Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 13, No. 3, Juli 2026

ISSN 0216 – 0544
e-ISSN 2301– 6914

Comparative analysis of CatBoost and LightGBM algorithms for predicting dental caries risk based on lifestyle factors

Chayadi Oktomy Noto Susanto [a], Muhajirah Ulfah [b], Bimo Aditya Pangestu [c], Slamet Riyadi [d]

[a,b,c,d] Department of Information Technology, Universitas Muhammadiyah Yogyakarta
E-mail: chayadions@ft.umy.ac.id, muhajirah.ulfah.ft22@mail.umy.ac.id, bimo.aditya.ft22@mail.umy.ac.id, riyadi@umy.ac.id

Abstract

Dental caries is a common oral infection but significantly impacts life. This study aims to compare the performance of the CatBoost and LightGBM algorithms in predicting dental caries risk based on health and lifestyle data from over 55,000 South Korean NHIS participants. Imbalanced data was addressed using the SMOTE technique, while hyperparameter tuning was performed using Optuna to maximize the F1-score. The models were evaluated using the confusion matrix, accuracy, logloss, ROC Curve, AUC, Precision, Recall, and F1-Score metrics. The results showed that CatBoost was slightly superior with an accuracy of 90.4%, an F1-score of 89.7%, and an AUC of 0.95, compared to LightGBM with an accuracy of 90.1%, an F1-score of 89.2%, and an AUC of 0.94. In addition, CatBoost also produced fewer false negatives, making it more sensitive in detecting caries cases. Therefore, CatBoost is recommended as the primary model in a caries risk prediction system to support early detection and disease prevention more effectively.

Key words: CatBoost, Dental and oral health, Dental caries, LightGBM, Machine Learning.

INTRODUCTION

Dental caries is widely recognized as one of the most prevalent and longstanding diseases affecting the human population [1]. The history of dentistry dates back to 5000 BC. The term "dental caries" was first documented in the literature circa 1634, originating from the Latin word caries, signifying decay [2]. It is estimated that there are approximately 1.8 billion cases of dental caries annually, indicating that lack of dental care can lead to increased tooth loss [3]. The impact is not only on physical health, such as loss of appetite, but also affects mental health, such as work concentration and emotional stability [4, 5].

Before the development of modern Machine Learning (ML) algorithms, various approaches were used to detect dental caries, such as logistic regression and Random Forest. However, these methods have limitations in terms of accuracy and sensitivity, especially when applied to complex and imbalanced health data [6, 7]. With technological adv ances, the use of artificial intelligence (AI) and machine learning (ML) is now beginning to be implemented as solutions to improve dental health services [8, 9]. This study aims to compare two gradient boosting algorithms, namely CatBoost and LightGBM, in predicting dental caries risk. The dataset used includes health examination results. This study combines the Synthetic Minority Oversampling Technique (SMOTE) technique to balance the data and hyperparameter optimization using the Bayesian optimization approach using the optuna library [10, 11].

The CatBoost model excels in natively handling categorical features, using ordered boosting techniques to produce a more stable and accurate model, and is able to reduce overfitting [12]. Meanwhile, LightGBM uses a histogram-based tree approach and leaf-wise leaf growth, making it very fast and memory-efficient, ideal for large dental caries datasets [12].

Dental caries is one of the most common oral infections, yet it significantly impacts daily life. Therefore, Machine Learning (ML) technology is being utilized to aid in the early detection of dental caries. In developing ML-based prediction systems, various algorithms have been used, such as Random Forest and Logistic Regression, as well as gradient boosting models like CatBoost and LightGBM, which are known to excel in accuracy and efficiency [13].

CatBoost, a gradient boosting algorithm introduced by Yandex, is specifically designed to handle categorical data efficiently, obviating the need for conventional preprocessing steps such as one-hot encoding [14]. This is relevant in the medical domain, which often involves non-numerical data [15]. Furthermore, CatBoost uses a symmetric tree structure, meaning each branch grows equally. This helps maintain model stability and prevents overfitting [16]. CatBoost also implements ordered boosting, a step-by-step training method that prevents the model from relying too heavily on target data during learning [17].

LightGBM is a decision tree algorithm-based framework developed by Microsoft in 2017 [18]. LightGBM has several key features that make it superior in training speed. One of these is Gradient-based Rather than treating all data points equally, GOSS prioritizes samples with substantial gradients by keeping them entirely, while samples with minimal gradients are sampled randomly at a reduced rate. This approach helps speed up training without compromising accuracy [19, 20]. Another key component of LightGBM is Exclusive Feature Bundling (EFB), which merges mutually exclusive features into a single group, since such features seldom occur together with non-zero values. This reduction in feature dimensionality contributes to faster computation without sacrificing model performance [21, 22]. Research conducted by [23] compared the performance of four machine learning algorithms for predicting dental caries in early childhood. The results showed that LightGBM, XGBoost, Random Forest, and logistic regression were the tested models. LightGBM performed best with an Area Under the ROC Curve (AUROC) of 0.785, followed by XGBoost with an AUROC of 0.783, Random Forest with a value of 0.780, and logistic regression with a value of 0.774. This indicates that boosting-based algorithms, particularly LightGBM, have an advantage in predictive accuracy compared to conventional methods such as logistic regression, although the difference between the models is not significant.

Furthermore, compared different gradient boosting algorithms on medical diagnosis data [24]. They found that CatBoost and LightGBM consistently outperformed conventional models such as logistic regression and deep learning approaches, especially for tabular data with heterogeneous features and imbalanced data.

Used CatBoost and LightGBM to perform tabular-based medical data fraud detection. The results showed that CatBoost achieved an AUC of 0.8825, higher than LightGBM, which achieved an AUC of 0.8514, especially when used on data with complex categorical features [25]. This research shows that the CatBoost and LightGBM algorithms provide high accuracy and are capable of handling complex data. This is because both algorithms use a boosting approach, which gradually improves prediction errors with each iteration of model training [26]. Overall, both CatBoost and LightGBM are robust and efficient ML algorithms for dental caries detection. However, based on experimental results and a review of the existing literature, CatBoost is recommended as the primary choice in this case due to its higher accuracy, F1-score, and better prediction stability [10].

Apart from research comparing different algorithms, attention has also been directed toward the impact of data preprocessing techniques on machine learning model performance, particularly in the context of imbalanced medical datasets. A study examining the effect of SMOTE on tabular health data demonstrated that synthetic oversampling substantially improved recall and F1-score for minority class prediction, particularly in disease classification tasks where positive cases are underrepresented

relative to negative cases [26]. Hyperparameter optimization has likewise received considerable attention, with techniques such as grid search, random search, and Bayesian optimization commonly employed to fine-tune gradient boosting models. Research indicates that Bayesian-based tools like Optuna are generally more efficient at identifying optimal parameter settings than exhaustive search methods, particularly in settings with limited computational resources [3]. These findings suggest that the combination of resampling techniques and automated hyperparameter tuning may further enhance the predictive performance of boosting algorithms beyond what has been reported in prior dental caries prediction studies.

Although the literature on machine learning applications for dental caries prediction continues to grow, most existing studies have either concentrated on a single algorithm or assessed model performance without sufficiently accounting for class imbalance. This issue is particularly relevant in clinical datasets, where caries-positive cases are typically underrepresented relative to caries-negative cases. Furthermore, limited attention has been given to the integration of automated hyperparameter optimization frameworks, such as Optuna, in conjunction with resampling techniques like SMOTE, specifically within the context of comparing CatBoost and LightGBM for dental caries risk classification. This study seeks to address this gap by systematically evaluating the impact of SMOTE-based class balancing and Optuna-driven hyperparameter tuning on the predictive performance of CatBoost and LightGBM, thereby providing a more comprehensive and methodologically robust comparison between the two algorithms in the domain of dental health prediction.

MATERIAL AND METHODS

This study used a quantitative machine learning-based approach. The goal was to compare two classification algorithms, Catboost and LightGBM, in predicting dental caries risk based on lifestyle and health factors.

Dataset

The dataset used is the publicly available body signal of smoking, accessible through the data.go.kr portal. This study utilizes data derived from South Korea's National Health Insurance Service (NHIS), which oversees the country's national health screening program. The dataset includes health and lifestyle data from over 55,692 individuals. It contains 27 columns [27].

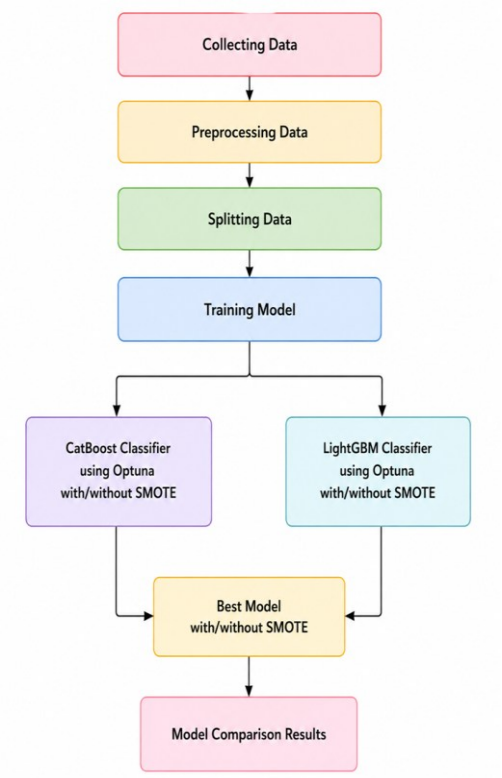

Fig. 1. Research methodology

Preprocessing Data

The preprocessing pipeline illustrated in Figure 2 was carried out to cleanse and prepare the dataset for subsequent use in machine learning (ML) model training. A novel feature designated as age group was engineered through a binning technique, which served to discretize the continuous age variable into categorical groups, thereby facilitating more interpretable analysis. Categorical variables, namely gender and smoking status, were subsequently transformed into numerical representations via Label Encoding so that the machine learning algorithms could process them computationally. Furthermore, columns deemed irrelevant to the prediction task, such as identifier fields, were removed to preclude any adverse interference during the model training phase. Collectively, these preprocessing steps are essential to ensure that the data utilized in

subsequent analyses is clean, well-structured, and fully compliant with the input requirements of the selected ML algorithms [18], [23].

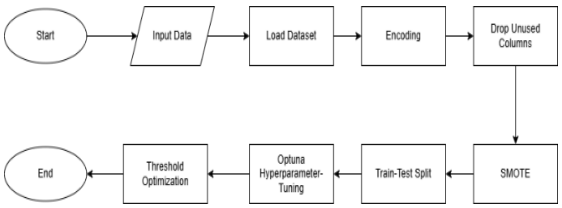

Fig.2. Preprocessing data

Training and Testing Data

Table 1 shows the distribution of data before and after applying SMOTE for the CatBoost and LightGBM models. Before applying for SMOTE, the training data set was 44,553 and the test data set was 11,139. After applying SMOTE, the training data set increased to 70,097, while the test data set became 17,525.

Table 1. Training and testing data

Model	Before SMOTE		After SMOTE	
	Training	Testing	Training	Testing
CatBoost	80%	20%	80%	20%
LightGBM	80%	20%	80%	20%

CatBoost Architecture

Two models were trained separately. The architecture in Figure 3 shows a decision tree-based CatBoost model specifically designed to handle categorical features directly without one-hot encoding. CatBoost also has the advantage of avoiding overfitting through sequence processing techniques.

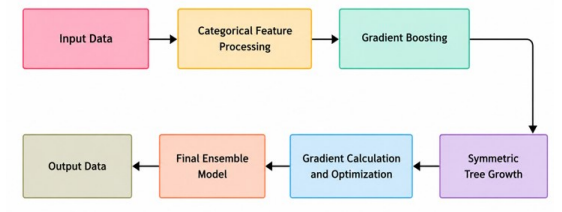

Fig. 3. Catboost architecture

LightGBM Architecture

Figure 4 depicts the model's architecture, highlighting how its speed and efficiency stem from the combination of histogram-based learning and a leaf-wise approach to tree growth. LightGBM is known to excel in handling large-scale and sparse datasets. To address the class imbalance observed in the dental caries dataset, the training data was rebalanced using SMOTE, a technique that generates synthetic samples for the minority class to promote a more equitable class distribution. To maximize the F1 score, each model's parameters were optimized through Optuna, a Bayesian optimization-based framework for hyperparameter tuning [24].

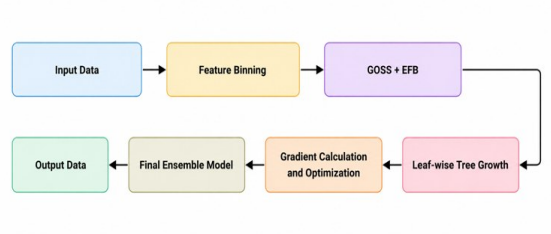

Fig. 4. LightGBM architecture

Model Evaluation

The trained model is evaluated on test data using the following classification matrices:

Confusion Matrix

The number of correct and incorrect predictions for each class is presented in a 2×2 table. This matrix provides a detailed overview of model errors, such as the number of caries cases missed (false negatives) and cases incorrectly classified as caries (false positives).

Accuracy

The percentage of correct predictions against all test data for positive and negative classes is used as an initial indicator of model performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \qquad (1)$$

Logloss

A log loss (logarithmic loss curve) graph is a graph that depicts the change in the log loss value (also called binary cross-entropy) during the classification model training process, usually on training data and validation data, as the number of iterations or boosting rounds increases.

$$\text{Logloss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1-y_i) log\ (1-p_i)] \qquad (2)$$

ROC Curve and AUC

The Receiver Operating Characteristic (ROC) Curve is a graph illustrating the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) as the classification threshold is varied. A model's ability to

distinguish between positive and negative classes is captured by the AUC metric, where values nearing 1 signify superior classification performance.

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

$$FPR = \frac{FP}{FP + FN} \quad (4)$$

Precision

Indicates how many of the predicted caries cases are truly caries, expressed as a proportion of all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FN}\ x\ 100\% \quad (5)$$

F1-Score

The F1-Score, calculated as the harmonic mean of precision and recall, offers a single value that captures the balance between these two metrics, making it particularly useful under imbalanced data conditions.

$$F1 - Score = 2\ x \frac{Precission\ x\ Recall}{Precission + Recall}\ x\ 100\% \quad (6)$$

One of the key metrics derived from the ROC Curve is the Area Under the Curve (AUC), which ranges from 0 to 1, with values closer to 1 reflecting a stronger ability of the model to distinguish between positive and negative classes.

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (7)$$

Implementation

Hyperparameter optimization was performed using Optuna, a Bayesian optimization framework employing the Tree-structured Parzen Estimator (TPE) algorithm to efficiently identify optimal parameter configurations. For each model, an objective function was defined to maximize the F1-score across 30 independent trials (n_trials = 30).

For the CatBoost model, the search space comprised the learning rate (0.01–0.2), tree depth (4–10), L2 regularization coefficient (1×10^{-3}–10.0), random strength (1×10^{-3}–10.0), and bagging temperature (0.0–1.0), with the number of iterations fixed at 500. For the LightGBM model, the search space comprised the learning rate (0.01–0.2), number of leaves (20–100), maximum tree depth (4–12), minimum child samples (10–50), subsample ratio (0.6–1.0), and column subsampling ratio per tree (0.6–1.0).

Upon completion of the trials, Optuna automatically selected the parameter configuration yielding the highest F1-score for each model. Compared with manual grid search, this Bayesian approach enabled more efficient and adaptive exploration of the hyperparameter space, reducing computational cost while improving the likelihood of convergence toward near-optimal configurations.

RESULT AND DISCUSSION

Beyond numerical metrics such as accuracy and F1-score, model performance evaluation also incorporates visual representations, including ROC curves and confusion matrices. Such visual representations make it easier to grasp how well the model distinguishes between positive cases (patients with dental caries) and negative cases (patients without dental caries).

CatBoost Confusion matrix

Figure 5 presents the confusion matrix generated from the classification results before SMOTE and Optuna optimization were applied. Of the 8,692 actual negative instances (label 0), the model correctly classified 8,621 instances, while 71 instances were misclassified. Conversely, of the 2,376 actual positive instances (label 1), only 135 were correctly identified, whereas the remaining 2,241 instances were erroneously classified as negative. These findings indicate that the model exhibited a pronounced classification imbalance, whereby its capacity to detect the minority class (positive) remained substantially limited. Figure 6 presents the confusion matrix depicting how the CatBoost model performed on the test dataset following the application of SMOTE and Optuna optimization.

The model successfully classified 7,308 out of 8,762 actual caries-positive cases as true positives. The false negative count of 1,454 was comparatively lower than that observed in the LightGBM model, indicating that CatBoost demonstrated superior sensitivity in identifying patients at risk of dental caries, thereby

affirming its greater clinical utility in minimizing missed diagnoses.

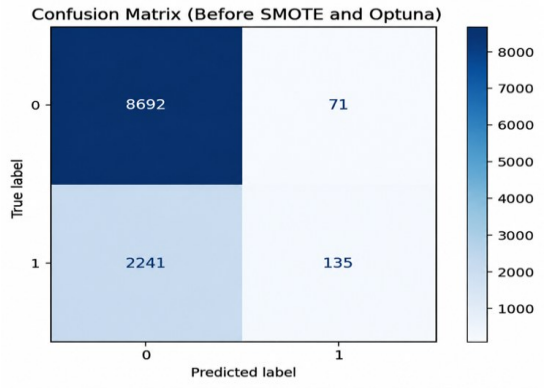

Fig. 5. Confusion matrix of the CatBoost model before SMOTE and Optuna

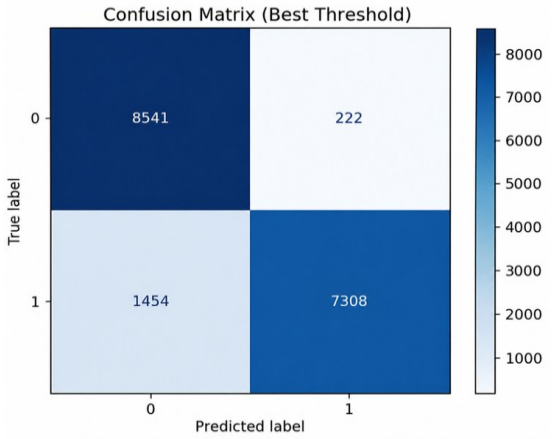

Fig. 6. Confusion matrix of the CatBoost model after SMOTE andOptuna

Evaluation

Figure 7 depicts how model accuracy evolved during training prior to applying SMOTE and Optuna optimization, with boosting rounds plotted on the x-axis and accuracy values on the y-axis. The training accuracy (blue line) climbed steadily before leveling off at approximately 0.82, while the validation accuracy (orange line) showed little change, hovering around 0.79. The widening gap between training and validation accuracy suggests early signs of overfitting, whereby the model learns the training data too well but fails to perform adequately on unseen validation data.

Figure 8 presents an accuracy curve reflecting the CatBoost model's effective learning process, with strong performance observed on both the training and validation datasets. Nevertheless, a marginal divergence between the training and validation accuracy observed beyond the 150th iteration indicates early signs of mild overfitting. Consequently, the incorporation of regularization techniques such as early stopping or more stringent regularization constraints may prove beneficial in mitigating this discrepancy, thereby preserving the model's generalization capability across unseen data.

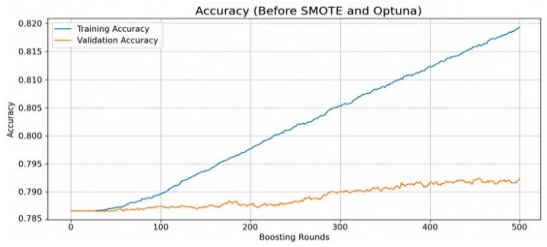

Fig. 7. CatBoost accuracy before SMOTE and Optuna

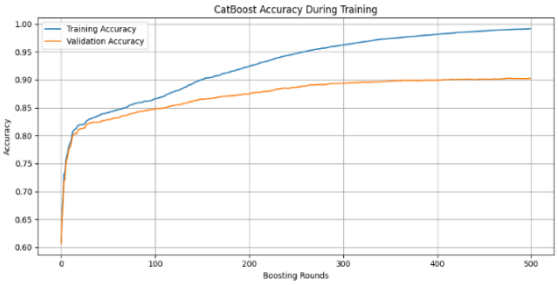

Fig. 8. CatBoost accuracy after SMOTE and Optuna

Logloss

Figure 9 illustrates how log-loss evolved throughout the training process before SMOTE and Optuna optimization were implemented, with the number of boosting rounds shown on the x-axis and the corresponding log-loss values on the y-axis. Training log-loss, shown by the blue line, decreased continuously throughout training, whereas validation log-loss, represented by the orange line, dropped during the initial iterations before leveling off at a fairly constant value. As the gap between the two curves continues to widen, it signals the presence of overfitting—the model's training performance kept improving while validation performance remained stagnant, pointing to a decline in its ability to generalize.

The logloss curve presented in Figure 10 demonstrates that the CatBoost model successfully captured underlying data patterns at a rapid rate during the early stages of training and sustained consistently strong performance throughout the remaining iterations. Despite the

visible gap between training and validation log-loss, the difference was small enough to remain within an acceptable range, falling short of severe overfitting. Nevertheless, to further optimize model performance and prevent excessive model complexity, the adoption of early stopping or fine-tuning of regularization parameters, particularly *l2_leaf_reg*, warrants consideration as a viable avenue for improvement.

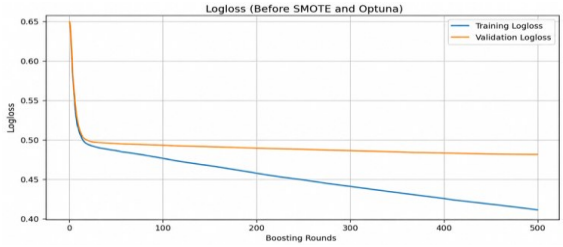

Fig. 9. Logloss curve of the CatBoost model before SMOTE and Optuna

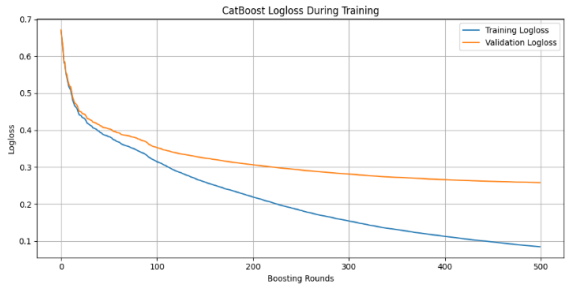

Fig. 10. Logloss curve of the CatBoost model after SMOTE and Optuna

ROC Curve and AUC

Figure 11 illustrates the model's Receiver Operating Characteristic (ROC) curve prior to SMOTE and Optuna optimization, capturing the trade-off between True Positive Rate (sensitivity) and False Positive Rate (1 − specificity) as the classification threshold varies. The model's ability to discriminate between classes is depicted by the blue line, while the orange dashed line serves as a reference for random chance performance (AUC = 0.50). With an AUC value of 0.68, the model demonstrated a relatively limited capacity to discriminate between the positive and negative classes, performing only slightly better than random guessing. These results underscore the necessity for further model refinement to achieve a more clinically and statistically meaningful level of discriminative ability.

Compared to the CatBoost model's ROC curve in Figure 12 (AUC = 0.95), the LightGBM model exhibited slightly lower performance, particularly regarding its discriminative ability for the positive class. Consistent with results from other classification metrics, this modest AUC difference reflects CatBoost's slightly stronger recall and F1-score, which reinforces its standing as the more discriminative classifier within the context of this study.

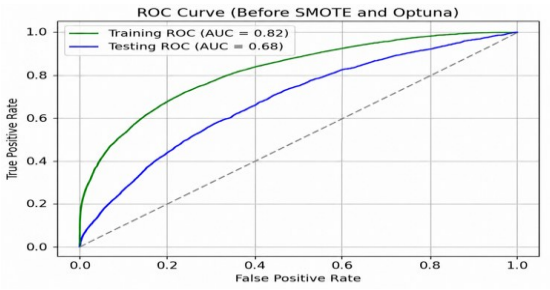

Fig. 11. ROC curve and AUC of the CatBoost model before SMOTE and Optuna

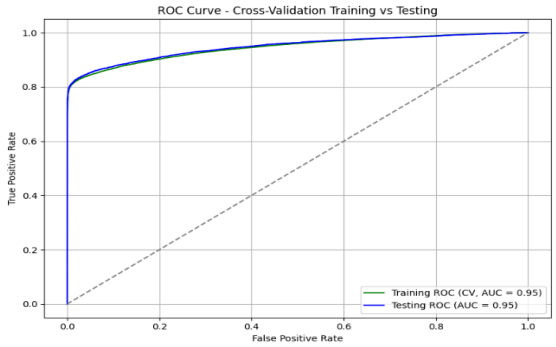

Fig. 12. ROC curve and AUC of the CatBoost model after SMOTE and Optuna

LightGBM Confusion Matrix

Figure 13 presents the confusion matrix obtained prior to the application of SMOTE and Optuna optimization within the classification model. The results reveal that 8,718 instances belonging to class 0 (non-caries) were correctly classified as true negatives, while 45 class 0 instances were erroneously assigned to class 1, constituting false positives.

Conversely, 2,303 instances of class 1 (Caries) were misclassified as class 0, representing false negatives, with only 73 class 1 instances correctly identified as true positives. Taken together, these findings indicate a pronounced bias in the model toward the majority class, reflecting a severely constrained capacity to identify minority class instances.

Such a deficiency carries considerable implications in the context of medical diagnosis, where the failure to accurately identify positive cases may result in clinically consequential misclassification outcomes.

Figure 14 shows the classification performance of the LightGBM model under the same experimental conditions. The model demonstrated a comparably high level of accuracy, yielding 7,242 true positive instances and only 219 false positives. Nevertheless, the false negative count remained relatively elevated at 1,520 cases in comparison to the CatBoost model, indicating that a non-trivial proportion of actual caries cases were not optimally detected by the LightGBM classifier. This indicates that despite LightGBM's strong overall accuracy, its ability to detect the minority class remains comparatively weak, which may pose a significant limitation in clinical diagnostic settings where reducing false negatives is essential.

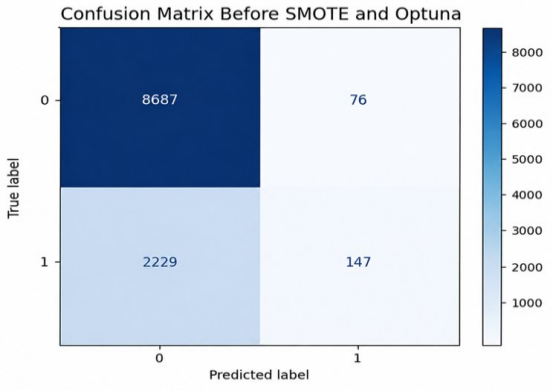

Fig. 13. Confusion matrix of the LightGBM model before SMOTE and Optuna

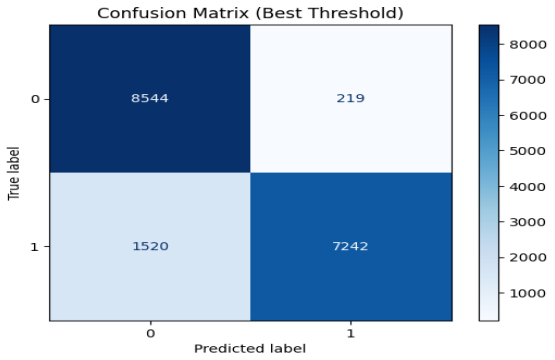

Fig. 14. Confusion matrix of the LightGBM model after SMOTE and Optuna

Evaluation

Figure 15 shows the training and validation accuracy curves before applying SMOTE. The training accuracy steadily increased to approximately 0.83, while the validation accuracy remained nearly constant at around 0.79 throughout the training process. The persistent gap between the two curves indicates potential overfitting and limited generalization performance, likely caused by class imbalance in the dataset. This issue motivated the application of SMOTE to improve class distribution and model performance.

Figure 16 shows that the LightGBM model achieved stable convergence with high accuracy on both the training and validation datasets. Although a small gap between the curves suggests mild overfitting, the validation accuracy improved steadily and remained stable, indicating good generalization and strong predictive performance without requiring major model adjustments.

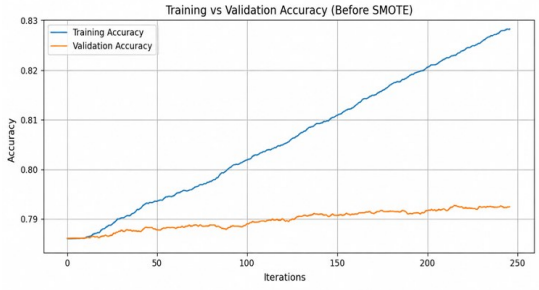

Fig. 15. Accuracy curve of the LightGBM model before SMOTE and Optuna

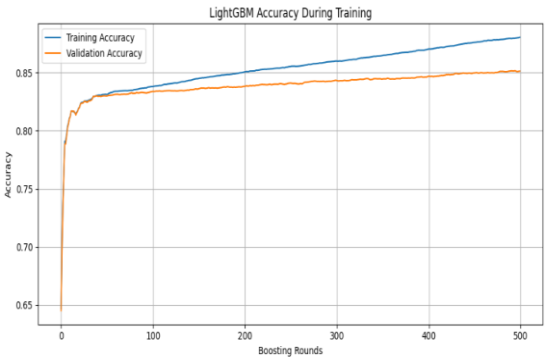

Fig. 16. Accuracy curve of the LightGBM model after SMOTE and Optuna

Logloss

Figure 17 shows the training and validation log-loss curves before applying SMOTE. The training log-loss steadily decreased to approximately 0.38, while the validation log-loss plateaued at around 0.48. This persistent gap indicates overfitting, where improvements on the training data did not translate to the validation set, highlighting the need for

SMOTE to address class imbalance and improve model generalization.

Figure 18 shows that the LightGBM model achieved stable learning, with both training and validation log-loss decreasing consistently. Although a small gap emerged after the 100th iteration, it remained limited, indicating good generalization. To further reduce the risk of overfitting, early stopping or regularization tuning may be beneficial.

Fig. 17. LogLoss curve of the LightGBM model before SMOTE and Optuna

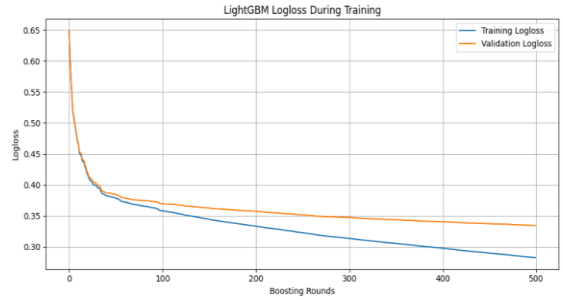

Fig. 18. LogLoss curve of the LightGBM model after SMOTE and Optuna

ROC Curve and AUC

Figure 19 presents the Receiver Operating Characteristic (ROC) curves prior to the application of SMOTE and Optuna optimization, providing a comparative assessment of model performance across the training and testing datasets. The ROC curve corresponding to the training data (green line) yielded an AUC of 0.87, indicative of highly proficient classification performance within the training set. However, the ROC curve associated with the testing data (blue line) demonstrated a substantially lower AUC of 0.70, reflecting a considerable degradation in the model's generalization capacity when evaluated on previously unseen data. The pronounced discrepancy in AUC values between the training and testing phases constitutes compelling evidence of overfitting, wherein the model exhibited an excessive adaptation to the training data, thereby compromising its predictive efficacy when applied to independent test instances.

Figure 20 presents the ROC curve for the LightGBM model, revealing a consistent AUC value of 0.94 across both the training (cross-validation) and testing datasets. This uniformity in discriminative performance across both evaluation phases indicates that the LightGBM model demonstrated a high degree of classification proficiency and stability, with no substantial evidence of overfitting, thereby confirming its robust generalization capability to unseen data instances.

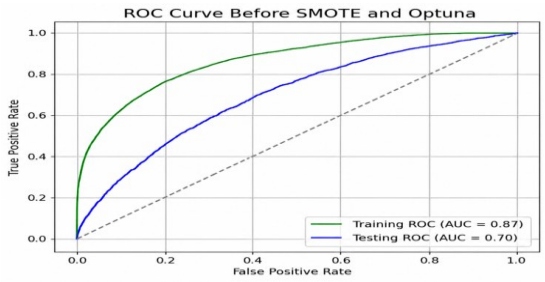

Fig. 19. ROC curve and AUC of the LightGBM model before SMOTE and optuna

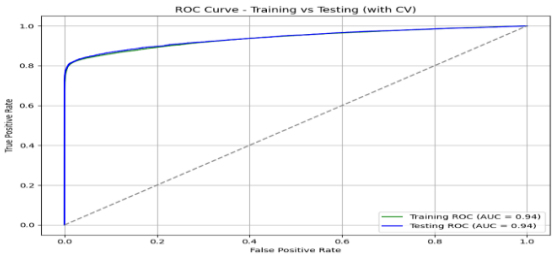

Fig. 20. ROC curve and AUC of the LightGBM model after SMOTE and Optuna

Without SMOTE and Optuna

Table 2 presents the preliminary evaluation results of the LightGBM and CatBoost models without SMOTE application. Although both models exhibited a capacity to learn from the training data, overall accuracy remained below 80%. Notably, both models showed markedly weak performance for label 1 (caries-positive cases), particularly in terms of recall and F1-score, indicating an inability to reliably identify individuals with dental caries. This limitation is attributable to the inherent class imbalance in the dataset, which biased the models toward the majority class during training.

Table 2. Evaluation of CatBoost and LightGBM models before SMOTE and Optuna

Model	Label	Accuracy	Precision	Recall	F1 Score
LightGBM	0	0.7892	0.80	0.99	0.88
	1		0.66	0.06	0.11
CatBoost	0	0.7902	0.79	0.99	0.88
	1		0.63	0.04	0.07

Incorporating SMOTE and Optuna

Table 3 shows that combining SMOTE with Optuna-based hyperparameter tuning substantially improved classification performance, with overall accuracy exceeding 90% and consistent gains across all other metrics.

Based on the Confusion Matrix, ROC curve, AUC, accuracy curve, and log-loss analysis, both CatBoost and LightGBM demonstrated strong performance in predicting dental caries risk. CatBoost showed a slight advantage, yielding more true positives and fewer false negatives, indicating higher sensitivity in detecting caries-positive cases. Its ROC curve achieved an AUC of 0.95, marginally outperforming LightGBM's 0.94, reflecting stronger discriminative capability for the positive class.

Table 3. Evaluation of CatBoost and LightGBM models after SMOTE and Optuna

Model	Label	Accuracy	Precision	Recall	F1 Score
LightGBM	0	0.9008	0.85	0.97	0.91
	1		0.97	0.83	0.90
CatBoost	0	0.9044	0.85	0.97	0.91
	1		0.97	0.83	0.90

The accuracy and log-loss curves confirmed effective learning and good generalization for both models, though CatBoost exhibited mild overfitting after the 150 epoch. While the pre-balancing evaluation showed poor performance on the minority class, applying SMOTE with Optuna-based tuning yielded substantial improvements across all metrics, including F1-score and recall. CatBoost achieved an F1-score of 89.7% and accuracy of 90.4%, slightly outperforming LightGBM's F1-score of 89.2% and accuracy of 90.1%.

Consequently, CatBoost is demonstrated to be the more reliable classifier for dental caries risk classification, particularly in the context of early detection of rare caries cases.

CONCLUSION

This study demonstrates that the CatBoost and LightGBM algorithms can be effectively implemented for predicting dental caries risk based on health and lifestyle-related data. The integration of SMOTE to address class imbalance and Optuna for automated hyperparameter optimization yielded a significant positive impact on overall model performance. Although both models exhibited commendable predictive capability, CatBoost is recommended as the superior classifier, owing to its marginally higher accuracy, F1-score, and AUC values, as well as its comparatively lower false negative rate — a critical consideration in the context of early disease detection. These findings underscore the substantial potential of machine learning-based models in supporting clinical decision-making systems within the domain of dental healthcare.

REFERENCES

[1] R. Amalia, *Karies gigi: Perspektif terkini aspek biologis, klinis, dan komunitas*. Ugm Press, 2021.

[2] A. Warreth, "Dental caries and its management," *International journal of dentistry,* vol. 2023, no. 1, p. 9365845, 2023. https://doi.org/10.1155/2023/9365845.

[3] I.-A. Kang, S. Ngnamsie Njimbouom, K.-O. Lee, and J.-D. Kim, "DCP: prediction of dental caries using machine learning in personalized medicine," *Applied Sciences,* vol. 12, no. 6, p. 3043, 2022. https://doi.org/10.3390/app12063043 .

[4] R. S. S. Suhadi and A. Mauludin, "Scoping Review: Rokok Sebagai Faktor Risiko terhadap Kejadian Karies Gigi," in *Bandung Conference Series: Medical Science*, 2023, vol. 3, no. 1, pp. 77-83. https://doi.org/10.29313/bcsms.v3i1.5626.

[5] V. I. Sunarko, E. Y. Puspaningrum, R. R. Widiastuty, S. Hadi, M. K. Awang, and I. G. S. M. Diyasa, "Convolutional layer exertion on few-shot learning for brain tumor classification," *Jurnal Ilmiah Kursor,* vol. 13, no. 2, pp. 98-110, 2025. https://doi.org/10.21107/kursor.v13i2.430 .

[6] Y. Wu, M. Jia, Y. Fang, D. Duangthip, C. H. Chu, and S. S. Gao, "Use machine learning to predict treatment outcome of early childhood caries," *BMC Oral Health,* vol. 25, no. 1, p. 389, 2025. https://doi.org/10.1186/s12903-025-05768-y.

[7] W. Fernando, D. Jollyta, D. Priyanto, and D. Oktarina, "The Influence Of Data Categorization And Attribute Instances Reduction Using The Gini Index On The Accuracy Of The Classification Algorithm ModeL," *Jurnal Ilmiah Kursor,* vol. 12, no. 3, pp. 111-122, 2024. https://doi.org/10.21107/kursor.v12i3.372.

[8] A. Kalalo, R. Rosmini, and A. Anto, "Klasifikasi Penyakit Karies Gigi Menggunakan Algoritma Modified K-Nearest Neighbor," *Journal of Big Data Analytic and Artificial Intelligence,* vol. 7, no. 2, pp. 48-54, 2024. https://doi.org/10.71302/jbidai.v7i2.60 .

[9] Y. Tanuwijaya and T. H. Rochadiani, "Komparatif Studi Model Deep Learning Untuk Deteksi Karies Gigi," *The Indonesian Journal of Computer Science,* vol. 14, no. 2, 2025. https://doi.org/10.33022/ijcs.v14i2.4857.

[10] M. Imani and H. R. Arabnia, "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis," *Technologies,* vol. 11, no. 6, p. 167, 2023. https://doi.org/10.3390/technologies11060167.

[11] F. N. R. Putri, R. R. Isnanto, and A. Sugiharto, "Skin Rash Classification System using Modified Densenet201 Through Random Search for Hyperparameter Tuning," *Jurnal Ilmiah Kursor,* vol. 12, no. 4, pp. 179-190, 2024. https://doi.org/10.21107/kursor.v12i4.418.

[12] K. Ileri, "Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks," *International Journal of Machine Learning and Cybernetics,* vol. 16, no. 9, pp. 6937-6956, 2025. https://doi.org/10.1007/s13042-025-02654-5.

[13] A. Odeh, Q. A. Al-Haija, A. Aref, and A. A. Taleb, "Comparative study of CatBoost, XGBoost, and LightGBM for enhanced URL phishing detection: a performance assessment," *Journal of Internet Services and Information Security,* vol. 13, no. 4, pp. 1-11, 2023. https://doi.org/10.58346/JISIS.2023.I4.001.

[14] M. Gonca, B. B. Gul, and M. F. Sert, "How successful is the CatBoost classifier in diagnosing different dental anomalies in patients via sella turcica and vertebral morphologic alteration?," *BMC Medical Informatics and Decision Making,* vol. 24, no. 1, p. 237, 2024. https://doi.org/10.1186/s12911-024-02643-8.

[15] J. Hancock and T. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review. J Big Data 7 (1): 94," ed, 2020. https://doi.org/10.1186/s40537-020-00369-8.

[16] Y. Cai, Y. Yuan, and A. Zhou, "Predictive slope stability early warning model based on CatBoost," *Scientific reports,* vol. 14, no. 1, p. 25727, 2024. https://doi.org/10.1038/s41598-024-77058-6.

[17] R. G. Farahani, A. Zarrabi, and P. Ghazanfari, "A Report on CatBoost: unbiased boosting with categorical

features," *Accessed: Aug,* vol. 11, 2025.

[18] Y. Wang and T. Wang, "Application of improved LightGBM model in blood glucose prediction," *Applied Sciences,* vol. 10, no. 9, p. 3227, 2020. https://doi.org/10.3390/app10093227.

[19] Y. Wu, M. Luo, S. Ding, and Q. Han, "Using a light gradient-boosting machine–shapley additive explanations model to evaluate the correlation between urban blue–green space landscape Spatial patterns and carbon sequestration," *Land,* vol. 13, no. 11, p. 1965, 2024. https://doi.org/10.3390/land13111965.

[20] S. S. Sabila and H. P. A. Tjahyaningtyas, "The multiple brain tumor with modified DenseNet121 architecture using brain MRI images: Classification multiclass brain tumor using brain MRI images with modified DenseNet121," *Jurnal Ilmiah Kursor,* vol. 12, no. 3, pp. 147-158, 2024. https://doi.org/10.21107/kursor.v12i3.379.

[21] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems,* vol. 30, 2017.

[22] W. Swastika, K. R. Prilianti, P. L. T. Irawan, and H. Setiawan, "Comparative analysis of random forest and deep learning approaches for automated acute lymphoblastic leukemia detection using morphologicaland textural features," *Jurnal Ilmiah Kursor,* vol. 13, no. 1, pp. 24-31, 2025. https://doi.org/10.21107/kursor.v13i1.427.

[23] Y.-H. Park, S.-H. Kim, and Y.-Y. Choi, "Prediction models of early childhood caries based on machine learning algorithms," *International Journal of Environmental Research and Public Health,* vol. 18, no. 16, p. 8613, 2021. https://doi.org/10.3390/ijerph18168613.

[24] A. Y. Yıldız and A. Kalayci, "Gradient boosting decision trees on medical diagnosis over tabular data," in *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, 2025: IEEE, pp. 1-8. https://doi.org/10.1109/ICAD65464.2025.11114069.

[25] J. T. Hancock and T. M. Khoshgoftaar, "Gradient boosted decision tree algorithms for medicare fraud detection," *SN Computer Science,* vol. 2, no. 4, p. 268, 2021. https://doi.org/10.1007/s42979-021-00655-z.

[26] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review: JT Hancock, TM Khoshgoftaar," *Journal of big data,* vol. 7, no. 1, p. 94, 2020. https://doi.org/10.1186/s40537-020-00369-8.

[27] 공공데이터포털, Accessed: Jun. 10, 2025. [Online]. Available: https://www.data.go.kr/index.do.