# Performance Comparison of Machine Learning Algorithms Using EfficientNetB0 Feature Extraction on Dental Disease Classification

**Mohammad Fa'iq Ruliff Mustafa [1]*, Ajie Kusuma Wardhana [2]***
* Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta
mohammadfaiq03@students.amikom.ac.id [1], ajiekusuma@amikom.ac.id [2]

**Article Info**

***Article history:***

Received 2025-07-17
Revised 2025-08-01
Accepted 2025-08-10

***Keyword:***

*Convolutional Neural Networks,*
*Dental Disease Classification,*
*EfficientNetB0,*
*Machine Learning,*
*SMOTE.*

## ABSTRACT

Oral health conditions such as dental caries, calculus, gingivitis, and ulcers are prevalent globally and require accurate early detection to prevent further complications. Traditional diagnostic methods such as visual inspection and manual radiograph analysis often rely on subjective judgment, leading to inconsistencies, delayed treatment, and limited accessibility, particularly in underserved areas. This study proposes an intelligent classification framework for dental disease detection based on intraoral images. Deep features were extracted using EfficientNetB0, followed by classification through eleven machine learning algorithms, including SVM, XGBoost, and K-Nearest Neighbors. Preprocessing steps included image augmentation, SMOTE for class balancing, and feature normalization. Among all models, SVM achieved the highest accuracy of 92,9%, while XGBoost and LightGBM followed closely at 91.3%. Using K-Fold Cross Validation, KNN algorithm has an increasing value with accuracy of 91,24%. This indicate the KNN algorithm able to tackle generalization problem towards the classification. The results demonstrate that features extracted using CNNs, when classified using machine learning algorithms, can provide a scalable and effective alternative to conventional diagnostic practices. Hence, Machine Learning algorithms provide a promising result towards dental disease classification.

## I. Introduction

Global oral health faces significant challenges, affecting approximately 3.5 billion individuals worldwide who suffer from untreated dental conditions, constituting a major public health concern as highlighted by the World Health Organization [1]. Dental caries, one of the most widespread yet preventable oral diseases, continues to be inadequately addressed due to traditional detection methods and limited early diagnostic interventions [2]. It impacts up to 95.6% of adolescents in certain populations, particularly among those with lower socioeconomic status [3]. Similarly, dental calculus is prevalent, affecting over 73% of immunocompromised children, while gingivitis has been reported in nearly all adolescents in some studies [4]. Mouth ulcers, though less frequently discussed in population studies, are also common in vulnerable groups [5].

Before the integration of artificial intelligence (AI) into dental diagnostics, practitioners relied heavily on manual inspections and radiographic interpretation, which introduced variability and limited early detection [6]. These traditional approaches often lacked consistency and standardization, resulting in delayed diagnosis and treatment. The demand for more accurate and timely identification of dental conditions has spurred interest in AI technologies, which have shown potential in addressing these diagnostic limitations by providing consistent and automated solutions [7][8].

Conventional diagnostic methods are constrained by subjectivity and limited sensitivity, particularly in detecting subtle or early-stage conditions [9]. This has driven the development of computational approaches that leverage image-based data, where convolutional neural networks (CNNs) have proven especially effective [10]. CNNs are capable of extracting high-dimensional spatial features from intraoral images, making them well-suited for identifying dental anomalies [11][12]. These features can then be classified using machine learning (ML) algorithms such as Support Vector Machine (SVM) and XGBoost, which are known for their effectiveness in handling complex data structures [13][14].

Rather than designing an end-to-end diagnostic tool, the primary focus of this study is to evaluate the classification performance of several supervised machine learning algorithms using CNN-derived features. By applying EfficientNetB0 as the backbone for deep feature extraction, this study aims to assess the effectiveness of machine learning models in classifying four common dental conditions: caries, calculus, gingivitis, and ulcers based on the extracted feature representations. The results are intended to provide empirical evidence on the strengths and limitations of different algorithms, serving as a benchmark for future research in AI-assisted dental classification under data-constrained conditions.

## II. Methodology

The system flow of this study, which comprises dataset preparation, image preprocessing, deep feature extraction using EfficientNetB0, class balancing with SMOTE, training of eleven machine learning models, and evaluation using standard classification metrics, is illustrated in Figure 1.

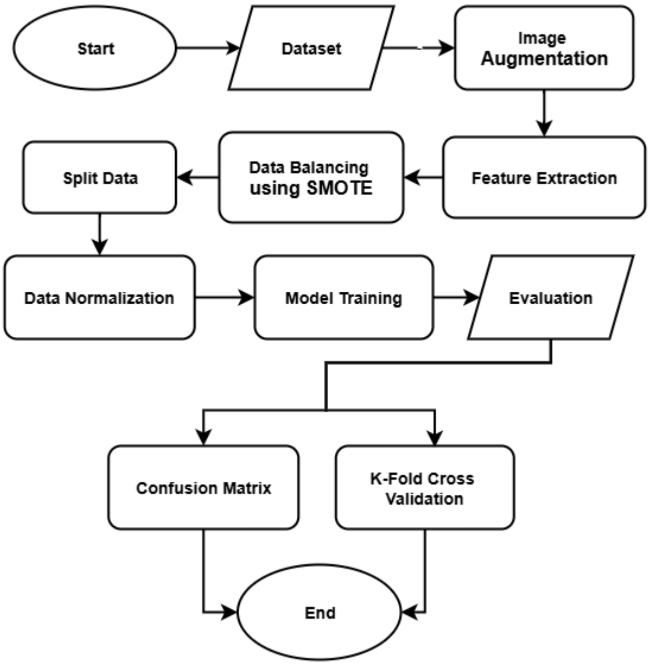

Figure 1. Flow Diagram

### A. Dataset

The dataset used in this study was obtained from the Oral Diseases dataset available on Kaggle, a publicly accessible platform for data science and machine learning research. This dataset provides a diverse collection of dental condition images, curated specifically to support automated classification and diagnostic tasks in the dental domain. It includes six distinct categories of oral diseases: caries, calculus, gingivitis, tooth discoloration, ulcers, and hypodontia. The dataset provider offers both the original and augmented versions of the data, complete with labeled annotations. Notably, data augmentation such as rotation, flipping, and scaling was applied only to the caries and ulcer classes, while the remaining categories consist of original images.

For the purpose of this study, only four disease categories were selected for multiclass classification experiments: caries, calculus, gingivitis, and mouth ulcer. The other two classes, Tooth Discoloration and Hypodontia, were excluded due to their limited representativeness in the current experimental scope. All images were collected from multiple reputable sources, including hospitals and dental knowledge bases, ensuring a high degree of reliability and authenticity in the visual representations of the conditions. An overview of the dataset samples across the selected classes is presented in Figure 2.

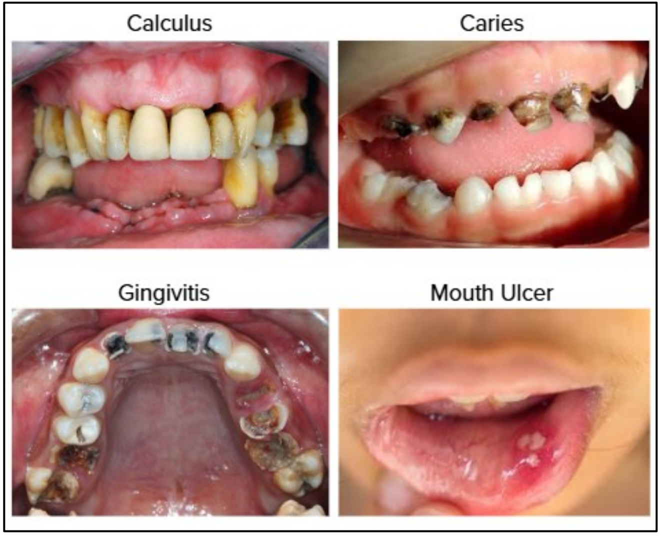

Figure 2. Image Sample Visualization

### B. Image Augmentation

To improve the robustness and generalization capability of the dataset, various image augmentation techniques were employed during the preprocessing phase. Standard augmentation operations such as horizontal and vertical flipping, random rotations, zooming, brightness adjustments, and noise addition were applied to artificially increase the dataset size and variability. These transformations simulate real-world variations in dental imagery and help the model become more invariant to positional and illumination changes. Furthermore, additional custom augmentations were applied specifically to the gingivitis and calculus classes, which initially exhibited poor classification performance due to limited and less diverse image samples. This targeted enhancement contributed significantly to balancing class representation and ultimately improved the model's classification accuracy, particularly for these two categories.

### C. Feature Extraction

In this study, Convolutional Neural Network (CNN)-based feature extraction was performed using the EfficientNetB0 architecture, a lightweight yet powerful model pretrained on the ImageNet dataset. EfficientNetB0 was employed without its top classification layers and configured with global average pooling to obtain compact and discriminative feature vectors. Each input image was resized to 224×224 pixels and preprocessed using the model's standardized normalization function to ensure compatibility with the pretrained weights. The extracted features, representing high-level spatial and semantic information from the dental images, were flattened into 1280-dimensional vectors. These vectors were later used as input for various machine learning classifiers, effectively decoupling feature learning from the classification phase and enabling a hybrid deep learning and traditional machine learning pipeline.

### D. Data Balancing

To address the issue of class imbalance within the dataset particularly the underrepresentation of certain conditions such as gingivitis and calculus the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE is an oversampling algorithm that generates synthetic examples in the feature space between minority class instances rather than simply duplicating them. By interpolating new samples using the k-nearest neighbors of a given minority class instance, SMOTE enhances the diversity and distribution of underrepresented categories [15]. This not only helps mitigate model bias toward the majority class but also improves overall generalization in classification tasks. The SMOTE algorithm can be represented in the form of Equation (1).

$$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i) \tag{1}$$

$x_{\text{new}}$ : Synthetic sample data
$x_i$ : Original sample from the minority class
$\delta$ : Random value between 0 and 1
$x_{zi}$ : One of the nearest neighbors of $X_i$

This process is repeated until the number of minority class samples reaches a better level of balance with the majority class. Thus, SMOTE helps address the problem of imbalanced data without simply duplicating the original data, thereby reducing the risk of overfitting and improving the performance of machine learning models [16].

### E. Models

To classify the extracted CNN features, a variety of supervised machine learning algorithms were employed to evaluate performance across different learning paradigms. To ensure a fair and consistent evaluation framework, all models were tested using their default hyperparameters. This decision was made to standardize the treatment across models and objectively observe each algorithm's baseline performance without the influence of hyperparameter tuning. The primary focus was placed on the Support Vector Machine (SVM), due to its proven effectiveness in handling high-dimensional feature spaces and its strong generalization capability. Additionally, XGBoost was selected for its optimized gradient boosting framework that combines scalability with state-of-the-art performance in structured data tasks. Logistic Regression was included as a baseline linear classifier, providing interpretable results and fast training, while Decision Tree and Random Forest were chosen to capture nonlinear decision boundaries and feature interactions.

Other ensemble-based classifiers such as Gradient Boosting, AdaBoost, and LightGBM were also implemented to assess the potential improvements gained from boosting techniques and model ensembling. Furthermore, simpler yet statistically robust models such as Naive Bayes and K-Nearest Neighbors (KNN) were tested to explore their effectiveness when applied to CNN-extracted features. Lastly, the Ridge Classifier was incorporated to evaluate the performance of regularized linear models in multi-class classification scenarios. This diverse collection of models allowed for a comprehensive comparison and helped identify the most suitable approach for dental condition classification based on pre-trained CNN representations.

### F. Evaluation Metrics

To assess the performance of each machine learning classifier, multiple evaluation metrics were employed, encompassing both overall accuracy and class-wise performance indicators. *Accuracy* measures the proportion of correctly predicted samples over the total number of instances and provides a general overview of model performance. This metric can be defined mathematically as equation (2). However, in multiclass imbalanced classification tasks, accuracy alone can be misleading. Therefore, additional metrics such as *Recall*, which quantifies the ability of the model to correctly identify positive instances, were also considered and can be formally defined as equation (3). *Precision* evaluates the proportion of true positive predictions among all positive predictions made by the model, as described in equation (4). To balance both recall and precision, the *F1-score*, a harmonic mean of the two, was also calculated, as defined in equation (5). Finally, confusion matrices were generated to visualize the classification results for each class and to identify patterns of misclassification, providing further insights into the model's behavior.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

## III. Results and Analysis

### A. Augmentation

The initial evaluation revealed that the gingivitis and calculus classes within the original dataset exhibited notably lower classification performance compared to other categories. Visual inspection and sample count analysis indicated that these classes suffered from limited variation and potentially redundant or low-quality images. To address this, the dataset underwent a refinement process involving the removal of non-representative or noisy images, followed by a targeted image augmentation strategy that can be observed on Figure 3.

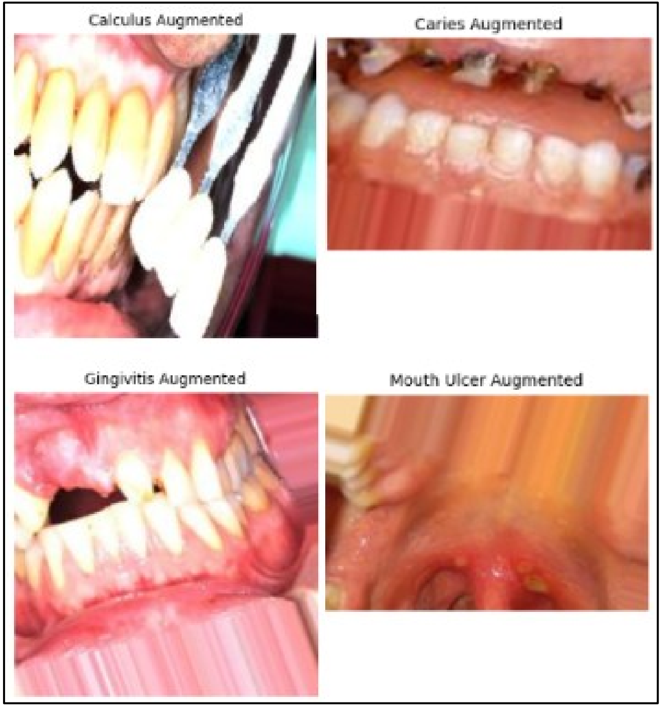

Figure 3. Image Augmented Sample

Specifically, augmentation was applied using a combination of geometric and photometric transformations designed to enhance data diversity while preserving diagnostic relevance. The augmentation parameters included random rotations, width and height shifts, zooming, horizontal flipping, brightness adjustments, and filling of empty pixels using the Nearest-Neighbor method. These transformations simulate real-world variability in patient imaging conditions and intraoral perspectives. The augmentation process not only balanced the dataset but also enhanced the model's ability to generalize to previously underperforming classes, contributing directly to the overall increase in system accuracy.

### B. Grad-CAM

To enhance model interpretability and gain insight into the decision-making process of the CNN feature extractor, Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to several augmented dental disease images. As depicted in Figure 4, the heatmaps highlight spatial regions in the intraoral images that were most influential in model predictions for each class: calculus, caries, gingivitis, and mouth ulcer.

Each visualization shows that the EfficientNetB0 model effectively localizes relevant anatomical features. For example, in the calculus and caries images, attention is focused around the tooth surface and occlusal grooves where lesions or buildup are visually apparent. Gingivitis activations concentrate near the gingival margins, while mouth ulcer images highlight irregular tissue textures and discolorations. These class-specific activation patterns suggest that the CNN has successfully learned to identify meaningful features aligned with symptoms, thus supporting its role as a robust and interpretable feature extractor for downstream classifiers.

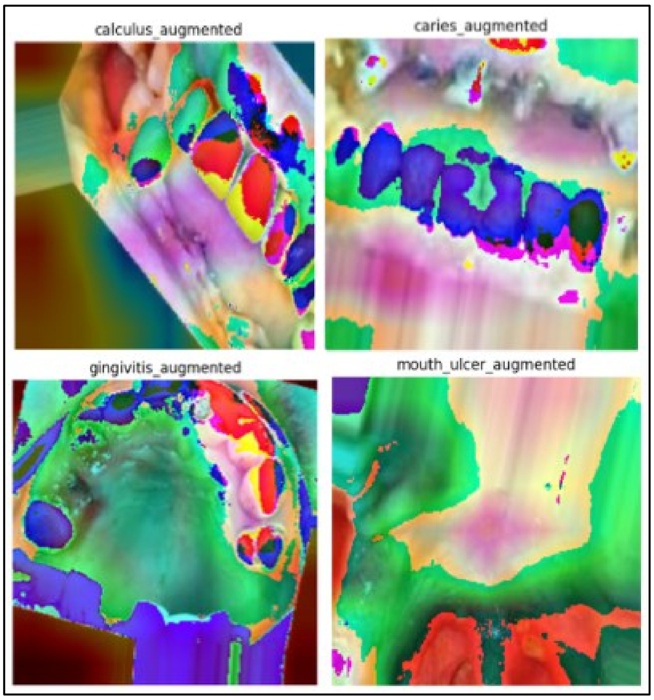

Figure 4. Grad-CAM Visualization

### C. Feature Extraction

Feature extraction was performed using the EfficientNetB0 model by removing its classification head and applying global average pooling, resulting in consistent 1280-dimensional feature vectors per image. These vectors encapsulate high-level information relevant to dental abnormalities. To visualize the effectiveness of these extracted features in distinguishing between disease types, dimensionality reduction was applied to project the high-dimensional data into two components, shown in Figure 5. The scatter plot reveals overlapping distributions among the classes, particularly with ulcers dominating a large area and partially masking other classes. However, some degree of clustering is still visible for caries and gingivitis, indicating that EfficientNetB0 captures certain discriminative patterns, though not fully separable across all categories without further refinement or class balancing.

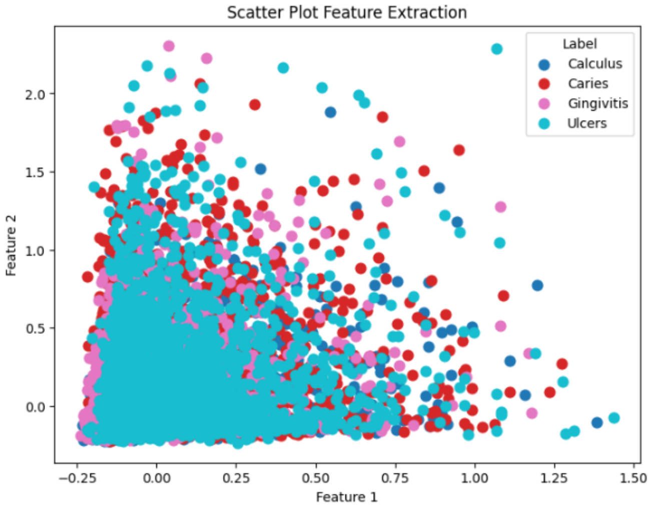

Figure 5. Scatter Plot After Feature Extraction

### D. Data Balancing SMOTE

Initial analysis of the dataset revealed a clear class imbalance, with the number of samples for certain dental conditions being significantly unequal. Specifically, the dataset initially contained 2,541 images for ulcers, 2,382 for caries, 2,219 for gingivitis, and only 2,216 for calculus. This imbalance posed a challenge for model training, as classifiers tend to be biased toward majority classes, leading to reduced performance on minority categories. To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the extracted feature dataset.

SMOTE generates new synthetic feature vectors by interpolating between existing samples of underrepresented classes in the feature space, thus expanding class representation without simply duplicating original data. After applying SMOTE, the dataset achieved perfect balance, with 2,541 samples for each of the four classes (calculus, caries, gingivitis, and ulcers). This rebalancing is visualized in Figure 6 and Figure 7, which illustrate the class distributions before and after SMOTE. Equalizing the sample counts ensured a fairer training process and significantly improved classification performance particularly for classes that were previously underrepresented by enhancing precision, recall, and overall model generalization.

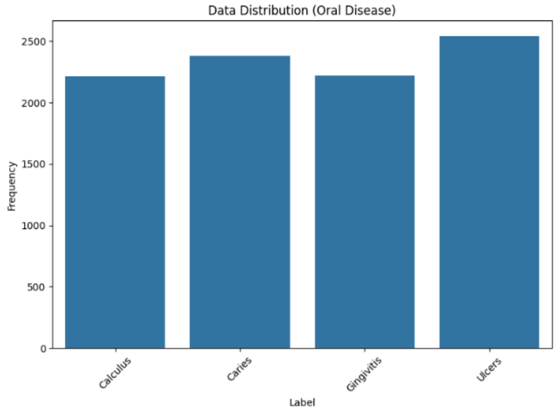

Figure 6. Data Before SMOTE

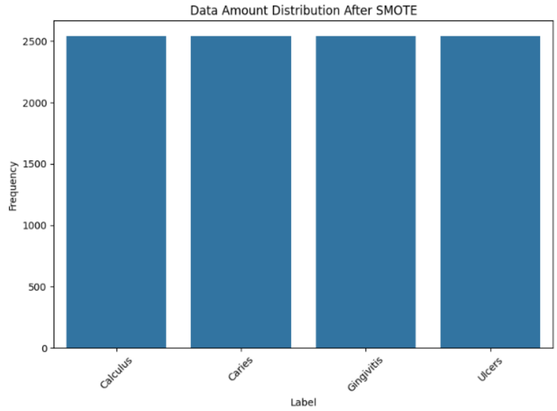

Figure 7. Data After SMOTE

### E. Splitting Data

To evaluate the generalization capability of the classification models, the original dataset consisting of 10,164 intraoral images was used as the input for preprocessing and balancing. After applying augmentation and the SMOTE technique to address class imbalance, the resulting balanced dataset was split into 7,114 training images and 3,050 testing images, following a 70:30 ratio. Stratified sampling was employed to maintain consistent class distribution across both subsets, ensuring a fair evaluation and avoiding bias during model assessment. This strategy allowed the models to be trained on a representative dataset while preserving the improvements achieved for minority classes. The detailed data distribution is presented in Table I.

TABLE I
TRAIN/TEST SPLIT

| **Information** | **Training Data** | **Testing Data** |
|---|---|---|
| **Proportion** | **70%** | **30%** |
| Calculus | 1778 | 763 |
| Caries | 1779 | 762 |
| Gingivitis | 1779 | 763 |
| Mouth Ulcers | 1778 | 762 |
| Total | 7114 | 3050 |

### F. Data Normalization

Prior to classification, the feature data underwent normalization using the Standard Scaler technique, which standardizes features by removing the mean and scaling to unit variance. This preprocessing step is particularly important for distance-based and gradient-sensitive algorithms, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and logistic regression, ensuring that all features contribute equally to the model's decision-making process. Without normalization, features with larger numerical ranges could disproportionately influence the model. By applying Standard Scaler, each feature in the dataset was transformed to have a mean of zero and a standard deviation of one. This transformation enhanced model stability, reduced training time, and contributed to the improved convergence and performance observed during the evaluation of the various classifiers.

### G. Evaluation

Evaluating various machine learning algorithms is essential to determine the most effective solution for multiclass classification of oral diseases based on intraoral images. In this study, eleven machine learning models were systematically assessed, ranging from conventional classifiers to advanced ensemble methods, to measure their effectiveness when applied to deep features extracted using EfficientNetB0. Such an approach enables a balanced comparison of model capabilities under identical conditions. Standard evaluation metrics including accuracy, precision, recall, and F1-score were utilized to capture multiple aspects of model effectiveness. The classification results for all models are presented in Table II.

TABLE II
EVALUATION MODEL COMPARISON [JANGAN DIBULETIN, HARUS KONSISTEN]

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 88,13 | 88,22 | 88,14 | 88,16 |
| Decision Tree | 70,3 | 70,44 | 70,3 | 70,36 |
| Random Forest | 88,89 | 88,87 | 88,89 | 88,88 |
| Gradient Boosting | 86,89 | 86,87 | 86,89 | 86,88 |
| **Support Vector Machine** | **92,95** | **92,98** | **92,95** | **92,96** |
| AdaBoost | 67 | 67,81 | 67 | 67,31 |
| Naive Bayes | 70,5 | 72,88 | 70,5 | 69,82 |
| K-Nearest Neighbors | 91 | 91 | 91 | 91 |
| Ridge Classifier | 90,1 | 90,2 | 90,1 | 90,1 |
| **XGBoost** | **91,3** | **91,4** | **91,3** | **91,3** |
| **LightGBM** | **91,3** | **91,3** | **91,3** | **91,3** |

Support Vector Machine (SVM), XGBoost, and LightGBM emerged as the top-performing models in this study. SVM demonstrated superior performance, effectively handling the high-dimensional feature space derived from EfficientNetB0 and achieving consistent predictions across all classes. XGBoost and LightGBM, both gradient boosting-based ensemble methods, also showed excellent classification capabilities by capturing intricate patterns in the extracted features with high stability. These three models stood out in terms of generalization, robustness, and their ability to accurately differentiate among multiple oral disease categories.

Meanwhile, other models such as K-Nearest Neighbors, Ridge Classifier, Random Forest, Logistic Regression, and Gradient Boosting also performed competitively, though slightly behind the leading trio. These models maintained strong accuracy and balanced metric scores, suggesting that with appropriate preprocessing and feature extraction, traditional classifiers can still yield promising results in medical image classification tasks. Ridge and Logistic Regression, in particular, demonstrated efficiency on linearly separable data, while ensemble-based Random Forest and Gradient Boosting contributed to stable yet slightly less optimal predictions.

In contrast, Decision Tree, AdaBoost, and Naïve Bayes recorded the lowest performance among the models evaluated. The Decision Tree model, while easy to interpret, was limited in its capacity to manage the feature complexity, resulting in lower classification accuracy. AdaBoost's reliance on weak base learners, such as shallow decision trees, proved inadequate when dealing with high-dimensional feature representations extracted by EfficientNetB0, resulting in elevated misclassification rates. Naïve Bayes, with its strong independence assumptions, struggled to model the dependencies present in CNN-derived features, thus underperforming across all metrics.

In comparison to prior research by Shang-Ting Hsieh et al. [14], which achieved an accuracy of 92.5% through a fusion-based approach combining five CNN architectures with SVM and Naive Bayes classifiers, our study offers a more efficient and practical solution for multiclass dental disease classification. Although their method benefited from a large dataset and complex feature integration, it also introduced significant computational overhead. Similarly, research by Mohammed Zubair Hussain et al [17], demonstrated competitive performance using multiple deep CNN architectures such as Xception, ResNet, and DenseNet achieving accuracies up to 85.19%. However, their approach required extensive training on augmented clinical images and relied heavily on end-to-end deep learning pipelines, which may limit deployment flexibility in resource-constrained environments.. In contrast, our approach utilizes EfficientNetB0 as a single CNN backbone, and model evaluation using SVM resulted in an improvement of 0.5%. A detailed performance comparison between both methods is presented in Table 3.

TABLE III
PERFORMANCE COMPARISON

| Study | Feature Extraction | Model | Accuracy |
|---|---|---|---|
| Shang-Ting Hsieh et al [14] | 5 CNNs Fusion | SVM/Naïve Bayes | 92.5% |
| This Study | EfficientNetB0 | SVM | 92.9% |
| Mohammed Zubair Hussain et al [17] | Single-model CNN | MobileNet (best) | 85.2% |

To gain a more comprehensive understanding of each algorithm's classification performance across the four dental disease categories, confusion matrices were employed as visual analysis tools. A confusion matrix presents the distribution of correct and incorrect predictions for each class, making it easier to identify patterns of misclassification and to evaluate the model's ability to distinguish between visually similar conditions. Based on the confusion matrix visualizations, the Support Vector Machine (SVM) model identified as the best-performing classifier was analyzed in detail to illustrate how accurately it classified the four oral disease categories. For this classification task, the label encoding was as follows: 0 for Calculus, 1 for Caries, 2 for Gingivitis, and 3 for Ulcers.

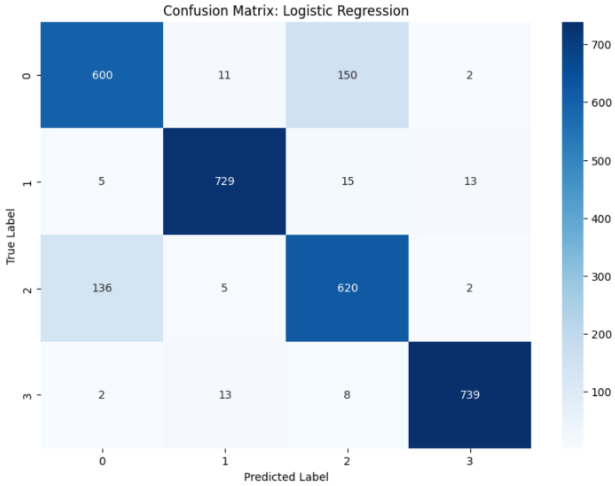

Figure 8. Matrix Logistic Regression

The confusion matrix in Figure 8 shows the performance of the Logistic Regression model in classifying the four types of dental diseases. This model performs best in the Ulcers class with 739 mostly accurate predictions, reflecting its ability to recognise the distinctive visual characteristics of this class. However, there are still some prediction errors between other classes, particularly between Calculus and Gingivitis, indicating overlapping features that are difficult for the model to distinguish. Overall, Logistic Regression performs competently in this classification task, especially in distinguishing classes that are more visually distinct.

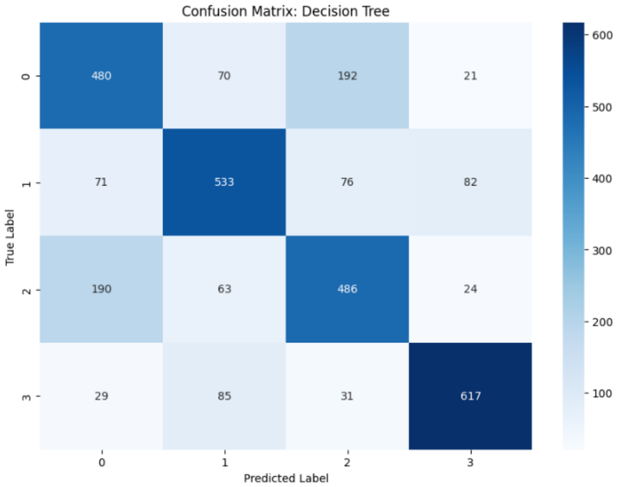

Figure 9. Matrix Decision Tree

The confusion matrix in Figure 9 illustrates the performance of the Decision Tree model in classifying dental disease categories. The highest correct prediction was achieved for class 3 (Ulcers), with 617 samples accurately classified. Despite this, the model struggled notably in differentiating between class 0 (Calculus) and class 2 (Gingivitis), where 192 calculus samples were misclassified as gingivitis, and 190 gingivitis samples were predicted as calculus. Additionally, class 1 (Caries) experienced moderate misclassification, especially toward ulcers and gingivitis. Overall, while the model shows some ability in identifying clear-cut cases, its performance is hindered by high misclassification rates in overlapping and visually similar conditions.

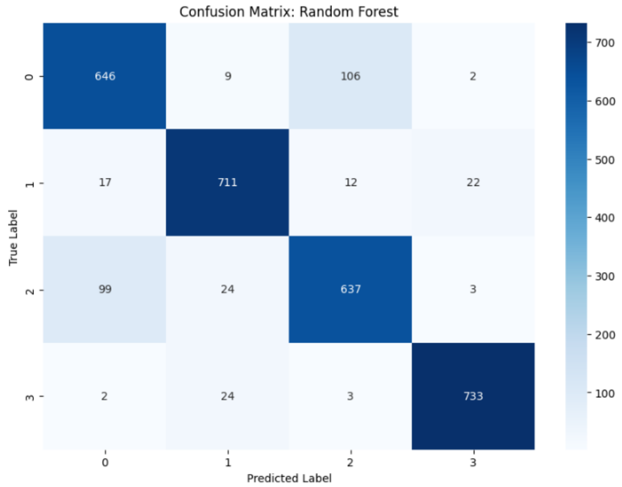

Figure 10. Matrix Random Forest

The confusion matrix in Figure 10 reveals that the Random Forest model achieved strong and balanced performance across all four dental disease classes. The highest correct predictions were recorded for class 3 (Ulcers) with 733 samples accurately classified, followed closely by class 1

(Caries) and class 0 (Calculus). However, some misclassifications were still observed, particularly between class 0 (Calculus) and class 2 (Gingivitis), where 99 gingivitis samples were incorrectly predicted as calculus. These results indicate that Random Forest is capable of generalizing well across complex feature spaces, although minor confusion remains between classes with overlapping characteristics.

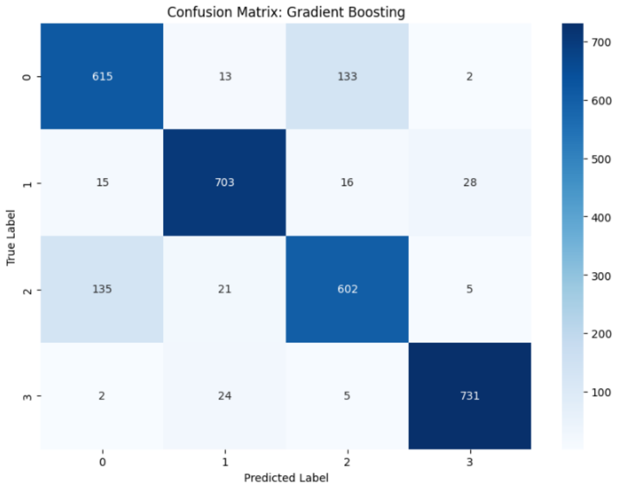

Figure 11. Matrix Gradient Boosting

The confusion matrix in Figure 11 shows that the Gradient Boosting model performed well overall, with the highest number of correct predictions found in class 3 (Ulcers), totaling 731 correctly classified samples. The model also demonstrated reliable performance for class 1 (Caries), with a relatively low number of misclassifications. However, notable confusion occurred between class 0 (Calculus) and class 2 (Gingivitis), where a significant portion of samples were misclassified between the two. Despite this, Gradient Boosting managed to capture important patterns in the data, making it a competitive choice for dental image classification, though slightly behind top-performing models in precision.

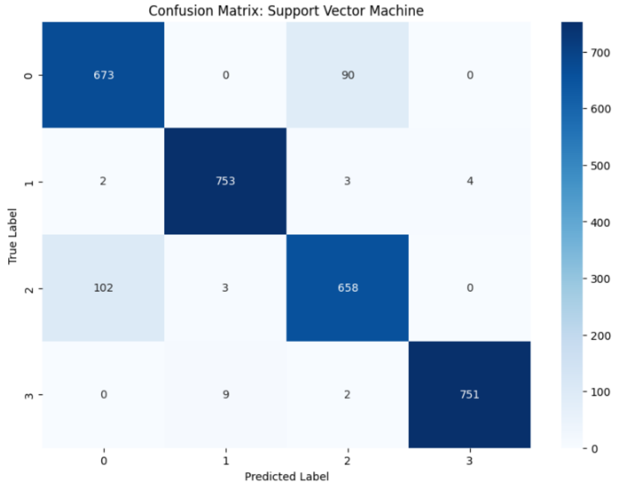

Figure 12. Matrix Support Vector Machine

The confusion matrix in Figure 12 highlights the outstanding performance of the Support Vector Machine (SVM) model, which achieved highly accurate predictions across all four dental disease classes. The model demonstrated exceptional precision in classifying class 1 (Caries) and class 3 (Ulcers), with 753 and 751 correct predictions respectively. Even in more challenging classes such as class 0 (Calculus) and class 2 (Gingivitis), SVM maintained strong results, showing minimal confusion. This balanced and accurate classification underscores the robustness of SVM when paired with deep feature representations, making it the top-performing model in this study.

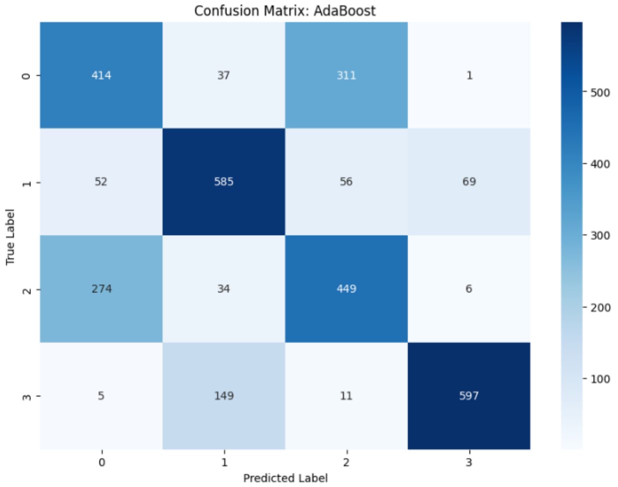

Figure 13. Matrix AdaBoost

The confusion matrix in Figure 13 reveals the weak performance of the AdaBoost model in this multiclass classification task. The model struggled significantly with class separation, particularly between class 0 (Calculus) and class 2 (Gingivitis), as evidenced by 311 Calculus samples misclassified as Gingivitis and 274 Gingivitis samples predicted as Calculus. Misclassification was also high for class 3 (Ulcers), with 149 samples incorrectly labeled as Caries. This outcome reflects the limitations of AdaBoost's reliance on weak base learners, such as shallow decision trees, which may lack the capacity to capture the complex patterns extracted by CNN-based feature representations. As a result, AdaBoost was the lowest-performing model in this study.

The confusion matrix in Figure 14 shows that the Naive Bayes classifier exhibited considerable misclassification across multiple classes, particularly between Calculus and Gingivitis. For example, 394 Calculus images were incorrectly predicted as Gingivitis, indicating a substantial overlap in feature representation that Naive Bayes failed to distinguish. Similarly, for Caries and Ulcers, the model misclassified several samples, highlighting its limited ability to handle complex and interdependent features.

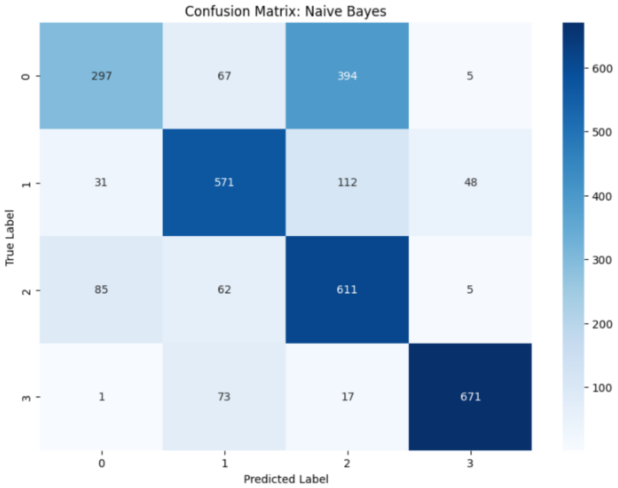

Figure 14. Matrix Naive Bayes

This performance can be attributed to the algorithm's strong assumption of feature independence, which does not hold for high-dimensional CNN-based representations. Consequently, Naive Bayes was among the lowest-performing models in this experiment.

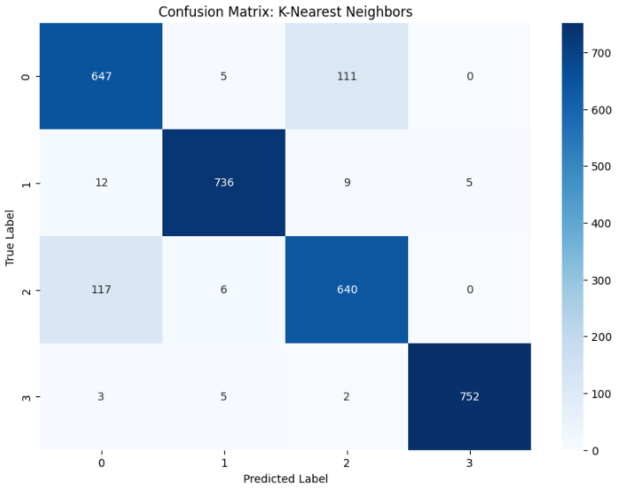

Figure 15. K-Nearest Neighbors

The confusion matrix in Figure 15 illustrates that the K-Nearest Neighbors (KNN) algorithm achieved strong classification performance, particularly in predicting Mouth Ulcers with 752 correct classifications and minimal errors across other classes. The model also performed well for Caries and Calculus, although a moderate number of Calculus samples (111) were misclassified as Gingivitis. This indicates that while KNN effectively leveraged neighborhood information in the feature space, it occasionally struggled to differentiate between visually similar conditions like Calculus and Gingivitis. Overall, KNN demonstrated robust generalization in this multiclass dental classification task.

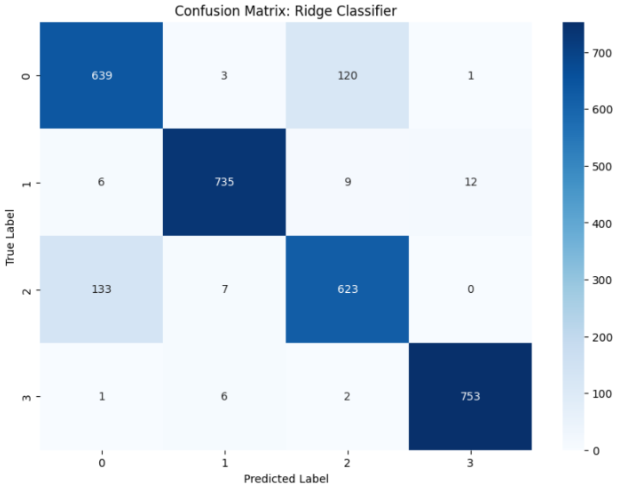

Figure 16. Matrix Ridge Classifier

The confusion matrix in Figure 16 highlights that the Ridge Classifier delivered strong performance across all classes, with particularly high accuracy in predicting Mouth Ulcers (753 correct predictions) and Caries (735 correct predictions). Despite this, the model showed some confusion between Calculus and Gingivitis, where 120 samples of Calculus were misclassified as Gingivitis and 133 Gingivitis samples were predicted as Calculus. These misclassifications suggest that while Ridge Classifier handles high-dimensional features well, it may still face challenges distinguishing between classes with overlapping visual characteristics. Nonetheless, its overall performance indicates effective generalization in the multiclass classification setting.

The confusion matrix in Figure 17 demonstrates that XGBoost achieved highly consistent classification performance across all dental disease classes. The model correctly identified the majority of instances, especially for Mouth Ulcers (742) and Caries (728), with minimal misclassifications.

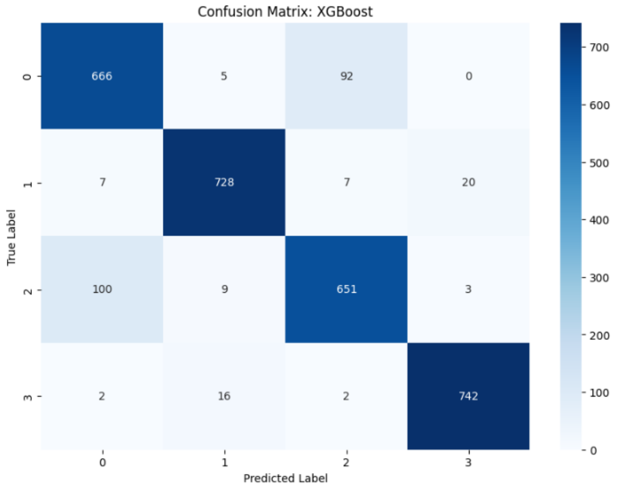

Figure 17. Matrix XGBoost

Although there were some errors between Calculus and Gingivitis common across most models XGBoost effectively distinguished between visually distinct categories. This reinforces its strength as a gradient-boosted ensemble method capable of capturing complex patterns from deep CNN-derived features, making it one of the most reliable models in this study.

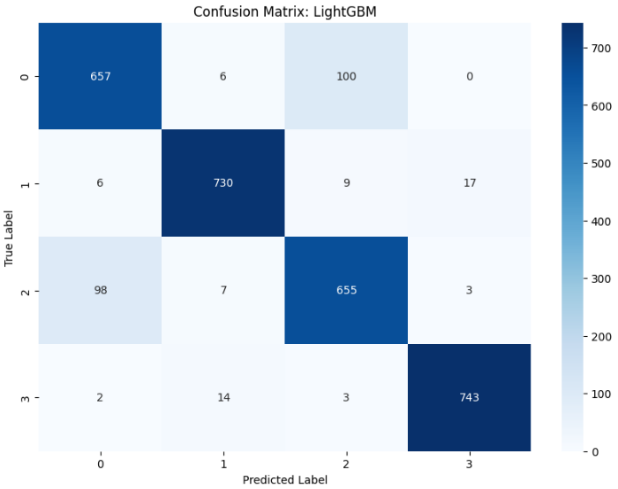

Figure 18. Matrix LightGBM

The confusion matrix in Figure 18 indicates that LightGBM achieved a strong classification performance, particularly in identifying Mouth Ulcers (743) and Caries (730) with high accuracy. While minor misclassifications were observed between Calculus and Gingivitis, the model generally maintained reliable predictions across all four classes. As a gradient boosting framework optimized for efficiency and speed, LightGBM effectively leveraged the high-level features extracted from EfficientNetB0, confirming its capability to handle complex dental image classification tasks with competitive precision.

Overall, the evaluation of 11 machine learning models using CNN-based features extracted from EfficientNetB0 revealed distinct differences in classification performance across algorithms. Support Vector Machine (SVM) emerged as the most accurate and consistent model, demonstrating strong generalization and minimal misclassification across all four dental disease categories. Its superior performance is likely attributed to its margin-maximizing capability, which allows it to effectively separate complex patterns within the high-dimensional feature space produced by the CNN extractor.

However, confusion between classes particularly between Calculus (label 0) and Gingivitis (label 2) remained a challenge. This limitation is visually apparent in the scatter plot representation Figure 3, where considerable overlap between these two classes indicates that their extracted features share similar distributions. These findings emphasize the critical importance of selecting classifiers that are well-suited to the underlying structure of the feature vectors, in order to achieve optimal performance in multiclass dental image classification.

## *H. Cross Validation*

To ensure the reliability and generalizability of the classification models, 10-Fold Cross Validation was applied during the training phase. This technique divides the dataset into five equal subsets, iteratively training the model on four subsets while validating on the remaining one. By averaging performance across all folds, this approach reduces the risk of overfitting and provides a more robust estimation of model effectiveness, especially when working with limited or imbalanced data. The detailed results of cross-validation for each algorithm are presented in Table 4.

TABLE IV
10-FOLD CROSS VALIDATION COMPARISON

| **Model** | **10-Fold Cross Validation** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Std Deviation** | **Precision** | **Std Deviation** | **Recall** | **Std Deviation** | **F1-Score** | **Std Deviation** |
| Logistic Regression | 0.8899 | 0.0094 | 0.8910 | 0.0096 | 0.8899 | 0.0094 | 0.8903 | 0.0095 |
| Decision Tree | 0.6832 | 0.0157 | 0.6840 | 0.0145 | 0.6831 | 0.0157 | 0.6831 | 0.0151 |
| Random Forest | 0.8805 | 0.0078 | 0.8805 | 0.0080 | 0.8805 | 0.0078 | 0.8803 | 0.0080 |
| Gradient Boosting | 0.8677 | 0.0085 | 0.8687 | 0.0083 | 0.8677 | 0.0085 | 0.8679 | 0.0085 |
| Support Vector Machine | **0.9279** | **0.0083** | **0.9284** | **0.0081** | **0.9279** | **0.0084** | **0.9280** | **0.0082** |
| AdaBoost | 0.6726 | 0.0125 | 0.6815 | 0.0131 | 0.6726 | 0.0124 | 0.6738 | 0.0130 |
| Naive Bayes | 0.7170 | 0.0123 | 0.7391 | 0.0097 | 0.7170 | 0.0124 | 0.7079 | 0.0123 |
| K-Nearest Neighbors | 0.9124 | 0.0115 | 0.9140 | 0.0111 | 0.9124 | 0.0115 | 0.9128 | 0.0114 |
| Ridge Classifier | 0.9033 | 0.0076 | 0.9038 | 0.0069 | 0.9033 | 0.0077 | 0.9034 | 0.0073 |
| **XGBoost** | **0.9103** | **0.0090** | **0.9109** | **0.0087** | **0.9103** | **0.0090** | **0.9103** | **0.0091** |
| **LightGBM** | **0.9093** | **0.0101** | **0.9103** | **0.0098** | **0.9093** | **0.0101** | **0.9095** | **0.0101** |

When compared to the initial evaluation results (Table II), several models experienced performance shifts during 10-fold cross-validation, highlighting the influence of data variability on model generalization. Models such as Logistic Regression, Naïve Bayes, AdaBoost, and particularly K-Nearest Neighbors (KNN) exhibited improved or stable performance during cross-validation. These models generally have simpler learning mechanisms and are less prone to overfitting, allowing them to benefit from exposure to more diverse subsets of the dataset. In the case of KNN, its performance surpassed LightGBM, which had previously achieved a higher accuracy in the initial evaluation. This improvement likely stems from KNN's instance-based learning approach, which becomes more effective when trained across all folds of the data, allowing it to better capture local structure and generalizable decision boundaries that were previously underrepresented.

Conversely, some high-performing models in the initial test set such as LightGBM, XGBoost, Gradient Boosting, and even SVM experienced slight declines in performance during cross-validation. This decline indicates that these models, although highly effective when trained on a fixed data split, might exhibit reduced consistency when exposed to varying subsets of the dataset. Their complex architectures, especially in ensemble methods and SVM, tend to adapt strongly to specific data characteristics, which can result in fluctuating outcomes when the data distribution changes. Nonetheless, SVM maintained the highest overall performance with minimal standard deviation, reinforcing its robustness and reliability across different data partitions.

Therefore, 10-fold cross-validation proves essential not only for evaluating model accuracy but also for assessing the stability and generalization ability of classifiers in real-world applications. Models that perform consistently across folds such as SVM, KNN, and Ridge Classifier demonstrate robustness and able to achieve generalization towards the classification.

## IV. Conclusion

This study demonstrated the effectiveness of using CNN-extracted features, specifically from EfficientNetB0, as input for supervised machine learning algorithms in multiclass classification of dental conditions. By utilizing the EfficientNetB0 model as a deep feature extractor, the system successfully captured intricate spatial patterns from intraoral images, resulting in high-dimensional representations suitable for various traditional classifiers. Among the evaluated models, Support Vector Machine (SVM) achieved the highest accuracy of 92.9%, followed closely by XGBoost and LightGBM, both attaining 91.3%. These findings affirm the robustness of the hybrid architecture in supporting automated dental diagnostics.

To enhance the performance, multiple data preprocessing techniques were employed. Data augmentation introduced variability into underrepresented classes, while SMOTE effectively addressed class imbalance by synthesizing new feature vectors for minority categories. Moreover, normalization using standard scaling ensured consistent feature ranges across models. These steps collectively contributed to improved recall, precision, and F1-scores for all classes, particularly in previous underperforming categories like gingivitis and calculus. K-Fold Cross Validation also employed to the algorithms which aim to enhance the previous training result. The results shows KNN algorithm is able to improve which imply its capability of achieving generalization.

In summary, this research validates the potential of integrating deep learning feature extractors with traditional classifiers for reliable, scalable, and efficient dental disease classification. Although the results are promising, future work should consider expanding the class scope with more balanced and diverse datasets, incorporating end-to-end deep learning pipelines, and deploying the model as a real-time diagnostic tool via mobile or cloud platforms. Such advancements would significantly enhance the practical applicability of AI in dentistry and promote equitable access to oral health services globally.

## References

[1] J. M. Cherian, N. Kurian, K. G. Varghese, and H. A. Thomas, "World Health Organization's global oral health status report: Paediatric dentistry in the spotlight," *J Paediatr Child Health*, vol. 59, no. 7, pp. 925–926, Jul. 2023, doi: 10.1111/jpc.16427.

[2] M. Abdelaziz, "Detection, Diagnosis, and Monitoring of Early Caries: The Future of Individualized Dental Care," *Diagnostics*, vol. 13, no. 24, p. 3649, Dec. 2023, doi: 10.3390/diagnostics13243649.

[3] A. R. Kareem and A. M. Alwaheb, "The Impact of the Socioeconomic Status(SES) on the Oral Health Status Among 15 Year-Old School Adolescents In Kerbala City/Iraq," *Bionatura*, vol. 8, no. CSS 1, pp. 1–8, Aug. 2023, doi: 10.21931/RB/CSS/2023.08.01.64.

[4] A. R. Kareem, A. M. Alwaheb, and N. F. Abdulhameed, "Dental caries and gingival health condition among secondary school adolescents in relation to the nutritional status in Kerbala City, Iraq," *Journal of Baghdad College of Dentistry*, vol. 36, no. 4, pp. 1–6, Dec. 2024, doi: 10.26477/jbcd.v36i4.3817.

[5] T. Chaiboonyarak, S. Chantarangsu, P. Gavila, M. Lao-Araya, N. Suratannon, and T. Porntaveetus, "Orodental health status of patients with inborn errors of immunity," *Int J Paediatr Dent*, vol. 34, no. 4, pp. 453–463, Jul. 2024, doi: 10.1111/ipd.13146.

[6] A. Sande, A. Mathur, R. Sapkal, and A. Tamboli, "Applications of AI-based Deep Learning Models for Detecting Dental Caries on Intraoral Images – A Systematic Review," *J Neonatal Surg*, vol. 14, no. 4S, pp. 523–533, Feb. 2025, doi: 10.52783/jns.v14.1827.

[7] S. Negi *et al.*, "Artificial Intelligence in Dental Caries Diagnosis and Detection: An Umbrella Review," *Clin Exp Dent Res*, vol. 10, no. 4, Aug. 2024, doi: 10.1002/cre2.70004.

[8] A. A. de Magalhães and A. T. Santos, "Advancements in Diagnostic Methods and Imaging Technologies in Dentistry: A Literature Review of Emerging Approaches," *J Clin Med*, vol. 14, no. 4, p. 1277, Feb. 2025, doi: 10.3390/jcm14041277.

[9] P. Kaushik and S. Khurana, "OralGuard: Harnessing Inception-ResNet-v2 for Cutting-Edge Oral Health Diagnostics," in *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, IEEE, Oct. 2024, pp. 882–887. doi: 10.1109/ICSSAS64001.2024.10760600.

[10] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artif Intell Rev*, vol. 57, no. 1, p. 11, Jan. 2024, doi: 10.1007/s10462-023-10631-z.

[11] A. S. Baquhaizel, M. Boumeddane, H. Gherram, and B. Alshaqaqi, "Enhancing Dental Caries Classification Through A VGG16-Based Transfer Learning," in *2023 International Conference on Electrical Engineering and Advanced Technology (ICEEAT)*, IEEE, Nov. 2023, pp. 1–4. doi: 10.1109/ICEEAT60471.2023.10426454.

[12] A. N, A. B, and V. P. S, "Study on Detecting the Teeth and Classifying the Teeth Structure using Machine Learning and CNN," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2024, pp. 1–7. doi: 10.1109/I2CT61223.2024.10543395.

[13] S. A. Shifani, M. S. Franklin Thamil Selvi, M. D. Suresh, M. Paramaiyappan, J. Giri, and M. Kanan, "An Automated Cavity Level Prediction based on Dental Imaging Sensors by using Enhanced AI Assisted Learning Principles," in *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2024, pp. 1233–1239. doi: 10.1109/ICECA63461.2024.10801154.

[14] S.-T. Hsieh and Y.-A. Cheng, "Multimodal feature fusion in deep learning for comprehensive dental condition classification," *Journal of X-Ray Science and Technology: Clinical Applications of Diagnosis and Therapeutics*, vol. 32, no. 2, pp. 303–321, Mar. 2024, doi: 10.3233/XST-230271.

[15] N. Sigeef, "An Oversampling Algorithm combining SMOTE and RF for Imbalanced Medical Data," *Int J Res Appl Sci Eng Technol*, vol. 11, no. 6, pp. 2429–2434, Jun. 2023, doi: 10.22214/ijraset.2023.54074.

[16] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, Oct. 2024, doi: 10.25126/jtiik.2024117989.

[17] M. Z. Hussain, S. Gupta, B. Hambarde, P. Parkhi, and Z. Karimov, "Multiclass Classification of Oral Diseases Using Deep Learning Models," in *The Impact of Algorithmic Technologies on Healthcare*, Wiley, 2025, pp. 189–207. doi: 10.1002/9781394305490.ch10.