STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–18.
Published online in International Academic Press (www.IAPress.org)

An Interpretable Deep Learning Framework for Multi-Class Dental Disease Classification from Intraoral RGB Images

Dawlat Abdulkarim Ali [1,2], Haval Tariq Sadeeq [2,*]

[1]*Information Technology Department, Technical College of Informatics, Akre University for Applied Science, Kurdistan Region, Iraq*
[2]*Artificial Intelligence Department, Technical College of Duhok, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq*

Abstract Dental anomalies and diseases are among the most prevalent health concerns world-wide, and their early and precise diagnosis is critical to ensuring effective treatment and improved patient outcomes. Traditional diagnostic approaches, particularly conventional radiography, are often time-consuming and may not provide sufficient diagnostic accuracy. To address these limitations, this study proposes a robust deep learning framework for the automated classification of dental conditions from intraoral RGB images. Three publicly available datasets—Oral Diseases (six classes), Oral Infection (six classes), and Teeth Dataset (five classes)—covering a broad spectrum of dental anomalies and exhibiting notable class imbalance were utilized. Five state-of-the-art convolutional neural network (CNN) architectures, namely Efficient-NetB3, EfficientNetB0, ResNet50, DenseNet121, and InceptionV3, were systematically evaluated using a unified transfer learning pipeline. Techniques such as stratified 5-fold cross-validation, ensemble inference, focal loss, class weighting, and label smoothing were employed to enhance generalization and mitigate class imbalance. EfficientNetB3 emerged as the optimal model, achieving accuracies of 95.4%, 89.9%, and 99.3% on the three datasets, with Kappa values reaching 0.989. Grad-CAM visualizations confirmed clinically meaningful feature localization, strengthening interpretability. The proposed framework demonstrates strong potential for integration into intelligent clinical decision-support systems, offering an optimal balance between diagnostic accuracy, computational efficiency, and transparency to assist dental practitioners in timely and reliable decision-making.

Keywords Dental Disease Classification, Medical Imaging, Intraoral Imaging, Explainable AI (XAI), Deep Learning, EfficientNetB3, Transfer Learning

DOI: 10.19139/soic-2310-5070-2880

1. Introduction

Oral health is a fundamental component of overall health and well-being; however, it continues to be deprioritized globally, particularly in low- and middle-income countries. Oral and dental conditions such as dental caries, gingivitis, hypodontia, tooth staining, and calculus deposition affect more than 3.5 billion individuals worldwide. When left untreated, these conditions may lead to pain, infection, tooth loss, and even systemic complications, thereby exerting a significant negative impact on quality of life [1, 2].

Conventional diagnostic methods are based on the manual examination or interpretation of radiographic reports [3, 4]. However, these methods suffer from subjectivity, intra-observer variability, and potential diagnostic errors [5, 6]. In contrast, RGB intraoral images can be obtained quickly and non-invasively without radiation exposure, making them highly suitable for routine examinations, early disease screening, and telehealth applications. Their ease of acquisition and cost-effectiveness further support their integration into clinical workflows and large-scale

*Correspondence to: Haval Tariq Sadeeq (Email: haval.tariq@dpu.edu.krd). Artificial Intelligence Department, Technical College of Duhok, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

dental care programs [7]. The development of artificial intelligence in dental field, especially deep learning (DL), has the potential to revolutionize dental diagnostics by means of automation, enhancement of accuracy, and acceleration of clinical decision-making process [8].

Among the most effective subsets of DL models are convolutional neural networks (CNNs), which are particularly valuable in medical image analysis due to their robust feature extraction capabilities. In dentistry, CNNs have demonstrated notable success in applications such as tooth structure segmentation [9], detection of caries and perio-dontal diseases [10], and interpretation of complex radiographs [11]. Despite these advances, the performance of automated classification systems continues to face significant challenges, including imbalanced class distributions, subtle inter-class visual differences, and variations in image quality [12].

In order to address these issues, researchers apply various methodologies such as transfer learning, data augmentation and multi-modal feature fusion. Pretrained CNN models like ResNet, DenseNet, and EfficientNet have achieved good predictive performances in dental imaging tasks without much labeled data [13, 14]. Furthermore, it was shown that the hybrid systems that embed the CNN-based feature extraction and the conventional classifiers such as SVM or the ensemble techniques (e.g., weighted XGBoost) have achieved better classification accuracy and robustness for dental radio-graph analysis [15, 16].

Beyond clinical and technical performance, deep learning methods hold significant potential for expanding access to dental care, particularly in underserved and remote regions. When effectively integrated into clinical protocols, AI-based systems can facilitate early disease detection, improve diagnostic consistency, and help reduce disparities in the delivery of oral health services [17].

However, the majority of previous studies on dental image classification only focus on binary or small-scale problems and they lack the same benchmark dataset, inter-pretability and practicality of the results. For example, Ikhwani et al. (2024) [18] applied transfer learning models based on sub-regions of dental datasets, but did not cover multi-class classification, model interpretability or ensemble prediction techniques.

In parallel with the advancement of predictive performance, the interpretability of artificial intelligence systems has emerged as a critical factor for their adoption in clinical practice. Explainable AI (XAI) provides transparency into the decision-making processes of complex deep learning models, allowing clinicians to understand the rationale behind predictions and to build trust in automated diagnostic tools [19, 20]. In dental imaging, XAI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) enable the visualization of discriminative regions in intraoral images, thereby aligning algorithmic outputs with clinically relevant features. This interpretability reduces the risk of misdiagnosis and facilitates the integration of AI systems into evidence-based dentistry [21].

In this paper, we introduce a fully end-to-end deep learning architecture for multi-class classification of dental anomalies and diseases from intraoral RGB images. Our contributions include:

• A comprehensive benchmarking of five state-of-the-art deep learning models was conducted on three diverse dental image datasets, leading to the identification of EfficientNetB3 as the optimal architecture;

• A comprehensive end-to-end pipeline was developed, integrating focal loss, mixed-precision training, and stratified 5-fold cross-validation to ensure robust and reliable learning;

• Ensemble predictions were used to enhance result stability, and Grad-CAM visualizations were applied to EfficientNetB3 to ensure clinical interpretability;

• EfficientNetB3 model achieved high classification accuracies: 95.4% (Oral Diseases), 89.9% (Oral Infection), and 99.3% (Teeth Dataset);

• The proposed framework provides a scalable, interpretable, and clinically applicable AI-based dental diagnostics system.

The structure of the paper is organized as follows: Section 2 reviews recent advancements in deep learning methodologies applied to dental diagnostics. Section 3 describes the proposed methodology in detail, including datasets, image preprocessing procedures, CNN model architecture, and experimental protocols. Section 4 presents experimental results, evaluating model performance across various dental conditions. Section 5 discusses the obtained results in the context of existing literature. Finally, Section 7 concludes the study, highlighting the practical implications and outlining future perspectives for AI integration in dental care.

2. Related Works

Recent advancements in deep learning have made substantial contributions to the field of dental diagnostics, especially in the context of image-based classification of diseases. A number of works have investigated the use of CNNs and hybrid deep learning models to automate dental anomaly and diseases detection.

Zhu et al. (2023) [6] developed a convolutional neural network (CNN)-based model for the detection of multiple dental diseases using panoramic radiographs. Their study demonstrated the feasibility of employing CNNs within radiology-oriented dental workflows, where panoramic imaging is often used for initial screening. By processing a wide range of oral conditions within a single framework, they showed that CNNs could support clinical decision-making. However, the approach relied primarily on 2D panoramic images and lacked cross-validation and interpretability mechanisms, which limits its reproducibility compared to our study that emphasizes robust validation and model explainability.

Pang et al. (2025) [7] proposed a CNN-based grading system for assessing tooth wear severity using RGB intraoral photographs. This method allowed the detection of subtle dental changes, such as early enamel wear, that are often overlooked in routine visual inspections. Their work highlighted the complementary role of intraoral RGB imaging in capturing surface-level deterioration alongside radiographic workflows and demonstrated that CNNs can assist clinicians in tracking progressive tooth wear. Nonetheless, their model was designed for a single anomaly type rather than a general multiclass disease classification framework, which is the focus of our research.

Li et al. (2021) [9] introduced an automatic gingivitis screening model that combined lesion localization with CNN-based classification. This two-step approach first identified inflamed gingival areas and then classified their condition, offering an in-terpretable framework for periodontal diagnosis. Clinically, this method is valuable because it supports early detection of gum disease that could be missed without magnified inspection. While the study confirmed the potential of deep learning for targeted periodontal detection, it did not explore comprehensive anomaly classification across multiple dental conditions, which our framework addresses.

Bhat et al. (2024) [11] designed a task-specific CNN architecture to improve the detection of dental caries and structural anomalies in panoramic radiographs. Their results showed that customized CNNs optimized for a particular imaging modality can outperform generic pre-trained networks in certain diagnostic scenarios. By focusing on dental caries and related structural irregularities, the study demonstrated the value of architecture optimization and domain-specific training. However, it primarily addressed binary or limited-class tasks and did not employ cross-dataset validation or ensemble techniques, whereas our work provides a multiclass, cross-validated, and ensemble-enhanced approach.

Alsakar et al. (2024) [14] developed a multi-label dental disorder classification framework that integrated MobileNetV2 with a Swin Transformer in a bagging ensemble. By combining convolutional and transformer-based feature extractors, their approach improved performance in handling complex, multi-label dental images. This work demonstrated the potential of hybrid CNN-Transformer pipelines to enhance model generalization. However, the complexity of such ensembles may limit their scalability in real clinical settings, while our framework remains lightweight and clinically deployable.

Li (2024) [15] introduced DA-Net, a classification-guided CNN tailored for detecting subtle dental anomalies in maxillofacial images. By guiding feature extraction with classification signals, DA-Net improved sensitivity to early pathological changes that could otherwise go unnoticed. This study emphasized the importance of designing network architectures specific to dental imaging to achieve higher diagnostic precision. Our work builds on this principle but extends it to a comprehensive, multi-class classification framework validated across multiple intraoral datasets.

Parkhi et al. (2025) [17] proposed a hybrid architecture using CNN, ResNet, Vision Transformers (ViT) with confidence weighting. Their model had an accuracy of 87.6% and was designed for clinical use. Their results confirm the potential of deep learning for dental workflows.

Ikhwani et al. (2024) [18] conducted a comparative study using the same three datasets employed in this work: Oral Diseases, Oral Infection, and Teeth. Their research benchmarked ResNet50V2 and EfficientNet, achieving classification accuracies ranging from 81.6% to 82.5%. However, their approach did not incorporate ensemble

learning, cross-validation, or interpretability techniques such as Grad-CAM, which are integral components of the present study.

Hassanein et al. (2024) [22] utilized SMOTE to solve class imbalance in dental data and examined different efficient net versions from B3 to B7. They found that EfficientB5 yielded best performance (F1-score: 0.87, precision: 0.90). These results underscore the importance of model scaling and enhanced training methodologies, a strategy consistent with ours that applies focal loss, label smoothing, and class weights to address such tasks.

Hsieh and Cheng (2024) [23] developed a multimodal feature fusion-based approach, using the deep models (i.e., EfficientNetB0, MobileNetV2, ResNet50, InceptionV3 and ResNet101) for feature extraction and a conventional SVM classifier was used for categorization. Their method reported 92.5% accuracy and 0.909 in Kappa index. While their approach is distinct in that it decouples feature extraction from classification, it is in line with our concentration on CNN-based robustness and generality.

Jian Liu et al. (2024) [24] presented YoCNET, a hybrid model combining YOLOv5 for automatic tooth segmentation and ConvNeXt for periapical lesion classification. Their model performance (AUC = 0.9757) indicated the influence of architectural adjustment for the task-specific model design. Although they worked on a binary diagnostic task in radiographic images, in this work we aim for a more general multiclass classification in RGB intraoral images.

Hussain et al. (2023) [25] proposed an ensemble classification structure to identify oral diseases from RGB images as a fusion of VGG16, MobileNet, and InceptionV3. Their system took advantage of the weighted ensemble prediction and reached 97% diagnostic accuracy, indicating that lightweight convolutional models can be used for accurate dental diagnosis. Instead, we compare more CNNs and focus on the interpretability of them and the ensemble performance on the cross-validated level.

Razmjouei et al. (2025) [26] presented NFR-EDL, a fuzzy rank-based ensemble deep learning model which combines VGG16, ResNet50, DenseNet169, and SqueezeNet, through non-linear fuzzy aggregation, for identification of oral and dental diseases from RGB intraoral images. It attained state-of-the-art accuracy on several public datasets (in the range 90.92%-97.08%), proving the power of ensemble models by uncertainty-aware decision fusion. In contrast, our work achieves similar diagnosis performance with a single model, EfficientNetB3, with an ensemble averaging scheme, stratified cross-validation, and Grad-CAM interpretability — representing an easier, scalable and more clinically interpretable option.

Although there has been substantial progress in recent works, most approaches presented in literature are either based on ensemble architectures which have a high computational cost or do not provide rigorous interpretability mechanisms. In addition, cross-dataset evaluation and end-to-end reproducibility are overlooked. In this paper we address these limitations by providing a unified framework that combine EfficientNetB3, robust cross-validation, and lightweight ensemble prediction's technique.

3. Materials and Methods

This section outlines the datasets, data preprocessing steps, model configuration, training strategies, evaluation metrics, and inference process on which the proposed method is based. It has been designed to provide an end-to-end, robust, stable, and reproducible feature extraction framework suitable for detecting dental abnormalities and diseases across diverse intraoral RGB image datasets.

3.1. The Data

Three publicly available dental image datasets obtained from Kaggle: Oral Diseases, Oral Infection, and Teeth Dataset, containing clinically obtained RGB intraoral images that portray various dental conditions were used in this study. The data sets were selected to mimic a real-world diagnostic challenge, in order to assist the development of an effective multi-class classification model.

• Oral Diseases Dataset: Comprising 11,653 annotated RGB images, this dataset includes six diagnostic categories: calculus, caries, discoloration, gingivitis, hypodontia, and ulcers. It was obtained from Kaggle https://www.kaggle.com/datasets/salmansajid05/oral-diseases [27].

• Oral Infection Dataset: Consisting of 5,563 RGB images across six conditions: caries, gingivitis, ulcers, oral cancer, discoloration, and hypodontia. It was retrieved from Kaggle https://www.kaggle.com/datasets/sizlingdhairya1/oral-infection [28].

• Teeth Dataset: Contains 5,048 RGB images categorized into five classes: caries, calculus, ulcers, discoloration, and hypodontia. It was obtained from Kaggle https://www.kaggle.com/datasets/rajapriyanshu/teeth-dataset [29].

The visual characteristics of each class, including color, shape, severity, and lesion location, were distinct and thus well-suited for CNN-based classification. Representative examples are shown in Figure 1. Unreadable or corrupted images were excluded using the Python Imaging Library (PIL), and non-image files were removed. Class labels were automatically extracted from the directory structure, converted to numeric form, and one-hot encoded using Keras utilities.

Figure 1. Representative RGB images from the Oral Diseases dataset: (A) Calculus, (B) Caries, (C) Discoloration, (D) Gingivitis, (E) Hypodontia, (F) Mouth ulcers.

3.2. Flowchart of Research

The structured workflow for the classification of dental anomalies and diseases based on deep learning presented in this study is shown in Figure 2. The pipeline was executed independently for each of the three dental image datasets. The steps are summarized as follows:

Step 1: Dataset Selection

The loading of one of the three available image datasets (Oral Diseases – Oral Infection – Teeth Dataset) initiates this procedure. These example color images are of practical clinical cases from different pathological categories suitable for training and testing of deep learning models.

Step 2: Data Preprocessing Pipeline

Before they are fed into neural networks, all images are resized to the resolution of 256×256 pixels for input size consistent. Pixel values are normalized to [0, 1] so as to regularize and speed up the training. This standardized pre-processing allows all models and data set to be consistent.

Step 3: CNN Model Training

Five pretrained CNNs (EfficientNetB3, EfficientNetB0, ResNet50, DenseNet121, and InceptionV3) are individually fine-tuned on the dental datasets. The preprocessed images are used to train each model independently.

Step 4: Training Optimization Techniques

To enhance classification performance and mitigate data imbalance, several optimization techniques were employed during training. These included Focal Loss, which assigns greater weight to misclassified examples; Class Weights, which address underrepresented classes; and Label Smoothing, which reduces model overconfidence and improves generalization.

Step 5: Cross-Validation and Evaluation

Each model was evaluated using stratified 5-fold cross-validation, which provides a more reliable performance assessment and reduces the risk of overfitting compared to standard k-fold cross-validation. For each fold, performance metrics including Accuracy, F1-Score, Cohen's Kappa, and ROC–AUC (class-wise, micro-average, and macro-average) were computed. This rigorous evaluation ensures fair model comparison and robust generalization.

Step 6: Ensemble Prediction Generation

The predictions obtained from the five folds were aggregated by averaging the softmax probabilities. This ensemble approach reduces variance and enhances generalization, leading to more stable and accurate final predictions.

Step 7: Ensemble-Based Final Evaluation

Final performance curves and classification metrics were reported by aggregating results across all folds. The ensemble of fold-specific predictions was retained for evaluation, ensuring stability and minimizing variance.

Step 8: Dataset-Wise Repetition

The complete process, from loading the data up to the selection of the best model, is iterated for each dataset separately. This approach ensured dataset-specific tuning and testing, thereby enabling a fair and reliable comparison of cross-dataset performance.

Figure 2. Workflow for dataset-specific dental disease classification using multiple CNN architectures and ensemble evaluation.

3.3. Image Preprocessing

Standard preprocessing reprocessing was performed for all data to keep consistency and achieve best-performance through CNN models. The following pipeline was implemented:

• Resizing: All images were resized to 256×256×3 pixels to keep the same image size for every dataset. • Normalization: Images were standardized using the standard ImageNet mean and standard deviation ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]).

• Data Augmentation with Albumentations: Real-time data handling and augmentation were implemented using a custom generator based on tf.keras.utils.Sequence, which managed image loading, resizing, and label encoding. To enhance generalization and robustness, augmentations were applied using the Albumentations library. The augmentation pipeline included random rotations (±10°), scaling (±5%), translations (±5%), and horizontal flipping (50% probability). Additional adjustments included gamma correction for contrast enhancement, variations in color, saturation, and brightness to simulate illumination changes, the addition of Gaussian noise to model imaging noise, and Coarse Dropout, which randomly masked a rectangular patch of up to 32×32 pixels (30% probability) to simulate partial occlusions.

These improvements have resulted in a better capability of dealing with variance and a lower overfitting risk. Such augmentations are intended to mimic the natural intraclass variability (different position shifts, different lighting conditions, image occlusions) and help to boost the performance and generalization of the model to real clinical scenarios.

3.4. Transfer Learning of CNNs

Transfer learning was applied using five pretrained CNN architectures — Effi-cientNetB3, EfficientNetB0, ResNet50, DenseNet121, and InceptionV3 — all initialized with ImageNet weights. The original classification head of each model was replaced with a customized architecture comprising a Global Average Pooling (GAP) layer, followed by two fully connected layers with 256 and 128 neurons, respectively, each activated by ReLU and regularized using an L2 penalty. Dropout layers with a rate of 0.5 were inserted after each dense layer to mitigate overfitting. The final layer consisted of a softmax classifier, outputting multi-class probabilities corresponding to the dental conditions.

Transfer learning is particularly useful in medical imaging applications where annotated data is scarce. It is based on pretrained model to achieve faster convergence and better generalization on small datasets.

Focal Loss was employed to cope with class imbalance and further emphasize hard-to-classify samples. a hybrid loss function was employed, combining Focal Loss with smoothed Categorical Cross-Entropy with a weighting parameter $\gamma = 2.0$ and balancing factor $\alpha = 0.25$. The loss is defined as:

$$FL(p_t) = \alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where p_t denotes the model's estimated probability for the true class.

Label smoothing ($\varepsilon = 0.1$) was also employed to mitigate overconfidence and further enhance generalization. The final loss was computed as a weighted sum of 70% focal loss and 30% smoothed cross-entropy as follows:

$$\text{Loss}_{\text{total}} = 0.7 \times \text{FocalLoss} + 0.3 \times \text{SmoothedCrossEntropy} \tag{2}$$

This weighted formulation balances the emphasis on hard-to-classify samples with the regularization benefits of label smoothing, thereby improving model robustness and stability. To further mitigate the effects of class imbalance, class weights were computed using the *compute_class_weight* function from scikit-learn, according to the following formula:

$$w_i = \frac{N}{k \times n_i} \tag{3}$$

where w_i denotes the weight for class i, N represents the total number of samples, k is the number of classes, and n_i indicates the number of samples in class i.

The efficiency of such a combined strategy has been proved in previous researches. For example, class-weighted focal loss was applied to enhance detection results in rare categories, while class-dependent weights in CNN training have shown to improve F1-scores on highly imbalanced datasets [30, 31, 32].

Training was performed using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 16, and 40 epochs. To enhance training efficiency, several callback functions were employed: `ModelCheckpoint` to save the best-performing model, `ReduceLROnPlateau` to lower the learning rate upon performance stagnation, and `EarlyStopping` to terminate training after five consecutive epochs without improvement.

This carefully designed training pipeline, which integrated transfer learning, Focal Loss, label smoothing, class weighting, and advanced data augmentation, was intended to ensure robustness, mitigate class imbalance, and enhance the generalization of the model across diverse dental conditions.

3.5. Validation of Performance

Model evaluation was conducted using both quantitative metrics and interpretability tools to ensure a comprehensive assessment of performance.

Quantitative Metrics:

• Accuracy: Proportion of correctly classified samples relative to the total number of samples.

• Precision, Recall, and F1-Score: Evaluated per class to measure detection quality.

• Cohen's Kappa: A statistical measure of agreement between predicted and true labels, adjusted for chance.

• Receiver Operating Characteristic (ROC) Curves and Area Under the Curve (AUC): Class-wise ROC analysis was conducted to evaluate discrimination, with micro-average providing a global measure across classes and macro-average reflecting balanced performance. Corresponding AUC values were reported for comprehensive assessment.

Interpretability and Visual Analysis:

• Confusion Matrix: Used to identify patterns of misclassification across classes.

• Grad-CAM: Applied to highlight the regions of each image most influential in the model's decision-making process.

• Training Curves: Depicted the learning dynamics across epochs, providing insights into model convergence and potential overfitting.

Cross validation was used to identify the best performing model, which was subsequently evaluated on the entire dataset. Final models were saved in Keras format, and all results were logged to ensure reproducibility. Model development was carried out using TensorFlow and Keras, with Albumentations and OpenCV applied for preprocessing and augmentation, while metric computations were performed using scikit-learn.

4. Results

The primary objective of this study was to evaluate the performance of CNNs in classifying dental diseases using three RGB image datasets: Oral Diseases, Oral Infection, and Teeth. The models assessed included ResNet50, DenseNet121, InceptionV3, EfficientNetB3, and EfficientNetB0.

All models were trained in a single pipeline, using transfer learning from ImageNet weights. The training was fine-tuned using the Adam optimizer and evaluated with stratified 5-fold cross-validation to ensure robustness and prevent overfitting. To enhance prediction stability, ensemble predictions were generated by averaging the softmax outputs across the 5-fold-specific models. Performance was evaluated using both global and per-class metrics, including Accuracy, Precision, Recall, F1-Score, AUC, and Cohen's Kappa. We present the results for each dataset separately in the next subsections.

4.1. Evaluation on the Oral Diseases Dataset

Table1 summarizes the overall performance of the evaluated CNN models on the Oral Diseases dataset. All five architectures - EfficientNetB3, EfficientNetB0, InceptionV3, ResNet50, and DenseNet121 - delivered consistent and near-identical outcomes, achieving an overall accuracy of 95.4%, Cohen's Kappa values above 0.94, and macro-F1 scores of approximately 0.945. AUC values exceeded 0.996, confirming excellent discriminative capability across all classes. Among them, EfficientNetB3 and DenseNet121 achieved the highest Kappa values (0.943), reflecting stronger agreement with the ground truth.

Detailed per-class results in Table2 show that all models performed nearly perfectly in detecting caries, discoloration, hypodontia, and mouth ulcers, with F1-scores approaching or reaching 0.99-1.00. In particular, caries and mouth ulcers were classified with the highest consistency, achieving almost flawless detection across all models. However, calculus and gingivitis remained more challenging, with F1-scores ranging between 0.79 and 0.88. For these classes, DenseNet121 exhibited slightly better balance between precision and recall compared with the other models.

4.2. Evaluation on the Oral Infection Dataset

Table3 presents the overall results for the CNN models on the Oral Infection dataset. All models achieved comparable accuracy of approximately 89.7–89.9%, with Cohen's Kappa values ranging from 0.838 to 0.849 and macro-F1 scores between 0.930 and 0.939. The AUC values were uniformly high (> 0.987), indicating strong discrimi-natory power. Among the evaluated architectures, EfficientNetB3 reported the highest macro-F1 (0.939), while DenseNet121 achieved the highest Kappa (0.849) and the best AUC (0.9875), highlighting them as the most reliable performers for this dataset.

Detailed class-wise performance metrics are shown in Table4. Consistent with findings from the Oral Diseases dataset, caries, discoloration, hypodontia, and mouth ulcers were classified with excellent precision and recall, yielding F1-scores close to or exceeding 0.95. Mouth ulcers in particular were detected flawlessly across all models (F1 =1.000). In contrast, calculus and gingivitis remained more challenging classes. Calculus exhibited the lowest F1-scores (0.771-0.803), reflecting the greatest difficulty in achieving both high precision and recall, while gingivitis produced moderate F1-scores (0.882-0.886).

4.3. Evaluation on the Teeth Dataset

Table5 presents the overall results for the CNN models on the Teeth dataset. All five models demonstrated outstanding performance, with overall accuracies exceeding 99.2% and Cohen's Kappa values above 0.986, reflecting near-perfect agreement with the ground truth. The macro-F1 scores ranged from 0.976 to 0.987, while micro-F1 values were consistently high, between 0.986 and 0.991, underscoring strong balance between precision and recall across classes. AUC values approached perfection (> 0.998), confirming excellent discriminatory capacity. Among the tested models, EfficientNetB3 achieved the highest macro-F1 (0.987) and micro-F1 (0.991), while ResNet50 and DenseNet121 also performed exceptionally well with macro-F1 values above 0.985.

The per-class results in Table6 further demonstrate this robustness. Discoloration, hypodontia, and mouth ulcers were detected with near-perfect precision and recall, yielding F1-scores of 0.993–1.000 across all models. Similarly, calculus was identified with very high precision and recall, with F1-scores consistently around 0.986. Although caries was comparatively more challenging, performance remained strong, with F1-scores between 0.960 and 0.987, and DenseNet121 producing the highest class-specific F1 for caries (0.987).

Overall, these results confirm that CNN architectures are highly reliable for Teeth dataset classification, with EfficientNetB3, ResNet50, and DenseNet121 emerging as the most effective models, offering superior precision, recall, and balanced F1-scores.

Table 1. Summary Performance Metrics of CNN Models on the Oral Diseases Dataset.

CNN Model	**Accuracy**	**Kappa**	**Macro F1**	**AUC**
Efficient-NetB3	0.954	0.943	0.945	0.9964
Efficient-NetB0	0.954	0.942	0.945	0.9965
InceptionV3	0.954	0.942	0.945	0.9965
ResNet50	0.954	0.942	0.945	0.9962
DenseNet121	0.954	0.943	0.945	0.9965

Table 2. Detailed Class-wise Precision, Recall, and F1-Score of CNN Models on the Oral Diseases Dataset.

Class	Metric	Efficient-NetB3	Efficient-NetB0	InceptionV3	ResNet50	DenseNet121
Calculus	Precision	0.753	0.740	0.746	0.757	0.766
	Recall	0.863	0.883	0.869	0.846	0.833
	F1 score	0.804	0.805	0.803	0.799	0.798
Caries	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	0.995	0.995	0.995	0.995	0.995
	F1 score	0.997	0.997	0.997	0.997	0.997
Discoloration	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	0.998	0.998	0.998	0.998	0.998
	F1 score	0.999	0.999	0.999	0.999	0.999
Gingivitis	Precision	0.915	0.927	0.920	0.908	0.902
	Recall	0.851	0.839	0.846	0.861	0.868
	F1 score	0.882	0.882	0.882	0.884	0.884
Hypodontia	Precision	0.992	0.992	0.992	0.994	0.992
	Recall	0.986	0.986	0.986	0.984	0.986
	F1 score	0.989	0.989	0.989	0.989	0.989
Mouth Ulcers	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	1.000	1.000	1.000	1.000	1.000
	F1 score	1.000	1.000	1.000	1.000	1.000

Table 3. Summary Performance Metrics of CNN Models on the Oral Infection Dataset.

CNN Model	Accuracy	Kappa	Macro F1	AUC
Efficient-NetB3	0.899	0.848	0.939	0.9871
Efficient-NetB0	0.897	0.847	0.930	0.9893
InceptionV3	0.897	0.842	0.932	0.9872
ResNet50	0.897	0.838	0.934	0.9874
DenseNet121	0.897	0.849	0.931	0.9875

4.4. Explainable AI: Grad-CAM Visualization for Model Interpretability

To enhance interpretability and visualize the regions that influenced the CNN's predictions, Gradient weighted Class Activation Mapping (Grad CAM) was applied to the EfficientNetB3 model. Grad CAM was generated for a subset of samples from all three datasets: Oral Diseases, Oral Infection, and Teeth. This technique has been widely adopted in medical imaging to highlight class discriminative regions [33, 34].

The resulting heatmaps revealed that the model primarily focused on diagnostic regions such as decayed teeth, swollen gums, and sore mucosa. In correctly classified cases, activations were highly localized, suggesting strong model confidence. Conversely, misclassified examples exhibited more diffuse and ambiguous activations, particularly in visually similar categories such as calculus and gingivitis.

Although Grad CAM was applied only to the EfficientNetB3 model and limited to selected illustrative cases, the visualizations provide valuable insights into the model's decision making process and support its potential clinical applicability. Figures 3–5 present representative Grad-CAM visualizations from the datasets, highlighting the model's focus on clinically relevant regions for both correctly and incorrectly classified cases. Green labels denote correct predictions, whereas red labels indicate misclassifications.

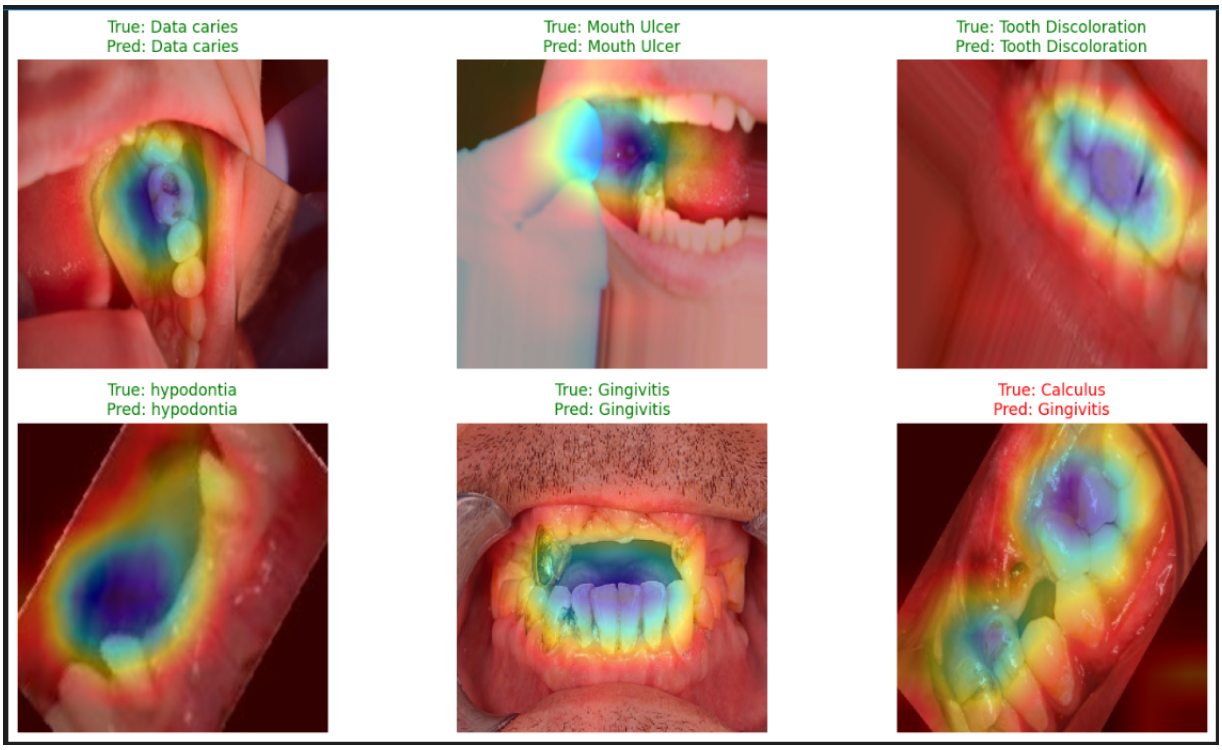

Figure 3. Grad-CAM visualization on the Oral Diseases dataset.

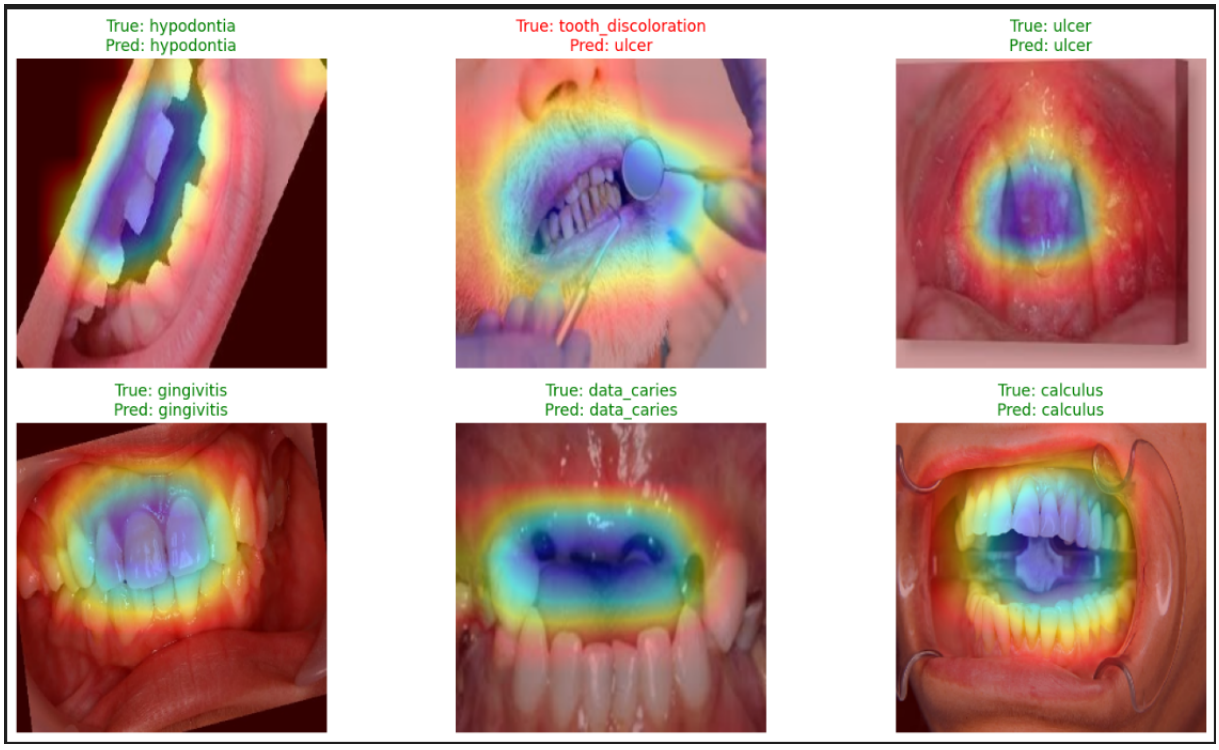

Figure 4. Grad-CAM visualization on the Oral Infection dataset.

Table 4. Detailed Class-wise Precision, Recall, and F1-Score of CNN Models on the Oral Infection Dataset.

Class	Metric	Efficient-NetB3	Efficient-NetB0	InceptionV3	ResNet50	DenseNet121
Calculus	Precision	0.753	0.734	0.746	0.757	0.752
	Recall	0.860	0.812	0.811	0.846	0.859
	F1 score	0.803	0.771	0.777	0.799	0.802
Caries	Precision	0.981	0.908	0.942	0.937	0.868
	Recall	0.954	0.986	0.968	0.945	0.991
	F1 score	0.967	0.945	0.955	0.941	0.925
Discoloration	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	0.978	0.978	0.978	0.978	0.978
	F1 score	0.989	0.989	0.989	0.989	0.989
Gingivitis	Precision	0.914	0.894	0.892	0.881	0.916
	Recall	0.854	0.874	0.880	0.888	0.850
	F1 score	0.883	0.884	0.886	0.884	0.882
Hypodontia	Precision	0.998	0.998	0.990	0.998	0.992
	Recall	0.980	0.981	0.979	0.980	0.986
	F1 score	0.989	0.989	0.984	0.989	0.989
Mouth Ulcers	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	1.000	1.000	1.000	1.000	1.000
	F1 score	1.000	1.000	1.000	1.000	1.000

Table 5. Summary Performance Metrics of CNN Models on the Teeth Dataset.

CNN Model	Accuracy	Kappa	Macro-F1	Micro-F1	AUC
Efficient-NetB3	0.993	0.990	0.987	0.991	0.998
Efficient-NetB0	0.992	0.986	0.976	0.986	0.999
InceptionV3	0.992	0.986	0.978	0.987	0.999
ResNet50	0.993	0.989	0.981	0.989	0.999
DenseNet121	0.993	0.988	0.985	0.990	0.998

Table 6. Detailed Class-wise Precision, Recall, and F1-Score of CNN Models on the Teeth Dataset.

Class	Metric	Efficient-NetB3	Efficient-NetB0	InceptionV3	ResNet50	DenseNet121
Calculus	Precision	0.978	0.982	0.976	0.979	0.981
	Recall	0.995	0.987	0.995	0.994	0.991
	F1 score	0.986	0.984	0.985	0.986	0.986
Caries	Precision	0.968	0.928	0.968	0.964	0.984
	Recall	0.973	0.995	0.959	0.977	0.991
	F1 score	0.970	0.960	0.963	0.970	0.987
Discoloration	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	0.998	0.998	0.998	0.998	0.998
	F1 score	0.999	0.999	0.999	0.999	0.999
Hypodontia	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	0.986	0.986	0.986	0.986	0.986
	F1 score	0.993	0.993	0.993	0.993	0.993
Mouth Ulcers	Precision	1.000	1.000	1.000	1.000	1.000
	Recall	1.000	1.000	1.000	1.000	1.000
	F1 score	1.000	1.000	1.000	1.000	1.000

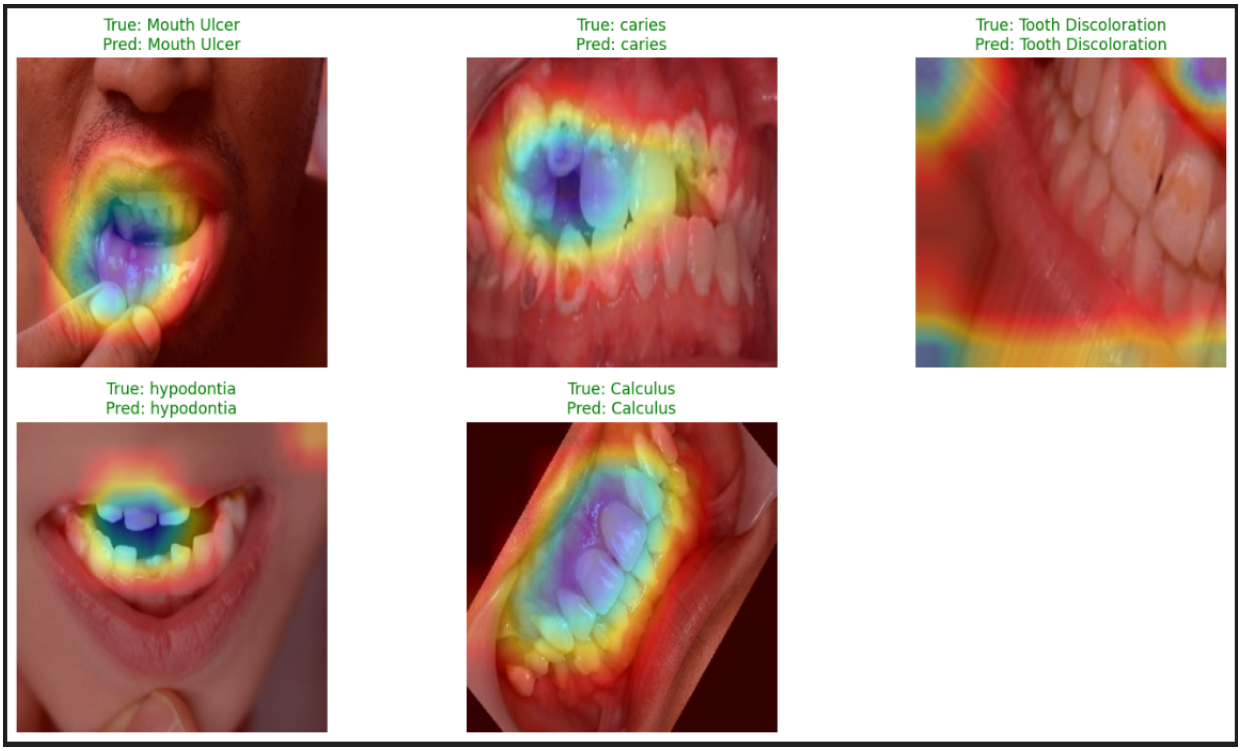

Figure 5. Grad-CAM visualization on the Teeth dataset.

4.5. Learning Curve Analysis

To further analyze the training behavior and generalization ability for the EfficientNetB3 model, learning curves were depicted for the three datasets (Oral Diseases, Oral Infection, and Teeth). Plotting these, we can see the model's average training loss and accuracy over several epochs and cross-validation folds.

As shown in Figure 6, the learning curves of the Oral Diseases dataset show steady convergence. The validation accuracy surpasses 90% and no severe overfitting is observed. It helped the model sustain a smooth and continuous loss decrease which confirms its stability in learning complex multi-class classification tasks.

For the Oral Infection dataset (Figure 7), the learning curves suggest some variance in the validation accuracy, especially after epoch 30. This indicates the existence of class skewed or visually overlapping features that are difficult to generalize. However, the trend overall shows progressive learning and decreasing loss.

Lastly, Figure 8, show learning curves for the Teeth dataset - EfficientNetB3 has an almost perfect training and validation accuracy and very low values on the end loss function. It demonstrates the stability of the model, which makes the model suitable for the simpler dental classification problem. These findings prove that the model had consistently high performance and generalization ability across datasets, which can attest to the effectiveness of our design in training pipeline, regularization, balancing, and ensemble evaluation.

4.6. Confusion Matrix Analysis

An even deeper analysis of the classification process, has been carried out by providing confusion matrices for each dataset, trained with the EfficientNetB3 model. These matrices uncovered some misclassification patterns particularly between visually similar classes. For example, it was noticed that there were confusing samples between calculus and gingivitis over several folds, which might be explained by their common textural and locational properties. There was also some overlap between caries and tooth color.

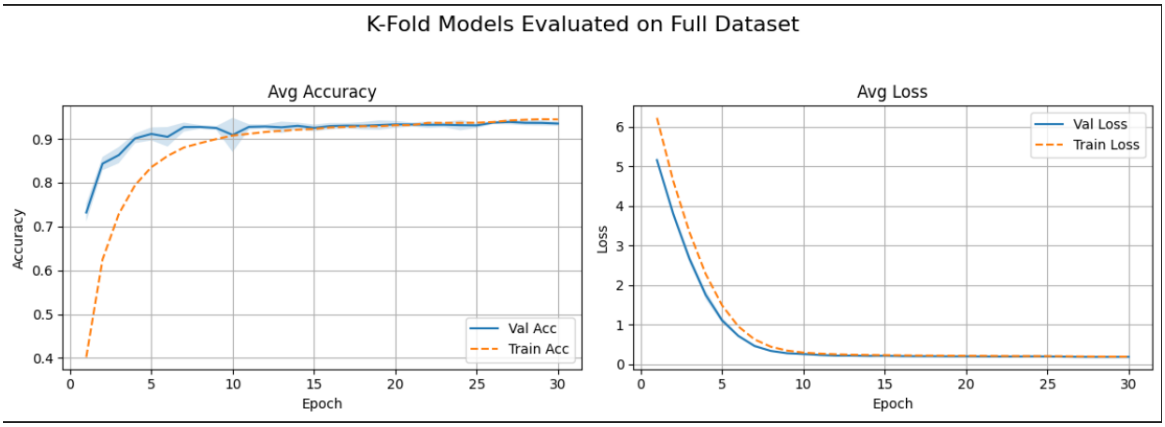

Figure 6. Learning curves (accuracy and loss) for EfficientNetB3 on the Oral Diseases dataset.

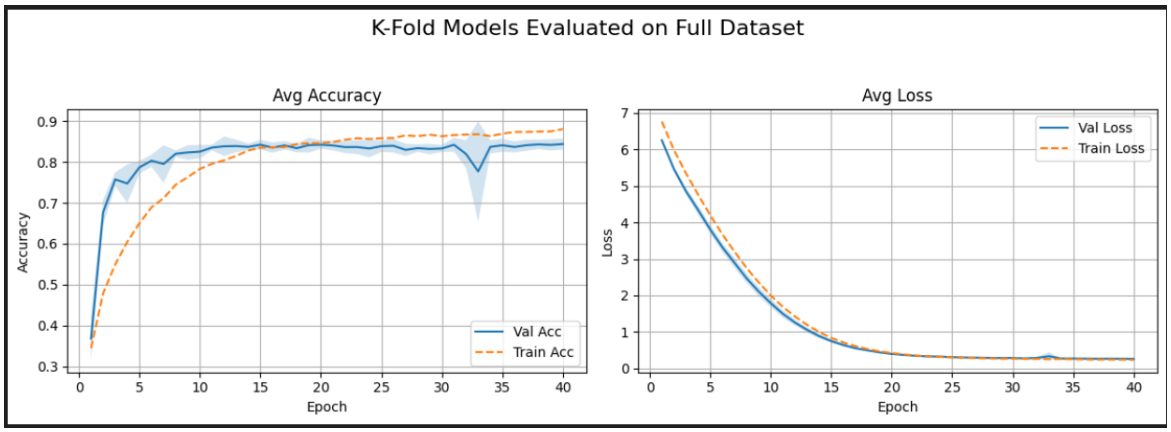

Figure 7. Learning curves (accuracy and loss) for EfficientNetB3 on the Oral Infection dataset.

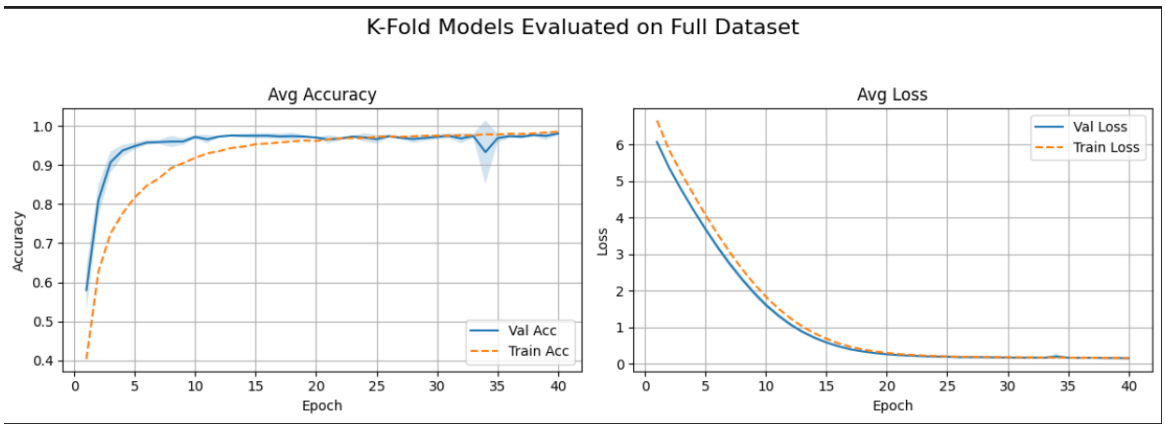

Figure 8. Learning curves (accuracy and loss) for EfficientNetB3 on the Teeth dataset.

Confusion matrices are more informative than just accuracy and can be used to diagnose where systematic misclassifications are being made by the model. They are further able to supplement quantitative measures by

exposing visually which classes are more easily confused. The confusion matrices for the Oral Diseases, Oral Infection, and Teeth datasets are presented in Figures 9, 10, and 11, respectively.

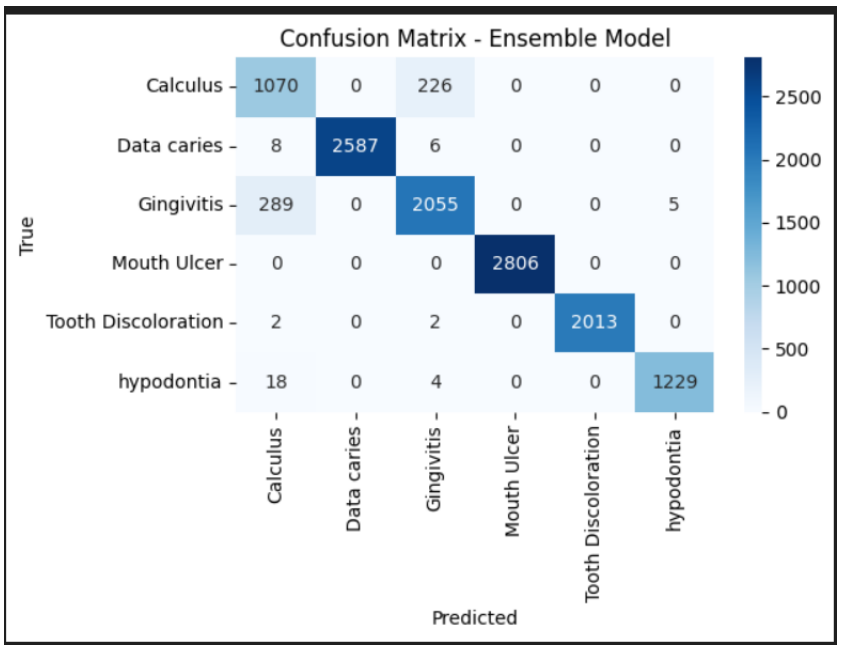

Figure 9. Confusion Matrix of the EfficientNetB3 Ensemble Model on the Oral Diseases Dataset.

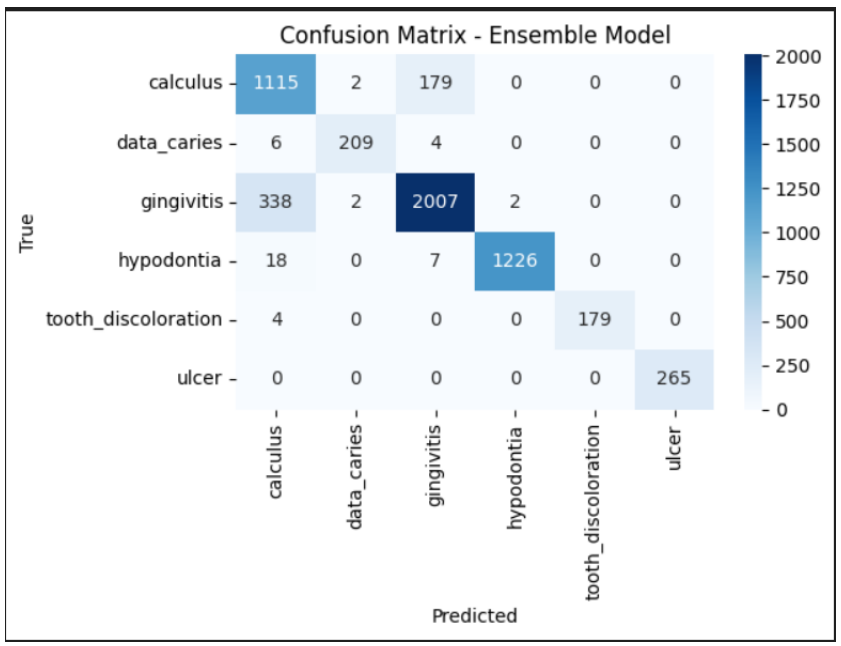

Figure 10. Confusion Matrix of the EfficientNetB3 Ensemble Model on the Oral Infection Dataset.

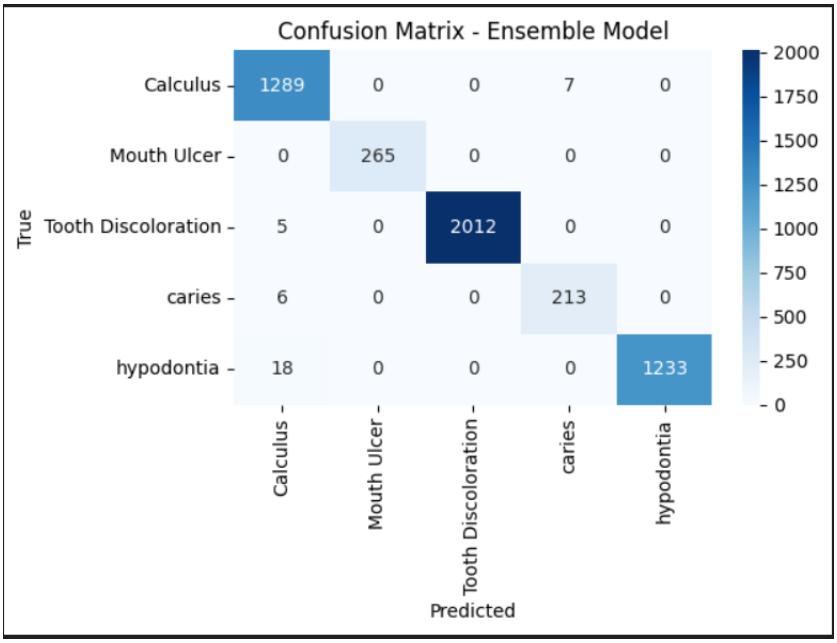

Figure 11. Confusion Matrix of the EfficientNetB3 Ensemble Model on the Teeth Dataset.

5. Discussion

This study investigated the application of transfer learning-based CNNs for multi-class dental disease classification using three RGB intraoral image datasets. The methodology was designed to provide practical insights for large-scale development of AI-based dental diagnostics. Among the evaluated backbones, EfficientNetB3 consistently achieved the highest ensemble accuracy, F1-Score, and Cohen's Kappa, either surpassing or matching alternative architectures. Its superior performance was further supported by test–retest reliability, as evidenced by consistent results across cross-validation folds and a reduced misclassification rate, particularly under visually complex or imbalanced class settings.

Despite these advantages, confusion matrix analysis revealed recurring misclassifications between visually similar conditions such as calculus and gingivitis, or caries and discoloration. These errors are largely attributable to overlapping visual features (e.g., shape, location, and shade) and the inherently subjective nature of image annotation. Moreover, underrepresented classes such as hypodontia showed reduced recall and precision, indicating that, although focal loss, class weighting, and label smoothing strategies alleviated imbalance to some extent, they did not entirely overcome the limitations imposed by insufficient sample representation.

The integration of stratified 5-fold cross-validation, focal loss, and ensemble inference contributed to robust and generalizable outputs, demonstrating the feasibility of applying deep learning models in dental diagnostics. These findings underscore both the promise of CNN-based frameworks for clinical application and the ongoing challenges of addressing subtle inter-class variability and sample imbalance.

5.1. Comparative Evaluation with Related Methods

Compared to Ikhwani et al., who focused primarily on binary classification and limited model diversity, this study employed a broader and more unified evaluation framework. EfficientNetB3 achieved ensemble accuracies of 0.955 on the Oral Diseases dataset, 0.899 on the Oral Infection dataset, and 0.993 on the Teeth dataset—surpassing prior benchmarks in both performance and reliability. The consistent use of cross-validation and focal loss contributed to improved generalization and reduced overfitting, particularly for imbalanced and visually similar classes such as calculus and gingivitis.

5.2. Importance of Interpretability

For transparency and clinical interpretability, Grad-CAM was applied to visualize the regions of attention during classification across the three dental datasets. Grad-CAM is a widely studied post-hoc method for generating class-specific activation maps in medical imaging. Consistent with findings from related studies [34], it provided an accessible means of understanding the model's decision-making, thereby bridging the gap between "black-box" predictions and clinical reasoning. In this study, the Grad-CAM heatmaps highlighted clinically relevant anatomical features such as inflamed gingiva, enamel caries, and missing teeth. Correctly classified cases demonstrated strong, localized activations over lesions, while misclassified cases exhibited diffuse or ambiguous attention, particularly in overlapping categories. These visual insights not only enhance interpretability and foster clinical trust but also support error analysis and diagnostic refinement.

5.3. Misclassification Insights

Although global metrics were strong, misclassifications persisted in categories such as calculus and gingivitis, where overlapping features in color, shape, and location contributed to ambiguity—especially near the gingival margin. Underrepresented classes, particularly hypodontia, also showed lower recall. These findings reaffirm the importance of class weighting, label smoothing, and focal loss in improving per-class performance, though they also highlight the need for additional strategies to address rare conditions.

5.4. Clinical Relevance and Practical Deployment

The use of intraoral RGB images, as opposed to radiographs, enhances the framework's feasibility for routine dental practice and telehealth applications. Unlike more complex hybrid approaches, the EfficientNetB3-based pipeline imposes minimal computational demands, making it particularly suitable for deployment in low-resource clinical environments. The consistent performance across diverse datasets underscores its potential integration into intelligent decision support systems for early diagnosis, patient triage, and screening programs.

5.5. Future Research Directions

Although attention modules such as the Convolutional Block Attention Module (CBAM) [35] and the Squeeze-and-Excitation (SE) block [36] were tested and did not improve accuracy in this study, future architectural refinements or integration strategies may enhance their effectiveness in different imaging contexts.

6. Conclusions

This study evaluated the performance of deep convolutional neural networks (CNNs) for the automated classification of dental anomalies and diseases using RGB intraoral images. A unified transfer learning approach was applied across three open dental datasets—Oral Diseases, Oral Infection, and Teeth—demonstrating that Effi-cientNetB3 substantially outperformed other CNN architectures, achieving up to 99.3% accuracy alongside superior F1-Score and Cohen's Kappa values. Key optimization strategies, including focal loss, label smoothing, class weighting, ensemble averaging, and stratified 5-fold cross-validation, collectively contributed to reliable and generalizable performance, even under conditions of significant class imbalance. Furthermore, Grad-CAM-based visualizations enhanced interpretability, providing critical insights into model decision-making and reinforcing clinical trust in AI-driven diagnostic systems.

Despite these promising results, certain limitations remain. Misclassifications were observed in visually similar conditions, such as gingivitis and calculus, underscoring the need for further optimization to improve the discrimination of overlapping pathologies. Additionally, while intraoral RGB images enhance accessibility and ease of acquisition, they lack the capacity to capture deeper structural details available in radiographic imaging.

For future research, it is proposed to explore attention-based modules such as the Convolutional Block Attention Module (CBAM) and Squeeze-and-Excitation (SE) blocks, as well as to extend the framework to include panoramic and radiographic datasets. Incorporating semi-supervised learning strategies and integrating patient metadata could

further strengthen the model's diagnostic capabilities. Most importantly, transitioning from experimental evaluation to real-world clinical validation remains essential to establish practical utility.

REFERENCES

1. E. Sivari, G. B. Senirkentli, E. Bostanci, M. S. Guzel, K. Acici, and T. Asuroglu, *Deep learning in diagnosis of dental anomalies and diseases: A systematic review*, Diagnostics, vol. 13, no. 15, pp. 2512, 2023, doi: 10.3390/diagnostics13152512.
2. K. Wang, S. Zhang, Z. Wei, X. Fang, F. Liu, M. Han, and M. Du, *Deep learning-based efficient diagnosis of periapical diseases with dental X-rays*, Image and Vision Computing, vol. 147, pp. 105061, 2024, doi: https://doi.org/10.1016/j.imavis.2024.105061.
3. H. T. Sadeeq, S. Y. Ameen, and A. M. Abdulazeez, *Cancer Diagnosis based on Artificial Intelligence, Machine Learning, and Deep Learning*, 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2022, pp. 656–661, 2022, doi: 10.1109/3ICT56508.2022.9990784.
4. J. Xin, M. Khishe, D. Q. Zeebaree, L. Abualigah, and T. M. Ghazal, *Adaptive habitat biogeography-based optimizer for optimizing deep CNN hyperparameters in image classification*, Heliyon, vol. 10, no. 7, 2024, doi: 10.1016/j.heliyon.2024.e28147.
5. S. Sukegawa, K. Yoshii, T. Hara, F. Tanaka, Y. Taki, Y. Inoue, K. Yamashita, F. Nakai, Y. Nakai, R. Miyazaki, T. Ishihama, and M. Miyake, *Optimizing dental implant identification using deep learning leveraging artificial data*, Scientific Reports, vol. 15, no. 1, pp. 3724, 2025, doi: 10.1038/s41598-025-87579-3.
6. J. Zhu, Z. Chen, J. Zhao, Y. Yu, X. Li, K. Shi, F. Zhang, F. Yu, K. Shi, Z. Sun, N. Lin, and Y. Zheng, *Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: a preliminary study*, BMC Oral Health, vol. 23, 2023, doi: 10.1186/s12903-023-03027-6.
7. Y. Pang, Z. Yang, L. Zhang, X. Liu, X. Dong, X. Sheng, J. Tan, X. Mao, and M. Liu, *Establishment and evaluation of a deep learning-based tooth wear severity grading system using intraoral photographs*, Journal of Dental Sciences, vol. 20, no. 1, pp. 477–486, 2025, doi: https://doi.org/10.1016/j.jds.2024.05.013.
8. A. AL-Ghamdi, M. Ragab, S. AlGhamdi, A. Asseri, R. Mansour, and D. Koundal, *Detection of Dental Diseases through X-Ray Images Using Neural Search Architecture Network*, Computational Intelligence and Neuroscience, vol. 2022, pp. 1–7, 2022, doi: 10.1155/2022/3500552.
9. W. Li, Y. Liang, X. Zhang, C. Liu, L. He, L. Miao, and W. Sun, *A deep learning approach to automatic gingivitis screening based on classification and localization in RGB photos*, Scientific Reports, vol. 11, no. 1, pp. 16831, 2021, doi: 10.1038/s41598-021-96091-3.
10. A. Imak, A. Celebi, K. Siddique, M. Turkoglu, A. Sengur, and I. Salam, *Dental Caries Detection Using Score-Based Multi-Input Deep Convolutional Neural Network*, IEEE Access, vol. 10, pp. 18320–18329, 2022, doi: 10.1109/ACCESS.2022.3150358.
11. S. Bhat, G. Birajdar, and M. Patil, *Enhanced Diagnostic Accuracy for Dental Caries and Anomalies in Panoramic Radiographs Using a Custom Deep Learning Model*, Cureus, vol. 16, no. 8, pp. e67315, 2024, doi: 10.7759/cureus.67315.
12. E. T. Yasin, M. Erturk, M. Tassoker, and M. Koklu, *Automatic mandibular third molar and mandibular canal relationship determination based on deep learning models for preoperative risk reduction*, Clinical Oral Investigations, vol. 29, no. 4, pp. 203, 2025, doi: 10.1007/s00784-025-06285-6.
13. M. Tan, and Q. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, International conference on machine learning, pp. 6105–6114, 2019.
14. Y. M. Alsakar, N. Elazab, N. Nader, W. Mohamed, M. Ezzat, and M. Elmogy, *Multi-label dental disorder diagnosis based on MobileNetV2 and swin transformer using bagging ensemble classifier*, Scientific Reports, vol. 14, no. 1, pp. 25193, 2024.
15. J. Li, *DA-Net: A classification-guided network for dental anomaly detection from dental and maxillofacial images*, Journal of King Saud University - Computer and Information Sciences, vol. 36, no. 9, pp. 102229, 2024, doi: https://doi.org/10.1016/j.jksuci.2024.102229.
16. T. Kabir, C. Lee, L. Chen, X. Jiang, and S. Shams, *A comprehensive artificial intelligence framework for dental diagnosis and charting*, BMC Oral Health, vol. 22, no. 1, pp. 480, 2022.
17. P. Parkhi, S. Harjal, A. Sahu, P. Agrawal, H. Shingne, Y. Bobde, and A. Padole, *A Comprehensive Deep Learning Framework for Dental Disease Classification*, Journal Européen des Systèmes Automatisés, vol. 58, pp. 511–521, 2025, doi: 10.18280/jesa.580309.
18. Y. Ikhwani, E. Noersasongko, P. Purwanto, and M. Soeleman, *Comparative Performances of the Convolutional Neural Network based Transfer Learning Models for Classification of Dental Disease*, 2024.
19. Q. Sun, A. Akman, and B. W. Schuller, *Explainable artificial intelligence for medical applications: A review*, ACM Transactions on Computing for Healthcare, vol. 6, no. 2, pp. 1–31, 2025.
20. C. Amannah, K. F. Attai, and F. Uzoka, *A Data-Driven Intelligent Methodology for Developing Explainable Diagnostic Model for Febrile Diseases*, Algorithms, vol. 18, no. 4, 2025, doi: 10.3390/a18040190.
21. T. Rai, R. Malviya, and S. B. Sridhar, *Explainable Artificial Intelligence and Responsible Artificial Intelligence for Dentistry*, Explainable and Responsible Artificial Intelligence in Healthcare, pp. 145–163, 2025.
22. M. Hasnain, Z. Ali, M. Maqbool, and M. Aziz, *X-ray Image Analysis for Dental Disease: A Deep Learning Approach Using EfficientNets*, VFAST Transactions on Software Engineering, vol. 12, pp. 147–165, 2024, doi: 10.21015/vtse.v12i3.1912.
23. S. Hsieh, and Y. Cheng, *Multimodal feature fusion in deep learning for comprehensive dental condition classification*, Journal of X-Ray Science and Technology, vol. 32, pp. 1–19, 2024, doi: 10.3233/XST-230271.
24. J. Liu, X. Liu, Y. Shao, Y. Gao, K. Pan, C. Jin, H. Ji, Y. Du, and X. Yu, *Periapical lesion detection in periapical radiographs using the latest convolutional neural network ConvNeXt and its integrated models*, Scientific Reports, vol. 14, no. 1, pp. 25429, 2024, doi: 10.1038/s41598-024-75748-9.
25. S. M. Hussain, S. A. Zaidi, A. Hyder, and M. M. Movania, *Integrating Ensemble Learning into Remote Health Monitoring for Accurate Prediction of Oral and Maxillofacial Diseases*, 2023 25th International Multitopic Conference (INMIC), pp. 1–6, 2023, doi: 10.1109/INMIC60434.2023.10465788.

26. P. Razmjouei, E. Moharamkhani, S. S. Aryanezhad, M. Shokouhifar, M. Hosseinzadeh, and B. Zadmehr, *NFR-EDL: Non-linear fuzzy rank-based ensemble deep learning for accurate diagnosis of oral and dental diseases using RGB color photography*, Computers in Biology and Medicine, vol. 192, pp. 110279, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.110279.
27. S. Sajid, *Oral Diseases*, 2025.
28. Sizlingdhairya1, *Oral Infection*, 2025.
29. Rajapriyanshu, *Teeth Dataset*, 2025.
30. M. Wang, X. Xu, and H. Liu, *A Semi-Supervised Object Detector Based on Adaptive Weighted Active Learning and Orthogonal Data Augmentation*, Sensors, vol. 25, no. 6, 2025, doi: 10.3390/s25061798.
31. Y. Liu, C. Ma, Z. He, C. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, *Unbiased teacher for semi-supervised object detection*, arXiv preprint arXiv:2102.09480, 2021.
32. N. Aljohani, A. Fayoumi, and S. Hassan, *A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations*, Journal of Information Science, vol. 49, pp. 016555152199102, 2021, doi: 10.1177/0165551521991022.
33. E. Tjoa, and C. Guan, *A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI*, IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.
34. B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, *Explainable artificial intelligence (XAI) in deep learning-based medical image analysis*, Medical Image Analysis, vol. 79, pp. 102470, 2022, doi: https://doi.org/10.1016/j.media.2022.102470.
35. S. Woo, J. Park, J. Lee, and I. S. Kweon, *Cbam: Convolutional block attention module*, Proceedings of the European conference on computer vision (ECCV), pp. 3–19, 2018.
36. J. Hu, L. Shen, and G. Sun, *Squeeze-and-excitation networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141, 2018.