw text data in GBM models improves classification accuracy. This technique simplifies the data processing pipeline and maintains excellent accuracy, making it a feasible alternative for real-world applications [36].

LightGBM Algorithm. Microsoft created the machine learning algorithm LightGBM, which is based on decision trees. One of the main advantages of LightGBM is its speed and efficiency in dealing with large datasets. The algorithm uses a leaf-wise approach in building the decision tree, which allows for increased model training speed [37]. LightGBM also has the ability to handle categorical features directly, without the need to convert them into numerical representations [38], [39], [40].

A machine learning approach based on decision trees is called gradient boosting. Decision trees are implemented using LightGBM, where the tree grows leaf-wise, splitting only one leaf each time. LightGBM excels in handling bigger datasets. It is possible to prevent LightGBM's sensitivity and overfitting of small datasets by restricting the tree's depth. The accuracy is the primary focus of the LightGBM algorithm [41].

The main advantage of LightGBM in airline review classification is its speed and efficiency in processing large datasets. It is able to tackle multi-class problems and handle categorical features well. In addition, LightGBM has the ability to handle imbalanced data and minimize overfitting. Thus, the use of LightGBM in airline review classification can produce accurate and efficient models [42].

LightGBM offers a number of advantages in airline review classification. Firstly, LightGBM can cope with large and complex datasets quickly and efficiently. Secondly, the leaf-splitting algorithm used by LightGBM makes it possible to build more complex and accurate models in a relatively short time. Thirdly, LightGBM is able to handle categorical features without the need to convert them into numerical representations, which is often required by other classification methods. These advantages make LightGBM an attractive option for airline review classification.

Extreme Gradient Boosting (XGBoost). Algorithm One boosting mechanism is the XGBoost algorithm. It's a well-known algorithm for supervised machine learning (ML). XGBoost implements the gradient-boosted trees technique in an effective manner. It can be used to perform both regression and classification on large datasets. XGBoost employs sequentially generated decision trees to produce trustworthy findings while preventing overfitting [43].

Random Forest Algorithm. The well-known machine learning algorithm Random Forest uses supervised learning. It is possible to solve regression and classification issues with the Random Forest technique. There are several decision trees in it. The Random Forest classifier, which does not rely on a single decision tree, takes predictions from each tree and uses the forecasts of the majority of the trees to determine the final output [44].

The model will be trained on the previously processed training data. The number of iterations, tree depth, and other model parameters will all be adjusted during the training process. The purpose of training is to produce an ideal model capable of accurately and efficiently classifying reviews.

Hyperparameters are essential for regulating the behavior of machine learning algorithms, and the ideal design can result in significant performance increases [45].

GridSearchCV's hyperparameter tuning architecture aims to improve the performance of machine learning models through parameter optimization. The procedure begins by importing the processed dataset and dividing it into two sections: a training set for model training and a testing set for final evaluation. The model to be utilized is defined, as well as a parameter search space with possible values for each hyperparameter. GridSearchCV is then used with cross-validation, which divides the data into folds. Each parameter combination in the grid is tested on different data folds to determine its performance using specific metrics like accuracy, precision, recall, or F1 score.

This procedure ensures that the ideal parameter combination is chosen based on the average performance of the validation data, lowering the danger of overfitting. After determining the best combination, the model is retrained on all training data using those parameters. The optimized model is then run on the test data to see how it performs on the new data. The final results contain the optimal hyperparameter values and the generated model performance. Figure 2 shows the hyperparameter tuning flowchart. The number of folds is set for cross-validation. Split the data into k subsets: One for validation and k-1 for training.

Test data will be used to evaluate the model's accuracy, precision, recall, and F1-score. These metrics will indicate how accurately the model can classify customer reviews. Equation (2) can be used to describe accuracy, which is the percentage of correct predictions divided by the total number of samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (2)$$

Recall is defined as the proportion of expected data in a particular class. This is theoretically expressed as equation (3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

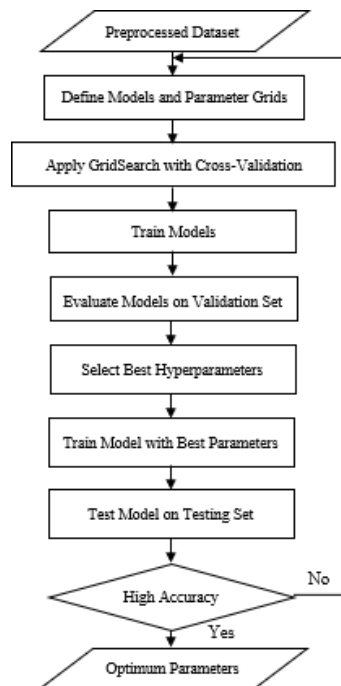
Equation (4) calculates precision, which is the capacity to reliably detect instances of authentic events.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

The F1-Score calculates the harmonic mean of precision and recall, revealing information regarding testing accuracy. Equation (5) is a mathematical representation.

$$\text{F1\_Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

TP, FN, TN, and FP represent true positives, false negatives, true negatives, and false positives, respectively (Muzakir et al., 2023).



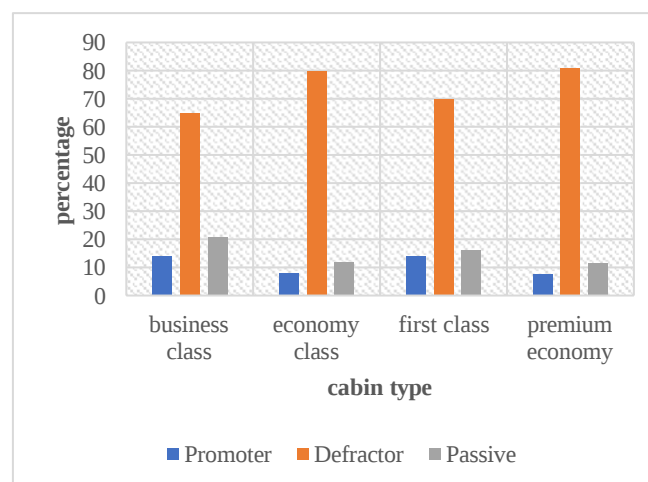
**Fig. 2** Hyperparameter tuning flow chart with gridsearch

## RESULT AND DISCUSSION

This study used four test scenarios to gather results. The first goal is to understand consumer loyalty and how delighted customers are with the company's offerings and customer experience using the NPS. Second, GridSearchCV is used to optimize hyperparameter tuning. Third, since the number of true recommendations is 5,449, which is not equal to the number of false recommendations, which is 10,304, the effect of resampling data using oversampling techniques will be evaluated. Fourth, the classification results will be compared to those of the ensemble learning boosting and random forest classification algorithms. The results and discussion are presented in the following sections.

### Results

Figure 3 describes the percentage of the overall column value used to calculate NPS by passenger cabin. Equation 1 is used to calculate NPS for business class:  $14\% - 65\% = -51\%$ ; NPS for economy class:  $8\% - 80\% = -72\%$ ; NPS for first class:  $14\% - 70\% = -56\%$ ; NPS for premium economy class:  $7.5\% - 81\% = -73.5\%$ .



**Fig. 3** Percentage of promoters, defractors, and passives

Table 2 contains the results of the hyperparameter tuning process using GridSearchCV for the ensemble learning, boosting and random forest algorithms. Each algorithm has a specific set of parameters that are optimized to improve the model's performance in airline customer review classification. Below are the details of the parameters and the best values obtained.

#### Random Forest:

- **max\_depth:** Determines the maximum depth of each tree in the forest. The greatest value discovered was 8, which reduces overfitting by restricting the complexity of trees.
- **max\_features:** The maximum number of features considered for sharing in each node. The optimal value is 8, which maintains a balance between variation and accuracy.
- **min\_samples\_split:** The minimum number of samples required to split internal nodes. The best value is 5, ensuring the tree is not too fragmented.
- **n\_estimators:** The number of trees in the forest. The optimal value is 100, providing a balance between training time and accuracy.

#### XGBoost (Extreme Gradient Boosting):

- **max\_depth:** Each tree's maximum depth. The optimal value is 5, which reduces the likelihood of overfitting.
- **learning\_rate:** The model update rate at each iteration. The optimal value is 0.01, providing small but steady updates for better training.
- **colsample\_bytree:** The proportion of randomly selected features for training each tree. The optimal value is 0.7, reducing the risk of overfitting.
- **n\_estimators:** The total number of trees in the model. The best value is 500, resulting in a more robust model.

#### Gradient Boosting Machine (GBM):

- **learning\_rate:** The learning rate at each iteration. The best value is 0.08, providing stable model updates.
- **max\_depth:** Maximum depth of the tree. The optimal value is 3, reducing model complexity.
- **n\_estimators:** Number of trees in the model. The best value is 50, enough to capture patterns without overfitting.
- **subsample:** The proportion of samples used to train each tree. The optimal value is 0.5, helping to reduce overfitting.
- **min\_samples\_split:** The minimum number of samples to split the nodes. The optimal value is 15, providing a balance between too frequent and infrequent splits.
- **min\_samples\_leaf:** The final leaf has a minimal number of samples. The optimal number is 17, which allows the model to capture key patterns without overfitting.
- **max\_features:** The maximum number of features considered for node division. The optimal value is 13, maintaining the diversity of the model.

#### LightGBM:

- **boosting\_type:** The type of boosting used. The optimal value is gbdt (Gradient Boosting Decision Tree), due to its high efficiency and accuracy.
- **learning\_rate:** The learning rate of the model. The best value is 0.1, providing a balance between convergence speed and model stability.
- **colsample\_bytree:** The proportion of features used for training each tree. The optimal value is 1, indicating the use of all features.
- **max\_depth:** Each tree's maximum depth. The optimal value is 10, which provides enough complexity to capture patterns while avoiding overfitting.
- **n\_estimators:** Total number of trees in the model. The optimal value is 500, providing strong predictive ability.

The third test aims to investigate the influence of data resampling on the classification of suggestions from online airline passenger reviews using boosting and random forest ensemble learning algorithms optimized with GridSearchCV. Table 3 shows the accuracy results obtained by resampling the data.

According to the results in Table 3, LightGBM with GridSearchCV + SMOTE has a greater accuracy after resampling, reaching 99.10% compared to 99.05% before resampling.

A detailed comparison of the no-tuning classifier and the proposed system, which uses the GridSearchCV method for parameter tuning and SMOTE, is depicted in Table 4.

**Table 2.** Best hyperparameters of ensemble learning boosting and random forest with gridsearchCV

| Model         | Hyperparameter    | Hyperparameter Value   | Best Hyperparameter |
|---------------|-------------------|------------------------|---------------------|
| Random Forest | max_depth         | [5,8]                  | 8                   |
|               | max_features      | [4,8,"auto"]           | 8                   |
|               | min_samples_split | [2,5,8]                | 5                   |
|               | n_estimators      | [100,200,500]          | 100                 |
| XGBoost       | max_depth         | [5,6]                  | 5                   |
|               | learning_rate     | [0.1, 0.01]            | 0.01                |
|               | colsample_bytree  | [0.7, 1]               | 0.7                 |
|               | n_estimators      | [100,200,500]          | 500                 |
| GBM           | learning_rate     | [0.07, 0.08]           | 0.08                |
|               | max_depth         | [1, 2, 3]              | 3                   |
|               | n_estimators      | [10,20,30,40,50]       | 50                  |
|               | subsample         | [0.5, 0.6]             | 0.5                 |
| LightGBM      | min_sample_split  | range[12,16]           | 15                  |
|               | min_sample_leaf   | Range[14,19]           | 17                  |
|               | max_features      | [7,10,13]              | 13                  |
|               | boosting_type     | ['gbdt','dart','goss'] | gbdt                |
|               | learning_rate     | [0.01, 0.1, 1]         | 0.1                 |
|               | colsample_bytree  | [0.7, 1]               | 1                   |
|               | max_depth         | [5,6,7,8,9,10]         | 10                  |
|               | n_estimators      | [100,200,500]          | 500                 |

**Table 3.** Results accuracy using data resampling

| Model            | GridSearchCV before resampling | GridSearchCV after resampling |
|------------------|--------------------------------|-------------------------------|
| Random Forest    | 0.9628                         | 0.9634                        |
| XGBoost          | 0.9635                         | 0.9647                        |
| GradientBoosting | 0.9620                         | 0.9617                        |
| LightGBM         | 0.9905                         | 0.9910                        |

**Table 4.** Classification report comparison

| Study                                       | Model            | Accuracy | Precision | Recall | F1-score |
|---------------------------------------------|------------------|----------|-----------|--------|----------|
| no tuning<br>(Murugesan et al.,<br>2024)    | Random Forest    | 0.9480   | 0.9500    | 0.9410 | 0.9450   |
|                                             | XGBoost          | 0.9480   | 0.9460    | 0.9450 | 0.9460   |
|                                             | GradientBoosting | 0.9480   | 0.9460    | 0.9450 | 0.9460   |
|                                             | LightGBM         | 0.9700   | 0.9700    | 0.9600 | 0.9600   |
| Proposed model +<br>GridSearchCV +<br>SMOTE | Random Forest    | 0.9634   | 0.9368    | 0.9565 | 0.9465   |
|                                             | XGBoost          | 0.9647   | 0.9388    | 0.9582 | 0.9484   |
|                                             | GradientBoosting | 0.9617   | 0.9362    | 0.9522 | 0.9441   |
|                                             | LightGBM         | 0.9910   | 0.9818    | 0.9922 | 0.9870   |

## DISCUSSION

The current work provides a detailed evaluation of machine learning methodologies for classifying airline reviews, with a focus on the effectiveness of hyperparameter tuning using GridSearchCV and resampling techniques. The use of ensemble approaches such as LightGBM, XGBoost, GBM, and Random Forest leads to considerable improvements in model performance, especially when used to a highly imbalanced dataset.

The adoption of LightGBM as the primary model demonstrates its ability to handle huge, complicated datasets. The leaf-wise tree growth algorithm built into LightGBM adds to its quick training time and ability to detect intricate patterns in data. Table 3 shows that the suggested technique, which includes GridSearchCV and SMOTE, outperformed baseline models in terms of accuracy, precision, recall, and F1-score.

One important finding is the effect of SMOTE on class imbalance. The post-resampling accuracy of LightGBM increased somewhat from 99.05% to 99.10%, demonstrating the importance of resolving data imbalance. This improvement, while minor, shows the model's increased ability to generalize predictions for underrepresented classes.

Table 4 shows that, while all models benefited from hyperparameter modification, LightGBM consistently outperformed them. Random Forest and XGBoost, despite their robust ensemble frameworks, lagged slightly because to their greater sensitivity to overfitting in imbalanced situations. The Gradient Boosting Machine produced competitive results, but it required more processing resources for marginal benefits.

These findings are consistent with previous research [42], which supports for the adoption of gradient-boosted models in large-scale, text-rich datasets. The addition of SMOTE is consistent with studies that show enhanced classification metrics in imbalanced situations [46].

The implications of this study extend beyond methodological enhancements. By accurately classifying customer reviews, airlines can gain actionable insights into passenger satisfaction and service quality. For instance, the low Net Promoter Scores (NPS) for economy and premium economy classes (Table 2) suggest the need for targeted interventions to enhance passenger experiences in these segments. Such data-driven strategies are critical in fostering brand loyalty and improving market competitiveness.

Despite its strengths, the study is not without limitations. Although the dataset is huge, it primarily consists of reviews in English, which may limit its application to a global client base. Future study could integrate multilingual datasets and advanced natural language processing techniques like transformer-based models to improve classification accuracy. The precision value can drop after SMOTE, because SMOTE adds synthetic data to the minority class to handle the class imbalance problem. The model may misclassify samples from the majority class as part of the minority class, thus increasing False Positives (FP) which directly impacts precision.

## CONCLUSION

This study successfully demonstrates the utility of machine learning techniques, particularly ensemble learning models like LightGBM, in classifying airline customer reviews with high accuracy and efficiency. By employing GridSearchCV for hyperparameter tuning and SMOTE for addressing data imbalance, the proposed methodology achieves notable improvements in classification performance. The LightGBM model, optimized through these techniques, outperforms traditional approaches such as Random Forest, GBM, and XGBoost, achieving a classification accuracy of 99.10%.

The findings highlight the crucial impact of hyperparameter adjustment in improving model performance, as well as the importance of resolving class imbalance in datasets with unequal distributions. Furthermore, the study highlights the potential of customer feedback analysis as a strategic tool for the airline industry, offering actionable insights into passenger satisfaction and areas requiring improvement, particularly in economy and premium economy classes.

Despite its strengths, the research has certain limitations, including its focus on English-language reviews and reliance on static datasets and precision decreases because there are more false positives.

Future studies should explore multilingual datasets and use more advanced SMOTE methods such as Borderline-SMOTE or ADASYN to generate more relevant synthetic samples and real-time feedback systems to enhance model generalizability and applicability. Additionally, integrating advanced optimization methods, such as Bayesian or genetic algorithms, for hyperparameter tuning could further streamline the process and improve computational efficiency.

In conclusion, this study provides a robust framework for leveraging machine learning to analyze and classify customer reviews, contributing valuable insights to the airline industry and advancing the application of data-driven decision-making in service quality enhancement.

## REFERENCES

- [1] M. Siering, A. V Deokar, and C. Janze, "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews," *Decis Support Syst*, vol. 107, pp. 52–63, 2018.
- [2] P. K. Jain, R. Pamula, S. Ansari, D. Sharma, and L. Maddala, "Airline recommendation prediction using customer generated feedback data," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, 2019, pp. 376–379.
- [3] S. Farzadnia and I. R. Vanani, "Identification of opinion trends using sentiment analysis of airlines passengers' reviews," *J Air Transp Manag*, vol. 103, p. 102232, 2022.
- [4] X. Wang, G. Xu, Z. Zhang, L. Jin, and X. Sun, "End-to-end aspect-based sentiment analysis with hierarchical multi-task learning," *Neurocomputing*, vol. 455, pp. 178–188, 2021.
- [5] R. Nagamanjula and A. Pethalakshmi, "A machine learning based sentiment analysis by selecting features for predicting customer reviews," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2018, pp. 1837–1843.
- [6] P. K. Jain, G. Srivastava, J. C.-W. Lin, and R. Pamula, "Unscrambling customer recommendations: a novel lstm ensemble approach in airline recommendation prediction using online reviews," *IEEE Trans Comput Soc Syst*, vol. 9, no. 6, pp. 1777–1784, 2022.
- [7] S. Sazed, "Identifying neutral reviews from unlabeled data: An exploratory study on user ratings and word-level polarity scores," in *Proceedings of the 33rd ACM conference on hypertext and social media*, 2022, pp. 198–202.
- [8] C. Yan, Z. Wu, H. Gan, and Z. Zhu, "Airline Customers Strategic Decision Analysis and Formulation," 2022.
- [9] P. Kunekar, M. Deshpande, A. Gharpure, V. Gokhale, A. Gore, and H. Yadav, "Evaluating the predictive ability of the LightGBM Classifier for assessing customer satisfaction in the airline industry," in *2023 International Conference for Advancement in Technology (ICONAT)*, IEEE, 2023, pp. 1–6.
- [10] D. Sartori and J. Teodoro, "NET PROMOTER SCORE: BIBLIOMETRIC REVIEW OF THEORY AND PRACTICE.," *Consumer Behavior Review (CBR)*, vol. 8, no. 1, 2024.
- [11] J. G. Dawes, "Net promoter and revenue growth: An examination across three industries," *Australasian Marketing Journal*, vol. 32, no. 1, pp. 4–18, 2024.
- [12] J. Iristian and S. Irdiana, "Buying Experience: The Impact Of Trust And Satisfaction Of Grab Food Customers In Surabaya," *Ekspektra: Jurnal Bisnis dan Manajemen*, vol. 3, no. 2, pp. 157–164, 2019.
- [13] P. K. Jain and R. Pamula, "Content-based airline recommendation prediction using machine learning techniques," in *Machine learning algorithms for industrial applications*, Springer, 2020, pp. 185–194.
- [14] P. O. Bamgboye et al., "Text Classification on Customer Review Dataset Using Support Vector Machine," in *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2*, Springer, 2023, pp. 407–415.
- [15] E. Deniz, H. Erbay, and M. Coşar, "Multi-label classification of e-commerce customer reviews via machine learning," *Axioms*, vol. 11, no. 9, p. 436, 2022.
- [16] N. Sunil and F. Shirazi, "Customer review classification using machine learning and deep learning techniques," in *International Conference on Human-Computer Interaction*, Springer, 2023, pp. 581–597.
- [17] T. Hasan, A. Matin, and M. S. R. Joy, "Machine learning based automatic classification of customer sentiment," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2020, pp. 1–6.
- [18] B. Noori, "Classification of customer reviews using machine learning algorithms," *Applied Artificial Intelligence*, vol. 35, no. 8, pp. 567–588, 2021.
- [19] N. Rane, S. P. Choudhary, and J. Rane, "Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions," *Studies in Medical and Health Sciences*, vol. 1, no. 2, pp. 18–41, 2024.
- [20] N. Srinu, J. Paramesh, and K. Bala, "A Systematic Review on Customer Feedback Classification Employing Deep Learning Methodologies for Recommend Systems," in *2022 1st International Conference on Computational Science and Technology (ICCST)*, IEEE, 2022, pp. 408–411.
- [21] F. M. Alotaibi, "A machine-learning-inspired opinion extraction mechanism for classifying customer reviews on social media," *Applied Sciences*, vol. 13, no. 12, p. 7266, 2023.
- [22] D. Kim, "Research on text classification based on deep neural network," *International Journal of Communication Networks and Information Security*, vol. 14, no. 1s, pp. 100–113, 2022.
- [23] S. Lyu and J. Liu, "Convolutional recurrent neural networks for text classification," *Journal of Database Management (JDM)*, vol. 32, no. 4, pp. 65–82, 2021.
- [24] X. Peng, "A comparative study of neural network for text classification," in *2020 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, IEEE, 2020, pp. 214–218.
- [25] S. Tammina and S. Annareddy, "Sentiment analysis on customer reviews using convolutional neural network," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, 2020, pp. 1–6.
- [26] C. Zhang, "Text Classification Using Deep Learning Methods," in *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, IEEE, 2022, pp. 1327–1332.
- [27] Y. Zhou, "A review of text classification based on deep learning," in *Proceedings of the 2020 3rd international conference on geoinformatics and data analysis*, 2020, pp. 132–136.
- [28] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, "A comparative review on deep learning models for text classification," *Indones. J. Electr. Eng. Comput. Sci*, vol. 19, no. 1, pp. 325–335, 2020.
- [29] F. F. Rachim, A. Damayanti, and E. Winarko, "Classification of Review Text using Hybrid Convolutional Neural

- Network and Gated Recurrent Unit Methods,” *Contemporary Mathematics and Applications*, 2022.
- [30] D. López, C. M. Alaíz, and J. R. Dorronsoro, “Modified grid searches for hyper-parameter optimization,” in *Hybrid Artificial Intelligent Systems: 15th International Conference, HAIS 2020, Gijón, Spain, November 11–13, 2020, Proceedings 15*, Springer, 2020, pp. 221–232.
- [31] Zarathustra Goertzel, J. Jan, J. Mikolás, and K. Cezary, “LightGBM Hyperparameter Optimization for Clause Classification in Theorem Proving,” 2022.
- [32] H. Yang, G. Qin, Z. Liu, Y. Hu, and Q. Dai, “LightGBM robust optimization algorithm based on topological data analysis,” in *Proceedings of the 2024 International Conference on Computer and Multimedia Technology*, 2024, pp. 574–582.
- [33] E. P. A. Akhmad, K. Adi, and A. P. Widodo, “Machine learning approach to customer sentiment analysis in twitter airline reviews,” in *E3S Web of Conferences*, EDP Sciences, 2023, p. 02044.
- [34] R. Annamalai, S. A. Rasool, S. Deena, K. Venkatraman, and Y. Soundaram, “Sentiment Analysis using VADER: Unveiling Customer Sentiment and Predicting Buying Behavior in the Airline Industry,” in *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*, IEEE, 2024, pp. 277–282.
- [35] P. Reedy, “Interpol review of digital evidence for 2019–2022,” *Forensic Sci Int*, vol. 6, p. 100313, 2023.
- [36] L. Xu and X. Zhang, “Extracting key drivers of sky ratings and evaluating air passenger’s satisfaction classification model through online review analysis,” in *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)*, SPIE, 2024, pp. 94–101.
- [37] G. Ke *et al.*, “Lightgbm: A highly efficient gradient boosting decision tree,” *Adv Neural Inf Process Syst*, vol. 30, 2017.
- [38] J. Hancock and T. M. Khoshgoftaar, “Leveraging lightgbm for categorical big data,” in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, 2021, pp. 149–154.
- [39] Y. Xia, K. Cheng, Z. Cheng, Y. Rao, and J. Pu, “GBMVis: Visual analytics for interpreting gradient boosting machine,” in *Cooperative Design, Visualization, and Engineering: 18th International Conference, CDVE 2021, Virtual Event, October 24–27, 2021, Proceedings 18*, Springer, 2021, pp. 63–72.
- [40] Y. Zhou, H. Li, and M. Chen, “LGBM-CBFS: A Heuristic Feature Sampling Method Based on Tree Ensembles,” *Security and Communication Networks*, vol. 2022, no. 1, p. 5156086, 2022.
- [41] S. Xu, S. Liu, H. Wang, W. Chen, F. Zhang, and Z. Xiao, “A hyperspectral image classification approach based on feature fusion and multi-layered gradient boosting decision trees,” *Entropy*, vol. 23, no. 1, p. 20, 2020.
- [42] P. K. Jain, G. Srivastava, J. C.-W. Lin, and R. Pamula, “Unscrambling customer recommendations: a novel lstm ensemble approach in airline recommendation prediction using online reviews,” *IEEE Trans Comput Soc Syst*, vol. 9, no. 6, pp. 1777–1784, 2022.
- [43] J. Chen, F. Zhao, Y. Sun, and Y. Yin, “Improved XGBoost model based on genetic algorithm,” *International Journal of Computer Applications in Technology*, vol. 62, no. 3, pp. 240–245, 2020.
- [44] R. M. Mohana, C. K. K. Reddy, P. R. Anisha, and B. V. R. Murthy, “WITHDRAWN: Random forest algorithms for the classification of tree-based ensemble,” 2021, *Elsevier*.
- [45] L. Franceschi *et al.*, “Hyperparameter Optimization in Machine Learning,” *arXiv preprint arXiv:2410.22854*, 2024.
- [46] E. P. A. Akhmad, K. Adi, and A. P. Widodo, “Machine learning approach to customer sentiment analysis in twitter airline reviews,” in *E3S Web of Conferences*, EDP Sciences, 2023, p. 02044.