

Independent University

Bangladesh (IUB)

IUB Academic Repository

Computer Science and Engineering

Undergraduate Thesis

2025-12

Deep Learning on Dental RGB and OPG Images for Public Health and Clinical Decision Support

Smaron, J. M. Sadik-UI Islam

Independent University, Bangladesh

<https://ar.iub.edu.bd/handle/11348/1254>

Downloaded from IUB Academic Repository



Deep Learning on Dental RGB and OPG Images for Public Health and Clinical Decision Support

December 2025

Prepared by:

J. M. Sadik-Ul Islam Smaron

ID: 2220221

Mustaqueem Alam

ID: 2220769

Mohammed Tashfiq Islam

ID: 2221651

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by

Ashraful Islam, PhD

Assistant Professor

Department of Computer Science and Engineering
Independent University, Bangladesh

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project, which have been properly cited following the international standards and proper guidelines.

Author Name:

Signature:

Author Name:

Signature:

Author Name:

Signature:

Evaluation Committee

Supervision Panel

.....
Dr. Ashraful Islam
Supervisor

.....
Dr. Rashedur Rahman
Co-supervisor

Examiners

.....
Dr. AKM Mahbubur Rahman (Internal) Mr. Sanzar Adnan Alam (Internal)

.....
Mr. Asif Md. Akhlak (External Examiner)

Office Use

.....
Dr. AKM Mahbubur Rahaman
Program Coordinator

.....
Dr. Saadia Binte Alam
Head of the Department

.....
Prof. Dr. M. Arshad Momen
Director, Office of the Graduate Studies, Research and Industry Relations

.....
Library

Acknowledgement

We would like to express our sincere gratitude to our respected supervisor, Dr. Ashraf Islam, and co-supervisor Dr. Rashedur Rahman from Independent University, Bangladesh. Their dedicated guidance, insightful suggestions, remarkable patience, and the generous investment of his valuable time played a pivotal role in the successful completion of our project. Our Professor's expertise and mentorship greatly enhanced the quality and depth of our work, and we are truly appreciative of his unwavering support throughout our academic journey.

Additionally, our heartfelt thanks go to the Center for Computational & Data Sciences (CCDS) for their extraordinary help and consistent support. The collaborative and resourceful environment provided by CCDS significantly contributed to the development and execution of our project. Moreover, the encouragement and assistance have been instrumental in shaping our academic and research fields. We are grateful for the opportunities and resources extended to us during our research, which undoubtedly enriched our learning experience.

We acknowledge and appreciate the contributions of all individuals who supported us, both directly and indirectly, during this endeavor.

Abstract

The rapidly increasing rates of unrecognized dental issues, along with the ongoing barriers to timely dental health monitoring, have raised the need for the creation of Machine Learning (ML) based dental systems that are accurate and can be used by both patients and dental professionals. In an effort to bridge the circuit, a combined Deep Learning (DL) architecture analysis is implemented, which takes advantage of two imaging sources that are very different yet complementary: Intraoral Red-Green-Blue (RGB) images for the monitoring of dental health in general purposes, and panoramic Orthopantomogram (OPG) radiographs for the support of clinical decisions. The approach is composed of two specifically designed frameworks that include an RGB-based classification system that also includes the detection of early cases of dental caries, which is for teeth health monitoring for public use, and an OPG-based system for conceiving and supporting dentists in radiograph interpreting as a supporting tool in case of complex patterns in OPG radiographs.

Powerful ablation studies were conducted on all scaffolding, showing the influence of architectures, preprocessing steps, augmentation strategies, and hyperparameter settings on the performance of models across both modes of image datasets. Initially, the ResNet101 model proved to be the strongest performer among the multiple Convolutional Neural Network (CNN) models examined in the RGB pipeline, achieving an F1-Score of 94% and 97% test accuracy, while still having difficulty with dental caries identification (F1-Score of 82%). This led to further investigations along with merging the initial dataset with another RGB dental caries dataset in which the Neural Search Architecture Network-Mobile (NASNet-Mobile) model performed superior among all the models to address the issue of dental caries, achieving 95% F1-Score on the dental caries class. After that, the RGB Intraoral Caries dataset was tested independently with lightweight variants of the You Only Look Once (YOLO) object detection model and the dataset achieved more than 92% Intersection over Union (IoU) at Mean Average Precision (mAP)@50 and more than 73% IoU at mAP@50-95 across all the models.

The classification and detection models indeed exhibited a high degree of precision in the location of the lesion, especially in the case of dental caries, but among the evaluated models, YOLO was chosen for the suggested public health screening solution due to its superiority in being able to assure the presence of tooth decay in a less technological and clinical setting without prior diagnosis of the patient, which is not the case with standard classification models. For the case of OPG radiographs, all ten lightweight YOLO models (YOLOv8-YOLOv12 in nano, tiny, and small versions) achieved a high level of precision, consistent with the performance differences between variants being very small. The score was around 80-92% IoU at mAP@50 and 55-67% IoU at mAP@50-95 in all the lightweight YOLO models. The complementary multi-label CNN experiments on the OPG dataset helped us to find the balance between the classification-only and detection-based approaches and were included in the ablation studies. In light of these results, we concluded that the lightweight variants of the YOLO12 models, along with the idea of two practical applications, emerge: an RGB-based automated screening tool for general dental health monitoring and an OPG-based clinical support system for improving diagnostic efficiency for dentists. In addition, the research also unveils key data and methodological limitations and outlines future directions to strengthen clinical robustness and deployment readiness.

Contents

1	Introduction	12
1.1	Background and Motivation	12
1.1.1	Image-Based Dental Datasets	12
1.1.2	Need for Early Dental Disease Detection	12
1.1.3	Advances in AI for Dental Diagnostics	12
1.1.4	Challenges in Real-World Dental Imaging	13
1.1.5	Model Benchmarking Motivation	13
1.2	Problem Statement	13
1.3	Research Objectives and Questions	14
1.3.1	Research Objectives	14
1.3.2	Research Questions	14
1.4	Significance of the Study	15
1.4.1	Technical Significance	15
1.4.2	Economic Significance	15
1.4.3	Social Significance	15
1.4.4	Environmental Considerations	15
1.4.5	Alignment with UN Sustainable Development Goals (SDG 2030) . .	16
1.5	Report Outline	17
2	Related Works	20
2.1	Imaging Approaches for Dental Diagnostics	20
2.2	Imaging Modalities: Reflective vs Penetrative Dental Images	20
2.3	Traditional (Non-Deep Learning) Approaches for Dental Image Analysis . .	21
2.4	Deep Learning in Dental Imaging	21
2.5	CNN Architectures Used in Medical Image	21
2.6	Model Comparison Strategies in Medical AI	22
2.7	Object Detection and Segmentation Techniques in Dentistry	22
2.8	YOLO Family Models for Object Detection	23
2.9	Explainable AI in Medical Imaging	23
3	Dataset Materials and Preprocessing	24
3.1	Overview of Data Sources	24
3.2	Background study of the dataset classes	25
3.2.1	Healthy Teeth	25
3.2.2	Caries	25
3.2.3	Impacted Teeth	25
3.2.4	Broken-down Crown/Root	26
3.2.5	Infection	26
3.2.6	Fractured Teeth	26
3.2.7	Calculus	26

3.2.8	Hypodontia	27
3.2.9	Mouth Ulcer	27
3.2.10	Tooth Discoloration	27
3.3	OPG Dataset for Classification	28
3.3.1	Dataset Overview	28
3.3.2	Multi-Label Dataset Splitting Strategy	28
3.3.3	Preprocessing and Data Cleaning Pipeline	30
3.3.4	Data Augmentation	30
3.3.5	Contrast Enhancement	30
3.4	OPG Dataset for Object Detection	32
3.4.1	Dataset Overview and Annotation Criteria	32
3.4.2	Class Distribution and Lesion Characteristics	33
3.4.3	Dataset Preparation	33
3.4.4	Contrast Enhancement	34
3.4.5	Dataset Configuration	34
3.4.6	Multi-Label Dataset Splitting Strategy	34
3.5	RGB Teeth Dataset for Classification	35
3.5.1	Dataset Overview	35
3.5.2	Contrast Enhancement	35
3.5.3	Dataset Augmentation	36
3.5.4	Single-Label Dataset Splitting Strategy	36
3.6	Intraoral Caries Dataset for Object Detection	37
3.6.1	Dataset Overview and Annotation Criteria	37
3.6.2	Dataset Preparation	37
3.6.3	Dataset Configuration	38
3.6.4	Preprocessing and Data Cleaning Pipeline	38
4	Materials and Methodology	39
4.1	Proposed Framework Overview	39
4.2	CNN Architectures for Disease Classification	39
4.2.1	Residual Learning Frameworks	39
4.2.2	Dense Connectivity Network	41
4.2.3	Lightweight Architectures	41
4.2.4	Multi-Scale Feature Learning	42
4.2.5	Compound Scaling Architecture	43
4.3	Background of the YOLO Framework for Object Detection	43
4.3.1	Evolution of YOLO Models (v8-v12)	44
4.3.2	Comparison of YOLO Architectural Components	45
5	Experimental Setup	47
5.1	Classification	47
5.1.1	Classification Models Evaluated on the OPG Dataset	47
5.1.2	Functions and Optimizers for Multi-Label Classification	47
5.1.3	Classification Models Evaluated on the RGB Teeth Dataset	49
5.1.4	Functions and Optimizers for Single Label Classification	49
5.2	Object Detection Experimental Setup	51
5.2.1	Loss Functions, Activation Functions, and Optimization Strategy	51
5.2.2	Object Detection Models Evaluated on both OPG and RGB Intraoral Caries Datasets	52
5.2.3	Training process of the YOLO models	53

5.3	Evaluation Metrics	54
5.3.1	Evaluation Metrics for Classification	54
5.3.2	Evaluation Metrics for Object Detection	55
5.4	Computational Environment	55
6	Results and Analysis	57
6.1	Classification Results on OPG Dataset	57
6.1.1	Class-wise Performance Interpretation (OPG)	57
6.1.2	Comparison of Classification Architectures (OPG)	58
6.1.3	Best Model Performance Analysis (OPG)	59
6.1.4	Confusion Matrices and Error Analysis (OPG)	60
6.1.5	Confidence Threshold Adjustment (OPG)	62
6.1.6	Grad-CAM Visualizations and Explainability (OPG)	63
6.2	Classification Results on RGB Teeth Dataset	65
6.2.1	RGB Teeth Dataset Confusion Matrices and Error	66
6.2.2	RGB Teeth Dataset Class-wise Performance Interpretation	67
6.2.3	RGB Teeth Dataset Model Performance Analysis	69
6.2.4	RGB Teeth Dataset Grad-CAM Visualizations and Explainability .	71
6.2.5	RGB Teeth Dataset Comparison of Classification Architectures . .	71
6.2.6	RGB Teeth Dataset Strengths and Weaknesses Across Classification Models	72
6.2.7	RGB Teeth Dataset AUC-ROC Curves of the Best Performing Models	72
6.2.8	Handling Class Imbalance in Caries Classification on RGB Teeth Dataset	73
6.3	Object Detection Performance on OPG Dataset	78
6.3.1	Precision-Recall Curves and Confidence Analysis	78
6.3.2	mAP@50 and mAP@50-95 Comparison	79
6.3.3	Localization and Error Cases	80
6.3.4	Strengths and Weaknesses Across Models	82
6.4	Object Detection Performance on RGB Intraoral Caries Dataset	83
6.4.1	RGB Intraoral Caries Dataset Precision-Recall Curves and Confi- dence Analysis	83
6.4.2	RGB Intraoral Caries Dataset mAP@50 and mAP@50-95 Comparison	84
6.4.3	RGB Intraoral Caries Dataset Localization and Error Cases	85
6.4.4	RGB Intraoral Caries Dataset Strengths and Weaknesses Across Models	87
7	System Implementation	88
7.1	Methodology and Overall System Architecture	88
7.1.1	Technology Stack	90
7.1.2	Technical Considerations and Challenges	90
7.2	OPG-Based Dental Decision Support System	91
7.2.1	System Overview	91
7.2.2	Functional Objectives	91
7.3	RGB Teeth Health Monitoring System	93
7.3.1	System Overview	93
7.3.2	Functional Objectives	93
7.4	Integration and Deployment Environment	93

8	Discussion of Key Findings	94
8.1	Classification Experiments Summary	94
8.2	Object Detection Experiments Summary	94
8.3	Interpretation of Performance Trends	95
8.4	Impact of Data Quality and Preprocessing	96
8.5	Real-World Applicability of the Models	96
8.6	Addressing the Research Questions	97
9	Threats to Validity	98
9.1	Internal Validity	98
9.1.1	Dataset Bias	98
9.1.2	Impact of Preprocessing	98
9.1.3	Data Splitting and Leakage	99
9.2	External Validity	99
9.2.1	Generalization to Real-World Settings	99
9.2.2	Device and Lighting Variations	100
9.2.3	Population Diversity	100
9.3	Theoretical Validity	101
9.3.1	Appropriateness of Metrics	101
9.3.2	Label Definitions and Class Boundaries	101
9.3.3	Clinical Alignment of Outputs	101
9.4	Statistical Conclusion Validity	102
9.4.1	Statistical Robustness	102
9.4.2	Model Comparisons	102
9.4.3	Stability Across Runs	103
9.5	System Limitations	103
10	Conclusion and Future Work	104
	Bibliography	104

List of Figures

3.1	Sample images from the datasets showing labeled classes for both RGB and OPG	25
3.2	Label frequency distribution after dataset cleaning of the OPG dataset for classification task	29
3.3	Dataset split overview of the OPG dataset for classification	30
3.4	Image preprocessing pipeline of OPG classification	31
3.5	Sequential stages of the OPG data preprocessing pipeline	32
3.6	Sample visualization of training a image ID from the augmented OPG dataset	33
3.7	Sequential stages of the RGB data preprocessing pipeline	35
3.8	Sequential stages of the Intraoral RGB dataset preprocessing pipeline . . .	37
4.1	Proposed Methodology	40
4.2	Comparative summary of classification model architectures in this study .	42
4.3	Generalized workflow of YOLO-Based Object detection models	43
4.4	Comparative summary of the latest YOLO-based object detection models in this study	45
6.1	Performance summary of the CNN models on the OPG dataset	58
6.2	Model accuracy progression of ResNet-18 on OPG dataset	59
6.3	Model loss convergence of ResNet-18 on OPG dataset	60
6.4	Confusion Matrix for Multi-Label OPG classification using ResNet-18 . . .	61
6.5	F1-Score across varying Confidence Thresholds	62
6.6	Grad-CAM for correct and incorrect OPG samples	63
6.7	Performance summary of the CNN models on the RGB Teeth dataset . . .	65
6.8	Confusion Matrix of ResNet-101 on the RGB Teeth dataset	66
6.9	Confusion Matrix of NASNet-Mobile on the RGB Teeth dataset	67
6.10	Model accuracy and loss convergence of NASNet-Mobile on RGB Teeth dataset	69
6.11	Model accuracy and loss convergence of ResNet-101 on RGB Teeth dataset	69
6.12	Grad-CAM heatmap analysis on the RGB Teeth dataset showing the activated regions that contribute most to model decision-making on the test dataset	70
6.13	AUC/ROC curve of RGB Teeth dataset on ResNet-101 model	72
6.14	AUC/ROC curve of RGB Teeth dataset on NASNet-Mobile model	73
6.15	Confusion matrix of NASNet-Mobile after oversampling on the RGB Teeth dataset	75
6.16	Confusion matrix of NASNet-Mobile after merging Intraoral Dataset & RGB Teeth dataset	75
6.17	Performance summary of the YOLO models on the OPG dataset	78
6.18	Box precision and recall curve of YOLO12s on the OPG dataset	78

6.19	F1 confidence curve of YOLO12s on the OPG dataset	79
6.20	Ground truth vs Predicted labels on validation set of the OPG dataset . .	81
6.21	Confusion matrix of YOLO12s on the OPG dataset	82
6.22	Performance summary of the YOLO models on the RGB Intraoral Caries dataset	83
6.23	Box precision and recall curve of YOLO12n on the RGB Intraoral Caries dataset	84
6.24	F1 confidence curve of YOLO12n on the RGB Intraoral Caries dataset . .	84
6.25	Confusion matrix of YOLO12n on the Intraoral Caries dataset	85
6.26	Ground truth vs Predicted labels on validation set of the RGB Intraoral Caries dataset	86
7.1	Methodology of proposed solution	88
7.2	Technical simulation of the proposed system	89
7.3	UI Snippets of the implemented system in web browser view	92

List of Tables

3.1	Split criteria based on signature size of the OPG dataset for classification task	29
3.2	Class-wise distribution across train, validation, and test sets in the OPG augmented dataset for object detection	34
3.3	Dataset class distribution and data splits of the RGB Teeth dataset	36
5.1	Comparison of different model architectures and their classification heads on the OPG dataset	47
5.2	Fine-tuning configuration and training duration for each model on the OPG dataset	48
5.3	Comparison of different model architectures and their classification heads on the RGB Teeth dataset	49
5.4	Fine-tuning configuration and training duration for each model on the RGB Teeth dataset	50
5.5	Comparative architectural summary of YOLO variants employed in this study's object detection pipelines	53
6.1	Label-wise accuracy comparison across different CNN models	58
6.2	Comparative performance analysis of CNN models	59
6.3	Impact of adaptive confidence thresholding on F1 Scores by disease class (Δ indicates change)	62
6.4	Performance metrics of CNN models on the RGB Teeth dataset	68
6.5	Test performance comparison across DL models on RGB Teeth dataset	71
6.6	Per-class performance of NASNet-Mobile after oversampling	76
6.7	Performance comparison of before and after oversampling	76
6.8	NASNet-Mobile performance on test data for five dental classes	76
6.9	Performance comparison of before and after merging the dataset	76
6.10	Performance comparison of YOLO models across different dental classes on the OPG dataset	80
6.11	Comparison of precision, recall, and mAP metrics across different YOLO models	82
6.12	Comparison of precision, recall, and mAP metrics across different YOLO models	87
7.1	Technology Stack used in System Implementation	90

1 Introduction

1.1 Background and Motivation

1.1.1 Image-Based Dental Datasets

The field of dentistry has experienced a significant boost from the application of Artificial Intelligence (AI), primarily due to the increasing availability of high-quality and well-annotated image datasets. The researchers who are trying to minimize the discrepancy between the private clinical data and the publicly available resources have contributed greatly, among whom a few are worth mentioning. For instance, recent research has shown that the detection of oral calculus can be very accurately done with just standard smartphone cameras capturing simple Red Green Blue (RGB) images, which proves that effective screening does not necessarily need expensive radiographic equipment [1]. In addition, the use of heterogeneous imaging techniques in such a way that it ranges from regular intraoral photographs to dental radiographs has become more and more necessary for the creation of explainable diagnostic systems. The access to such multimodal datasets is a big advantage in the training of artificial intelligence systems that could be capable of delivering transparent, clinically meaningful explanations for multi-class disease predictions, ultimately strengthening trust and adoption within dental practice [2].

1.1.2 Need for Early Dental Disease Detection

The issue of dental diseases is especially notable in developing areas such as Bangladesh, which makes the need for early detection systems more urgent than ever. Findings from a study in Bangladesh indicate that dental caries constitute a major health problem among pre-school children, and the situation is worse since most of these children have not received any treatment [3]. Besides, the problem is made worse by the fact that even educated people lack awareness of preventive care; e.g, a survey among university students showed that the students' knowledge and attitudes toward oral health were insufficient, and thus, late diagnosis was a result of not only clinical inaccessibility but also behavioral factors as well [4]. In this regard, automated early detection tools could play a very significant role in striking a balance on the accessibility side by providing a constant and uniform screening that is not exclusively dependent on the patient's proactive approach towards dental care or the immediate availability of a specialist.

1.1.3 Advances in AI for Dental Diagnostics

The medical imaging field is inversely transformed with the help of Deep learning (DL), as it is moving from relying on human interaction to getting machines that can do accurate and precise analysis automatically. A general review of digital image processing states that the traditional approach used handcrafted algorithms, and the modern DL models are

the ones that can get to know the representations of the features in a complicated manner since they automatically learn and make use of such representations for complex tasks like segmentation and classification [5]. In the narrow context of dentistry, a literature review shows that these Machine Learning (ML) models have quite literally exceeded the 90% accuracy mark in several use cases within the region of dental caries detection, periodontal disease classification, and age estimation [6]. Consequently, this has been occurring in synergy with the holistic trend of computer-aided detection in medicine, where deploying state-of-the-art hundred architecture incorporating You Only Look Once (YOLO) based frameworks and Vision Transformers (ViT) has conspicuously enhanced the velocity and precision of locating anatomical anomalies, thereby, unwittingly, making the possibility of the real-time clinical decision support systems very easily attainable [7].

1.1.4 Challenges in Real-World Dental Imaging

Even with the technical advancements, the real-world dental application of AI still faces considerable difficulties, mostly in terms of accessibility of equipment and variability in images. Traditional diagnostic techniques mainly depend on X-ray images, which are usually not present in remote or developing areas. To tackle this problem, the latest studies have introduced light-weight ML models that are suitable for smartphones and can detect conditions like oral calculus by using standard RGB images. Thus, it will be possible to offer telemedicine and also reduce the amount of radiation exposure [1]. Nevertheless, the processing of colored images has its own difficulties when compared to the high-contrast radiographs. For instance, a research on periodontal disease pointed out the intricacy of classifying the health status of gums from colored images of teeth, which called for advanced Convolutional Neural Networks (CNNs) to effectively tell healthy tissues from diseased ones, given the variations in lighting and the presence of intraoral obstructions [8].

1.1.5 Model Benchmarking Motivation

The necessity to ensure clinical reliability urges the need for benchmarking various DL architectures over the complex multi-label dental datasets. Model training with a single architecture frequently leads them to fail due to the complex dependencies present in the dental imagery. Therefore, the researchers introduced the hybrid dual-stream framework that brought together MobileNetV2 for local feature extraction, and Swin Transformers for global context, and this combination was able to dramatically enhance the X-ray multi-label disorder diagnosis [9]. Similarly, overcoming the constraints of CNN’s receptive fields in huge pictures such as Orthopantomogram (OPG), the researchers have been able to implement pure VViT on capturing long-range pixel relationships for precise classification [10].

1.2 Problem Statement

Dental Caries is a widespread oral health issue in Bangladesh, and access to timely diagnosis is limited by the shortage of trained dental specialists, insufficient screening facilities, and a lack of preventive care awareness. Dental imaging methods such as OPG radiographs and clinical RGB photographs may provide crucial diagnostic information, but their interpretation remains subjective and highly dependent on the availability of an expert. Recently, DL models have demonstrated a strong capability in the detection of dental diseases; however, the datasets used for training the models are often small and not

representative, or the models are optimized for specific populations that do not generalize well to South Asian or Bangladeshi patients. Moreover, there is a very small number of studies that have done a systematic comparison of classification and object detection pipelines on multi-label OPG datasets under real-world conditions, where noise, occlusion, and small lesions are the generating challenges. The lack of interpretable AI techniques that can provide explanations to healthcare professionals is also an obstacle to clinical acceptance [11]. Hence, a systematic benchmarking of multiple state-of-the-art DL architectures, including lightweight YOLO models and modern classification CNN models, on locally sourced dental datasets is evidently needed for broad and impactful research in the future and practical implementation. This research aims to identify the most robust, accurate, and clinically interpretable method for early-stage dental caries detection and other dental conditions from both OPG and RGB images, with a real-world deployment plan in two separate pipelines for public Health and clinical decision support for dentists.

1.3 Research Objectives and Questions

1.3.1 Research Objectives

The objectives of this study are as follows:

- To develop and evaluate a robust detection and classification pipeline for OPG radiographs that can accurately localize dental abnormalities and verify their clinical relevance through classification models.
- To perform comprehensive ablation studies on both OPG and RGB image models, examining the influence of architectures, preprocessing strategies, augmentation, and hyperparameters on system performance.
- To design and optimize a framework for Intraoral RGB images aimed at reliable detection of dental conditions for general use for health monitoring applications.
- To compare and identify the best performing DL model for each image domain using multiple evaluation metrics to ensure robust selection.
- To propose two practical dental diagnostic solutions: one clinical decision support system for dentists using OPG images and one screening tool using RGB images for general-purpose health monitoring, based on the findings of both pipelines.

1.3.2 Research Questions

- **RQ1:** How accurately can YOLO-based object detection models localize OPG radiographs, and which configuration provides the most reliable detection performance across evaluation metrics?
- **RQ2:** How effectively can CNN-based classification models verify and categorize detected abnormalities in OPG images, and what network variants or training strategies lead to the highest diagnostic accuracy and robustness?
- **RQ3:** How accurately can CNN-based models classify common dental conditions from RGB intraoral images, and which model or preprocessing choices maximize generalizability?

- **RQ4:** How does the application of Gradient-weighted Class Activation Mapping (Grad-CAM) enhance the interpretability of CNN-based classification models, and does it demonstrate that the models focus on clinically meaningful or visually relevant regions?
- **RQ5:** How can the experimental findings from the OPG detection-classification pipeline and the RGB classification pipeline be translated into two distinct AI-driven dental assessment solutions, one for clinical use and one for public health monitoring?
- **RQ6:** What are the similarities and differences in model performance and error patterns between reflective RGB datasets and penetrative OPG radiographs?

1.4 Significance of the Study

1.4.1 Technical Significance

This work is a technical contribution in that sense that it systematically evaluates the leading-edge DL architectures on both external and South asian: Pakistani and Bangladeshi dental images. The RGB Intraoral Dataset belongs to Pakistani patients, whereas the OPG dataset belongs to the Bangladeshi population. The comparison between the classification pipeline and the object detection pipeline grants critical insights into the robustness, generalization, and lesion localization of the models under real-world imaging conditions. The addition of Explainable AI (XAI) continues to enhance the transparency of the model, and thus, its acceptance as a support tool in clinical decision-making has also increased.

1.4.2 Economic Significance

The early identification of dental caries and other dental conditions can lead to a considerable decrease in treatment costs and also has the consequence in restoring the teeth less often. Clinics in remote areas would find AI-based diagnostic tools using the YOLO and MobileNet models running on cheap hardware as a great and economical option. This allows for the creation of large-scale and inexpensive oral health screening programs in Bangladesh.

1.4.3 Social Significance

Different dental problems, including cavities that result from untreated dental caries, are a problem everywhere, since they affect a persons overall quality of life, the ability to eat, and even speak. One of the goals of this research is to improve oral health in children and marginalized groups through the application of early, automated diagnosis, which is made possible by the study. The role of interpretable AI is also that of a facilitator; it builds trust between the patient and the professional, which in turn helps to level out the differences in dental care accessibility.

1.4.4 Environmental Considerations

Every time a traditional radiographic examination is performed, the patient is exposed to some amount of radiation, and the process should be considered energy-consuming. The research proposes the use of diagnostic methods that could be applied more with non-radiographic RGB images, thus promoting eco-friendly healthcare. Moreover, lightweight

models that can be processed on mobile devices further decrease the carbon footprint of large computation systems that are usually quite power-hungry.

1.4.5 Alignment with UN Sustainable Development Goals (SDG 2030)

The extent and the societal effect of this research not only correspond but also heavily support some of the United Nations Sustainable Development Goals (SDG 2030), especially those which emphasize health, new and more effective methods, fairness, careful technological integration, and joint development. The research has a global impact by providing AI-based detection systems for dental diseases that employ both RGB intraoral pictures and OPG radiographs, thus making dental diagnosis accessible, fast, and fair. Below, we elaborate on the correspondence with particular SDGs:

- **SDG 3: Good Health and Well-Being** With the help of the proposed classification and detection systems, dental diseases can be detected early, which will enable the patients to undergo interventions on time and, thus, decrease the number of people with untreated oral health conditions. The incorporation of low-cost and accessible imaging sources, such as smartphone cameras, leads the research to support preventive care, lessen diagnostic delays, and improve health monitoring on the community level.
- **SDG 9: Industry, Innovation, and Infrastructure** The project pushes digital healthcare innovations by merging modern DL, computer vision, and tiny AI models that are ready for deployment on real-world platforms. The building of scalable diagnostic pipelines is a major driver in the tech modernization of healthcare, and the upsurge of smart systems in the dental practices continues.
- **SDG 10: Reduced Inequalities** The proposed solutions intend to get rid of the inequities in oral healthcare access, more so towards the backward, neglected, or low-income communities. Mobile-friendly RGB-based screening, along with AI-assisted OPG analysis, cuts down the role of expensive machines and supports radiologists, thus making the health monitoring and diagnostic skills for patients and dentists better.
- **SDG 12: Responsible Consumption and Production** The research promotes digital health habits that are good for the environment by recommending the use of efficient, low-resource AI models and the application of the imaging methods that are most commonly accessible. The focus on low-radiation OPG imaging and the utilization of the existing clinical data additionally contribute to the reduction of the unnecessary generation of clinical waste and the effective use of resources.
- **SDG 17: Partnerships for the Goals** The cooperation of universities, dental practitioners, diagnostic centers, and health authorities is the main prerequisite for the successful implementation of the proposed systems in clinical practice. Besides, such interdisciplinary cooperation brings about knowledge transfer, speeds up innovation, and makes the long-term, large-scale implementation that is in line with national health strategies possible.

1.5 Report Outline

Chapter 1: Introduction

This chapter reviews DL applications to aid in the diagnosis of dentistry, looking at the development of this area over time and the reasons behind its growth today, including the increasing prevalence of dental disease, macro-and micro-level factors regarding the early detection of dental disease as an essential part of improving outcomes for the population as well as contributing to the growth of departmental imaging datasets, the importance of developing comprehensive and reliable dental imaging as part of an AI model’s evaluation to support diagnostic pipelines and research aims and questions that guide this research project. These research aims and questions include an overview of the social, technical, and financial dimensions of this area, which support current trends in global health, as cited by the UN SDG 203.

Chapter 2: Related Works

A thorough analysis of previous research is offered that includes imaging methods in dentistry, both reflective and penetrative image modes. Besides, traditional non-deep learning methods are described, alongside a detailed account of CNNs, object detection frameworks, and YOLO variants used in medical imaging. Furthermore, the chapter presents XAI methods for demonstrating model understanding, thereby creating a basis for model selection and appraisal in the current study.

Chapter 3: Dataset Materials and Preprocessing

The different datasets used in the study are briefly described in this chapter. The datasets include the RGB Teeth dataset, the OPG dataset, and the Intraoral Caries Detection dataset. Each of the datasets is defined in terms of its classes, annotations, and acquisition sources, along with the descriptions in great detail, the data preprocessing methods, image resizing, normalization, denoising, augmentation, and splitting for single-label and multi-label scenarios, etc., that were applied to maintain the consistency and reliability of the data for the next modeling tasks.

Chapter 4: Materials and Methodology

The comprehensive overview provided in this section includes both the Multi-Class Classification Pipelines and the Object Detection Pipelines in our framework. It mainly evaluates how the architecture of CNNs can positively impact the development of a model’s overall accuracy and focuses on the development of YOLO v8 to YOLO v12 for object detection. This chapter focuses on model selection, feature extraction, and architectural decisions, and merges the various methods of preprocessing into a single end-to-end pipeline.

Chapter 5: Experimental Setup

In this chapter, the details of experimental strategies for classification and object detection have been presented. Experimental Design includes training, validation, model optimization, loss function parameters, and computing resources used in combination with specified configurations of datasets, along with hyperparameter tuning methods.

This chapter also includes descriptions of performance measures that will be used for each of the following types of experiments: Classification (i.e., accuracy, precision,

recall, and F1 score), as well as Object Detection (i.e., mAP, precision-recall, and confidence). These performance assessments are replicable and standardized for comparative assessments of performance.

Chapter 6: Results and Analysis

The results of the image classification as well as object detection experiments performed with RGB and OPG images will be discussed.

The analysis covers a variety of different topics; this includes all metrics available for evaluating neural networks (i.e., confusion matrices, average performance for each class, Grad-CAM visualizations, Precision-Recall curves, Mean Average Precision (mAP), localization accuracy, and error analysis).

Additionally, there will be comparisons made between different CNN models from several datasets, and final comments regarding both the strengths/weaknesses of the various CNNs, with some suggestions for CNN model selection and their potential real-world applicability.

Chapter 7: System Implementation

Two web-based systems are implementation strategies: (1) the OPG-Based Dental Decision Support System for dentists and (2) the RGB Teeth Health Monitoring System for general users.

Both systems are built on the YOLO12s models and have a modular, service-oriented architecture consisting of a frontend interface, backend server, and ML inference module. Besides, the integration of classification and detection models into a responsive web interface has been discussed along with technical considerations, challenges faced, and limitations such as the inability to include healthy-class predictions in RGB classification for public use.

Chapter 8: Discussion of Key Findings

Classification and detection experiments were the main source of the key findings that were summarized through a trend interpretation of the performance, preprocessing impacts, and data quality. The chapter presents the practical implications for the dental diagnostics world, such as the model generalizability, operational environments, and clinical importance. It also discusses the insights into the trade-off between the multi-label classification and object detection strategies for efficient and interpretable dental disease detection. The initial research questions have also been answered in this section.

Chapter 9: Threats to Validity

The assessment is carried out for potential sources of threat to internal, external, construct, and conclusion validity. Issues raised are the dataset bias, the preprocessing impacts, the data splitting leakage, the population diversity, the device and lighting variations, the metric appropriateness, and the statistical robustness. Furthermore, this section highlights the collapsed strategies and limitations that are implemented to ensure the reliability and reproducibility of the research results by providing transparency and instilling confidence.

Chapter 10: Conclusion and Future Work

This last chapter recaps the study, reinforcing the message that the analysis of multiple deep learning frameworks for dental disease classification and detection has been successful.

The report reiterates the contributions, the practical applications, and the performance highlights. As for the future directions, they include the making of more comprehensive datasets, the strengthening of the methodological robustness, the pondering of real-time deployment, and the merging of XAI techniques to facilitate the clinical adoption and patient-centered health monitoring.

2 Related Works

2.1 Imaging Approaches for Dental Diagnostics

AI-powered medical imaging has come a long way, and its making a real difference in how dentists spot problems. These days, deep learning models like Visual Geometry Group (VGG) and Residual Neural Network (ResNet) aren't just catching cavities, calculus, and tooth discoloration; they're actually beating the old-school visual exams for accuracy and consistency [12]. Transformer-based models push things even further, helping dentists pick up on issues like dental caries and hypodontia, and they come with explainable AI features that make it easier for clinicians to trust what they see [13]. When it comes to checking for bone loss around teeth, CNN models can now perform as well as professional dentists, as they cut down on disagreements between experts and make screening more reliable, especially with panoramic X-rays [14]. Moreover, deep learning models such as No New U-Net (nnU-Net) and Densely Connected Convolutional Network 121 (DenseNet-121) also stand out, as they match expert dentists when it comes to spotting and classifying dental caries lesions on panoramic radiographs [15]. Of course, it's not all smooth sailing. There are still big hurdles, which include keeping patient data private and secure, making these AI systems less of a black box, sorting through legal and ethical issues, dealing with messy or limited data, and actually getting this tech to fit into everyday dental practice, but the progress so far is pretty remarkable [16].

2.2 Imaging Modalities: Reflective vs Penetrative Dental Images

According to modern dental diagnostics, there are basically two modalities to diagnose dental problems: Reflective imaging and Penetrative imaging. Reflective imaging uses regular light to take pictures of what a dentist can see, like using a normal camera. But Penetrative imaging uses X-rays to look inside the teeth and jaw to find hidden issues.

Reflective tools like standard RGB photography have become more important lately, especially for non-invasive screening and the telemedicine sector. Garg et al. built a lightweight machine learning model that spots oral calculus just from smartphone photos. There is no need for a clinic visit or any radiation [1]. Ahmed et al., on the other hand, pushed things further by releasing the first annotated dataset of intraoral photos for public use so that any researchers can train AI models to detect dental caries from standard intraoral images that have five types of views [17]. When it comes to spotting things happening under the surface, e.g., root infections or impacted teeth, penetrative imaging is far superior. Panoramic X-rays, or OPGs, are a staple here. Adnan et al. tapped into this by designing a deep learning framework using the YOLOv5 model, which can automatically classify tricky dental problems straight from OPGs [18]. Besides, Haghani-far et al. have shown how panoramic X-rays can help automatically pick out individual

teeth and catch deep dental caries; all of this credit goes to the strong contrast that is obtained from X-ray imaging [19]. In the end, these two approaches cover all the bases as Reflective images give us a quick, non-invasive look at the surface, while penetrative images dive deep for a thorough internal check.

2.3 Traditional (Non-Deep Learning) Approaches for Dental Image Analysis

Before the era of deep learning in medical imaging, dental diagnoses mainly depended on the skills of dentists, supported by older machine learning methods. This common method is based on dentists regularly assessing their work and thinking carefully to understand what they see. This is a standard way of teaching and practicing in the field. In the shift toward automation, older computer approaches used simpler learning methods, which needed features to be pulled out by hand [20]. Efitli et al. offer an example of this with their use of Random Forest methods, a common ensemble approach, to accurately guess tooth color from mouth images, showing that good diagnosis can be done without deep neural networks [21]. Furthermore, comprehensive reviews by Wang et al., Najeeb et al., and Fatima et al. further describe this change, saying that while older methods like Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) were the initial wave of AI tools for fixing and diagnosing issues, their reliance on set features limits them when compared to how modern deep learning systems automatically learn features [16, 22, 23].

2.4 Deep Learning in Dental Imaging

DL has changed dental imaging analysis by automating difficult diagnostic work on both reflective and penetrative methods. In panoramic radiography, Jaiswal et al. used DL to tell the difference between child and adult teeth, which is a basic step for age estimation and treatment planning [24]. Building on this penetrative way to study, Almalki et al. used transfer learning models on OPG images to sort different dental problems with high precision [25]. In more complex X-ray diagnosis datasets, involving multiple labels, Alsakar et al. presented an innovative dual-stream setup that uses MobileNetV2 and Swin Transformers. This setup grabs both local details and the overall picture [9]. When it comes to finding specific diseases, there have been gains, too. Chen et al. worked on spotting periodontitis and dental caries in X-ray images [26]. AlSayed et al. made progress in reflective imaging, using a group of CNNs to enhance dental caries classification accuracy using standard oral photographs [27].

2.5 CNN Architectures Used in Medical Image

CNNs now sit at the heart of medical image analysis. What started as simple classifiers has grown into a whole ecosystem of advanced detection systems. Early methods leaned heavily on handcrafted algorithms, but deep learning models have changed the game as they pulled out complex features on their own, which is exactly what is needed for things like denoising, segmentation, and classification. In this rapidly changing field, object detection has moved forward much faster [5].

Weaving in ViT and hybrid attention mechanisms, and building them right into core architectures to better handle different data types, like Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans, is being broadly experimented with nowadays

[7]. The YOLO series deserves a special mention here as these models have become essential for real-time analysis, and recent studies track how each new version tweaks the balance between speed and accuracy, which is something absolutely crucial in clinical environments [28]. But the progress doesn't stop at plugging in new architectures. Research teams are now digging even deeper, tweaking the inner workings of these networks to squeeze out even better performance. One of those approaches combines CNNs with bio-inspired computing; specifically, an improved tunicate swarm algorithm to fine-tune hyperparameters for oral cancer detection [29].

Others take a different route, changing network mechanics themselves, e.g., Common Vector Approach Pooling (CVApool) is a new pooling method that taps into the null-space of CNN weights, replacing average pooling. The result is much better feature retention for tooth disease classification [30].

2.6 Model Comparison Strategies in Medical AI

AI models in healthcare have come a long way. Early on, people just checked if the models worked, but now the focus has shifted to detailed statistical and operational evaluations. When researchers look at model design, they pay close attention to the balance between detection accuracy and inference speed, especially when comparing single-stage detectors like YOLO to more complex two-stage detectors, e.g., Region-based Convolutional Neural Network (R-CNN) variants like Fast R-CNN and Faster R-CNN [31]. Studies also dig into how each version of YOLO, from v1 to v8, has improved to keep up with the demands of real-time use in clinics, as they perform better in some cases than R-CNN variants [32, 33].

But it's not just about picking the right architecture. Researchers now combine methods such as pairing "Transformers" with Genetic Algorithms (GA) to push feature extraction and classification even further [34]. Despite all this technical progress, the best model isn't just the one with the highest accuracy; rather, real-world healthcare needs more than that. The Medical Machine Learning Operations (MedMLOps) framework points out that what matters most is how these models fit into clinical workflows over time, prioritizing ongoing monitoring, safety, and reliability, instead of relying on one-time metric checks [35].

2.7 Object Detection and Segmentation Techniques in Dentistry

The advancement of computer vision techniques has led to the development of very precise object detection and segmentation. Granular analysis of dental pathologies is now possible due to this shift in technique from simple classification to computer vision techniques. The YOLO framework, among others, has been a key player during this transition thanks to its real-time performance; for instance, in one study, the use of YOLO architectures for the automatic detection of dental plaque and counting oral hygiene, using the O'Leary Index, was reported [36].

Likewise, the contemporary YOLOv8 model was employed in another study to reliably recognize a vast number of dental diseases [37].

The approach to complex radiography was further refined by showing the efficiency of single-stage detectors in locating exact dental anomalies in panoramic images. Instance segmentation not only provides bounding boxes but also offers pixel-level information that is essential for surgical planning [38].

In a similar vein, researchers used Mask R-CNN on Cone-Beam Computed Tomography (CBCT) scans to identify teeth and simultaneously extract dental caries lesions [39]. Also, this segmentation feature applies to panoramic X-rays, wherein the development of automated systems aimed at the extraction of individual teeth and deep-seated dental caries detection has finally led to a convergence of 2 Dimension(2D) and 3 Dimension (3D) precision in diagnostics [19].

2.8 YOLO Family Models for Object Detection

The trajectory of the YOLO framework has been incredibly significant for real-time medical object detection in terms of not only speed but also accuracy. The detailed reviews of the architecture disclose the journey of the models from the first versions to the most updated ones like YOLOv11, which have not only constantly improved their performance but also tailored it to the needs of resource-limited medical settings [28, 40].

This non-stop progress has made it possible for the dentistry industry to accept it wholeheartedly, where the efficiency of the process is a top consideration. For e.g., a systematic review and meta-analysis has conducted a critical evaluation of the YOLO framework's performance in dental caries diagnosis that sensitivity was indeed comparable for all modalities, but the radiographic ones provided significantly greater specificity than the photographic ones [41].

Models have been further customized for specific applications, with one research investigating the use of different YOLO variants to detect a variety of tooth diseases and thus proving the framework's strength in the area of pathology identification [42]. Besides that, it has been the recent usage of the newest version, like YOLOv8, that has allowed the development of the low-cost and real-time screening tools capable of distinguishing dental caries, cracks, and healthy teeth from intraoral camera images, consequently, bringing the high-end AI diagnostics closer to the everyday clinical practices [43].

2.9 Explainable AI in Medical Imaging

The deployment of deep learning models in the clinical setting has risen to the point where the demand for XAI and stringent ethical standards is at the same height. The literature on the subject indicates that the non-interpretability of "black box" models is the most important barrier to turning XAI into a major trend for the acceptance of automated diagnosis [6]. Their solution was to resort to the visualization techniques, such as Grad-CAM. Grad-CAMs ability to generate medical image explanations was reviewed critically in a recent research by Suara et al. [44], while Ali et al. describe the successful application of the method within an interpretable framework, demonstrating the reason behind the multi-class dental disease classifications [2]. Technical transparency is one side of the coin, the other being ethical deployment, which is equally important. Studies also point to the ethical problems in pathology AI, and they also mention the necessity of accountability and bias mitigation as measures to ensure that health inequalities do not arise [45]. Simultaneously, protecting the patients' data is still a concern, and recent research shows how privacy-preserving machine learning systems can be designed that would ensure data security all the way from training to inference [46].

3 Dataset Materials and Preprocessing

3.1 Overview of Data Sources

This study incorporates three external secondary datasets to enable comprehensive experimentation in dental image classification and object detection. The sample images for all dataset classes are shown in Figure 3.1.

- **OPG Dataset**

The first dataset is the OPG dataset by Rahman *et al.*, which consists of 232 anonymized panoramic radiographs that were collected under expert supervision from three dental clinics in Dhaka, Bangladesh [47]. Although this dataset supports both classification and object detection tasks and is one of the few curated, annotated OPG benchmark datasets available for dental informatics research, it has several limitations, one of which is that it is incorrectly annotated for classification purposes. It is originally a multi-label dataset, but the classification dataset was designed for single-label classification. As a result, it poses a threat to the validity of our research study while experimenting with single-label classification models, and thus might lead to multiple interpretations.

- **RGB Teeth Dataset**

The second dataset of the study, the RGB Teeth Dataset [48], includes 5,048 RGB intraoral photographs categorized into five common dental conditions: dental calculus, dental caries, hypodontia, mouth ulcer, and tooth discoloration. The source of this dataset is still unverified.

- **Intraoral Caries Detection Dataset**

The third dataset, the Intraoral Caries Detection Dataset [17], contains 6,313 intraoral clinical images collected from individuals aged 10-24 years in Mithi, Sindh, Pakistan. All the images in this dataset were professionally annotated for dental caries using LabelMe and exported in YOLO, Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), Visual Object Classes (VOC), and Common Objects in Context (COCO) formats. Therefore, it provides diverse imaging modalities, e.g., RGB intraoral photography and clinical intraoral imaging, and serves as a complementary resource for evaluating deep learning approaches across varying dental contexts.

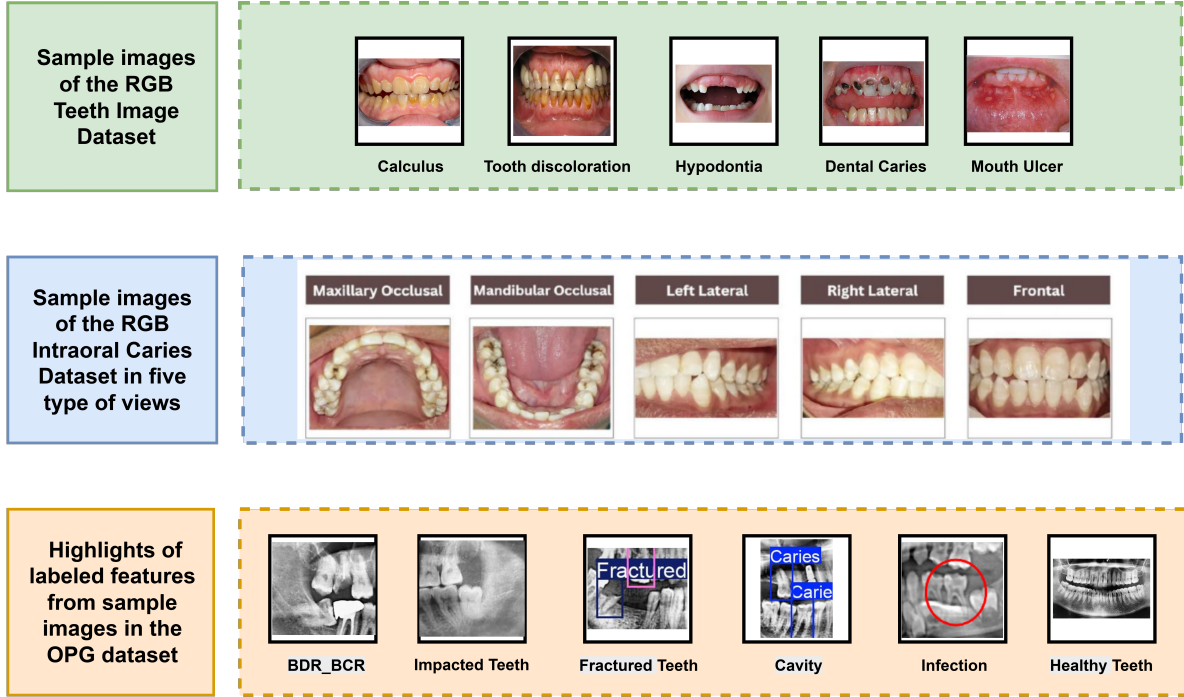


Figure 3.1: Sample images from the datasets showing labeled classes for both RGB and OPG

3.2 Background study of the dataset classes

3.2.1 Healthy Teeth

Healthy teeth are the best example of dental anatomy and the standard for diagnosing. They have sound enamel, good positioning, and no active problems, and in general, a healthy tooth is very smooth with a translucent white to ivory color and no visible cracks, holes, or stains. The gum around the tooth is hard, pink (coral pink), and this shows that there is no periodontal disease. Properly defining this "normal" class is essential for anomaly detection models to have a lower rate of false positives [47].

3.2.2 Caries

The term dental caries refers to the lesions (cavities) created when the tooth enamel gets demineralized by the acids from the oral bacteria. It is an irreversible process that, if not treated, will eventually reach the dentin and pulp. Early dental caries may be seen in RGB images as chalky white spots (demineralization) or dark brown/black cracks on the occlusal surfaces. Along with the decay, the formation of cavitation (holes) becomes evident. The presence of dental caries in radiographic images is indicated by dark spots (radiolucent) within the surrounding whitish (radiopaque) enamel and dentin. Detecting dental caries at an early stage is crucial to avoid large-scale restorative treatments [47].

3.2.3 Impacted Teeth

Impacted teeth are the ones that have not been able to erupt into their proper functional position because of various reasons, such as a shortage of space, being obstructed by other teeth, or having abnormal development paths. Third molars (wisdom teeth) are of most common cases of this situation. In RGB images, impaction might be depicted as

a tooth that is partially erupted and has associated gum inflammation (pericoronitis) or the crowding of nearby teeth. However, X-rays clearly show them, revealing the tooth that is buried under the bone or gum, and often is at an abnormal angle in relation to the jaw [47].

3.2.4 Broken-down Crown/Root

The term Broken-down Crown (BDC) or Broken-down Root (BDR) is a classification for teeth that have experienced a major loss of structure. This condition arises out of neglectful management of very large carious lesions, multiple failed restorations, or very severe trauma, thus making the tooth non-restorable with the conventional methods. These are the teeth that are often seen as extremely weak, and they are frequently nothing but gum-level parts, or stumps of roots that are not very well defined. A large part of the crown has gone, and the little bit left is usually dark in color because of decay. On the X-ray, these teeth look like broken pieces with probably inflamed areas at the apex [47].

3.2.5 Infection

Infection in the tooth, among other dental diseases, takes place when the bacteria first attack the pulp, then the apical area of the alveolar bone and soft tissues. Such a condition is called a dental abscess, and baric therapy is one of the options for treatment. Swelling and redness are usually the signs of infections in clinical RGB images, besides the parulis (gum boil) or draining sinus tract from which pus is coming. In radiographs, infections show up as radiolucent areas (dark shadows) around the apex of the root, meaning that there is either bone resorption or abscess formation [47].

3.2.6 Fractured Teeth

When a tooth has been fractured, a break in enamel, dentin, or root structure has occurred. Fractured teeth can be caused by trauma (such as a fall) or excessive biting forces on a compromised tooth, and there is a characteristic appearance with very conspicuous lines or fragments missing from the crown of the tooth. The fracture can occur in varying degrees of severity, ranging from superficial craze lines (also known as hairline fractures) to deep fractures that expose the inner dentin and/or pulp. These fractures appear as sharp, linear breaks on the surface of the tooth. Radiographs are used to help determine the depth of the fracture, particularly if beyond the tooth's gumline [47].

3.2.7 Calculus

Dental calculus, also known as tartar, is an accumulation of plaque that hardens and forms a solid mass due to the mineralization of the bacterial biofilm. Among other things, it is mainly responsible for the lack of good oral hygiene that allows soft plaque to grow. In standard RGB images, calculus is portrayed as hard, crusty deposits located either at the gum line (where the gums and teeth meet) or in between teeth. It usually appears as a white, yellow, or brownish material that is visible against the smooth and shiny surface of the natural tooth enamel. In extreme cases, supragingival calculus can be observed as dark brown or black spots, often, at the same time, there is concurrent gingival inflammation (redness) [2, 12].

3.2.8 Hypodontia

Hypodontia is an inherited condition that is marked by the total non-development of one or several teeth (except the third molars). This is a very common dental anomaly and can occur in both primary and permanent dentition. In clinical photography, the condition is detected by the presence of noticeable spaces known as diastema in the dental arch where certain teeth are anatomically expected to be, but are not. Unlike spaces left by extraction or trauma, the surrounding gum tissue is typically not only healed but also smooth. The condition leads to the drifting or spacing of the adjacent teeth, which, in turn, provides the dentist with a secondary visual clue [2, 12].

3.2.9 Mouth Ulcer

Mouth ulcer, also known as aphthous stomatitis, is a very uncomfortable sore that forms on the soft tissues of the mouth or at the gingival line. They are not contagious but may point to a few different problems, such as trauma, stress, or even more generalized health issues. An ulcer is noticed to be round or oval in size, a sore. A typical ulcer has a central necrotic or fibrin slough area that is white, yellow, or gray. This central part is surrounded by a red, inflamed area, which is typically quite distinct, erythematous (red). These features make the ulcers contrast very strongly against the pinkish shade of the healthy oral mucosa [2, 12].

3.2.10 Tooth Discoloration

Tooth discoloration is understood as any unexpected change in the hue, color, or translucency of a tooth. The causes include a wide variety of superficial staining (like tobacco and coffee) and intrinsic factors (like fluorosis, tetracycline staining, or pulp necrosis). Discoloration in RGB datasets is recognized by assessing the variances from the natural ivory or white spectrum of healthy teeth. Extrinsic stains appear mainly as yellow or brown lines on the outer surface of the enamel. Intrinsic discoloration may be represented as grey and dull yellow or very pronounced horizontal bands fusing the tooth structure. AI technologies spot the shifts of the specific colors among the shadow artifacts or bad lighting [2, 12].

3.3 OPG Dataset for Classification

3.3.1 Dataset Overview

The OPG dataset [47] consists of 231 panoramic dental radiographs having six diagnostic classes and containing both multilabel and single-label samples. The Dental OPG X-rays, with the help of a dentist, were taken from three different private dental clinics, which are Prescription Point Ltd., LabAid Specialized Hospital, and Ibn Sina Diagnostic and Imaging Center, located in Banani and Dhanmondi, Dhaka, Bangladesh, respectively. The usage of safety measures was ensured during data collection, and patient identities were completely anonymous.

A high-resolution Android phone camera was used for the image-capturing process that resulted in a uniform 16:9 aspect ratio and a median resolution of 1176 (E 603 pixels). The dataset was organized into six groups: healthy teeth, dental caries, infection, impacted teeth, fractured teeth, and BDC/BDR. The dataset was, for experimental purposes, divided into two main components: an object detection dataset containing 232 original and 604 augmented images with corresponding JavaScript Object Notation (JSON) labels, and a classification dataset organized into six separate class folders. All images are in Joint Photographic Experts Group (JPG) format, and the augmented samples are further split into training, validation, and testing subsets. The complete dataset size is 197 Mega Bytes (MB) and is made accessible as a zip file that has been compressed for easy distribution.

3.3.2 Multi-Label Dataset Splitting Strategy

In this study, the dataset presents a challenging multi-label classification scenario in which each image may belong to one or more dental conditions. The overall label frequency is shown in Figure 3.2. A particular complication arises because the label combinations are highly inconsistent; there is no stable or repeated pattern of multi-label associations across most images. Many images share some labels but differ in others, and in some cases, the presence of *Healthy Teeth* appears alongside disease labels, which is logically contradictory.

Because of this irregular label structure, a standard randomized split can not preserve the distribution of multi-label patterns. Therefore, the dataset required a carefully designed deterministic splitting strategy based on the following principles:

1. **Logical label correction:**

If an image contains *Healthy Teeth* together with any disease class, the *Healthy* label is removed so that the sample reflects only the disease characteristics.

2. **Signature based grouping:**

Each image is grouped by a "signature", defined as the sorted list of its cleaned labels. Images sharing the exact same signature represent a consistent pattern and must be split together.

3. **Deterministic allocation:**

Since the number of images per signature varies widely (some patterns appear once, some twice, others many times), the dataset is split non-randomly but logically, ensuring the following key factors:

- All images from each signature are assigned deterministically,

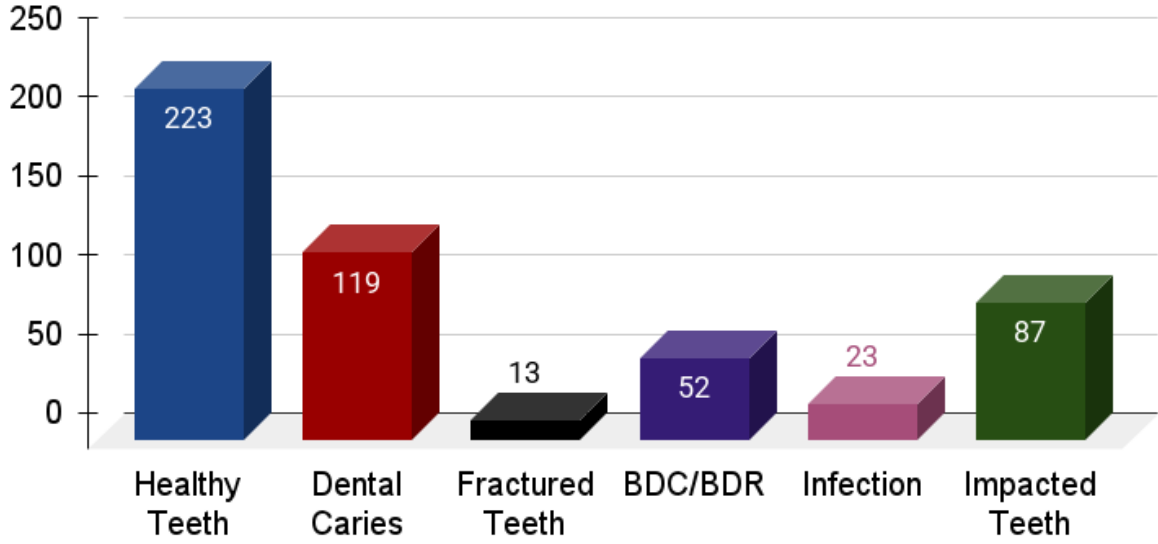


Figure 3.2: Label frequency distribution after dataset cleaning of the OPG dataset for classification task

- Every signature contributes to training and validation where possible,
- Extremely rare label patterns (i.e., appearing only once) are placed in `test_rare` to avoid bias.

4. Accuracy-oriented balancing:

For signatures with many samples, the split approximates 80% training, 10% validation, and 10% testing. This ensures broad representation in the training set while keeping test coverage meaningful.

Following these criteria, the data were partitioned into training and testing sets. The corresponding splitting scheme, which also adheres to the specified strategy, is presented in Table 3.1

Table 3.1: Split criteria based on signature size of the OPG dataset for classification task

Images	Split Allocation	Rationale
1	test_rare (1)	Unique pattern; placed only in rare-test to avoid overfitting.
2	train (1), validate (1)	Minimal balanced distribution for learning and validation.
3	train (2), validate (1)	Enough samples for learning while preserving validation stability.
4	train (3), validate (1)	Most samples support training; one ensures validation representation.
5 or more	train $\approx$ 80%, validate $\approx$ 10%, test $\approx$ 10%	Larger groups follow 80/10/10 split while keeping signature representation across partitions.

And as shown in Figure 3.3, after implementing these steps, the resulting ratios remain very close to the 80-10-10 criterion. This procedure ensures the highest possible reliability for multi-label training while respecting the logical and structural constraints of the dataset.

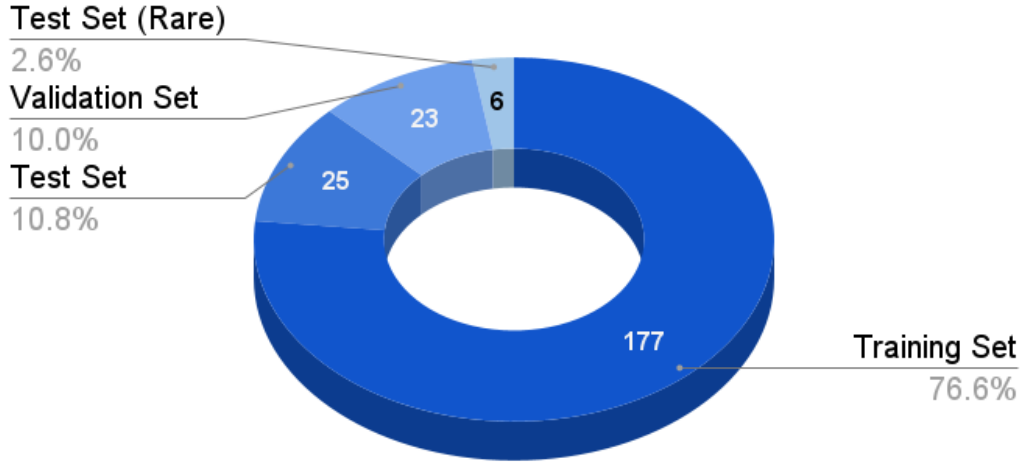


Figure 3.3: Dataset split overview of the OPG dataset for classification

3.3.3 Preprocessing and Data Cleaning Pipeline

During the preprocessing stage, the dataset initially appeared to consist of six separate folders, each representing a single-class structure. However, after a detailed analysis, it became evident that many images were duplicated across multiple folders, indicating that the dataset followed a multilabel pattern rather than a strictly single-label format. To resolve this, a comprehensive Comma Separated Values (CSV) file was created containing each image name along with its corresponding list of labels. Duplicate images were found and removed to create an accurate ground truth. All unique images were consolidated into a single cleaned folder to ensure consistency for further processing, illustrated in Figure 3.4.

3.3.4 Data Augmentation

To improve the generalization capability of the model and prevent overfitting, a specific data augmentation pipeline was established using the Albumentations library. This process introduced variations to the training dataset to simulate diverse clinical scenarios typically found in dental radiography. The geometric transformations included random horizontal flipping and slight rotations up to fifteen degrees to account for differences in patient positioning during X-ray acquisition. The system randomly changed the brightness and contrast and added Gaussian noise to simulate different sensor settings and quality. A coarse dropout method was also used to randomly block out small rectangular parts of each image. This helps the network focus on strong anatomical features instead of memorizing exact pixel patterns. Lastly, all images were normalized using the standard mean and standard deviation values from the ImageNet dataset to keep the inputs consistent for the neural network.

3.3.5 Contrast Enhancement

A standardized preprocessing system has been implemented for the raw radiograph images. The main enhancement step used Contrast Limited Adaptive Histogram Equalization (CLAHE), applied to each color channel with a set clip limit and grid size. This improved local contrast and edge clarity in dental structures while avoiding extra noise. After

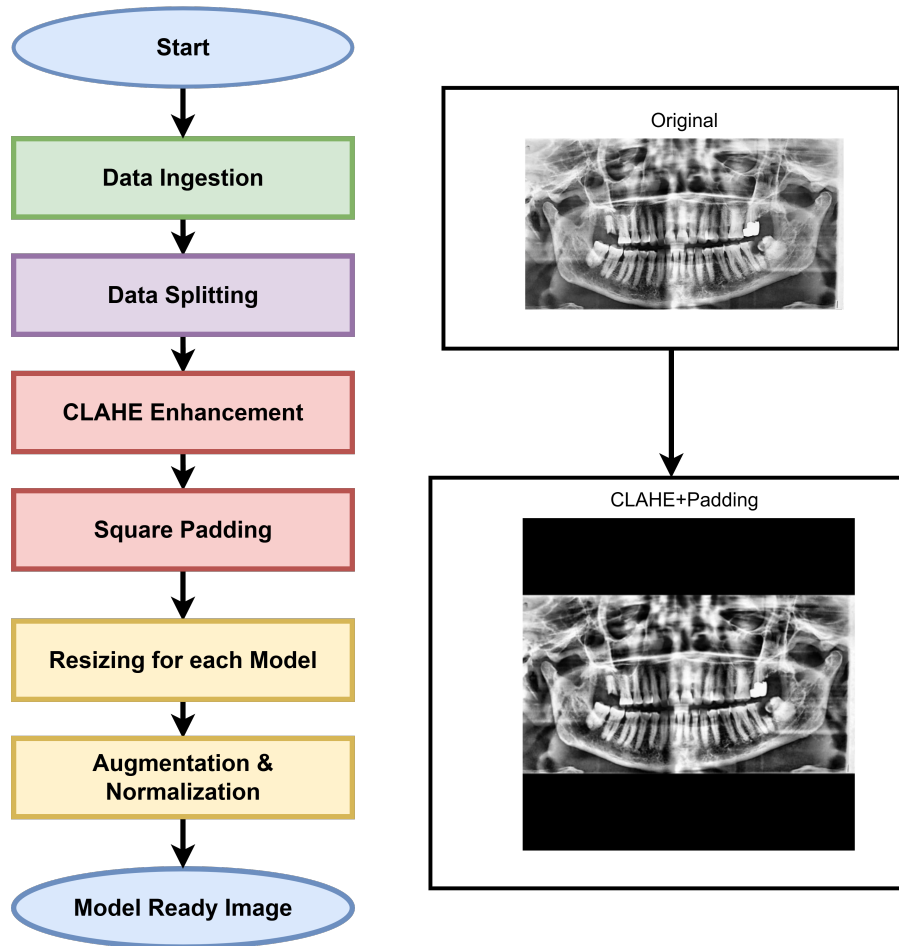


Figure 3.4: Image preprocessing pipeline of OPG classification

enhancement, we used black-padding to keep aspect ratios consistent by placing each image on a square black canvas. The image was scaled to make its longest side fit the canvas, and the shorter side was centered to preserve the natural proportions of OPG Images.

3.4 OPG Dataset for Object Detection

3.4.1 Dataset Overview and Annotation Criteria

For object detection, the OPG dataset includes the original 232 unique radiographs, along with expert-validated annotations generated with the Computer Vision Annotation Tool (CVAT). A dentist with 16 years of clinical experience performed the initial annotations, which a senior dentist subsequently reviewed with 32 years of experience. The dataset also contained a benchmark of 604 augmented images generated through rotations, flips, shearing, exposure adjustments, and mosaic augmentation. The benchmark scores of the augmented dataset were shared by the authors without mentioning the name of the model they used for the training process. So we had to evaluate the same dataset on several YOLO modes to understand its limitations. All annotations correspond to the six diagnostic categories used in the dataset. We used the augmented images for our experiments. However, there is a minor limitation: the test set contains no Fractured Teeth labels, which might affect the interpretation of the overall Mean Average Precision (mAP@50) and (mAP@95) in test-set results.

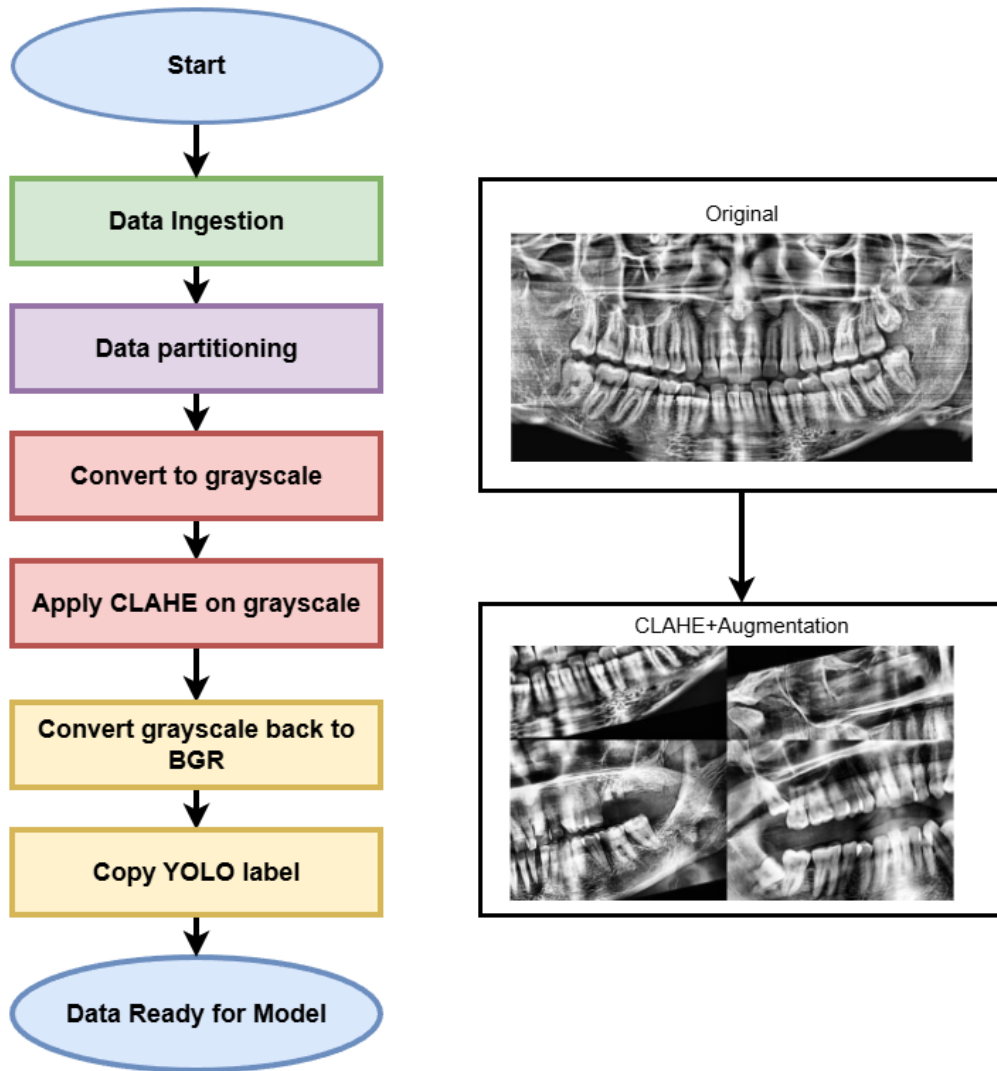


Figure 3.5: Sequential stages of the OPG data preprocessing pipeline

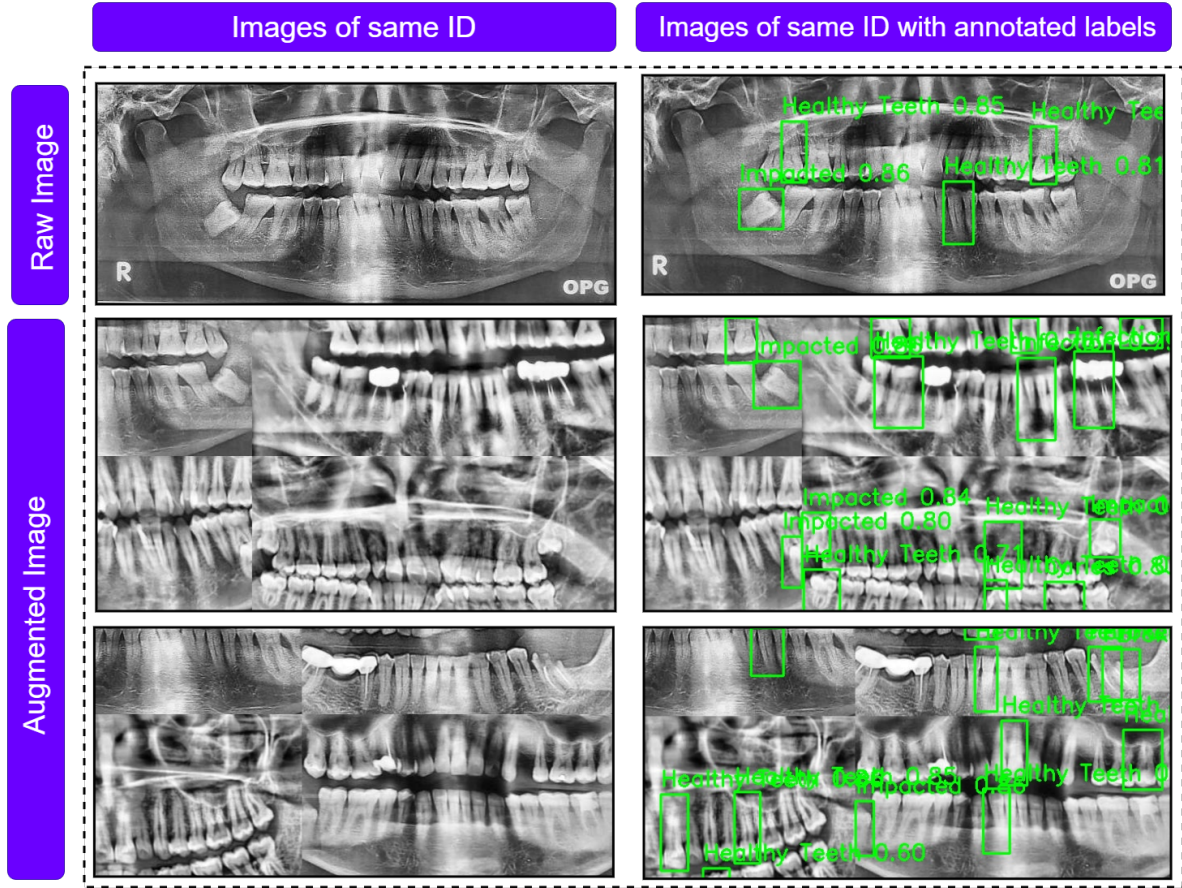


Figure 3.6: Sample visualization of training an image ID from the augmented OPG dataset

3.4.2 Class Distribution and Lesion Characteristics

The object detection component provides bounding-box labels for all six disease categories: Caries, Infection, Impacted Teeth, Fractured Teeth, Broken-Down Crown/Root, and Healthy Teeth. Lesions range from small radiolucent dental caries to large structural breakdowns such as fractures and BDC/BDR cases. Across raw and augmented images, the dataset includes a total of 1,204 annotated instances, providing a diverse and clinically meaningful benchmark for multi-class dental lesion detection.

3.4.3 Dataset Preparation

The dataset used in this study was paired with a YOLO-format annotation file containing bounding box coordinates and class labels for six diagnostic categories. To maintain compatibility with the YOLO framework, the dataset was then structured into three subsets: training, validation, and testing, where each of the subsets contained separate "images" and "labels" directories, where the "images" folder held the radiographs and the "labels" folder contained the corresponding annotation text files. This standardized organization ensured a consistent workflow for training, validating, and evaluating the YOLO models. The steps are summarized in Figure 3.5 and a sample image of ID: 147 from the training set is visualized in Figure 3.6.

3.4.4 Contrast Enhancement

To enhance contrast and improve visualization of dental structures, CLAHE was applied to all images in the dataset. Each radiograph was first converted to grayscale to isolate luminance information, and then CLAHE was applied with a clip limit of 2.0 and a tile grid size of 8×8 to enhance local contrast while preventing excessive noise amplification. Following the enhancement, the processed grayscale images were converted back to three-channel BGR format to ensure compatibility with the YOLO model input requirements. The original YOLO annotation files remained unchanged, as the preprocessing operations did not alter the spatial resolution or geometry of the images.

3.4.5 Dataset Configuration

A YOLO-compliant configuration file, ".yaml" file, was created to facilitate training. The configuration file played a crucial role in enabling the YOLO models to correctly interpret the dataset structure, access the appropriate image folders, and map prediction outputs to their corresponding diagnostic categories.

3.4.6 Multi-Label Dataset Splitting Strategy

To ensure balanced model training and unbiased performance evaluation, the dataset was by default divided into three subsets: training, validation, and testing. Table 3.2 summarizes the number of images in each set, for the default augmented dataset that was initially prepared by the dataset publishers for object detection purposes.

Table 3.2: Class-wise distribution across train, validation, and test sets in the OPG augmented dataset for object detection

Class	Total Images	Training	Validation	Test
Caries	119	97	8	14
Infection	87	82	3	2
Impacted Teeth	52	38	7	7
Fractured Teeth	23	20	3	0
Broken-Down Crown/Root	13	3	6	4
Healthy Teeth	310	408	22	23
Total	604	558	23	23

3.5 RGB Teeth Dataset for Classification

3.5.1 Dataset Overview

The RGB Teeth Dataset [48] consists of 5,048 reflective RGB images captured under natural lighting conditions. These images are categorized into five diagnostic classes: dental calculus, dental caries, hypodontia, mouth ulcer, and tooth discoloration. The dataset captures a range of color variations, surface reflectivity, and intraoral anatomical differences, making it suitable for evaluating CNN-based dental image classification models. The mixture of these groups encompasses a huge set of dental images that captures pathological and abnormal dental conditions in teeth, which is a good foundation for the development of a robust classification model.

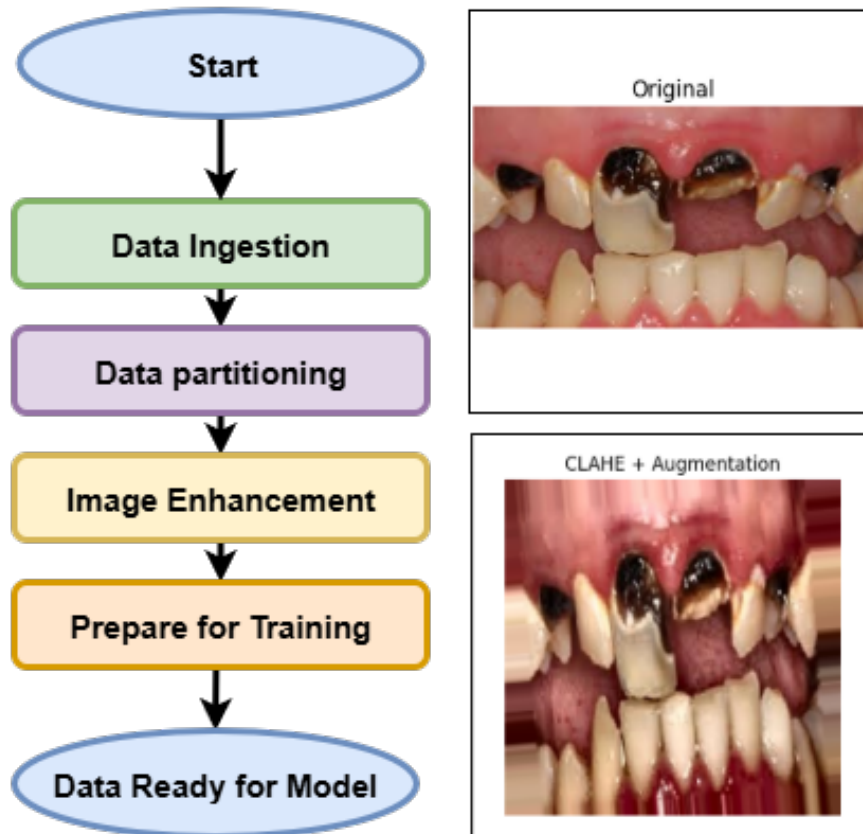


Figure 3.7: Sequential stages of the RGB data preprocessing pipeline

3.5.2 Contrast Enhancement

Dental images have poor contrast, which obscures important diagnostic features such as dental caries. To overcome this, we applied CLAHE, as shown in Figure 3.7, to the L-channel of L (lightness), a (green-red axis), and b (blue-yellow axis) (LAB) color space, thereby enhancing the local contrast without sacrificing global color detail of the tooth. To this process, a clip limit of 3.0 was employed to control noise amplification, and a 4×4 tile grid size was employed, dividing every image into sixteen regions for local equalization. However, it is better to be aware that while structural visibility is improved through the application of CLAHE, the use of contrast can alter the visual perception of color and may impact classification performance in those categories where a discriminative feature is color, for example, tooth staining. Previous radiographic work has suggested that

Table 3.3: Dataset class distribution and data splits of the RGB Teeth dataset

Class	Total Images	Training	Validation	Testing
Calculus	1,296	907	194	195
Caries	219	153	33	33
Hypodontia	1,251	875	188	188
Mouth Ulcer	265	185	40	40
Tooth Discoloration	2,017	1,411	303	303
Total	5,048	3,531	758	759

brightness and contrast manipulation neither compromises diagnostic validity for dental caries diagnosis nor produces images that are unacceptable for this purpose, although it does affect observer preference and visual perception. In contrast to manual adjustment of contrast/brightness, CLAHE provides automatic and spatially adaptive enhancement, which may potentially have a more significant impact on color-dependent operations and thus requires meticulous examination when discoloration is an intended effect.

3.5.3 Dataset Augmentation

In order to improve the training process of high-resolution RGB dental images, a complete preprocessing and data augmentation pipeline was developed. This pipeline led to the enhancement of feature extraction and stability of the model. The initial step taken for every image was to normalize it, this was done through the rescaling of pixel intensities from the standard $[0, 255]$ range to a floating-point range of $[0, 1]$, which virtually standardized the red, green, and blue color channels and at the same time prevented the explosion of the gradient during backpropagation. The dataset was then subjected to a series of stochastic augmentations intended to mimic real-world clinical variations. These augmentations addressed the class imbalance and improved the spatial invariance. The pipeline not only performed simple rescaling but also included random horizontal flips to reflect the bilateral symmetry inherent in dental anatomy. A zoom range of 5% was also employed to effectively simulate variations in sensor proximity and patient positioning. Also, random rotations of up to 10° were added to the model’s training in order to make the model invincible to slight head tilts or misaligned captures. The aforementioned geometric transformations were carefully selected to keep the morphological characteristics of dental pathologies yet at the same time artificially expand the diversity of the training set. This significantly increases its generalization and reliability on unseen RGB clinical data.

3.5.4 Single-Label Dataset Splitting Strategy

The dataset was initially organized into five class-specific folders. The images with five classes were randomly shuffled in the sub-folders. For maintaining an unbiased and rigorous evaluation of the classification model, the dataset was split into training, validation, and testing sets. Specifically, 3,353 images (70%) were left for training, 758 images (15%) for validation of model development, and 759 images (15%) for final testing. The class distribution of the images across the five classes is shown, and splitting is detailed in Table 3.3.

3.6 Intraoral Caries Dataset for Object Detection

3.6.1 Dataset Overview and Annotation Criteria

The Annotated Intraoral Caries Dataset [17] comprises 6,313 intraoral images that include various intraoral viewpoints with and without cheek retractors, representing both mixed and permanent dentitions.

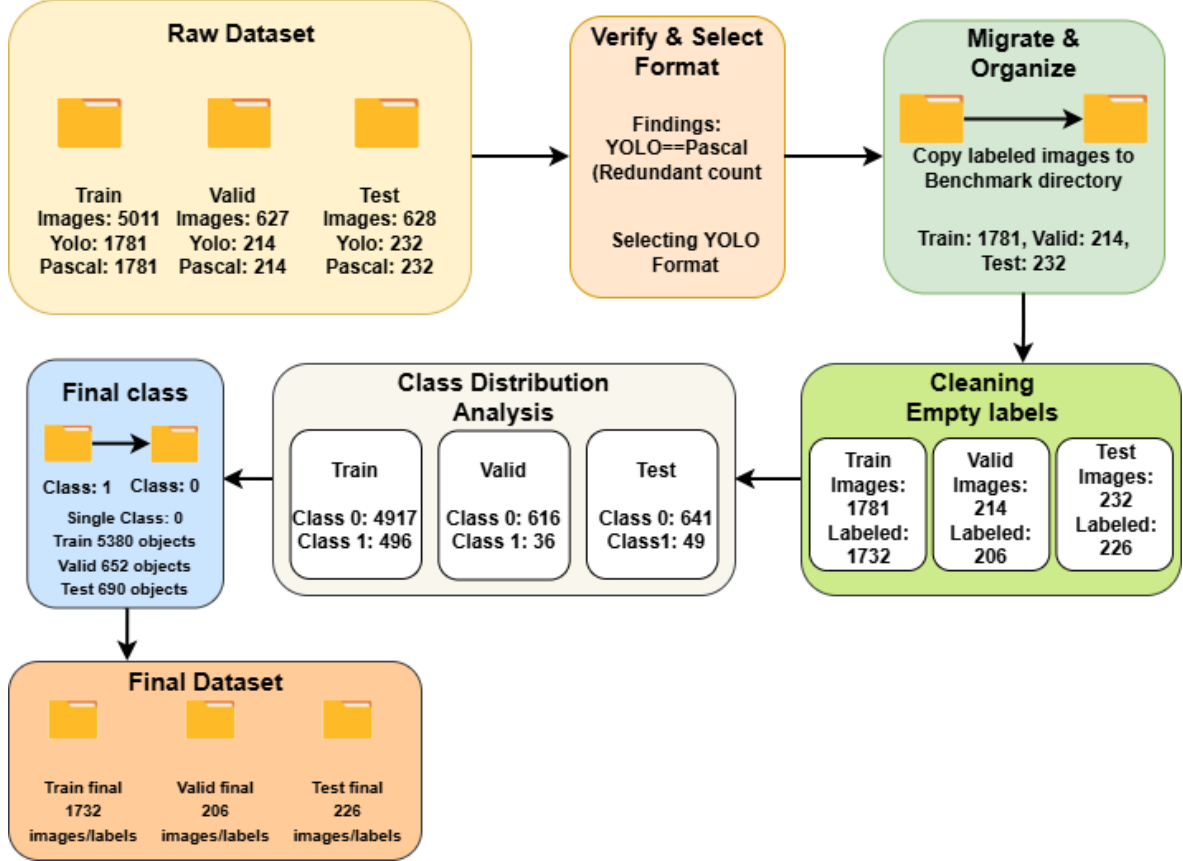


Figure 3.8: Sequential stages of the Intraoral RGB dataset preprocessing pipeline

This dataset focuses exclusively on dental caries. Each bounding box precisely localizes carious lesions across a range of severities and tooth surfaces. The carefully curated bounding-box annotations provide high-quality training examples for small-object and low-contrast intraoral lesion detection.

3.6.2 Dataset Preparation

In Figure 3.8, the various processes of dataset preprocessing are highlighted. During the initial stage of the study, a multi-step preprocessing pipeline was aimed at changing a mixed and uneven lot of images into a single and trustworthy dataset that would then be used for training a model competent enough to detect objects with high performance. The unrefined data was sorted out into three main folders - Training, Validation, and Testing, which contained thousands of image files more than the number of annotations made. The training folder alone had more than 5,000 image files, but a preliminary check revealed that only a few of them had corresponding labels pointing to the truth. To ensure the quality of the data, a thorough cross-checking process was carried out, and it was revealed that the number of files for the YOLO and Pascal annotations was exactly the same in

all splits, especially 1,781 labeled files in the training set. This coincidence confirmed that the bounding box data was identical across the different file types; thus, a very decisive standardization strategy was adopted. Hence, the YOLO format was selected as the single annotation standard for the rest of the pipeline on account of its simple text-based structure and direct compatibility with the target detection architecture, while the redundant Pascal VOC files were systematically cleaned up to make the ingestion process easier and eliminate parsing uncertainties.

3.6.3 Dataset Configuration

When the annotation formats were standardized, the next step was a data migration and cleaning process that was to be a highly synchronized dataset. To do this, there was a selective migration script used to locate only the images that had a YOLO annotation file that matched, and those images were then copied from the source to a special benchmarking folder. But only copying files was not enough to ensure reliable training, and thus a more thorough integrity check was done to discover "orphaned" imagesfiles that were physically in the labeled directory but were paired with missing or empty text files. In supervised object detection, keeping such images without valid bounding coordinates can unintentionally hurt the model during training by treating them as background negatives, which is a way of adding label noise. So, a filtering algorithm went through the new benchmarking folders one by one and checked if every label file existed and if the content was valid. This process led to the discovery and removal of 49 orphaned images from the training set, 8 from the validation set, and 6 from the test set. This thorough cleaning kept the final dataset to consist of valid 1:1 image-label pairs only, and thus it raised the learning algorithm's signal-to-noise ratio to the maximum possible level.

3.6.4 Preprocessing and Data Cleaning Pipeline

An exploratory analysis of the validated labels pointed to an initial binary classification scheme with two classes: Class 0 and Class 1. However, the distribution was extremely unbalanced, with the instances of Class 0 outnumbering the Class 1 ones by a large margin in all the dataset partitions. This extreme class imbalance is a major challenge in deep learning and often results in the models being biased toward the majority class and, thus, unable to generalize well to the minority class. Since the main goal of this research is the generic object localization, that is, identifying the presence and determining the coordinates of an object regardless of its exact subtype, the class structure was unified. A transformation script was run to change all the instances of Class 1 to Class 0, thus turning the problem into a single-class detection task.

The convergence of the model was not only facilitated by the elimination of the imbalance issue but also by the removal of the inter-class discrimination requirement. The combination of these three phases: standardization, cleaning, and unification, which have been very carefully applied, produced a final dataset that was composed of 1,732 training pairs, 206 validation pairs, and 226 testing pairs, which were all standardized to one class and were ready for the best model performance.

4 Materials and Methodology

4.1 Proposed Framework Overview

The research methodology, as illustrated in Figure 4.1 is a structured five-phase workflow that involves the acquisition of three datasets as the first step: an OPG dataset used for both classification and object detection, an RGB Teeth dataset for classification, and an RGB intraoral teeth/caries detection dataset for object detection.

In the next phase, a unified preprocessing pipeline, which involves image resizing or padding, normalization, augmentation, denoising, and train-test splitting, is applied to all datasets.

After the preprocessing, the samples will be sorted into two experimental pipelines according to their modality in the next phase. The OPG pipeline will use multiple deep CNN architectures, e.g., ResNet-18, ResNet-50, DenseNet-169, Inception-v3, EfficientNet-B3, and MobileNetV2, along with the lightweight models of YOLO (YOLOv8 to YOLOv12 in nano, small, and tiny variants).

Parallely, the RGB pipeline will use ResNet-101, ResNet-50, DenseNet-121, Inception-ResNet-v2, MobileNetV2, NASNet-Mobile, and the same lightweight models of YOLO for classification and detection tasks.

Each model is being trained and validated independently, and the metrics are being calculated; afterwards, the results from both pipelines are subjected to a comprehensive comparison through cross-model and cross-task evaluation, which allows us to find out the best models for each modality.

This final analysis assists in the creation of two proposed solutions: an RGB-based dental health monitoring system for general users and an OPG-based dental decision support system for clinics, thus making sure that the findings are turned into efficient and practical applications.

4.2 CNN Architectures for Disease Classification

4.2.1 Residual Learning Frameworks

ResNet18

ResNet-18 can be thought of as a simple deep residual network that was presented by He et al. [49]. The model has 18 layers that are assisted by the skip connections, which lessen the gradient vanishing issue by enabling the identity mappings to carry the data directly through the layers. Since its depth is limited, ResNet-18 is able to provide inference at high speed, and it is less demanding on computer resources while still extracting features that are good enough for medical imaging tasks.

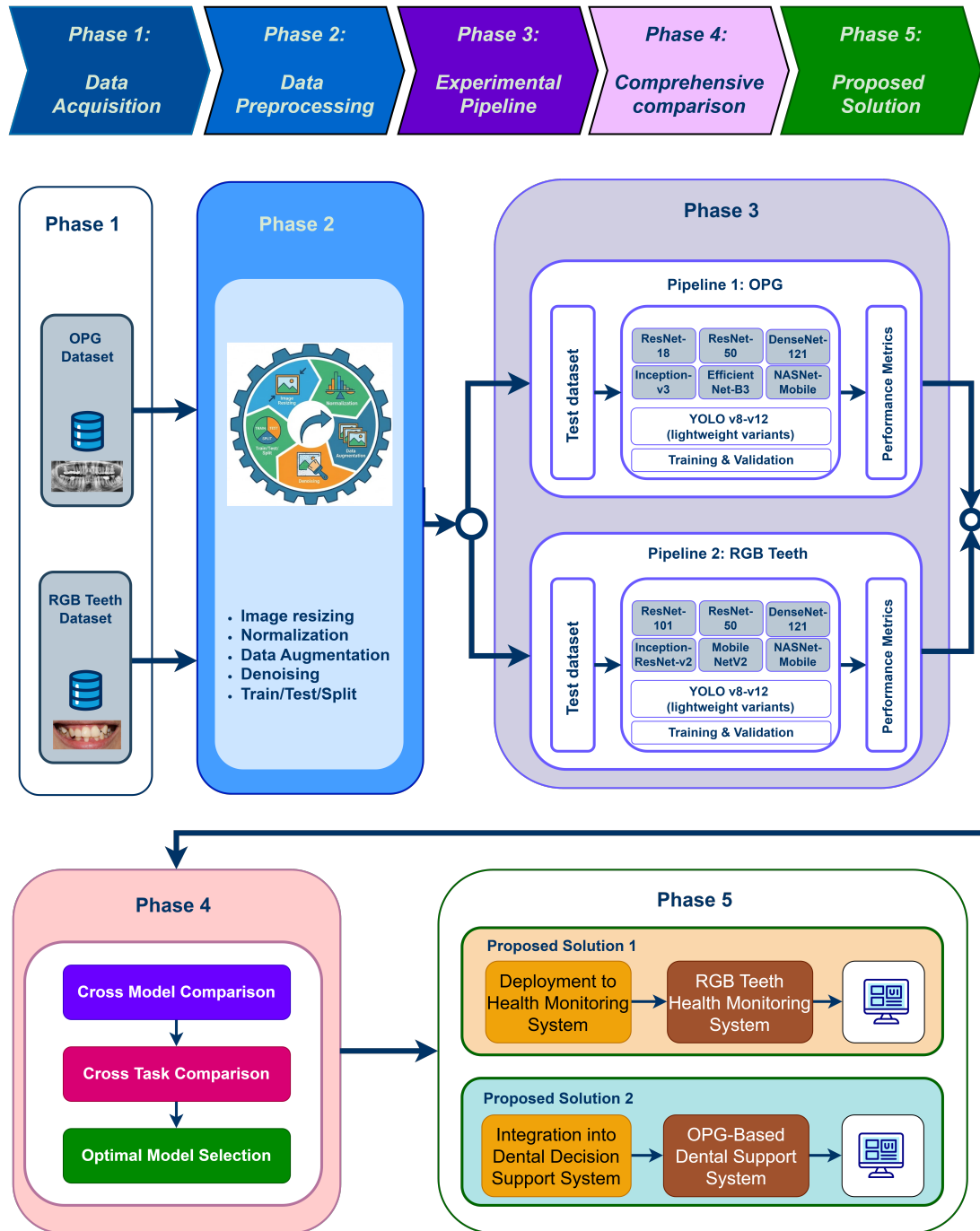


Figure 4.1: Proposed Methodology

ResNet-50

ResNet-50, while proposed by He et al. [49] as well, is a version with more layers of 50, which are created using bottleneck residual blocks. The bottleneck technique is for utilizing particular layers to compress the incoming data into a lower-dimensional representation that can be used for subsequent activities. The main goals of this method are to lower the number of parameters, simplify computation, and motivate the network to learn effective, abstract feature representations, which cuts down the required computations to a great extent, hence allowing the creation of deeper networks without spoiling their performance. The higher capability to express makes the model good enough for the difficult visual recognition tasks, like classifying dental diseases.

ResNet-101

ResNet-101 is a version of the ResNet family that has 101 layers and relies a lot on three-layer bottleneck blocks [49]. With the model being so deep, it can obtain multi-level semantic features that are very hard to discriminate from others. It is heavier in terms of computational power, but it is still the right choice for making deep hierarchical feature learning in very hard tasks.

4.2.2 Dense Connectivity Network

DenseNet-121

DenseNet121 was created by Huang et al. [50]. Each layer is connected to all the previous layers in a feed-forward manner, which leads to better re-use of features, quicker gradient flow, and a significant decrease in the number of parameters. DenseNet121's small footprint makes it very suitable for medical imaging applications where stability and low overfitting are required.

4.2.3 Lightweight Architectures

MobileNetV2

MobileNetV2, introduced by Sandler et al. [51], is a very effective convolutional architecture designed primarily for mobile and embedded systems. The model utilizes inverted residuals along with linear bottlenecks, which enable it to reach high accuracy levels with much lower parameter count. The resource-friendliness of the architecture is an important quality that makes it fit for dental diagnostic tools with restricted digitized resources.

NASNet-Mobile

NASNet-Mobile is the architecture Zoph et al. [52] got through audaciously employing Neural Architecture Search (NAS). It consists of the learned normal and reduction cells optimized to bolster accuracy under the constraints of mobile size. The model itself is the best performer when it comes to performanceefficiency trade-offs, thus making it the optimal choice for lightweight real-time inference applications.

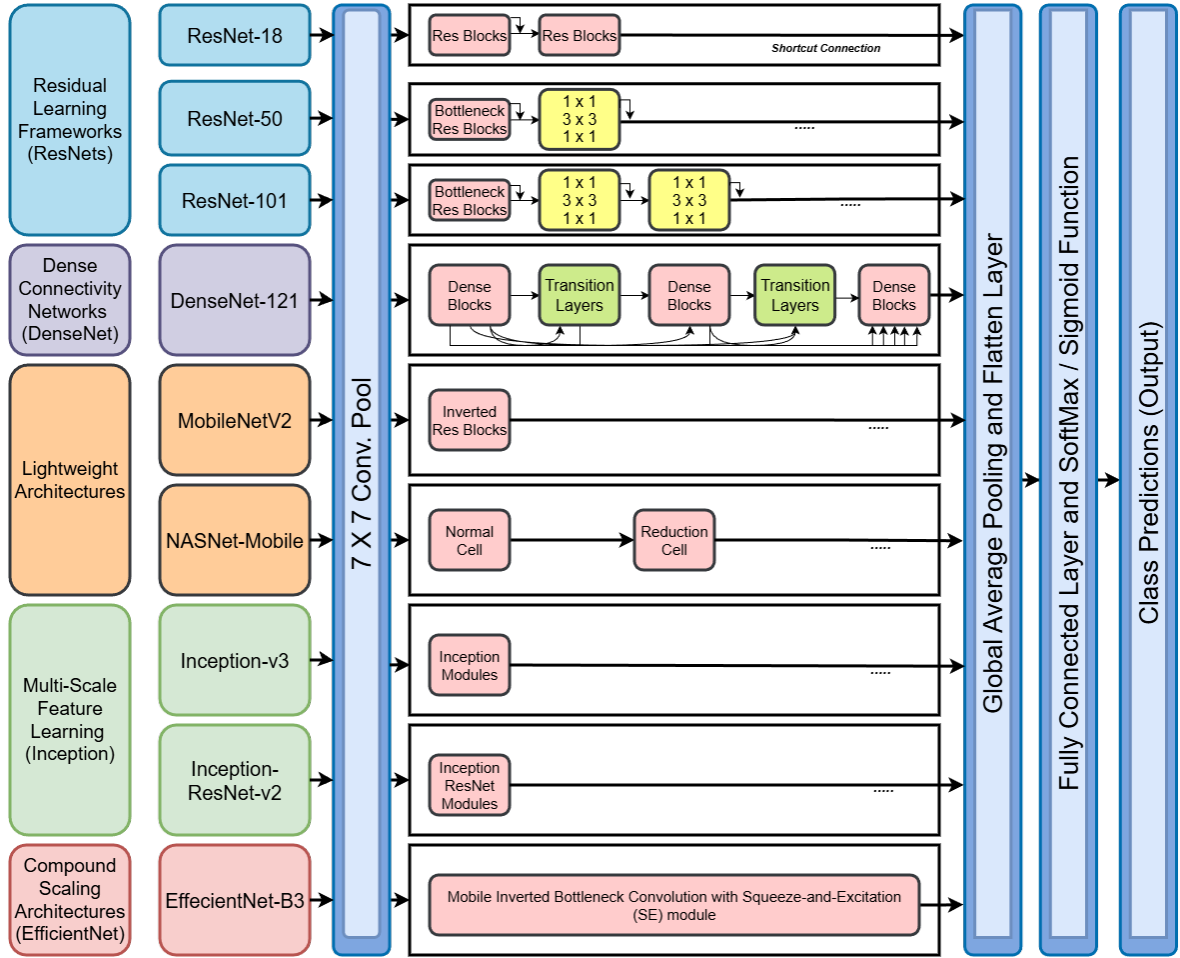


Figure 4.2: Comparative summary of classification model architectures in this study

4.2.4 Multi-Scale Feature Learning

Inception-v3

Inception-v3, proposed by Szegedy et al. [53], takes advantage of factorized convolutions, very aggressive dimensionality reduction, and a multi-branch architecture in combination to deliver a very efficient capture of multi-scale visual patterns. It is the viewpoint of the authors that the strong representational capacity with controlled computational cost has made it effective for the analysis of different texture-rich dental images.

InceptionResNet-v2

InceptionResNetV2 is the product of the combination of the Inception modules and the residual connections that the authors claim to have yielded the training stability and higher accuracy they wanted [54]. The synthesis of multi-scale feature extraction and identity mappings has been a breakthrough in the field, as it has made almost unbounded the depth and power of the representations, thus making it the go-to for high-resolution medical image analysis. The comparative differences of all the classification models in this study are visualized in Figure 4.2.

4.2.5 Compound Scaling Architecture

EfficientNet-B3

Tan and Le [55] presented EfficientNet-B3 as a result of using a compound scaling methodology that uniformly scales depth, width, and input resolution. By employing mobile inverted bottleneck convolutions (MBConv) and squeeze-and-excitation blocks, the model attains top-notch efficiency and accuracy. It offers a very good compromise between numerically intense operations and performance for dental imaging applications.

4.3 Background of the YOLO Framework for Object Detection

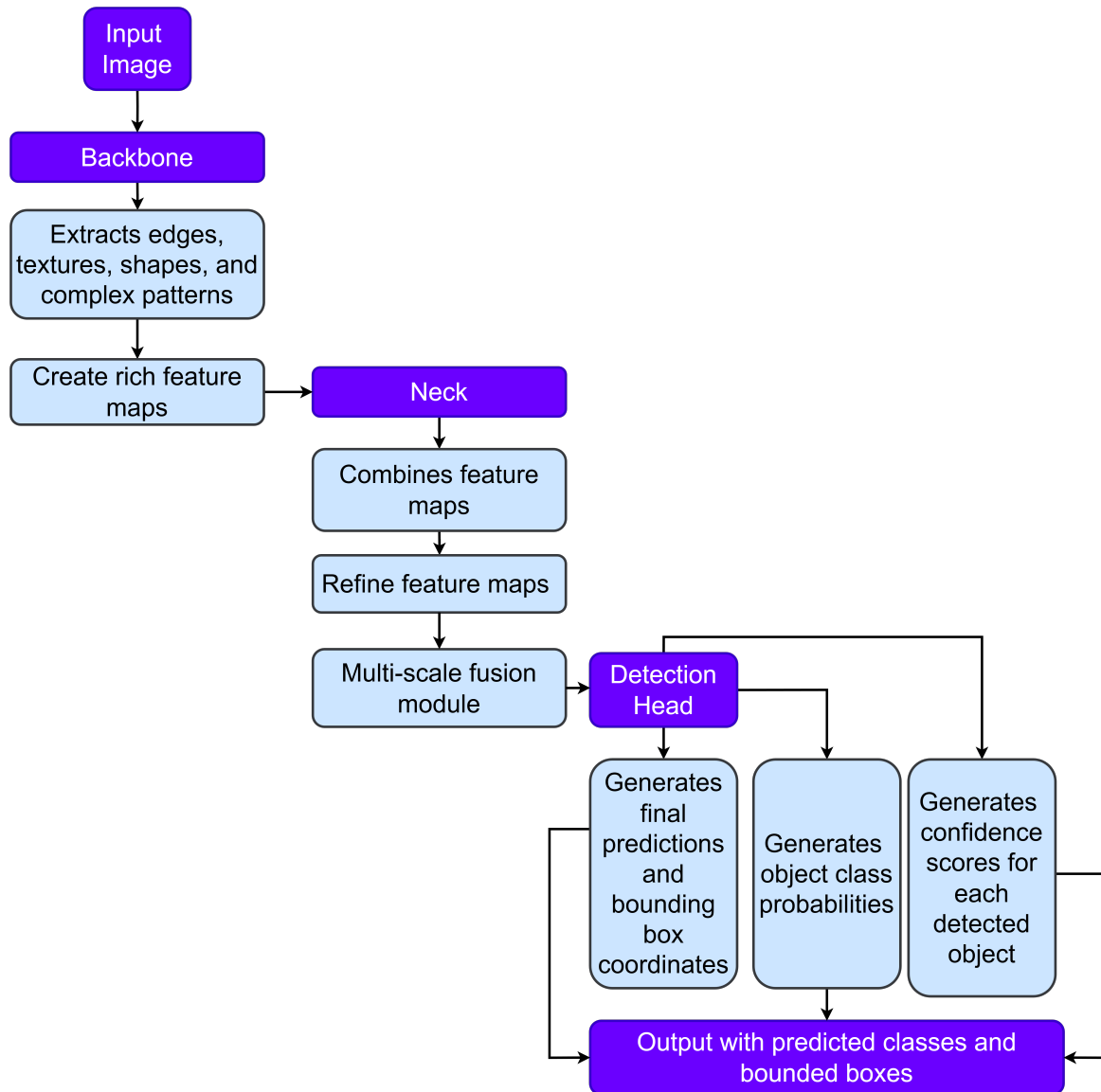


Figure 4.3: Generalized workflow of YOLO-Based Object detection models

4.3.1 Evolution of YOLO Models (v8-v12)

YOLO framework is a group of real-time object detection models that handle classification and localization all at once. Over time, these models have just kept getting faster and smarter, with better ways to spot features and more efficient designs that work on all sorts of devices. With each new version, YOLO has set the standard for hitting that sweet spot between accuracy, speed, and efficiency in computer vision.

In January 2023, Ultralytics launched the YOLOv8 object detection model, featuring an enhanced Cross-Stage Partial Darknet (CSPDarknet) architecture-tasked backbone, optimized for real-time feature extraction, and an anchor-free, decoupled detection head that separates the classifications and bounding-box regression, respectively, which increases the speed and accuracy of convergence. It achieves impressive COCO performance while remaining suitable for edge deployment [56].

In February 2024, Wang et al. introduced a newly revised YOLO framework model that focused specifically on the preservation of information and efficiency of gradient via the newly proposed YOLOv9. This model can be characterized through two main innovations: (1) the Programmable Gradient Information (PGI) mechanism that helps to ensure that information input is preserved more faithfully throughout the neural network, and (2) the Generalized ELAN (GELAN) module, a feature aggregation neck along the EfficientNet design principle and designed specifically to improve parameter efficiency. Both of the innovations of the proposed model have been shown to improve lightweight model performance and reduce information loss, all without incurring large additional computation increases and complexity [57].

In May 2024, the Tsinghua University researchers introduced the new version of the YOLO framework, YOLOv10. Its principal contribution is its *Non-Maximum Suppression (NMS)-free* detection paradigm, which is a dual-label assignment strategy (one-to-many and one-to-one matching) that trains the model to suppress redundant detections without requiring post-processing Non-Maximum Suppression NMS. NMS is a post-processing algorithm that refines object detection results by removing redundant, overlapping bounding boxes for the same object. The design is based on a CSP-based backbone that has been optimised, and a neck that is similar to the Path Aggregation Network (PANet). To support context modelling, it also includes large kernel convolutions and partial attention modules, which reduce latency whilst maintaining competitive mAP [58].

In September 2024 the Ultralytics released the YOLOv11, a new model that had refined the previous YOLOv8 models design with a more optimized backbone and neck configurations and an even more latency-efficient anchor-free decoupled head. According to the Ultralytics reports, the substantial parameter reductions, e.g., YOLOv11m contains approximately 22% fewer parameters than YOLOv8m while achieving higher COCO accuracy, reflecting improved architectural efficiency across tasks such as detection, segmentation, and classification [59].

Finally, the YOLOv12 model represented a shift towards an attention-centric design. It introduced the concept of the Area Attention for capturing long-range spatial dependencies with reduced cost and a Residual Efficient Layer Aggregation Network (R-ELAN) for advanced multi-scale fusion. Scaled residual connections and optional FlashAttention variants further decrease memory usage during training. YOLOv12 achieves state-of-the-art performance, surpassing YOLOv10/11 at the cost of marginally slower inference [60].

The architecture's comparative differences of YOLO (v8-v12) can be seen in Figure 4.3 and Figure 4.4 along with the Backbone, which serves as the feature extractor that converts the raw imager input to maps of different levels of features and captures the main patterns such as edges, textures, and shapes. The neck is the part that connects and combines the features from various scales, e.g., resolutions, so the model can use

the different detailed spatial information together with the global context to detect both small and large objects. The last layer of detection Head, which is the final prediction layer that processes and outputs the class probabilities and bounding box coordinates for each detected object, is the fused features.

4.3.2 Comparison of YOLO Architectural Components



Figure 4.4: Comparative summary of the latest YOLO-based object detection models in this study

YOLOv8 architecture

YOLOv8 sticks with the classic backbone, neck, and detection head setup. The backbone relies on CSPDarknet, which helps gradients flow better using cross-stage partial connections. For the neck, it uses a PANet-inspired design to pull in information from different spatial scales. The detection head is decoupled and anchor-free, so it predicts classes and bounding boxes directly. This setup speeds up training, helps the model lock onto objects faster, and boosts accuracy during inference [61].

YOLOv9 architecture

YOLOv9 keeps the CSP-style backbone but swaps out the usual neck for a GELAN module. GELAN does a better job at pulling features together, thanks to a smart layer-aggregation strategy built on gradient path optimization. During training, the model uses the PGI mechanism, but at inference, it sticks to an anchor-free, decoupled head—pretty similar to what YOLOv8 does. With richer multi-scale features, YOLOv9 ends up performing even better.

YOLOv10 architecture

YOLOv10 builds on a CSP backbone with more power to capture details, and pairs it with a PANet neck for fusing features across scales. The detection head stays decoupled and anchor-free, not much different from earlier designs. What really stands out is the

dual assignment training strategy. This change means that we don't need non-maximum suppression during inference, which cuts down inference time by a lot, without losing accuracy.

YOLOv11 architecture

YOLOv11 uses a CSP-inspired backbone and a PANet-style neck, but adds some Ultralytics-optimized modules to shrink the parameter count and speed up inference. The detection head is still anchor-free and decoupled, but it's been reworked to use less computation while keeping detection strong. Even with fewer parameters than YOLOv8, YOLOv11 manages to outperform it, thanks to these smart design tweaks.

YOLOv12 architecture

YOLOv12 takes things further by blending convolutional and attention-based mechanisms. The backbone now includes Area Attention, which does a better job of capturing global context. R-ELAN modules help gather features using residual paths, both in the backbone and neck. The detection head sticks to the anchor-free, decoupled style of previous YOLOs, but gets a boost from stronger feature representation, better normalization, and updated Multi-Layer Perceptron (MLP) layers.

5 Experimental Setup

5.1 Classification

5.1.1 Classification Models Evaluated on the OPG Dataset

This study aimed to automate the interpretation of dental OPG X-rays by applying a broad set of CNN models. The selected architectures, including ResNet-18, ResNet-50, EfficientNet-B3, MobileNetV2, Inception-v3, and DenseNet-121, were chosen to represent different design strategies, allowing the system to learn a wide range of features and improve the accuracy of X-ray analysis.

Instead of training these complex models from scratch, we used transfer learning to retain high-level feature extraction capabilities, while adapting the classification head structures and input dimensions to our specific dataset as outlined in Table 5.1. The whole network was fully fine-tuned so that its weights could be specifically adjusted for the dental image domain. To ensure the models learned actual diagnostic patterns rather than noise, the training process automatically stopped as soon as the validation performance peaked, with the specific epoch counts and fine-tuning details presented in Table 5.2.

5.1.2 Functions and Optimizers for Multi-Label Classification

To effectively train the models for multi-label classification, we applied a well-defined combination of loss functions, optimization algorithms, and regularization techniques. Each of these components was selected to align with the specific characteristics and constraints of the dataset. This approach ensured stable training, good generalization to new data, and proper guidance for learning within the multi-label prediction.

Since this is a multi-label classification problem where a single Dental OPG image may manifest multiple conditions simultaneously, standard categorical cross-entropy is unsuitable. Instead, we employ the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss). This function combines a Sigmoid layer and the Binary Cross-Entropy loss

Table 5.1: Comparison of different model architectures and their classification heads on the OPG dataset

Model Name	Classification Head Structure	Input Resolution (Pixels)
ResNet-18	Global Average Pooling (GAP) $\rightarrow$ Dropout(0.4) $\rightarrow$ Linear	448×448
EfficientNet-B3	Global Average Pooling $\rightarrow$ Dropout(0.4) $\rightarrow$ Linear	480×480
MobileNetV2	Features Out $\rightarrow$ Dropout(0.4) $\rightarrow$ Linear	224×224
ResNet-50	GlobalAveragePooling $\rightarrow$ Dropout(0.4) $\rightarrow$ Dense Layer	512×512
Inception-v3	Inception Features $\rightarrow$ Dropout(0.4) $\rightarrow$ Dense Layer	448×448
DenseNet-121	GlobalAveragePooling2D $\rightarrow$ Dropout(0.4) $\rightarrow$ Dense Layer	480×480

Table 5.2: Fine-tuning configuration and training duration for each model on the OPG dataset

Model Name	Fine Tuning Strategy	Maximum Training Epochs	Early Stop Available	Number of Epochs Reached
ResNet-18	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	24
EfficientNet-B3	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	42
MobileNetV2	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	21
ResNet-50	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	47
Inception-v3	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	28
DenseNet-121	End-to-End Fine-Tuning (All layers unfrozen)	50	Yes (Patience: 7)	42

in a single class, which is numerically more stable than using a plain Sigmoid followed by a BCELoss.

$$L = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C [y_{n,c} \cdot \log(\sigma(x_{n,c})) + (1 - y_{n,c}) \cdot \log(1 - \sigma(x_{n,c}))] \quad (5.1)$$

Where:

- $y_{n,c}$ is the ground truth label (0 or 1) for class c in image n .
- $x_{n,c}$ is the predicted logit for class c .
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the Sigmoid activation function.

We utilize the AdamW (Adaptive Moment Estimation with Weight Decay) optimizer. AdamW is chosen over standard Adam because it decouples weight decay from the gradient update, leading to better generalization performance, particularly in computer vision tasks. The optimizer is initialized with a learning rate (η) of 5×10^{-5} and a weight decay coefficient (λ) of 1×10^{-4} . The parameter update rule for AdamW at time step t is defined as:

$$\theta_t = \theta_{t-1} - \eta_t \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} + \lambda \theta_{t-1} \right) \quad (5.2)$$

Where:

- θ represents the model parameters.
- $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates.
- $\lambda \theta_{t-1}$ represents the decoupled weight decay term.

To ensure convergence and avoid getting stuck in local minima, we employ a ReduceLROnPlateau scheduler. If the validation loss fails to improve for a patience of 4 epochs, the learning rate is multiplied by a factor of 0.5.

This decay strategy can be formalized as:

$$\eta_{new} = \eta_{old} \times 0.5 \quad \text{if } \text{ValLoss}_t \geq \min(\text{ValLoss}_{t-\text{patience}...t}) \quad (5.3)$$

To mitigate overfitting and address data imbalance, three additional strategies are integrated into the pipeline:

- **Dropout:** A dropout layer with a probability of $p = 0.4$ is applied immediately before the final fully connected layer. During training, this layer randomly sets a subset of input activations to zero, thereby reducing neuron co-adaptation and enhancing model generalization.
- **Weighted Random Sampling:** Given the imbalance in disease classes, a *Weighted Random Sampler* is utilized. Sample weights are calculated inversely proportional to class frequencies, ensuring that rarer classes are sampled more frequently during training batches.
- **Early Stopping:** An early stopping mechanism with a patience of 10 epochs is employed. If the validation loss fails to improve for 10 consecutive epochs, training is stopped and the model parameters corresponding to the best validation performance are restored.

5.1.3 Classification Models Evaluated on the RGB Teeth Dataset

For the purpose of evaluating the effectiveness of different deep learning architectures on the dental image dataset, the training of six CNN architectures ResNet-50, ResNet-101, DenseNet-121, MobileNetV2, NASNet-Mobile, and Inception-ResNet-v2 was carried out. In all cases, the architectures were initialized with the weights pre-trained on the ImageNet dataset for feature extraction. The fine-tuning was applied successfully in the case of ResNet-50 and ResNet-101, during which the architectures were adapted to the dental dataset, initially by freezing the convolutional base and adding custom classification heads with Dense layers; thereafter, some layers of the convolutional base were unfrozen to augment the adaptation of the features. For DenseNet-121, MobileNetV2, NASNet-Mobile, and Inception-ResNet-v2, the models were set up as frozen feature extractors, so that all the convolution layers remained frozen while only the training of the custom classification heads was allowed. In general, the custom heads architecture was similar across all models, but a few differences were pointed out in terms of the number of neurons, the presence of Global Average Pooling (GAP) layers, and dropout rates. Fine-tuning strategies were determined based on architecture performance, and further details on the customized heads are presented in Table 5.3 and Table 5.4.

5.1.4 Functions and Optimizers for Single Label Classification

Both the Rectified Linear Unit (ReLU) activation function and the softmax function were utilized in the hidden and final output layers, respectively [62]. The ReLU activation and

Table 5.3: Comparison of different model architectures and their classification heads on the RGB Teeth dataset

Model	Classification Head
ResNet-50	GAP $\rightarrow$ Dense (1024) $\rightarrow$ Dense (512) $\rightarrow$ Dropout (0.3)
ResNet-101	Dense (1024) $\rightarrow$ Dense (512) $\rightarrow$ Flatten
DenseNet-121	GAP $\rightarrow$ Dense (512) $\rightarrow$ Dense (256) $\rightarrow$ Dropout (0.3)
MobileNetV2	GAP $\rightarrow$ Dense (512) $\rightarrow$ Dense (256) $\rightarrow$ Dropout (0.2)
NASNet-Mobile	GAP $\rightarrow$ Dense (1024) $\rightarrow$ Dense (512)
Inception-ResNet-v2	GAP $\rightarrow$ Dense (1024) $\rightarrow$ Dense (512) $\rightarrow$ Dropout (0.3)

Table 5.4: Fine-tuning configuration and training duration for each model on the RGB Teeth dataset

Model	Fine-Tuning Strategy	Initial	Fine-Tune
ResNet-50	Unfrozen from <code>conv5_block1</code> onward	30 Epochs	30 Epochs
ResNet-101	Unfrozen from layer 313	30 Epochs	30 Epochs
DenseNet-121	All layers frozen	30 Epochs	N/A
MobileNetV2	All layers frozen	30 Epochs	N/A
NASNet-Mobile	All layers frozen	30 Epochs	N/A
Inception-ResNet-v2	All layers frozen	30 Epochs	N/A

softmax activation functions are defined as:

$$f(x) = \text{ReLU}(x) = \max(0, x) \quad (5.4)$$

where x denotes the input to the activation function and $f(x)$ is the output. If $x > 0$, then $f(x) = x$; otherwise, $f(x) = 0$.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (5.5)$$

where z_i represents the i^{th} logit (raw score), K is the total number of classes, and the denominator $\sum_{j=1}^K e^{z_j}$ ensures that the outputs sum to 1, resulting in a valid probability distribution. The models were trained on images resized to 224×224 pixels, and each of the models was then optimized using the Categorical Cross-Entropy loss function along with the Adaptive Moment Estimation (Adam) optimizer and Learning Rate ranging from $(1 \times 10^{-2} - 1 \times 10^{-7})$ [62]. The Categorical Cross-Entropy loss can be defined as:

$$\mathcal{L} = - \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (5.6)$$

where N denotes the number of samples, C is the number of classes, $y_{i,c} \in \{0, 1\}$ represents the ground truth label of sample i for class c , and $\hat{y}_{i,c}$ is the predicted probability for class c obtained from the softmax output of the model. The Adam optimizer updates the weights by combining the advantages of momentum and adaptive learning rates. The parameter update rule is given as:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (5.7)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (5.8)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (5.9)$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad (5.10)$$

where θ_t denotes the model parameters at iteration t , g_t is the gradient of the loss with respect to θ_t , m_t and v_t are the first and second moment estimates, β_1 and β_2 are exponential decay rates for these moments, α is the learning rate, and ϵ is a small constant for numerical stability.

We employ categorical cross-entropy, which gauges how closely the predicted probability distribution resembles the actual one-hot encoded label, since our dataset has a single ground-truth label for every image across five classes. When the predicted probability for the correct class is low, the model is penalized more severely.

$$\text{CCE}(y, \hat{y}) = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (5.11)$$

$$C = \text{number of classes} \quad (5.12)$$

$$y_i = \begin{cases} 1, & \text{if class } i \text{ is the true label} \\ 0, & \text{otherwise} \end{cases} \quad (5.13)$$

$$\hat{y}_i = \text{predicted probability for class } i \quad (5.14)$$

5.2 Object Detection Experimental Setup

5.2.1 Loss Functions, Activation Functions, and Optimization Strategy

The object detection models were trained using a unified optimization strategy across various lightweight and latest YOLO families to ensure a fair performance comparison among them for the dental datasets, RGB Intraoral vs OPG. The optimizer used for every model in both pipelines was the AdamW optimizer, configured with a learning rate of 0.001 and a momentum of 0.9. AdamW decouples weight decay from the gradient update, improving convergence stability in deep convolutional [63]. The update rule for AdamW is given by:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (5.15)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (5.16)$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta \lambda \theta_t \quad (5.17)$$

where m_t and v_t are the exponential moving averages of the gradient and squared gradient, η is the learning rate, and λ is the decoupled weight-decay coefficient.

For bounding-box regression, all the YOLO models that were being experimented on had employed the Distribution Focal Loss (DFL) coupled with Complete IoU (CIoU) loss,

where the DFL improves localization precision by predicting continuous distributions over box offsets, and the CIoU incorporates overlap, center distance, and aspect-ratio consistency [?]. It is expressed as

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \quad (5.18)$$

where ρ is the Euclidean distance between the centers of the box, c is the diagonal length of the smallest enclosed box, v measures the similarity of the aspect ratio, and α is a stabilizing factor.

For classification, all models used Binary Cross-Entropy (BCE) loss with logits [64]. It is expressed as

$$\mathcal{L}_{BCE} = -[y \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))] \quad (5.19)$$

which is optimal for object detection settings.

All the experiments in both object detection pipelines were conducted using a fixed input resolution of 640*640 for both training and validation, by ensuring consistency across all YOLO model variants that were used in the experiment and facilitating a direct comparison of performance metrics.

5.2.2 Object Detection Models Evaluated on both OPG and RGB Intraoral Caries Datasets

In order to maintain a uniform approach and facilitate a comparative assessment that is not biased, both the OPG Dataset Pipeline and the Intraoral Caries Dataset Pipeline were developed with the same family of YOLO models, architectures, and pretraining setups. The number of models used was ten, and they were a mixture of four YOLO generations (YOLOv8, YOLOv9, YOLOv10, YOLO11, and YOLO12) controlled by the Ultralytics framework, covering the three tiers of nano (n), small (s), and tiny, having different computational complexities.

Every model started from its official pretrained weights, and then the same optimization settings, image resolutions, augmentation strategies, and early-stopping criteria were applied over the two detection tasks. The differences in performance that result from this uniform configuration can be attributed only to the characteristics of the dataset and not to architectural or hyperparameter variations.

In Table 5.5, a thorough overview of key architectural properties, among others, that have been considered for all YOLO versions, which have been used in our experiments have been presented.

In particular, (i) Total layers is the term that reflects the total number of all convolutional, normalization, activation, and detection layers that are responsible for the construction of the particular model. So, the model with more layers is deeper and thus can learn more complicated visual patterns, though, at the same time, it has to cope with more power and longer training time as well.

(ii) Number of parameters is a term that stands for the total number of trainable weights present in the neural network. Larger models are more resourceful in terms of recognizing and even abstracting features, but they also need more memory, data, and careful training to avoid getting overfitted, which is their main drawback.

(iii) Gradient-bearing layer tells about the number of layers that get gradient updates when backpropagation is applied. These layers, in fact, are the ones that learn from the

Table 5.5: Comparative architectural summary of YOLO variants employed in this study's object detection pipelines

Model	Total Layers	Parameters	Gradients	Transferred Weights
YOLOv8n	129	3,012,018	3,012,002	319/355
YOLOv8s	129	11,137,922	11,137,906	349/355
YOLOv9t	544	2,006,578	2,006,562	1303/1339
YOLOv9s	544	7,289,730	7,289,714	1333/1339
YOLOv10n	223	2,709,380	2,709,364	493/595
YOLOv10s	234	8,070,996	8,070,980	607/619
YOLO11n	181	2,591,010	2,590,994	448/499
YOLO11s	181	9,430,114	9,430,098	493/499
YOLO12n	272	2,569,218	2,569,202	640/691
YOLO12s	272	9,255,458	9,255,442	685/691

training data. If some layers do not get gradients (for instance, because they are frozen or excluded), they will not update their weights. Knowing how many layers are involved in learning gives an idea of the model's flexibility or non-flexibility during fine-tuning.

(iv) Transferred pretrained weights ratio is an indicator of the extent to which the original pretrained backbone of the official YOLO model is reused during the training process of our dental datasets. A high ratio means that the major part of the model is initialized with the knowledge learned from the large generic datasets like COCO, which helps the models to converge faster, improve accuracy, and require less training data. A lower ratio, on the other hand, indicates that more layers had to be randomly initialized, which leads to slower convergence and a greater requirement for training samples.

It's important to be aware of the architectural properties since they affect the speed of training, the speed of convergence, the accuracy of generalization to previously unseen images, and the speed of inference once the training has been completed. In addition, by using the same architecture and identical training conditions between all models used in both the OPG dataset and the intraoral caries dataset pipelines within this study, we can ensure that any performance differences will come exclusively from differences in either the architecture of the models or the type of data being used, rather than from differences in the way that the training pipelines were constructed. Thus, by using the same methodology to conduct both studies, we can be assured that this will lead to valid and reliable comparisons of performance from the different groups of models and that we will not have any extraneous variables affecting our experimental outcomes.

5.2.3 Training process of the YOLO models

The ten YOLO models (YOLOv8n, v8s, v9t, v9s, v10n, v10s, v11n, v11s, v12n, and v12s) were all trained using the transfer-learning technique that utilized the OPG and Intraoral Caries datasets to learn rich feature representations. Instead of building the models from scratch, which takes a lot of datasets, time, and risk of overfitting, the last DFL convolution layer in each model was kept unchanged to maintain its pretrained classification-regression mapping. This layer is the one that normally obtains the object-distribution priors, so that freezing it guarantees stability during adaptations. The training of the remaining layers was, therefore, allowed to continue, which effectively gave the backbone

and detection head the opportunity to learn features that were specific to the images. (Trainable layers = Total layers - Frozen layer) was the formula used to calculate the number of trainable layers for each model. The trick of fine-tuning most of the layers while at the same time locking the final DFL layer has allowed the models to draw heavily on the strong generalization from the pretrained weights and to adapt to the dental patterns, intensity variations, and anatomical structures very deeply. Such an approach strikes the perfect balance between retaining generic visual representations and acquiring domain-specific features, thus enabling the convergence to happen quickly, the accuracy to be enhanced, and the risk of overfitting to be reduced for all ten architectures.

5.3 Evaluation Metrics

5.3.1 Evaluation Metrics for Classification

The performance of the proposed models was assessed using standard evaluation metrics derived from the four possible prediction outcomes, which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy is used, which measures the overall proportion of correctly classified instances [65].

However, in the presence of class imbalance, accuracy alone can be misleading. To address this, additional metrics were considered, such as precision, which quantifies the proportion of correct positive predictions, reflecting the model’s exactness; recall measures the models ability to identify positive cases, thus accurately capturing completeness, and the F1-score, which balances precision and recall. This metric provides a harmonic mean that offers a more reliable measure of performance when both false positive and false negative are of concern. Together, these metrics ensure a comprehensive and clinically relevant evaluation of the models.

After the classification task, the prediction accuracy and the performance of the models were also evaluated utilizing an XAI approach. The approach that has been used is Grad-CAM, which was employed to create visual explanations for the predictions of the models. This provides a view of the internal workings of the CNNs by reflecting the specific patches in an input image that contributed the most to predicting a particular class.

The performance of the proposed methods was evaluated using four primary metrics: accuracy, precision, recall, and F1-score. Accuracy quantifies the proportion of correctly classified samples [66]. It is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.20)$$

Precision measures the models ability to correctly identify positive instances [66]. It is expressed as

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5.21)$$

Recall (also referred to as sensitivity) evaluates the proportion of actual positive cases that are correctly detected [66]. It is computed as

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5.22)$$

Finally, the F1-Score serves as the harmonic mean of precision and recall, providing a balanced representation of both metrics [66]. It is computed as

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.23)$$

5.3.2 Evaluation Metrics for Object Detection

The object detection performance was measured using a set of standard metrics that cover both how well the model classifies objects and how accurately it locates them. PrecisionRecall (PR) curves show how true positives stack up against false positives as we adjust the confidence threshold. These curves show whether the model maintains high precision while still capturing most of the real objects. We also look at confidence curves to see how the models confidence scores line up with how reliable its predictions actually are. That gives us a sense of whether the models confidence matches up with reality, and how sensitive it is to the thresholds we pick.

For the big picture, the focus is on Mean Average Precision (mAP). mAP@50 checks how well the model detects objects when we use a pretty forgiving IoU threshold of 0.50, so the boxes dont have to line up perfectly. mAP@95 takes it up a notch as it averages performance across a range of stricter IoU thresholds from 0.50 to 0.95, stepping by 0.05 each time. This gives us a tougher test of both accuracy and consistency. All together, these metrics give us a solid, well-rounded view of the models detection quality in all sorts of scenarios.

In object detection, the evaluation process involves measuring not only the correctness of the predicted class labels but also the spatial accuracy of the detected bounding boxes, where the fundamental component of this evaluation is the Intersection over Union (IoU), which then quantifies the overlap between a predicted bounding box and the corresponding ground-truth box [67]. IoU is defined as

$$\text{IoU} = \frac{\text{Area of } (B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area of } (B_{\text{pred}} \cup B_{\text{gt}})} \quad (5.24)$$

where a prediction is typically counted as a true positive if its IoU exceeds a specified threshold; building on this, the Average Precision (AP) metric summarizes the precision-recall tradeoff for a single object class by integrating the area under the precision-recall curve [67]. AP is defined as

$$\text{AP} = \int_0^1 p(r) dr \quad (5.25)$$

where $p(r)$ denotes precision as a function of recall. To generalize the performance across all classes, the Mean Average Precision (mAP) is computed as the mean AP over N object categories [67]. mAP is defined as

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i \quad (5.26)$$

with N representing the total number of classes. In practice, mAP is often reported at specific IoU thresholds; for instance, mAP@50 evaluates detection accuracy at an IoU threshold of 0.50, while mAP@[0.50:0.95] computes the mean AP across thresholds ranging from 0.50 to 0.95 in increments of 0.05, offering a more rigorous and comprehensive assessment of both the localization precision and detection reliability.

5.4 Computational Environment

Google Colab was used for all the experiments in this research, providing a cloud-based environment suitable for the training of DL models without local hardware. The runtime

was set on an NVIDIA Tesla T4 GPU with 16 GB of GDDR6 memory, which remarkably sped up training and inference times for YOLO-based object detection models and CNN classification models.

The software environment was established on Python 3, and it supported GPU acceleration through Compute Unified Device Architecture (CUDA) and the CUDA Deep Neural Network library (cuDNN). The deep learning frameworks were mainly PyTorch and TensorFlow; PyTorch was used for training custom classification models and YOLO models. Other libraries were also installed, including OpenCV, NumPy, Pandas, Matplotlib, and SciPy, to assist data preprocessing, augmentation, visualization, and performance evaluation, respectively.

Google Colab additionally offered easy integration with Google Drive, which made it possible to store datasets, save model checkpoints, and reproduce experiments efficiently. In conclusion, this computational setup provided a stable and powerful enough environment to conduct training, validation, and testing for all the object detection and classification models in this research.

6 Results and Analysis

6.1 Classification Results on OPG Dataset

6.1.1 Class-wise Performance Interpretation (OPG)

The results presented Figure 6.1 and Table 6.1 provide a comprehensive comparison of six distinct CNN architectures. By analyzing the performance across different dental conditions, specific patterns that highlight the strengths and weaknesses of each model.

Analysis of High-Performance Classes

One of the most important findings in this study is the similarity in detecting 'Fractured' teeth. Every single model, regardless of its internal complexity, achieved an identical accuracy score of 87%. This absolute consistency suggests that fractures present very clear, distinct visual features such as sharp lines or breaks in the bone structure, which are easily recognized by any traditional models.

Similarly, the models performed exceptionally well in identifying the 'Healthy' class. Most models achieved accuracy scores above 83%, with ResNet-18 standing out with the top performance. ResNet-18 achieved an accuracy of 93%, which is the highest individual score recorded in the entire table. This indicates that simpler architectures like ResNet-18 might be efficient enough to recognize the standard patterns of a normal tooth without "over-thinking" or looking for complex features that aren't there.

The Challenge of Caries Detection

Contrasting the very high success rates scored in fractures and healthy teeth, the class that proved the most challenging for all models was the 'Caries' class. Performance dropped in this category, with the accuracy ranging from a low of roughly 45% to only a peak of 61%. A fracture looks like a clear break on an X-ray, while dental caries looks like small, faint dark spots. The model has to catch very tiny details to find them.

- **Top Performers:** MobileNetV2 and Inception-v3 achieved the highest accuracy for dental caries (61%). Inception-v3 is designed to look at images at different scales simultaneously, which may help it spot these small decay spots better than other models.
- **Lowest Performer:** ResNet-50 performed the worst with a score of 45%. This suggests that deeper networks may smooth the image too much and lose the small details needed to spot tooth decay, even with a large pixel size as input.

While model performance varied for difficult classes like dental caries, the Infection class remained consistent, with four of the six models sharing the same score of 80%. Overall, the results show that no single model works best for every category. Identifying cavities is

a major challenge that may require more specialized feature extraction. Overall, ResNet-18 can be considered acceptable for the general task, even if it is not the best in every area.

Table 6.1: Label-wise accuracy comparison across different CNN models

Model	BDC-BDR	Caries	Fractured	Healthy	Impacted	Infection
MobileNetV2	0.7742	0.6129	0.8710	0.9032	0.7097	0.8060
Inception-v3	0.7419	0.6129	0.8710	0.8387	0.7742	0.8065
EfficientNet-B3	0.7742	0.5806	0.8710	0.8387	0.8710	0.8065
ResNet-18	0.8065	0.5806	0.8710	0.9355	0.8065	0.8065
ResNet-50	0.8387	0.4516	0.8710	0.8387	0.8065	0.7742
DenseNet-121	0.7742	0.5806	0.8710	0.8710	0.6774	0.8065

6.1.2 Comparison of Classification Architectures (OPG)

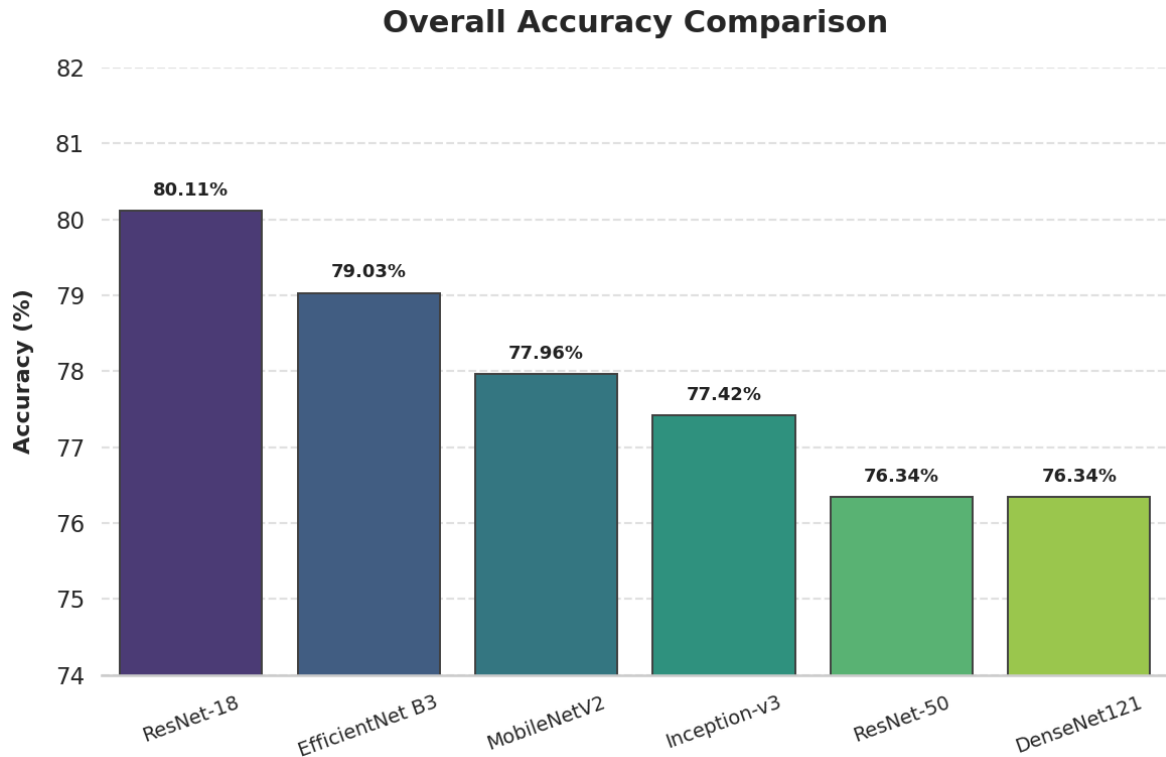


Figure 6.1: Performance summary of the CNN models on the OPG dataset

Table 6.2: Comparative performance analysis of CNN models

CNN Models	Overall Accuracy
MobileNetV2	77.96%
Inception-v3	77.42%
EfficientNet-B3	79.03%
ResNet-18	80.11%
ResNet-50	76.34%
DenseNet-121	76.34%

Performance metrics listed in Table 6.2 highlight a distinct advantage for ResNet-18, which attained the highest accuracy of 80.11%. Increasing the depth of the neural network did not lead to meaningful performance improvements. In fact, the deeper architectures ResNet-50 and DenseNet-121 performed worse than the lighter ResNet-18 model. Both deeper models reached only 76.34%, showing that adding more layers did not enhance accuracy and instead resulted in diminished returns. With an accuracy of 79.03%, EfficientNet-B3 showed notably strong generalization. Its performance exceeded that of MobileNetV2 and Inception-v3, highlighting its effectiveness in capturing relevant features across diverse samples. Therefore, ResNet-18 is identified as the classification model achieving the highest performance on this dataset.

6.1.3 Best Model Performance Analysis (OPG)

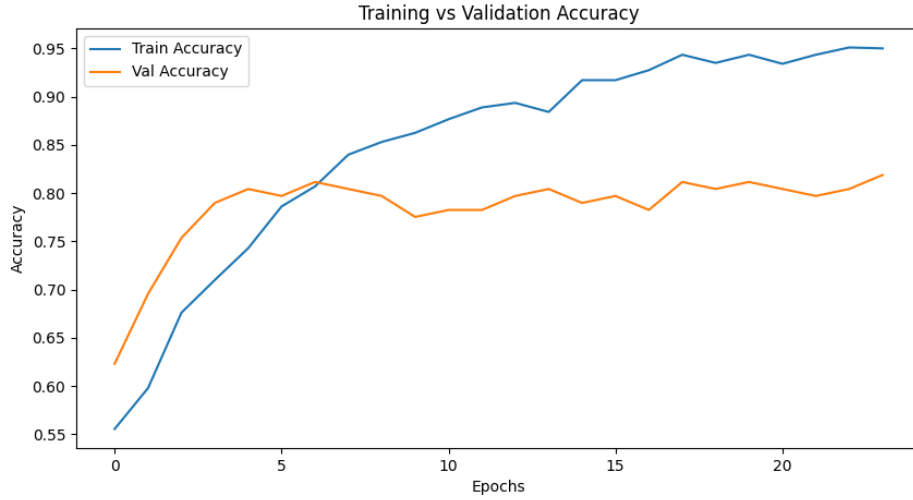


Figure 6.2: Model accuracy progression of ResNet-18 on OPG dataset

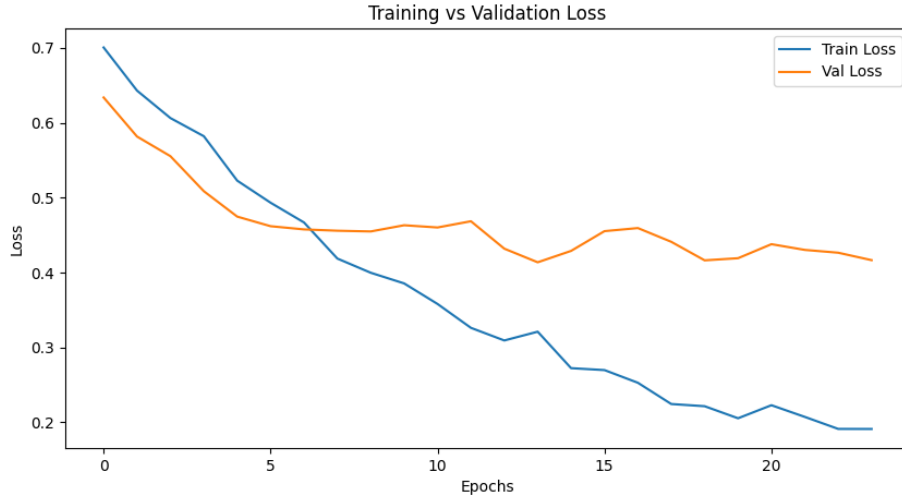


Figure 6.3: Model loss convergence of ResNet-18 on OPG dataset

Figure 6.3 illustrates the loss curve of the best-performing model, the ResNet-18 model, over 23 epochs. The training loss demonstrates a steady decline to approximately 20%, indicating effective learning, while the validation loss stabilizes between 40% and 45% after epoch 12.

Moreover, Figure 6.2 presents the accuracy performance. The model achieved a high training precision of 95%, with validation accuracy settling into a stable range of 80% to 82%. Although the gap between the training and validation metrics shows mild overfitting, the stability of the validation curve confirms the model’s robustness and strong convergence capabilities on unseen data.

6.1.4 Confusion Matrices and Error Analysis (OPG)

The classification performance of the ResNet-18 model across six distinct dental conditions is illustrated in the confusion matrices in Figure 6.4. Overall, the model demonstrates a strong ability to correctly identify negative cases (True Negatives) across all categories, indicating high specificity.

The model performed best in recognizing Healthy Teeth and Impacted Teeth, achieving a reasonable balance between correctly predicting the condition and ruling it out. For example, the ‘Healthy Teeth’ class had very minimal false positives and false negatives relative to the sample size. A high specificity could also be seen in the BDC-BDR class, although the model slightly struggled to detect the positive cases, hence giving some false negatives.

However, challenges remain in specific categories. The Caries class exhibited the highest level of confusion, with a significant number of both false positives and false negatives, suggesting that the visual features for dental caries may overlap with other classes or healthy tissue. Mainly, the model also faced significant difficulty with the Fractured Teeth and Infection classes. In both instances, the model failed to correctly predict any positive cases (0 True Positives), classifying all actual positive samples as negative. This indicates a potential limitation in the model’s sensitivity for these specific pathologies, possibly due to a lack of sufficient representative training data or the minimal features in OPG images.

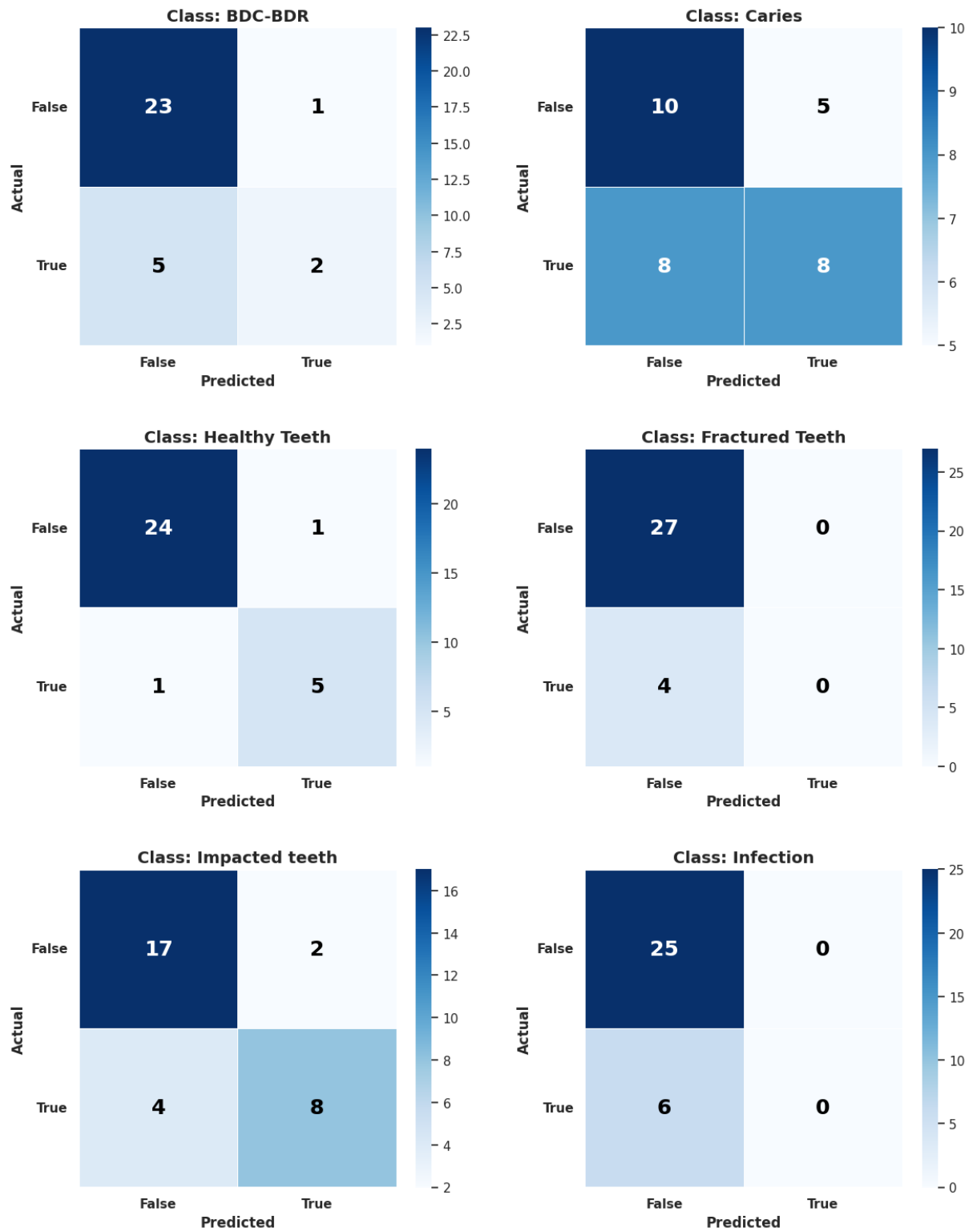


Figure 6.4: Confusion Matrix for Multi-Label OPG classification using ResNet-18

6.1.5 Confidence Threshold Adjustment (OPG)

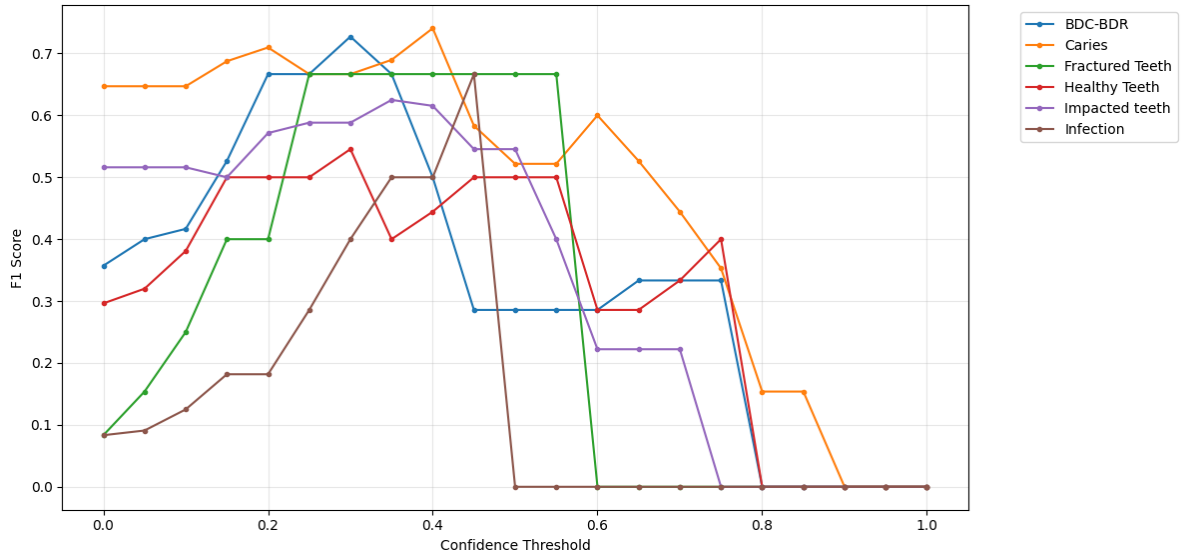


Figure 6.5: F1-Score across varying Confidence Thresholds

In the Figure 6.5, it is depicted the model performance sensitivity, measured through the F1 score, is depicted for six different dental classes at different confidence thresholds from 0.0 to 1.0. In the case of most of the classes, there is an increase in performance, and the confidence threshold, normally between 0.2 and 0.4, is the one at which the performance is the best before a steep decline is experienced. Overall, the classes like 'Caries' and 'BDC-BDR' show strong performance at high peak F1 scores but the more difficult and subtle class, 'Infection', is characterized by higher volatility in performance and dropping out of the race at lower thresholds which paints a picture of the complexity involved in applying strict confidence levels to detecting and hence, giving up on the rising rate of subtle anomalies detection.

Table 6.3: Impact of adaptive confidence thresholding on F1 Scores by disease class (Δ indicates change)

Disease Class	Confidence	F1	Optimized Confidence	F1 (Optimized)	Support
BDC-BDR	0.5	0.40	0.42 (Δ -0.08) ↓	0.50 (Δ +0.10) ↑	7
Caries	0.5	0.55	0.40 (Δ -0.10) ↓	0.67 (Δ +0.12) ↑	16
Fractured Teeth	0.5	0.00	0.15 (Δ -0.35) ↓	0.33 (Δ +0.33) ↑	4
Healthy Teeth	0.5	0.83	0.59 (Δ +0.09) ↑	0.80 (Δ -0.03) ↓	6
Impacted Teeth	0.5	0.73	0.33 (Δ -0.17) ↓	0.80 (Δ +0.07) ↑	12
Infection	0.5	0.00	0.29 (Δ -0.21) ↓	0.36 (Δ +0.36) ↑	6

In Table 6.3, the performance improvements that came with the adoption of class-specific optimized thresholds in place of the standard confidence threshold of 0.5 during the validation phase are shown. The reliance on the adaptive thresholding strategy resulted in the detection capabilities being especially enhanced for the Fractured Teeth and Infection classes that, under default settings, did not register any correct predictions at all ($F1 = 0.00$). Getting confidence requirements to the points of 0.15 and 0.29, respectively, for

the Fractured Teeth and Infection classes, the model managed to activate the features for these subtle anomalies, and thus the F1 scores for these two classes were raised to 0.33 and 0.36, respectively. This method brought about great improvements in the case of most categories of diseases. For instance, the case of 'Caries' saw an increase of 0.12, and thus it was shown that a static confidence score, one that fits all, is not good for the wide range of dental pathologies.

6.1.6 Grad-CAM Visualizations and Explainability (OPG)

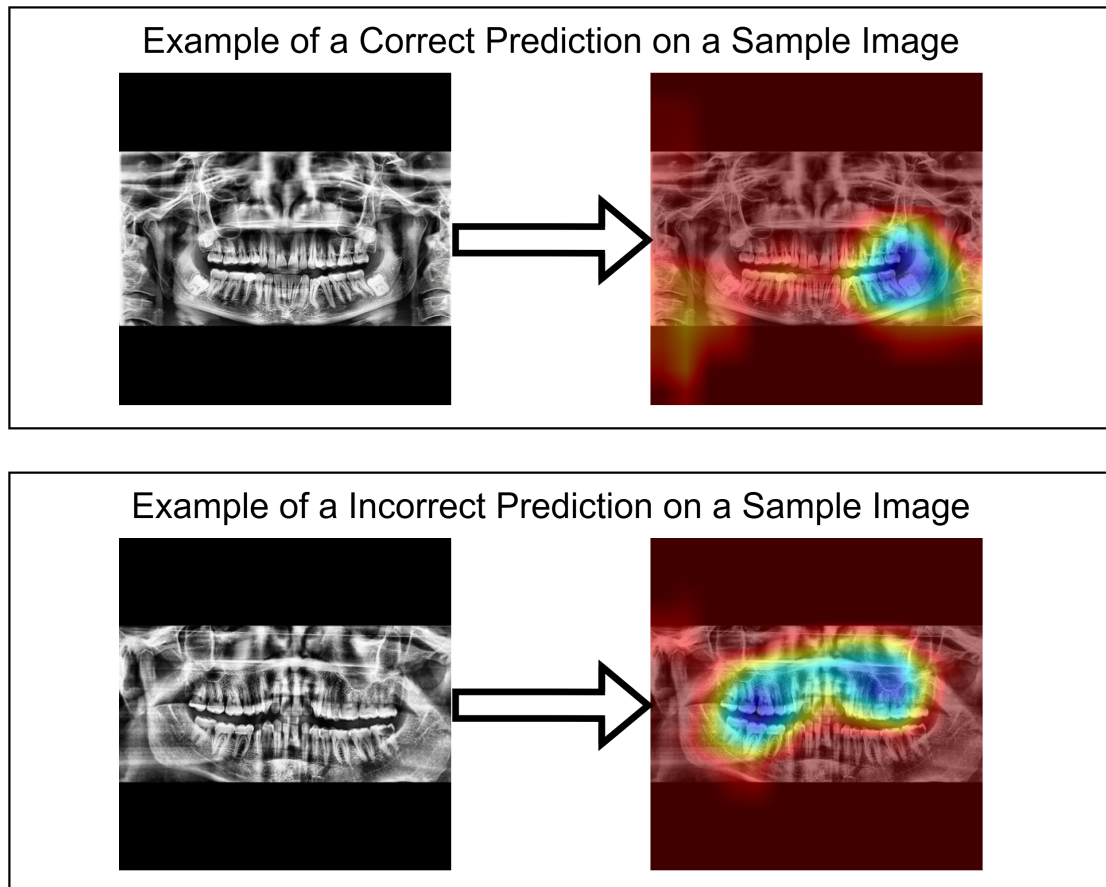


Figure 6.6: Grad-CAM for correct and incorrect OPG samples

The Figure 6.6 presents a visual analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) to evaluate the performance of the dental multi-label classification model. This technique overlays a heatmap onto the original OPG images, allowing us to see exactly which parts of the image the model focused on when making its decision. By comparing the heatmaps, we can determine if the model is looking at the correct dental features or if it is being distracted by irrelevant data.

The first image shows an example of a correct prediction. In this case, the patient's condition was actually an "impacted" tooth, and the model successfully predicted this label. The heatmap shows a strong, localized area of attention directly over the specific tooth region. This confirms that the model is performing well for this class, as it is learning from the actual pixels associated with the impacted tooth rather than guessing based on background noise or unrelated areas.

Similarly, the bottom image displays an example of an incorrect prediction. The ground truth for this OPG image included "BDC-BDR" and "dental caries", but the model failed to classify it correctly. The visualization reveals that the model's attention was not focused on the damaged areas. Instead, the heatmap highlights incorrect regions of the jaw, suggesting that the model struggled to identify the specific visual patterns required to find actual label, leading to a false prediction.

6.2 Classification Results on RGB Teeth Dataset

The classification experiments on the RGB Teeth Dataset showed very clearly that the deeper and more powerful neural networks always got better results than the lightweight architectures by far in all the main evaluation metrics. Among the models evaluated, ResNet-101 reached the highest accuracy and macro F1-score, thus being the most trustworthy architecture for detecting fine dental abnormalities, which often appear in the images as slight changes in structure or texture. The deeper residual connections in ResNet-101 enabled it to master more complicated hierarchical representations, thus vividly differentiating between dental conditions that are very close to each other visually. Inception-ResNet-v2 was another model that performed exceptionally well, very close to the accuracy of ResNet-101. Its hybrid structure, which combines inception modules with residual connections, made it possible to acquire features at multiple scales and levels, and thus, it could easily detect both local and global patterns in the dataset.

DenseNet-121 was also a strong performer, due to its densely connected structure that enhances feature reuse and reduces the effect of the vanishing gradient problem. The model's strong performance on the test set, even with the very small or poorly defined visual cues, was due to the majority of the data being supplied by that connectivity. However, the other light-weight models, MobileNetV2 and NASNet-Mobile, were not as reliable for minority class evaluations, though they were computationally efficient and achieved moderate accuracies. They could not capture the subtle pathological signals in small or less represented categories due to having fewer parameters and simplified architectures, leading to noticeable discrepancies in precision and recall for specific classes.

ResNet-50, essentially being the least deep of the deeper residual architectures that were subject to testing, got the least overall performance during fine-tuning. The limitation of its shallow depth led to its inability to pull out rich and discriminative features of a dataset marked by complex dental morphology and nuanced patterns. In general, the visual and structural complications of the dataset were in favor of the models that had the greater depth, richer connectivity, and improved feature extraction capacity. These deeper networks were not only superior in identifying subtle texture variations and structural irregularities across dental classes but also very robust and reliable in capturing clinically relevant variations in an extensive and meaningful manner.

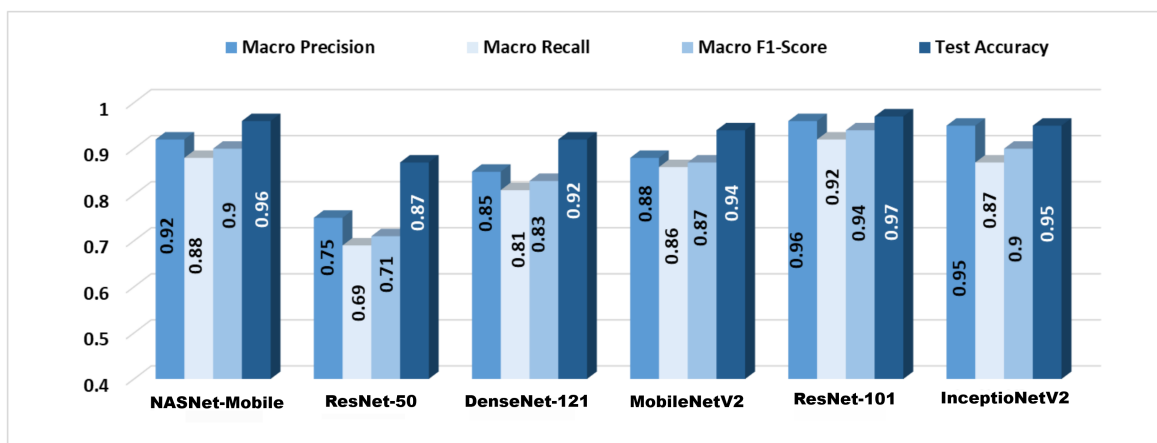


Figure 6.7: Performance summary of the CNN models on the RGB Teeth dataset

6.2.1 RGB Teeth Dataset Confusion Matrices and Error

A comprehensive comparison of six CNNs: NASNet-Mobile, ResNet-50, DenseNet-121, MobileNetV2, ResNet-101, and Inception-ResNet-v2 on a RGB image dataset is provided in Figure 6.7 and Table 6.4. The study covers five different dental conditions: Calculus, Mouth Ulcer, Tooth Discoloration, Caries, and Hypodontia. The evaluation of each model presents per-class metrics of precision, recall, and F1-score as well as the corresponding number of samples, which allows for a thorough evaluation of the model's performance in the multi-class classification task. DenseNet-121 and ResNet-101 were ranked as the best models in terms of overall precision and F1-scores, indicating that they are good at recognizing dental diseases. On the other hand, Caries was the least diagnosed condition for most of the models, probably because it had the fewest samples and looked more like other conditions than the others. NASNet-Mobile and MobileNetV2 produced very similar results with a little less precision but were equally good at recalling, suggesting they were not easily pushed over the edge for the lesser populous classes. Discriminating between Calculus and Tooth Discoloration proved to be especially easy for Inception-ResNet-v2, which was its strong point. This table, in general, maps out the pros and cons of various CNNs while at the same time giving pointers regarding model selection and tuning for the automation of dental image analysis and clinical decision support systems. The confusion matrix of the best performing mode with and without fine-tuning is shown in Figure 6.8 and Figure 6.9.

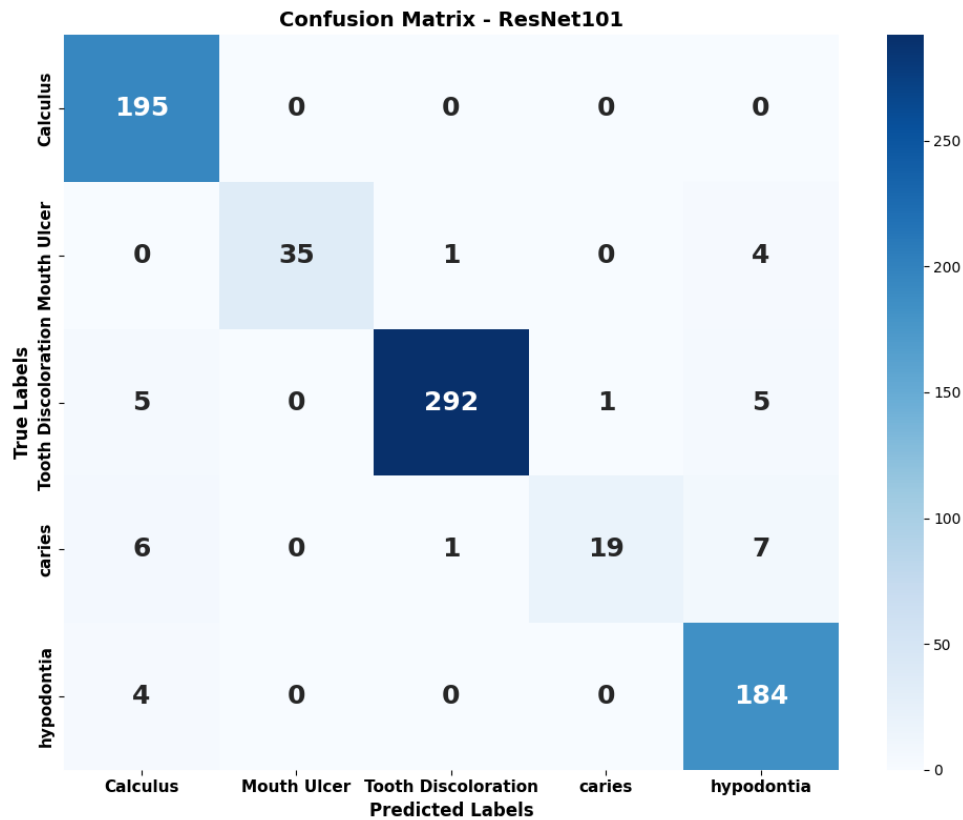


Figure 6.8: Confusion Matrix of ResNet-101 on the RGB Teeth dataset

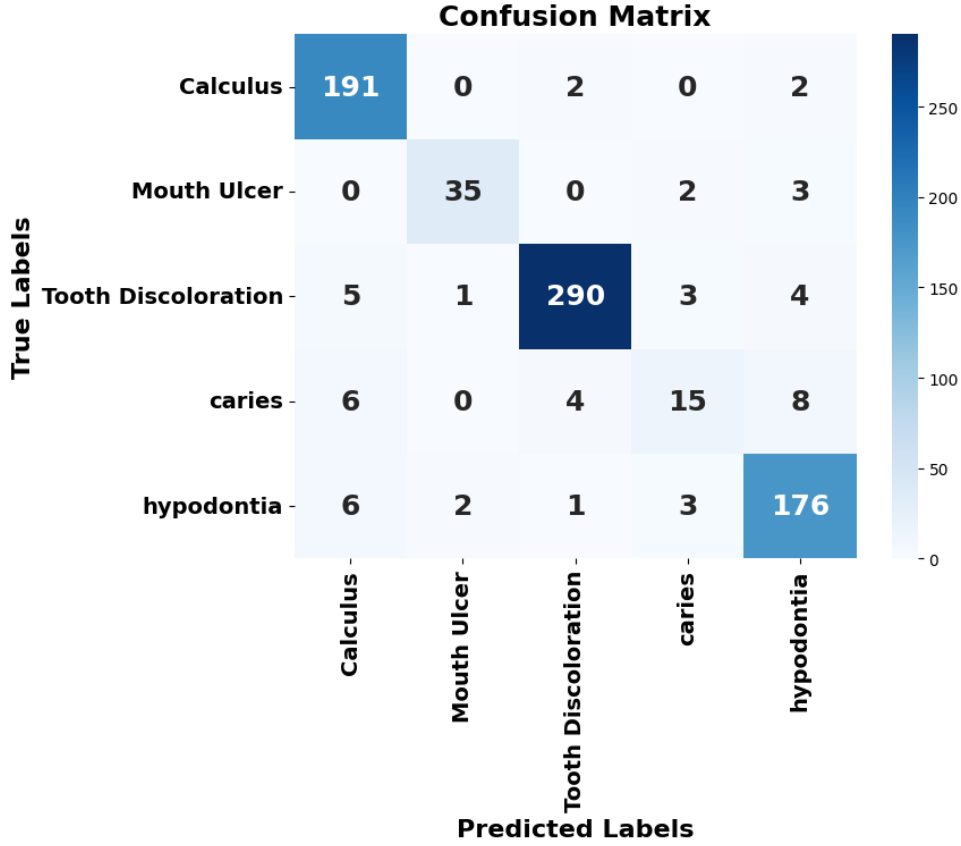


Figure 6.9: Confusion Matrix of NASNet-Mobile on the RGB Teeth dataset

6.2.2 RGB Teeth Dataset Class-wise Performance Interpretation

The analysis regarding each class from the RGB Teeth Dataset shows evident differences in the performance of the different CNN architectures on different oral disorders; the main factor being the visual complexity and sample distribution of each class. The majority of the models marked Calculus, Tooth Discoloration, and Hypodontia as the top conditions concerning precision, recall, and F1-scores. All these conditions have different patterned morphology or coloration, which, for instance, in the case of Calculus, the surface deposits are the least visible, while clear chromatic deviations are seen in discoloration and noticeable structural absence in hypodontia, thus allowing the deeper networks like DenseNet-121, ResNet-101, and Inception-ResNet-v2 to extract discriminative and stable representations. The large sample size for those classes also improved the models generalization ability, which is why the state-of-the-art architectures frequently had F1-scores more than 0.95.

Mouth Ulcer and Caries, on the other hand, received substantially lower scores across all models, indicating that these classes were more difficult for CNNs to classify accurately. Mouth ulcers can be recognized with high intra-class variability, and this is the primary reason why the precision and recall of lighter models like NASNet-Mobile and ResNet-50 have suffered.

Caries has been the most challenging class throughout the process, and consequently, several models couldn't even reach moderate recall scores. The main reason for this is the very small number of dental caries samples, only 33, and at the same time, it is very hard to detect the subtle and localized image patterns of early carious lesions.

The shallow and lightweight models did not have the ability to capture these fine-

grained texture variations, resulting in the extremely low recall for ResNet-50 and only moderate improvements in MobileNetV2. Even high-performing architectures like DenseNet-121 and ResNet-101, which had noticeably stronger results, were still under the influence of the limited dataset size and the visual ambiguity of dental caries lesions.

Table 6.4: Performance metrics of CNN models on the RGB Teeth dataset

Model	Class	Pre-cision	Recall	F1-score	Samples
ResNet-101	Calculus	0.93	0.99	0.96	195
	Mouth Ulcer	0.97	0.90	0.94	40
	Tooth Discoloration	0.99	0.97	0.98	303
	Caries	0.89	0.76	0.82	33
	Hypodontia	0.98	0.99	0.99	188
NASNet-Mobile	Calculus	0.89	0.97	0.93	195
	Mouth Ulcer	0.84	0.80	0.82	40
	Tooth Discoloration	0.98	0.96	0.97	303
	Caries	0.62	0.39	0.48	33
	Hypodontia	0.92	0.94	0.93	188
ResNet-50	Calculus	0.90	0.94	0.92	195
	Mouth Ulcer	0.54	0.47	0.51	40
	Tooth Discoloration	0.89	0.90	0.90	303
	Caries	0.54	0.21	0.30	33
	Hypodontia	0.88	0.93	0.90	188
DenseNet-121	Calculus	0.92	0.99	0.96	195
	Mouth Ulcer	0.95	0.90	0.92	40
	Tooth Discoloration	0.99	0.97	0.98	303
	Caries	0.79	0.58	0.67	33
	Hypodontia	0.97	0.97	0.97	188
MobileNetV2	Calculus	0.93	1.00	0.97	195
	Mouth Ulcer	0.85	0.85	0.85	40
	Tooth Discoloration	0.98	0.94	0.96	303
	Caries	0.67	0.55	0.60	33
	Hypodontia	0.96	0.97	0.97	188
Inception-ResNet-v2	Calculus	0.92	0.98	0.95	195
	Mouth Ulcer	0.92	0.88	0.90	40
	Tooth Discoloration	0.98	0.96	0.97	303
	Caries	0.65	0.45	0.54	33
	Hypodontia	0.91	0.94	0.92	188

The class-wise table, shown in Table 6.4, in general, shows that the architectures with the best performance are quite good in classes with enough data and sufficient morphological differences, but the situation is the other way around when the classes

have only slight differences, few annotated samples, or a lot of variation among the same class. The remarkable result of deeper models on hard classes, particularly ResNet-101 and DenseNet-121, indicates that rich feature extraction and dense connectivity are vital in dental image analysis, which is very robust. Nevertheless, the performance of the Caries class, which is still lower but not as much on Mouth Ulcer, points to the need for data balancing, augmentation, and possibly training techniques tailored to specific classes to address the complexity and poor visibility of these conditions.

6.2.3 RGB Teeth Dataset Model Performance Analysis

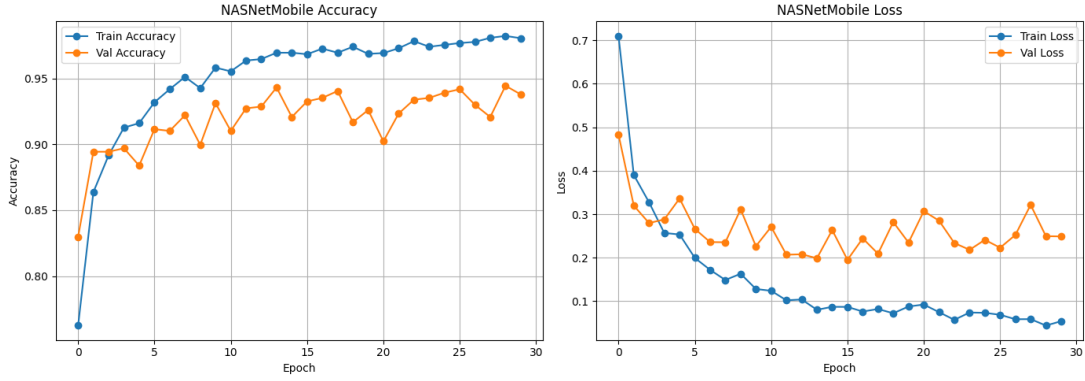


Figure 6.10: Model accuracy and loss convergence of NASNet-Mobile on RGB Teeth dataset

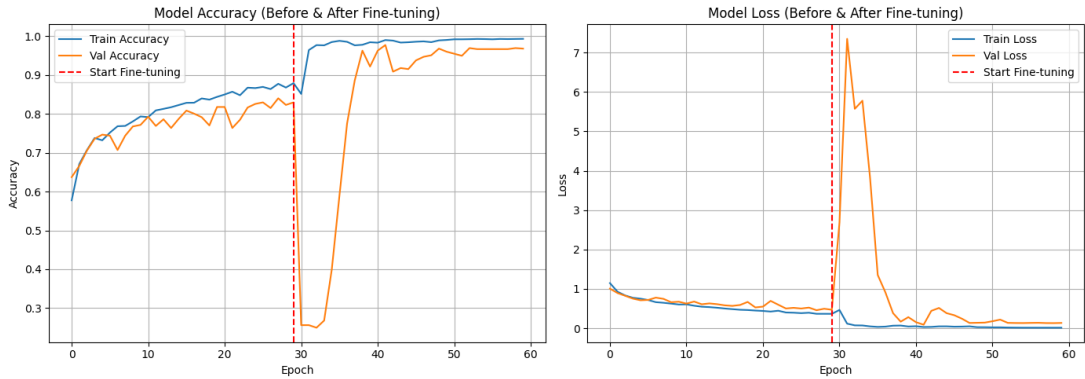


Figure 6.11: Model accuracy and loss convergence of ResNet-101 on RGB Teeth dataset

The NASNet-Mobile model was run for 30 epochs without any adjustments, relying solely on the pretrained feature extractor and newly initialized classification layers. Gradually, the models performance during the training process was increasing steadily, and the training accuracy was climbing to about 97-98%, whereas the validation accuracy was remaining in the 90-94% area. The training loss was going down steadily, which is a sign of effective optimization, while the validation loss was in the range of 20-30% with small fluctuations typical for pretrained models used as fixed feature extractors. In total, the 30-epoch NASNet-Mobile baseline revealed its excellent generalization capability, thus laying down a strong foundation for testing the fine-tuning strategies shown in Figure 6.10.

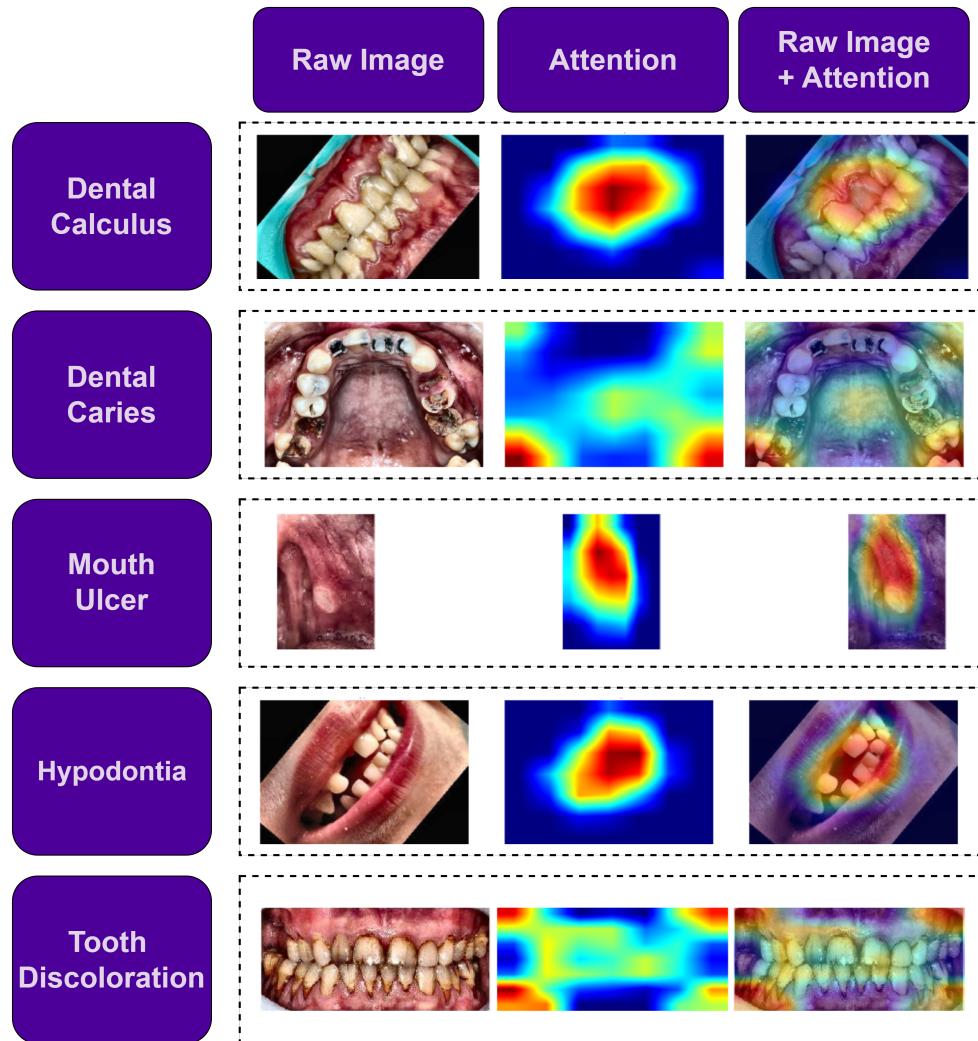


Figure 6.12: Grad-CAM heatmap analysis on the RGB Teeth dataset showing the activated regions that contribute most to model decision-making on the test dataset

The ResNet-101 model was subjected to a two-stage training procedure lasting a total of 60 epochs, divided into two parts: the first one with all layers frozen for 30 epochs, and then the second one for 30 epochs with fine-tuning, where layers from index 313 onwards were unfrozen to allow for deeper feature adaptation. In the beginning, when the layers were frozen, the model performed at a moderate level with training and validation accuracies remaining about 80.85% and losses showing gradual yet limited improvement due to no feature updates being done. After the deeper layers were unfrozen at epoch 30, the model experienced a temporary increase in validation loss; this is to be expected as the parameters that were previously kept static start to update, but the model was able to quickly stabilize and improve significantly. At the end of fine-tuning, the training accuracy reached very high levels, and the validation accuracy was above 95% while both training and validation losses went very low compared to their values at the start. This indicates that the two-stage process has proven the effectiveness of fine-tuning ResNet-101, particularly from the deeper layers, which resulted in building a model with a much greater capability of capturing dataset-specific patterns and thus surpassing the non-fine-tuned baseline as per Figure 6.11.

6.2.4 RGB Teeth Dataset Grad-CAM Visualizations and Explainability

In Figure 6.12, one can see Grad-CAM visualizations for different oral-health conditions the model uses to discriminate regions during the classification. For every sample, the original image is shown in the left panel, the Grad-CAM heatmap is shown in the center panel, and the right panel merges the heatmap into the models prediction. The network’s final decision is influenced most by the regions indicated by high-intensity red and yellow, which are also considered to be the most influential regions.

The classes of Calculus, Mouth Ulcer, and Hypodontia have strong correlations with the model, which focuses consistently on the clinically relevant structures, indicating a high degree of alignment with pathology-specific features. In the case of Caries, the activation is more diffuse, which is indicative of the more complex and variable nature of the disease. The overall visualizations confirm a scenario where the model is learning to discern meaningful spatial patterns and not just depending on irrelevant background cues, thereby supporting the interpretability and reliability of the classifier in real-world diagnostic situations.

6.2.5 RGB Teeth Dataset Comparison of Classification Architectures

The assessment of the performance among six deep-learning models: NASNet-Mobile, ResNet-50, DenseNet-121, MobileNetV2, ResNet-101, and Inception-ResNet-v2 reveals a distinctive ranking concerning the classification ability that is shown in Figure ??.

Among them, ResNet-101 is the model that gets the honor of being the champion as it achieves the top scores in nearly all metrics, such as Macro Precision of 96%, Macro Recall of 92%, Macro F1-Score of 94%, and a leading Test Accuracy of 97%. The competition is very tight between Inception-ResNet-v2 and NASNet-Mobile, as their performances in terms of precision and recall are almost similar, if not equal. Inception-ResNet-v2 has a score of 95% Macro Precision, 87% in Macro Recall, 90% in Macro F1, and 95% accuracy, while NASNet-Mobile is just around the corner, performing equally, however, with a little higher recall and accuracy.

MobileNetV2 exhibits remarkable outcomes given that its architecture is light-weight, gaining around 88% in the macro metrics and a very high accuracy of 94%. DenseNet-121 performs moderately and achieves middle-range values across all metrics, while ResNet-50 is at the bottom of the rank with a clear decline in precision, recall, and overall accuracy.

Table 6.5: Test performance comparison across DL models on RGB Teeth dataset

Model	Mac. Precision	Mac. Recall	Mac. F1-Score	Accuracy
NASNet-Mobile	0.92	0.88	0.90	0.96
ResNet-50	0.75	0.69	0.71	0.87
DenseNet-121	0.85	0.81	0.83	0.92
MobileNetV2	0.88	0.86	0.87	0.94
ResNet-101	0.96	0.92	0.94	0.97
Inception-ResNet-v2	0.95	0.87	0.90	0.95

6.2.6 RGB Teeth Dataset Strengths and Weaknesses Across Classification Models

The comparative evaluation in Table 6.5 exhibits the six DL models applied to the RGB Teeth dataset in their distinct strengths and weaknesses. Among the models, ResNet-101 outperformed the rest by getting the highest numbers for macro precision 96%, macro recall 92%, macro F1-score 94%, and overall accuracy 97%.

In other words, its capability to extract rich hierarchical features as well as to generalize effectively across all classes was demonstrated. Inception-ResNet-v2 and NASNet-Mobile were next in line, with precision that was nearly as strong and accuracy that was competitive. However, their slightly lower recall values indicate that they occasionally had difficulty in consistently identifying minority or visually subtle classes.

On the other hand, MobileNetV2 and DenseNet-121 exhibited a trade-off performance with modest F1-scores. They managed to achieve this with a resource-efficient feature reuse and a lightweight architectural design, making them appropriate for real-time or resource-limited deployments, even though their accuracy did not reach the top. ResNet-50, however, had the poorest performance overall, suggesting that its deep architecture did not capture the complex intra-class variations of dental images adequately. To sum up the results, they affirmed that deeper and more advanced architectures give superior generalization while lightweight models provide a balance between computational efficiency and classification robustness.

6.2.7 RGB Teeth Dataset AUC-ROC Curves of the Best Performing Models

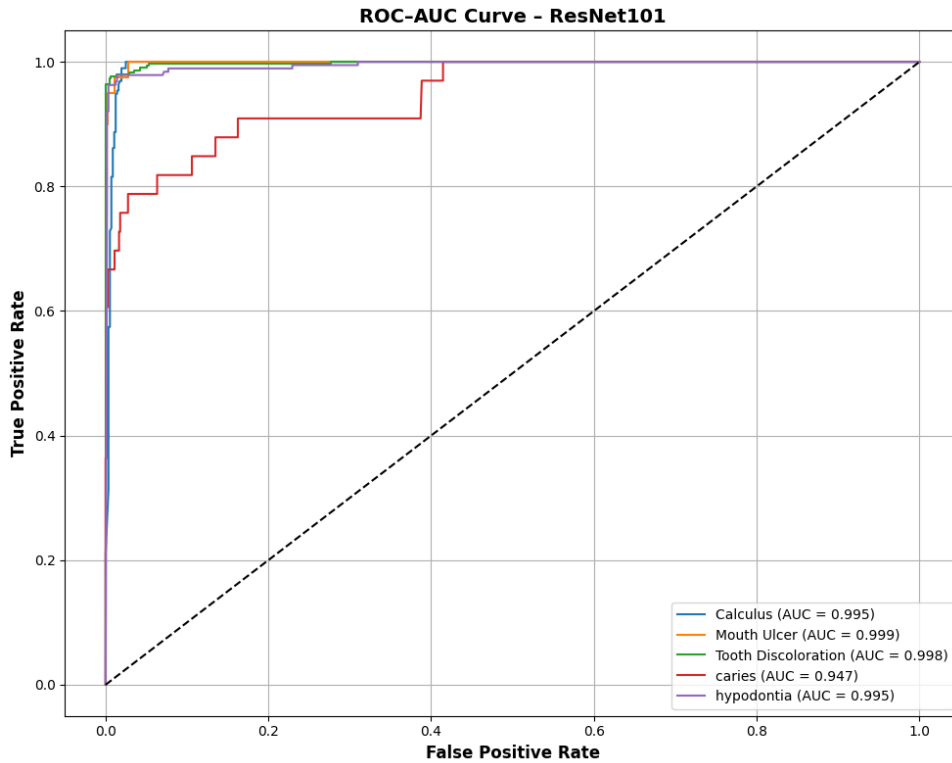


Figure 6.13: AUC/ROC curve of RGB Teeth dataset on ResNet-101 model

Figure 6.13 and Figure 6.14 present the multiclass Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) curves comparing ResNet-101 and NASNet-Mobile

for five oral-health classification categories. Both models demonstrate excellent discriminative performance, with AUC values exceeding 99% for Calculus, Mouth Ulcer, Tooth Discoloration, and Hypodontia. These curves show a steep rise toward the upper-left corner, indicating high true-positive rates with minimal false-positive errors. The only comparatively weaker class is dental caries, where both models achieve an AUC of approximately 94.7%; this suggests greater intra-class variability or visual similarity with other categories. Overall, the ROC plots highlight that both architectures are highly effective, with NASNet-Mobile performing on par with the deeper ResNet-101 while maintaining a lighter computational footprint making it suitable for deployment in resource-constrained environments.

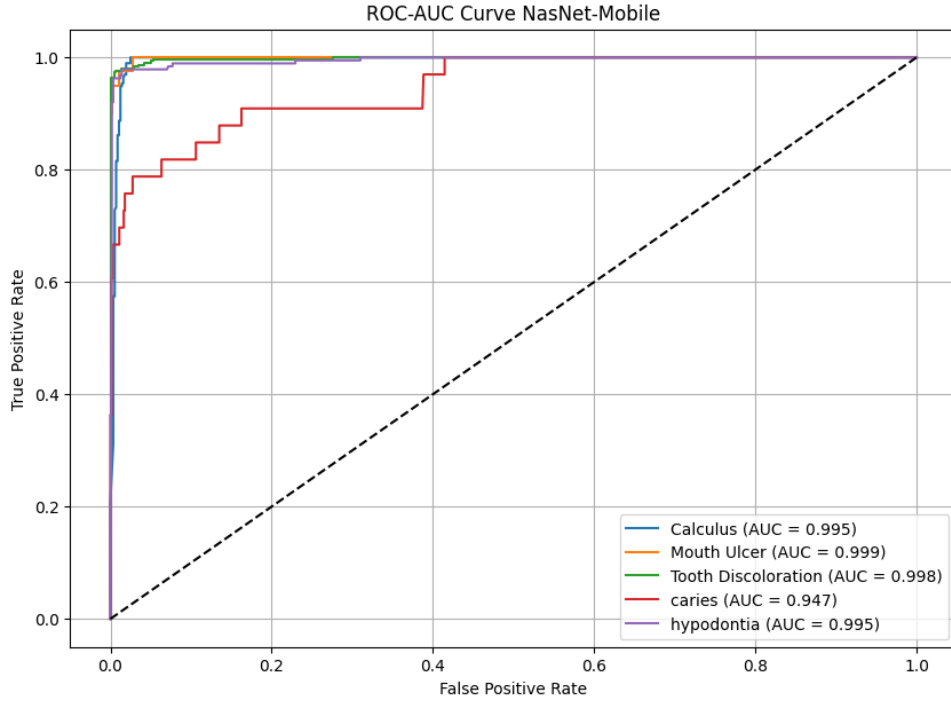


Figure 6.14: AUC/ROC curve of RGB Teeth dataset on NASNet-Mobile model

6.2.8 Handling Class Imbalance in Caries Classification on RGB Teeth Dataset

1. **Oversampling with Data Augmentation:** A significant class imbalance was present in the original dataset, especially for the Caries class, which had only 153 training images, while there were 907 for Calculus, 875 for Hypodontia, and 1411 for Tooth Discoloration. This kind of imbalance leads to a classifier being biased financially toward the majority classes and thus poor per-class performance for Caries. To solve this problem, the Caries class was oversampled in the training set through heavy data augmentation. The first step involved determining the largest class size in the training set (1411 images for Tooth Discoloration). To match the quantity of the Tooth Discoloration class, the Caries class needed 1258 additional synthetic images. The ImageDataGenerator module of TensorFlow/Keras was used to create an augmentation pipeline, which has applied controlled transformations such as small rotations, zoom, shifts, brightness variation, and horizontal flips. Thus, the transformations not only preserved the anatomical characteristics of carious lesions but also contributed a lot to the visual variety.

The analysis done through the comparison indicates that the Caries class, which

was a major underrepresented class in the dataset, benefited from oversampling significantly. The model was before the augmentation and balancing of the training data very few effective in recognizing dental caries. Its F1 score was only 48% due to very low recall that is 39%. With augmented data from the targeted method, the model performance across metric categories was significantly higher. Precision transitions from 62% to 80%, and F1 score transitions from 48% to 59% show that indeed the model was able to find more distinctive features for this minority class. Recall increase was not that dramatic, but still from 39% to 47% it can be seen that there is a great overshadowing. Generally, oversampling is seen to have lessened the bias towards majority classes and thereby improved the reliability of the Caries classification that is shown in Table 6.6. The comparison summary of experiments between before and after oversampling is shown in Table 6.7.

In Figure 6.15 the confusion matrix reveals that NASNet-Mobile, after the application of oversampling to balance the Caries category, still scored high in all five classes of dental diseases with a strong and reliable classification due to the clear diagonal dominance. Calculus deposits are delivered through the right process in 183 cases out of 195 and only slight interruption is reported towards Tooth Discoloration and Hypodontia. Mouth Ulcer comes out with a good performance as well, being able to identify 31 images from 41, where most wrong guesses were made against Tooth Discoloration on account of possible similarities in the appearance. Tooth Discoloration stands out with the highest accuracy, as 292 out of 304 samples are correctly identified with only a few misclassifications spread throughout the classes. Meanwhile, the Caries class that had a dramatic imbalance at the outset, now undeniably fewer cases of confusion after oversampling, with 16 out of 34 cases effectively classified, showing improvement in feature learning. Hypodontia does not waver and continues to be a top performer with 182 correct out of 189 guesses and little misclassification. The overall picture given by the confusion matrix is that minority-class Caries recognition received a moderate boost from oversampling but at the same time the change is very negligible.

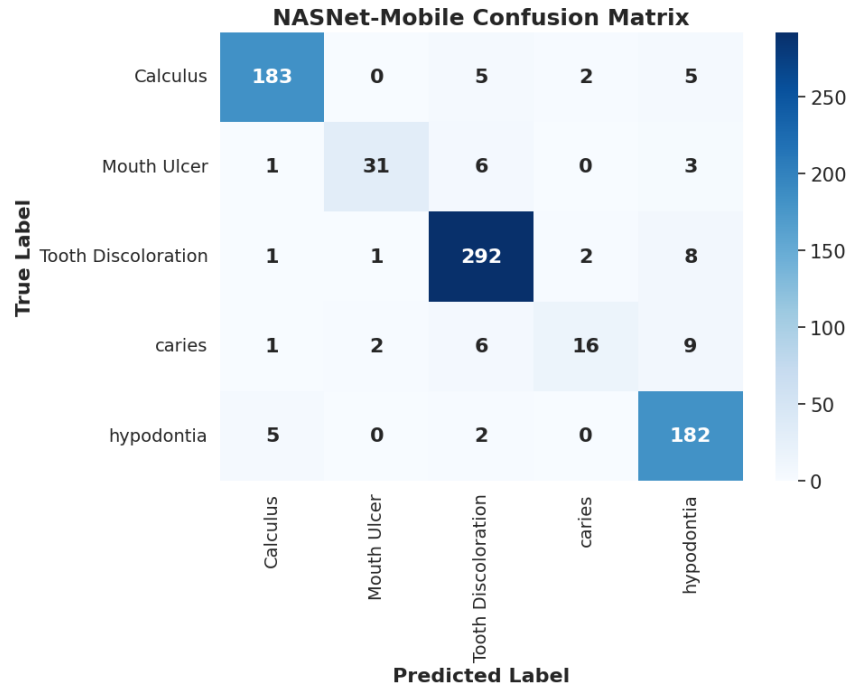


Figure 6.15: Confusion matrix of NASNet-Mobile after oversampling on the RGB Teeth dataset

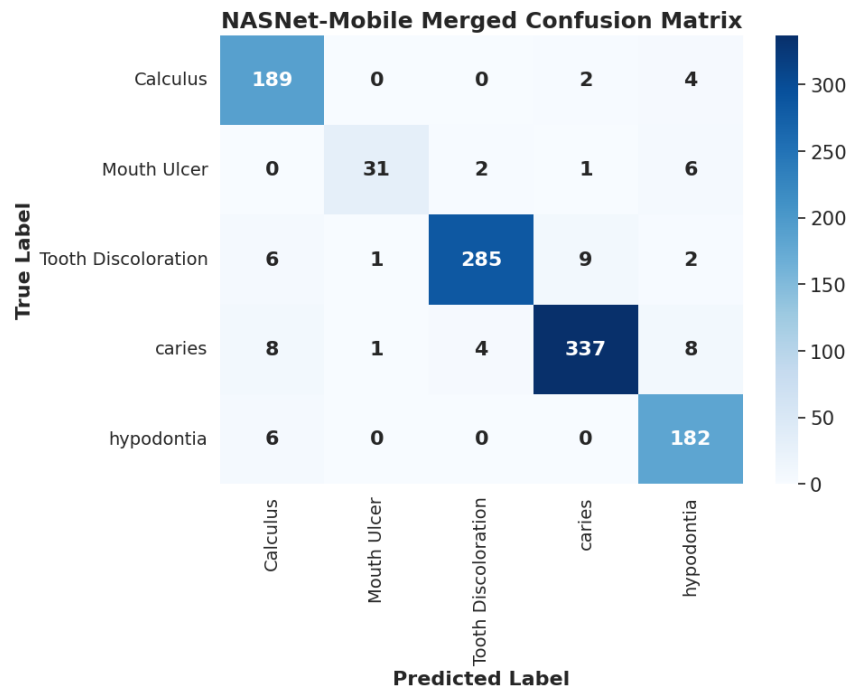


Figure 6.16: Confusion matrix of NASNet-Mobile after merging Intraoral Dataset & RGB Teeth dataset

Table 6.6: Per-class performance of NASNet-Mobile after oversampling

Class	Precision	Recall	F1-score	Support
Calculus	0.96	0.94	0.95	195
Mouth Ulcer	0.91	0.76	0.83	41
Tooth Discoloration	0.94	0.96	0.95	304
Caries	0.80	0.47	0.59	34
Hypodontia	0.88	0.96	0.92	189
Overall Accuracy	0.92 (763 test samples)			

Table 6.7: Performance comparison of before and after oversampling

Metric	Before Oversampling	After Oversampling
Precision	0.62	0.80
Recall	0.39	0.47
F1-score	0.48	0.59
Support	33	33

Table 6.8: NASNet-Mobile performance on test data for five dental classes

Class	Precision	Recall	F1-score	Support
Calculus	0.90	0.97	0.94	195
Mouth Ulcer	0.94	0.78	0.85	40
Tooth Discoloration	0.98	0.94	0.96	303
Caries	0.97	0.94	0.95	358
Hypodontia	0.90	0.97	0.93	188
Overall Accuracy	0.94 (1084 test samples)			

Table 6.9: Performance comparison of before and after merging the dataset

Metric	Before Merging Dataset	After Merging Dataset
Precision	0.62	0.97
Recall	0.39	0.94
F1-score	0.48	0.95
Support	33	358

2. Merging an external Dental Caries dataset:

To merge a foreign Caries dataset with our collection of dental images, we started by making a common folder structure that could hold all five classes: Calculus, Mouth Ulcer, Tooth Discoloration, Caries and Hypodontia. The number of Caries images was made larger by taking 2,164 images from the Intraoral dataset and 219 images of the existing RGB Teeth dataset were added; thus, the representation was good enough. After the merging, we double-checked all classes and correctly detected them and then the dataset was split into training, validation and test sets with 70%, 15% and 15% ratios, respectively. This led to 5,046 images for training, 1,082 images for validation, and 1,084 images for testing, where the class distribution was adequate for the splits (e.g. Caries: 1,668 train, 357 val, 358 test). To improve the

contrast in the images and to make the dental features clearer, we used CLAHE (Contrast Limited Adaptive Histogram Equalization) on all images in each split, replacing the original files. Subsequently, we instantiated ImageDataGenerators for each split where we applied data augmentation methods such as rotation, zoom, and horizontal flipping to the training set while rescaling was done for all images for normalization purposes.

A total of 1,084 dental images represented five categories: Calculus, Mouth Ulcer, Tooth Discoloration, Caries, and Hypodontia were used to test the NASNet-Mobile model. The overall accuracy of the model was 94%, and in fact, this accuracy was broadly recognized particularly, for the increase in accuracy for the classification of Caries was a direct benefit from including an external dataset of Caries in this study shown in Table 6.8. The addition of the external dataset led to an improvement in all three metrics: precision increased from 62% to 97%, recall increased from 39% to 94% and F1-score increased from 48% to 95%. Also, the number of Caries images increased substantially from 33 images to 358 images, which provides strong evidence of the positive impact of adding more dataset on model performance as shown in Table 6.9. Additionally, the model had good performance for Tooth Discoloration (98% precision, 94% recall) and Calculus (90% precision, 97% recall), whereas Mouth Ulcer had a lower score for recall (78%), meaning there were a few incorrect classifications; however, F1-scores for all categories were greater than 85% and therefore indicated reliable and consistent performance, even after merging the Caries images.

The confusion matrix depicted in Figure 6.16 reflects the performance of the NASNet-Mobile model in classifying the dental diseases after training on the combined dataset of Intraoral Caries and RGB Teeth. The model proves to be highly predictive in five distinct dental disease categories. The NASNet-Mobile model has made a correct identification of the greatest number of samples pertaining to the diseases of Calculus (189), Mouth Ulcer (31), Tooth Discoloration (285), Caries (337), and Hypodontia (182). This conveys the very good feature-learning capability of the model which is generalized even after the two datasets with different distributions have been merged. Misclassifications are very few and mostly occur among the clinically similar classes. For instance, there are nine instances of Tooth Discoloration samples being misclassified as Caries, and six Hypodontia cases being predicted as Calculus. NASNet-Mobile performed quite well after the merging approach.

6.3 Object Detection Performance on OPG Dataset

To establish the effectiveness of the leading-edge algorithms for object detection on dental radiography, a comparative analysis of five YOLO model versions: v8, v9, v10, v11, and v12 was conducted. Moreover, all tiny, nano (n), and small (s) variants were employed to examine the compromise between model complexity and detection accuracy. The performance outcomes are presented and visualized in Figure 6.17, while numerical details can be found in Table 6.10. It is very clear from the table that the models are evaluated across several classes, and there is a general observation of the trend that the 's' variants are always better than their 'n' and 't' counterparts when it comes to the complex classes like Caries and Broken Down Crown or Roots.

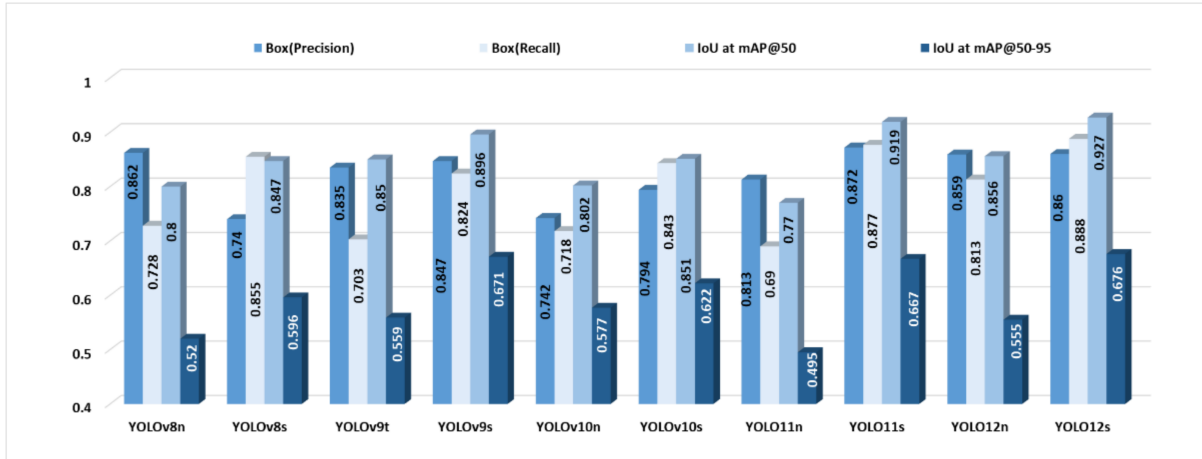


Figure 6.17: Performance summary of the YOLO models on the OPG dataset

6.3.1 Precision-Recall Curves and Confidence Analysis

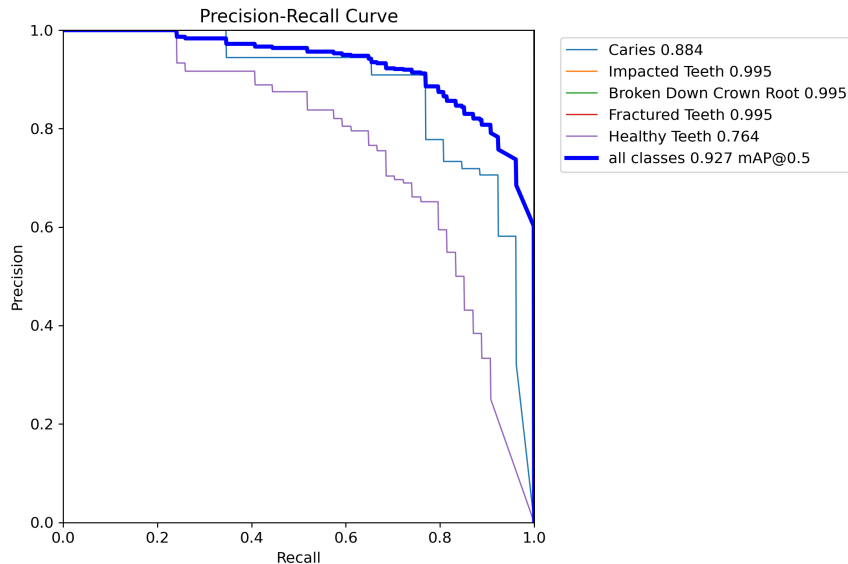


Figure 6.18: Box precision and recall curve of YOLO12s on the OPG dataset

The Precision (Box P) and Recall (R) metrics provided in Table 6.11 and Table 6.10 show different behaviors during the coming generations of the model. While precision stays

stable, recalls vary significantly. To illustrate, in the "Caries" category, the YOLOv8n model exhibits a recall of 44%, which corresponds to a large portion of missed detections classified in this subtle pathology. On the other hand, the YOLO11s and YOLO12s models respectively show Caries recalls of 84% and 76% without their precision dropping sharply. It is also interesting to note the "Broken Down Crown/Root" class, which scored a perfect recall of 100%, displaying this trait in a number of the models (YOLOv8n, YOLOv8s, YOLOv9t, YOLOv9s, YOLOv10s, YOLO12s). Such a constant recall score can be interpreted that the morphological features of fractures are so unique that even a weak confidence threshold would allow the YOLO network to classify the images properly. The output of the best model, YOLO12s, is demonstrated in Figure 6.18 and Figure 6.19.

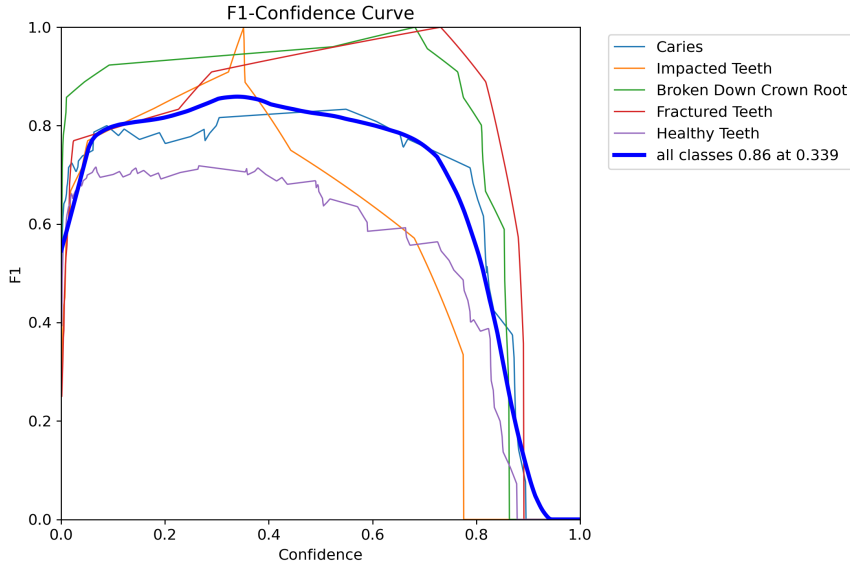


Figure 6.19: F1 confidence curve of YOLO12s on the OPG dataset

6.3.2 mAP@50 and mAP@50-95 Comparison

The Mean Average Precision (mAP) metrics in the result table have made it clear that the capability of the models has improved, which can be seen from the consistently strong mAP@50 scores, where "Broken Down Crown/Root" and "Impacted Teeth" are often above 0.95 in the case of the new models (YOLO11 and YOLO12). Nonetheless, the mAP@50-95 metric provides a more rigorous evaluation because it punishes poor localization. The data presented in Table 6.11 and Table 6.10 depict the scenario where "Caries" is still the hardest class to be localized and the mAP50-95 scores for this class are generally around 40-60%. The YOLO12s model is a winner in terms of composite performance, especially for the "Broken Down Crown Root" (mAP50-95 of 81%) and "Infection" (mAP50-95 of 70%), which shows excellent boundary regression over the ground versions; v8 and v9.

Table 6.10: Performance comparison of YOLO models across different dental classes on the OPG dataset

Models	Class	Images	Instances	Box(P	R)	mAP@50	mAP@50-95
YOLOv8n	Caries	14	26	0.881	0.44	0.655	0.429
	Infection	2	5	0.275	0.6	0.533	0.269
	Impacted Teeth	7	12	0.925	0.917	0.959	0.655
	Broken Down Crown/Root	4	5	0.755	1	0.925	0.72
	Healthy Teeth	23	54	0.77	0.655	0.764	0.524
YOLOv8s	Caries	14	26	0.762	0.769	0.755	0.523
	Infection	2	5	0.793	0.774	0.925	0.63
	Impacted Teeth	7	12	0.869	1	0.995	0.65
	Broken Down Crown/Root	4	5	0.734	1	0.962	0.743
	Healthy Teeth	23	54	0.845	0.731	0.594	0.406
YOLOv9t	Caries	14	26	0.802	0.577	0.751	0.473
	Infection	2	5	1	0.4	0.553	0.543
	Impacted Teeth	7	12	0.865	0.833	0.919	0.635
	Broken Down Crown/Root	4	5	0.825	1	0.995	0.632
	Healthy Teeth	23	54	0.679	0.704	0.734	0.51
YOLOv9s	Caries	14	26	0.833	0.731	0.895	0.646
	Infection	2	5	0.975	0.8	0.815	0.625
	Impacted Teeth	7	12	0.757	0.924	0.963	0.732
	Broken Down Crown/Root	4	5	0.829	1	0.995	0.764
	Healthy Teeth	23	54	0.811	0.667	0.807	0.556
YOLOv10n	Caries	14	26	0.858	0.615	0.764	0.492
	Infection	2	5	0.746	0.591	0.672	0.463
	Impacted Teeth	7	12	0.794	0.833	0.913	0.652
	Broken Down Crown/Root	4	5	0.7	0.939	0.935	0.795
	Healthy Teeth	23	54	0.612	0.611	0.724	0.45
YOLOv10s	Caries	14	26	0.759	0.862	0.875	0.62
	Infection	2	5	0.79	0.8	0.8	0.503
	Impacted Teeth	7	12	0.855	0.833	0.853	0.656
	Broken Down Crown/Root	4	5	0.745	1	0.895	0.75
	Healthy Teeth	23	54	0.793	0.722	0.804	0.55
YOLO11n	Caries	14	26	0.796	0.602	0.685	0.441
	Infection	2	5	0.95	0.4	0.603	0.25
	Impacted Teeth	7	12	0.893	1	0.959	0.682
	Broken Down Crown/Root	4	5	0.863	0.8	0.962	0.697
	Healthy Teeth	23	54	0.533	0.648	0.605	0.407
YOLO11s	Caries	14	26	0.847	0.849	0.881	0.631
	Infection	2	5	1	0.935	0.995	0.696
	Impacted Teeth	7	12	0.842	0.917	0.971	0.695
	Broken Down Crown/Root	4	5	0.855	1	0.995	0.765
	Healthy Teeth	23	54	0.755	0.655	0.753	0.544
YOLO12n	Caries	14	26	0.796	0.752	0.785	0.511
	Infection	2	5	1	0.754	0.837	0.485
	Impacted Teeth	7	12	0.876	0.917	0.954	0.634
	Broken Down Crown/Root	4	5	1	0.953	0.995	0.665
	Healthy Teeth	23	54	0.62	0.63	0.679	0.474
YOLO12s	Caries	14	26	0.875	0.762	0.854	0.523
	Infection	2	5	0.929	1	0.995	0.707
	Impacted Teeth	7	12	0.895	1	0.995	0.704
	Broken Down Crown/Root	4	5	0.882	1	0.995	0.819
	Healthy Teeth	23	54	0.751	0.67	0.764	0.556

6.3.3 Localization and Error Cases

In order to gain a deeper insight into the localization issues, we scrutinized the predicted label outputs depicted in Figure 6.20 for the best performing model, the YOLO12s. The overlapping features in panoramic radiographs, which are perceived as the main cause of error and correspond to the lower mAP50-95 scores for "Healthy Teeth" and "Caries"

categories, are the primary and main source of error. The regions of transition between healthy enamel and early-stage dental caries are quite often very subtle, resulting in lower Intersection over Union (IoU) scores. The Localization analysis reveals that, even though the models are successful in spotting the region of interest (high mAP@50), the accurate bounding box alignment for irregular shapes like dental caries (mAP50-95) continues to be a difficult task for the nano-scale models (YOLOv8n, YOLOv10n). The confusion matrix of the YOLO12s model is visualized in Figure 6.21, with the horizontal axis showing the actual labels and the vertical axis showing the predicted labels.

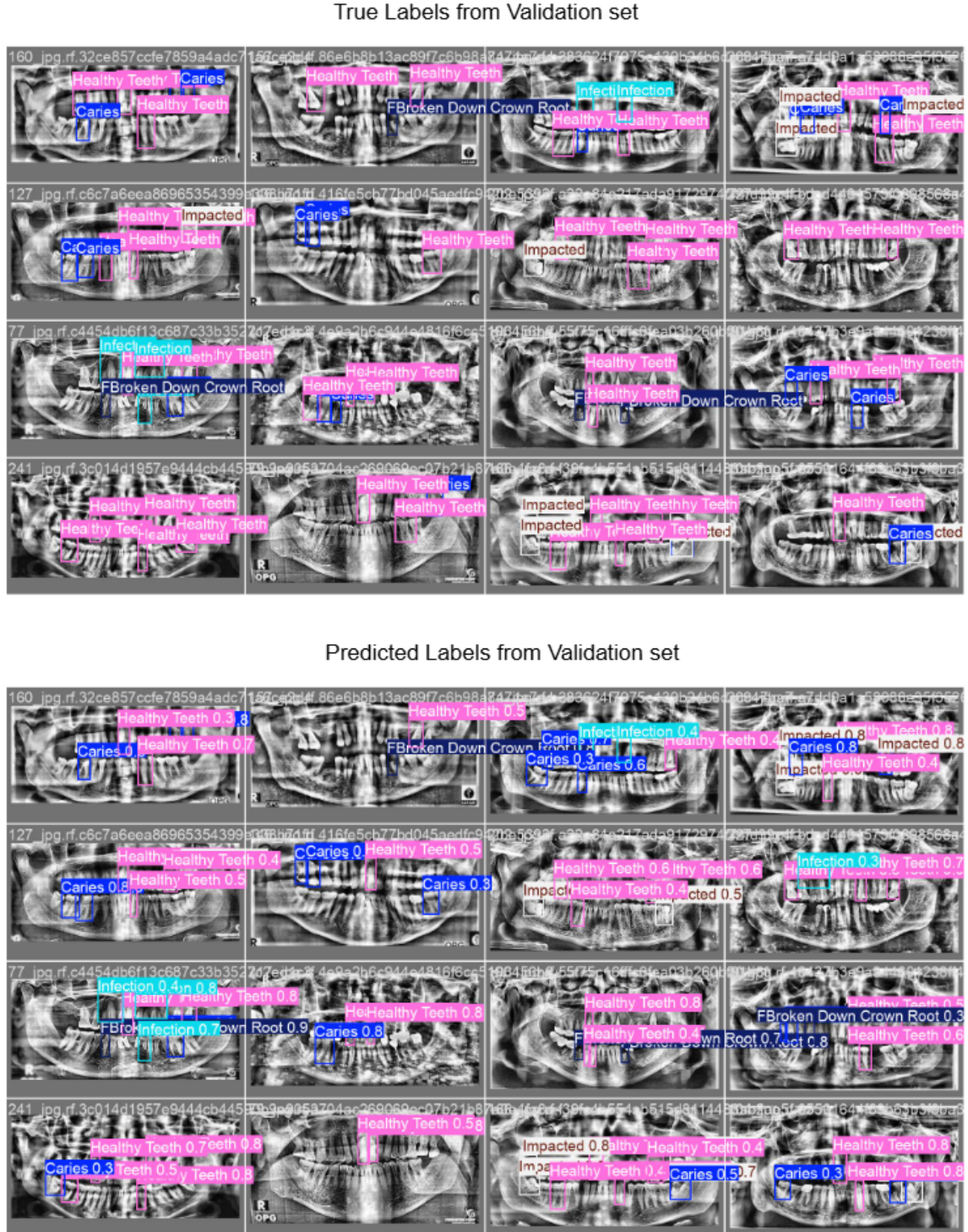


Figure 6.20: Ground truth vs Predicted labels on validation set of the OPG dataset

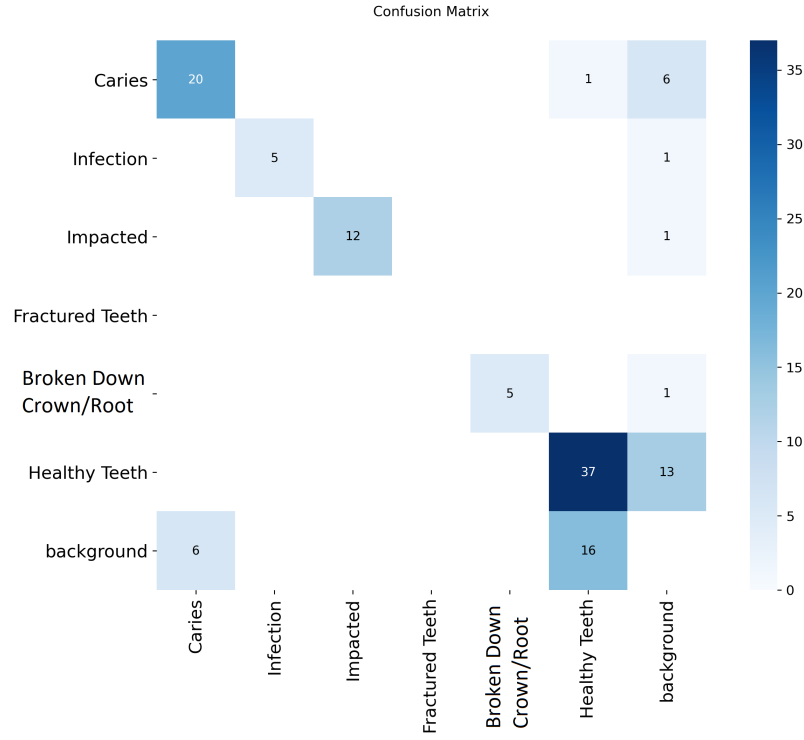


Figure 6.21: Confusion matrix of YOLO12s on the OPG dataset

6.3.4 Strengths and Weaknesses Across Models

The comparative analysis presented in Table 6.11 and Table 6.10 highlights certain aspects of clinical application related to strengths and weaknesses. The best thing about YOLO12s and YOLO11s models is that they are good at detecting structural anomalies, with models such as the YOLO12s model getting almost perfect results for the case of "Impacted Teeth" and "Broken Down Crown Roots." Nonetheless, a common issue shared by all generations is the continual problem with Caries detection, with a trade-off between recall and precision in the nano models. In addition, detection of "Healthy Teeth" has rather low mAP scores (e.g., YOLO12n at 47% mAP₅₀₋₉₅), probably caused by the high number of overlapping instances and visual confusion of OPG images. Thus, the 's' variants, especially the YOLO12s, have a very good performance for clinical decision support, where missing a diagnosis (low recall) is an important critical failure mode.

Table 6.11: Comparison of precision, recall, and mAP metrics across different YOLO models

Models	Precision	Recall	IoU at mAP@50	IoU at mAP@50-95
YOLOv8n	0.862	0.728	0.8	0.52
YOLOv8s	0.74	0.855	0.847	0.596
YOLOv9t	0.835	0.703	0.85	0.559
YOLOv9s	0.847	0.824	0.896	0.671
YOLOv10n	0.742	0.718	0.802	0.577
YOLOv10s	0.794	0.843	0.851	0.622
YOLO11n	0.813	0.69	0.77	0.495
YOLO11s	0.872	0.877	0.919	0.667
YOLO12n	0.859	0.813	0.856	0.555
YOLOv12s	0.86	0.888	0.927	0.676

6.4 Object Detection Performance on RGB Intraoral Caries Dataset

The RGB intraoral caries dataset was also analyzed with the help of ten cutting-edge YOLO models, ranging from light-weight Nano versions (YOLOv8nYOLOv12n) to the more powerful Small versions (YOLOv8sYOLOv12s), similar to the study in the OPG Dataset. A visual summary of the overall model behavior across metrics is provided in Figure 6.22, while a detailed report of the precise numerical values for Precision, Recall, and localization performance is given in Table 6.12 for IoU thresholds of mAP@50 and mAP@50-95. The performance of all models did not vary much, and it was always high; Precision values were between 81% and 88%, while Recall ranged from 83% to 89%. The mAP@50 values observed were between 91% and 92%, which means that the lesion localization was very good, whereas mAP@50-95 values indicated the comparative toughness of detecting small, irregular carious lesions across stricter IoU thresholds as they varied between 72% and 75%. Out of all the architectures tested, YOLOv12n was the best-performing lightweight model, as it performed equally well across all four metrics. YOLOv9s and YOLOv11s were also powerful in generalization, especially in Recall, which is very important from a clinical perspective since it helps reduce the number of missed detections.

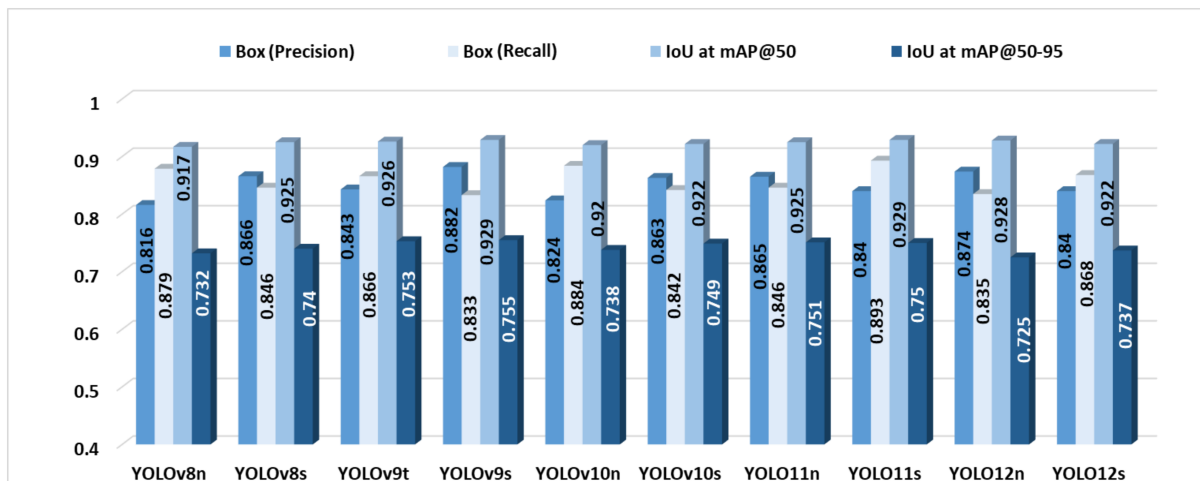


Figure 6.22: Performance summary of the YOLO models on the RGB Intraoral Caries dataset

6.4.1 RGB Intraoral Caries Dataset Precision-Recall Curves and Confidence Analysis

The behavior of the PrecisionRecall (PR) curve relating to the best model, YOLO12n, is presented in Figure 6.23, and it shows how the model manages false positives and false negatives at different confidence levels. The PR curve exhibits a very stable performance, where the precision was always higher than 80% while the recall was still increasing, which means good reliability of detection under various lesion sizes and light conditions. The F1 confidence curve in Figure 6.24, in fact, also shows the same trend of the precision-recall curve and has a peak that is smooth and is located around the optimal confidence region used in drawing the inferences. This implies that the confidence threshold selected for the final predictions is a judicious compromise between sensitivity and specificityan essential feature of a diagnostic screening tool, where high recall along with occasional

false positives is demanded.

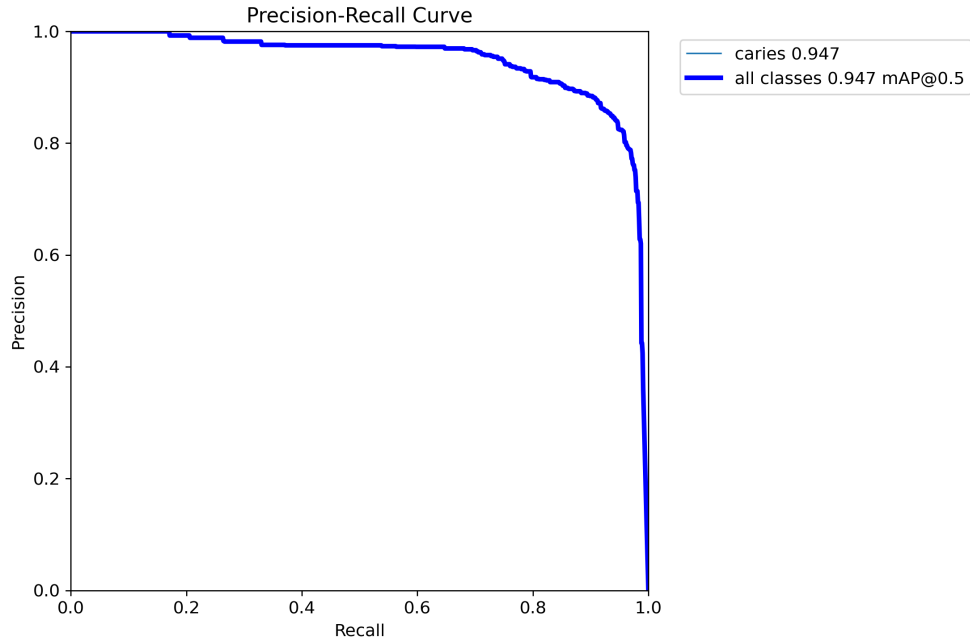


Figure 6.23: Box precision and recall curve of YOLO12n on the RGB Intraoral Caries dataset

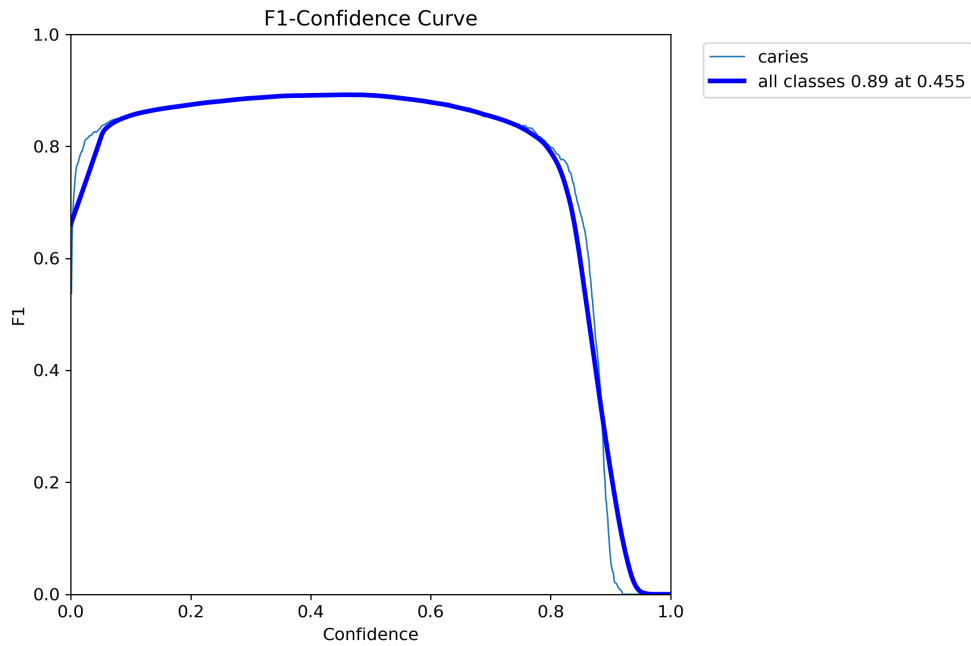


Figure 6.24: F1 confidence curve of YOLO12n on the RGB Intraoral Caries dataset

6.4.2 RGB Intraoral Caries Dataset mAP@50 and mAP@50-95 Comparison

The results displayed in Table 6.12 about mAP show two significant trends. The first one is that the mAP@50 among all model values is more than 91%, which means the detection of coarse lesions is always accurate for all versions of YOLO. This implies that

even if there were changes in angle, intensity, or oral conditions, the models would still point out the approximate area of the carious lesions perfectly.

The second trend is a significant drop in $mAP@50-95$ values, which is between 72% and 75%. The drop is anticipated due to the higher IoU thresholds, thus the requirement for very tight bounding box alignment. The lesions caused by dental caries are mostly subtle, small, and irregular; hence, accurate positioning is a hard task to accomplish. YOLOv9s and YOLOv12n have the highest $mAP@50-95$, which means that these two models are more than the others at handling the stricter localization demand. This points towards the better progressive performance in bounding-box regression as a result of the architectural innovations in these recent variants.

6.4.3 RGB Intraoral Caries Dataset Localization and Error Cases

In order to study the nature of the detection errors, the output predictions of YOLO12n were reviewed with some representative samples, and the results are depicted in Figure 6.26. Even though YOLO12n had the best overall performance among the lightweight criteria, certain localization error trends were noticed.

To begin with, in the instances of tiny or initial carious lesions, the model sometimes would generate a bounding box that was a little way off from the actual one, thus not surrounding the lesion area. The second source of the error was images containing glare, saliva reflections, or overlapping tooth structures being the most likely candidates for false negatives; hence, noise and occlusion remain difficult situations. Finally, occasional false positives were noticed on enamel surfaces that show high reflection, where the texture variations were similar to those of early demineralization. The confusion matrix of the YOLO12s model is shown in Figure 6.25.

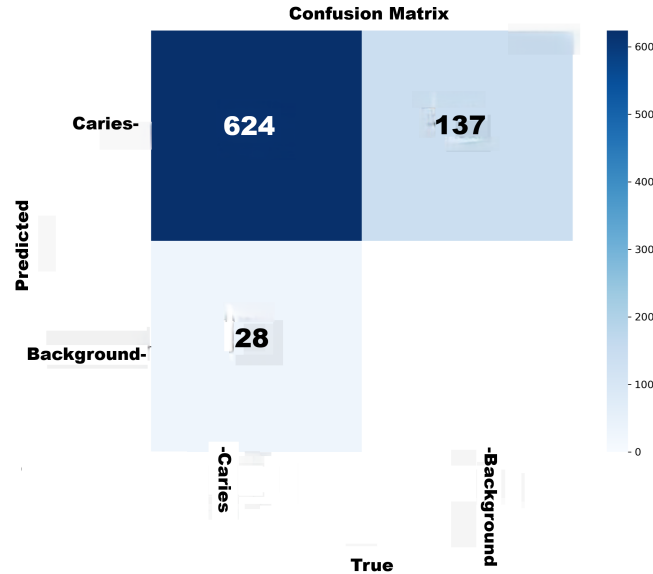


Figure 6.25: Confusion matrix of YOLO12n on the Intraoral Caries dataset

True Labels from Validation set



Predicted Labels from Validation set



Figure 6.26: Ground truth vs Predicted labels on validation set of the RGB Intraoral Caries dataset

6.4.4 RGB Intraoral Caries Dataset Strengths and Weaknesses Across Models

The comparative assessment in Table 6.12 uncovers a mix of strengths and weaknesses among the YOLO families that were reviewed. YOLOv9s and YOLOv12n are bringing the highest localization accuracy along with Precision and Recall, which proves their appropriateness for medical screening workflows. Their advanced feature extraction and improved regression heads are the reasons for better identifying small lesions.

Lighter models like YOLOv8n and YOLOv10n are still very efficient in terms of computation, but they come with the price of slightly lower precision and thus, a higher rate of false positives. On the other hand, YOLOv11s takes the lead in recall (89%), which is a great advantage when it comes to reducing the number of missed detections; however, it has to sacrifice a little bit of precision.

Therefore, all the models are very competitive, but the latest YOLO (v11 and v12) variants have much more reliable and clinically robust detection behavior, particularly when diagnostic applications require both high recall and acceptance of localization accuracy.

Table 6.12: Comparison of precision, recall, and mAP metrics across different YOLO models

Models	Precision	Recall	IoU mAP@50	IoU mAP@50-95
YOLOv8n	0.816	0.879	0.917	0.732
YOLOv8s	0.866	0.846	0.925	0.740
YOLOv9t	0.843	0.866	0.926	0.753
YOLOv9s	0.882	0.833	0.929	0.755
YOLOv10n	0.824	0.884	0.920	0.738
YOLOv10s	0.863	0.842	0.922	0.749
YOLOv11n	0.865	0.846	0.925	0.751
YOLOv11s	0.840	0.893	0.929	0.750
YOLOv12n	0.874	0.835	0.928	0.725
YOLOv12s	0.840	0.868	0.922	0.737

7 System Implementation

7.1 Methodology and Overall System Architecture

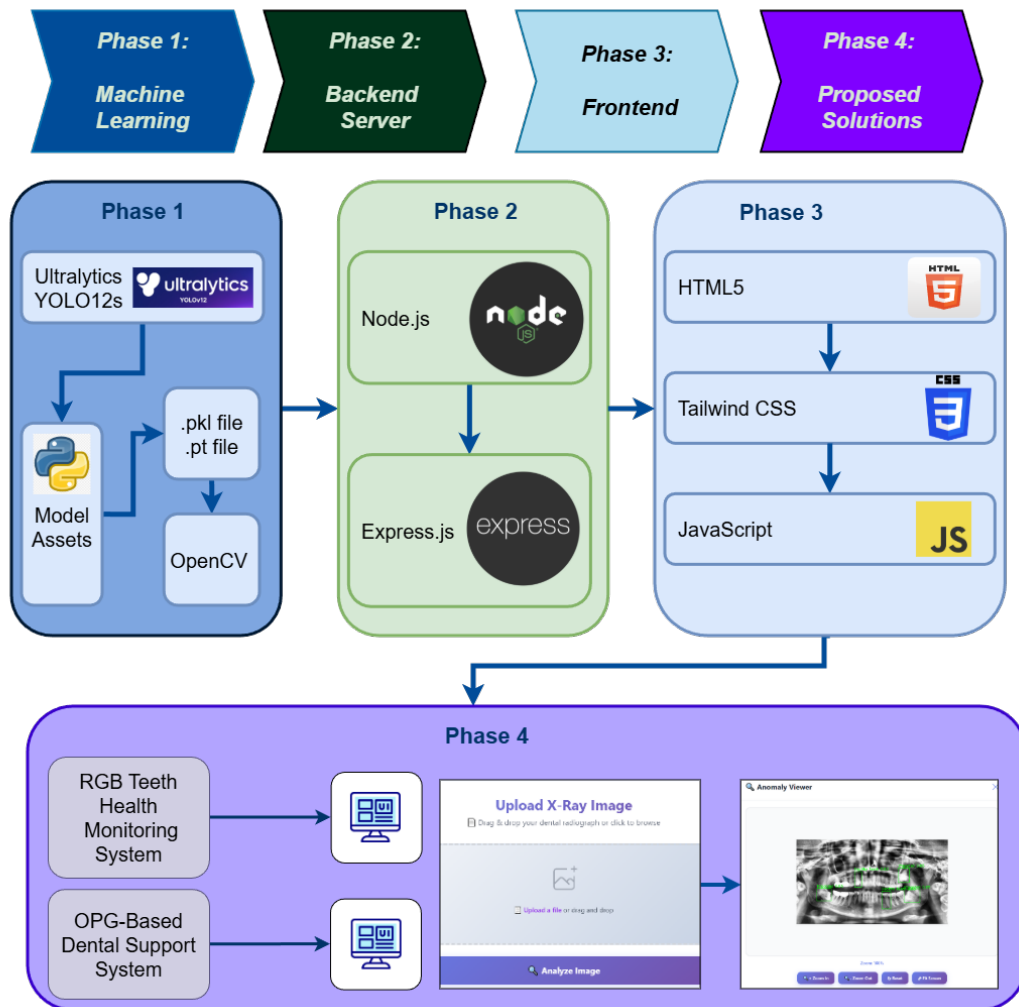


Figure 7.1: Methodology of proposed solution

Upon completing the result analysis and identifying the best models for both suggested systems the OPG-Based Dental Decision Support System and the RGB Teeth Health Monitoring System for general users, the YOLO12s object detection model was selected and implemented on both proposed systems. The two systems share similar architectural principles yet operate differently, catering to distinct users and environments. The OPG-based system is meant for hospital usage and aims to assist the dentist in diagnosing radiographs, whereas the RGB-based system is for the public and intends to provide a

basic dental caries screening and self-monitoring. As a result, the implementation was done in a modular, service-oriented manner that included a frontend interface, a backend server, and a machine learning inference module as the three core layers, as illustrated in Figure 7.1.

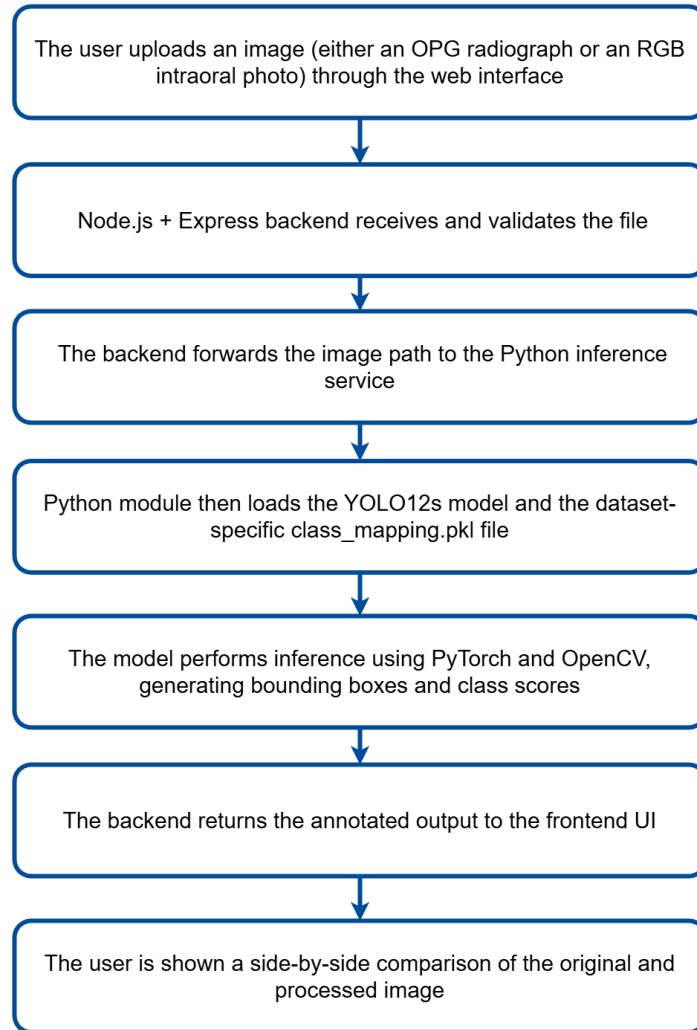


Figure 7.2: Technical simulation of the proposed system

Apart from that, the CNN-based classifier model that was first trained on the RGB data could not be part of the live system since the dataset lacks a “healthy” class, thus the classifier is not able to identify healthy teeth. Bringing such a model into a real-world web application would result in healthy uploads being categorized as diseased; hence, it would not be able to cater to the general public. That is why the RGB system received only the YOLO12s detection-based pipeline initially for the sake of ensuring the reliability and safety of user interaction.

Both systems in the workflow follow a similar pattern: the user makes an upload through the browser-based frontend, the server processes the request and transfers the image, the machine-learning inference module executes the model with the fine-tuned YOLO12s weights, and then sends the annotated predictions back to the user interface. The modular architecture makes it easier to scale, maintain, and upgrade the model or expand the system in the future. The simulation is illustrated in Figure 7.2.

7.1.1 Technology Stack

Table 7.1: Technology Stack used in System Implementation

Component	Technologies Used
ML	Py, PyTorch, Ultralytics YOLOv12, OpenCV
Backend Server	Node.js, Express.js
UI	HTML5, Tailwind CSS, JavaScript, SweetAlert2
Model Assets	.pt file and .pkl file

The deployment of both dental diagnostic systems, as summarized in Table 7.1, is based on a three-layered architecture that facilitates the merging of ML, backend processing, and a browser-based front-end UI (User Interface) together. The ML layer consists of using Python (Py), PyTorch, Ultralytics YOLOv12, and the OpenCV library, which supports a very fast inference for both OPG radiographs and RGB intraoral images.

In contrast, the backend layer is constructed with Node.js and the Express.js Framework, which are held responsible for carrying out the Hypertext Transfer Protocol (HTTP) requests from users, taking care of image uploads, and working with the Py inference modules. The front end is built with HTML5, Tailwind CSS, JavaScript, and SweetAlert2 libraries to create a neat, adaptive, and user-friendly interface. The models in use are relying on the files of the YOLO model in ".pt" extension, where the weights are together with a corresponding "mapping.pkl" file that guarantees the correct class interpretation for each dataset and system variant.

7.1.2 Technical Considerations and Challenges

During development, several technical challenges influenced the system design and implementation:

(a) Py-Node.js Interoperability

The backend is using Node.js, while model inference is done in Python. The inter-process messaging was thus required for communication between the two frameworks. This was done by using a child-process pipeline in which Node.js calls the Python script non-blockingly and waits for the structured JSON output.

(b) Model Loading and Latency

The loading of the YOLO model was done for each request, which caused a delay that was not acceptable. The problem was solved by allowing the Py inference script to load the model just once and keep it in memory so that it would be ready for subsequent detections very quickly.

(c) Handling High-Resolution OPG Images

Working with OPGs is certainly a challenging task because of their high resolution and various aspect ratios. However, to verify that the inference system's performance was consistent and not influenced by this, preprocessing techniques such as adaptive resizing and contrast normalization were utilized.

(d) Dataset-Specific Class Interpretation

The label sets vary from one system to another. The backend at runtime brings into play the right `class_mapping.pkl` file in order to prevent the two systems from misunderstanding the outputs of the detection process.

(e) Frontend Visualization

It is common for OPG pictures to have a lot of overlapping detections. The existing situation has not changed regarding this issue, and there is a risk that overlapping predictions will visually gather together in places where there are a lot of anatomical features. Even though the interface shows all the detections made by the model along with the respective labels and confidence percentages, no other post-processing or visualization improvement (like personalized suppression or bounding box merging) has been included so far. This drawback has been recorded for improvement in the future.

7.2 OPG-Based Dental Decision Support System

The first system focuses on assisting dental professionals by analyzing OPG radiographs. It is for clinical interpretability and high-resolution imaging environments. The UI snippets of the system are visualized in Figure 7.3.

7.2.1 System Overview

- The user, typically a dentist, uploads an OPG radiograph through the web interface.
- The fine-tuned YOLO12s-OPG model detects structures and pathologies such as fully healthy teeth, impacted teeth, BDC/BDR, fractured teeth, dental caries regions, and infection regions.
- The system then returns an annotated radiograph with bounding boxes and confidence levels of the predicted diseases.

7.2.2 Functional Objectives

- To provide rapid, automated screening of dental anomalies in OPG images.
- To support dentists by reducing manual review time.
- To improve diagnostic coverage, especially in settings with resource constraints.
- To offer a reproducible tool that can aid preliminary assessments.

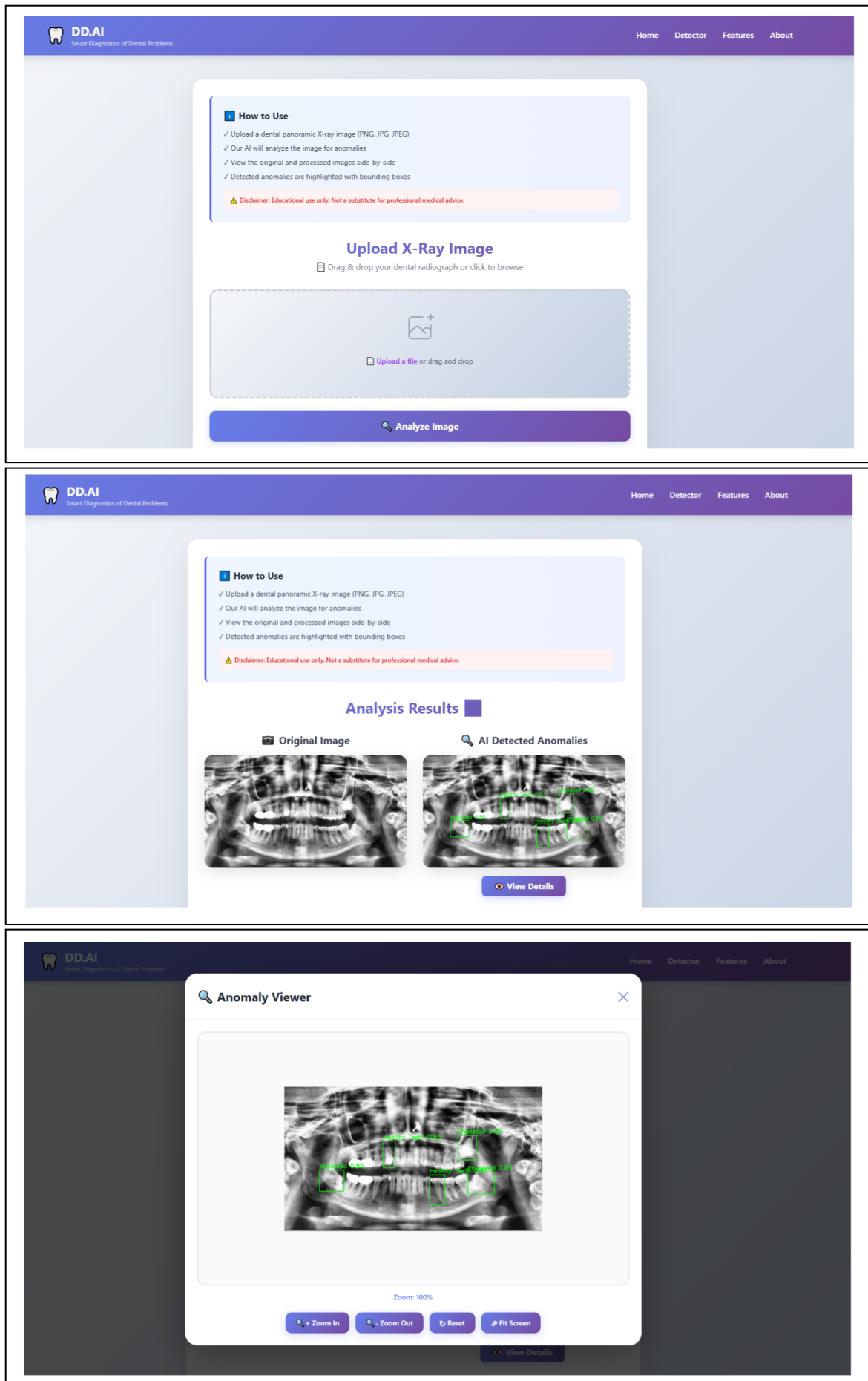


Figure 7.3: UI Snippets of the implemented system in web browser view

7.3 RGB Teeth Health Monitoring System

The second system is designed for the general population with an aim to promote self-screening and early detection of dental caries using RGB photographs captured via mobile devices.

7.3.1 System Overview

- The user uploads a close-range RGB image of their teeth.
- The YOLO12s-RGB model analyzes the image and detects instances of dental caries in their mouth.
- The processed result highlights detected regions with bounding boxes and confidence scores if it predicted any presence of cavity in the image.

7.3.2 Functional Objectives

- To promote dental hygiene and awareness among the general public.
- To provide a low-cost screening tool for remote or underserved areas.
- To serve as a pre-diagnostic aid that encourages timely dental consultations.

7.4 Integration and Deployment Environment

Both systems currently run on a local development environment using the following configuration:

- Static frontend delivered through HTML and Tailwind CSS.
- Node.js backend for handling HTTP requests, routing, and image upload management.
- A Py inference service running in a controlled virtual environment.
- Inter-process communication via asynchronous Node.js child processes.
- Browser access via localhost for testing and demonstration.

The backend service and the Py inference module will be hosted on a GPU-based cloud platform in the future deployment after optimizing the ML modules, which might be Google Cloud, HuggingFace Spaces, Railway, or Render, thus guaranteeing scalability and real-time performance.

8 Discussion of Key Findings

8.1 Classification Experiments Summary

The classification tests performed on the OPG dataset and the RGB Teeth dataset bring to light different model performances, which are mainly influenced by the data’s nature and the clinical characteristics underlying them. The RGB dataset, which is comprised of close-range images of individual teeth, is conducive to single-label classification and offers more apparent visual cues such as color, texture, and lesion boundaries. These attributes were a factor in the increase of accuracy and consistency in the predictions made by the model since the area of interest was localized and unaffected by the overlapping anatomical structures to a greater extent. On the other hand, the OPG dataset is complex by nature due to its panoramic view of the entire dental architecture, where multiple teeth and surrounding structures can be seen at once in one frame. The complexity makes it even more difficult for the convolutional networks to spot and categorize the disease-specific patterns.

When it comes to the OPG dataset, multilabel diagnostic purposes were what the original design aimed at, which brought conceptual inconsistencies and caused the models to simplify the intricate multi-condition cases. The dataset faced another issue of severe class imbalance, where some disease categories were represented to a very small extent. The imbalance had a drastic effect on precision and recall for minority classes since the models often predicted based on the majority classes. The structural differences among classes were still not distinct enough for deep networks to learn strong discriminative representations.

In general, the classification models yielded a higher performance on the RGB Teeth dataset consistently compared to the OPG dataset. The visual features were specific to the disease, the boundaries of the lesions were clearly marked, and the background noise was low in the RGB data. These findings indicate that single-focus clinical imaging, especially clean intraoral captures, is more suitable for automated classification tasks than broad panoramic radiographs by nature, but there is a limitation for training a model with single-label images in real-world deployment. The experiments have pointed out the significant impact of dataset design, label structure, and class distribution on the classification results, and this has also highlighted the necessity of task-appropriate imaging modalities for deep learning systems. Therefore, we need a broad amount of single-label images along with a small amount of multilabel images for each class in each class of the dataset, along with dealing with the inclusion of the healthy image class as a baseline.

8.2 Object Detection Experiments Summary

The object detection experiments also show a similar performance gap between the OPG and the RGB intraoral datasets, the latter being more stable and clinically meaningful

across all YOLO model families. Intraoral RGB imaging detects dental caries lesions at close range, which allows the YOLO architectures to detect small dental caries with very high precision and recall. The bounding boxes created across RGB images were mostly very close to the ground-truth annotations, whereas the high mAP@50 values (above 0.91 for all models) show that the models successfully localized lesions even under changes in lighting or tooth orientation. The clearer visual boundaries, reduced occlusion, and tighter framing of the region of interest were the major contributors to this improved detection performance.

On the other hand, the OPG dataset displayed few difficulties in object detection because of the panoramic nature of the radiographs used. Small or faint lesions, especially those hidden behind the overlapping anatomical structures, were hard for the models to detect. The wide view and the presence of several teeth in one picture caused highly complicated backgrounds, which made the feature extraction and bounding-box regression tasks considerably harder. Moreover, the OPG annotations frequently covered only very small lesion regions compared to the overall image size, which again impeded detection accuracy because of the difficulty of regressing exact bounding boxes on low-resolution patterns.

The performance of the RGB dataset was consistently superior to that of the OPG dataset across the YOLO models, not only regarding mAP values but also in the overall stability of PR curves and confidence thresholds. The newer YOLO variants (specifically YOLOv11 and YOLOv12) showed better ability with small-object detection, but even with such architectural advancements, the OPG dataset still proved to be significantly challenging. These results support the assertion that intraoral imaging, which confines the affected tooth, delivers more dependable and clinically actionable information for the automated object detection pipeline than panoramic radiography does.

8.3 Interpretation of Performance Trends

Analysis of performance trends in both classification and object detection experiments has signified a strong correlation between the input data’s structural quality and the model’s effectiveness. Regardless of the domain in the anatomical region of interest, the models have shown a clearly better performance with minimum background complexity.

The RGB images are the ones that have this relationship reflected in the results because these images have high-contrast features, color variations that are specific to certain diseases, and lesion boundaries that are well-defined. On the contrary, the image qualities of OPG, which are of a multistructural and multi-tooth nature, worsened the visibility of pathological features, thus making it more difficult for deep learning models to pick up and filter out relevant patterns.

Furthermore, a significant trend noted is the shrinking gap between the Nano and Small versions of the latest YOLO models. The YOLOv12n model, although designed to be lightweight, has performance that is often very near to or even higher than that of the Small variants. This is indicative that through the development of mechanisms like attention, feature pyramids, and regression heads in the architecture, the lightweight models have become less disadvantaged. In the same way, the classification models experienced a rise in performance when the training was done on data with balanced representation and clear visual features, thus confirming the deep learning systems’ sensitivity to the quality of the dataset.

The mAP@50-95 trends give an important insight: coarse detection is successful through datasets, but fine-grained localization is still challenging, and that is especially

the case with radiographic data. The sharp decline in mAP@50-95 for OPG images re-confirms that higher IoU thresholds make it even more difficult to fit the bounding boxes accurately in the subtle radiolucent areas. This set of trends is an indication that architecture is not the only limiting factor when it comes to the performance of models, but also that dataset design, annotation accuracy, and imaging modality characteristics have a strong impact.

8.4 Impact of Data Quality and Preprocessing

The data quality and preprocessing were decisive factors in all the classification and detection experiments' outcomes. The RGB images, which were of high quality, had very little preprocessing done to them; only normalization and resizing were performed, and the lesions were so clear that the models could easily get the features they needed with very little disruption. The lack of overlapping structures, better lighting conditions, and consistency in the framing were also some of the factors that led to the high performance of the models.

Nevertheless, the preprocessing pipeline for OPG images failed to counteract the basic insufficiencies of the data. Problems like variable brightness, uneven exposure levels, and the existence of overlapping anatomical structures were among the factors that greatly reduced the effectiveness of the models. Even after performing contrast enhancement, cropping, and noise-reduction techniques, the lesions remained small compared to the image size and often were visually ambiguous.

Moreover, the issue of drastic class imbalance was another factor that negatively affected both tasks, especially in the OPG dataset. Although oversampling and enhancement were helpful, they could not completely bridge the gap caused by the rarity of certain lesion categories. These results point out that preprocessing by itself cannot solve the problem of poor dataset design and that high-quality and specific to the task data acquisition is a must for achieving top performance.

8.5 Real-World Applicability of the Models

The results of this study give a very promising perspective on the corresponding use of deep learning models in dental screening workflows, especially for the YOLO-based detectors. The performance achieved with RGB intraoral images indicates that modern AI systems are capable of producing accurate and high-resolution images, which in turn makes them fit for a variety of application areas, such as chairside decision support, mobile dental health, and early screening tools in resource-poor regions.

Furthermore, lightweight YOLO models, for e.g., YOLOv12n or YOLOv12s, are the most appropriate for smartphone-based systems because of their incredible efficiency and outstanding accuracy.

Nonetheless, the findings also point to major restrictions regarding the use of these models for radiographic imaging, particularly OPG scans. The performance drop on panoramic radiographs signifies that, without a redesign of the dataset and an enhancement in the quality of annotations, the AI-based diagnostic analysis of OPG images might still be considered untrustworthy for detailed tasks like lesion detection or classification. To this end, the future of radiographic applications should aim at gathering OPG datasets that are high in resolution, balanced, and clinically validated, and that have multilabel annotations that truly capture the intricate clinical scenarios.

In spite of these difficulties, the research shows that AI models can be trusted to perform accurately, given that they are provided with intraoral data of the highest quality. Besides, the performance trends are very much a case for the incorporation of deep learning into preventive dentistry. The skills shown in lesion localization and classification imply that such models could provide substantial assistance to clinicians by making the early detection process easier, less variable diagnoses, and through mobile or tele-dentistry platforms, allowing more patients to attend, and thus, the outreach would be improved.

8.6 Addressing the Research Questions

The experimental results taken together provide a clear picture of the behavior of the models in the two imaging modalities.

(RQ1): The YOLO-based detection pipeline not only localized but also created a pretty clear picture of where the tooth decay, impacted teeth, and broken-down crown/root were. The performance of YOLO12s was the most reliable, achieving good scores in both precision and recall for the aforementioned classes.

(RQ2): The contrary was true for CNN-based classification of the OPG dataset, which was clearly inferior to the detection models because of the limitations of the dataset, such as the unverified labeling, small sample sizes, and different model architectures for classification and detection; thus, its clinical use is still dependent on the availability of larger datasets with professional-level annotation.

(RQ3): For the RGB intraoral images, the CNN-based models were able to set apart most dental conditions with high accuracy, and the preprocessing steps, such as CLAHE, were increasing the generalization, even though the subtle dental caries cases were still considered a challenge, which indicated that future work should focus on better attention mechanisms.

(RQ4): Grad-CAM's contribution in making both RGB and OPG classification tasks more interpretable was through marking the areas connected with clinically important structures, proving that the models generally were mostly looking at the right areas for diagnosis.

(RQ5): Altogether, this paved the way for the two experimental pipelines to be translated into two real-world systems: a dentist's clinical support system based on OPG, and an RGB-based dental caries-screening public health monitoring tool, each one tailored to its user group and imaging modality.

(RQ6): The comparative analysis showed clear performance variation and distinct patterns between two modalities: The first modality, using RGB images, revealed reflective surface-level defects which were easily classified and thus a more stable classification was achieved, while the second modality, using OPG radiographs, showed deeper and overlapping anatomical structures, thereby making the whole process of detecting more robust than classifying. The different aspects pointed out above strengthen the idea of the two modalities being complementary for a thorough dental evaluation.

9 Threats to Validity

9.1 Internal Validity

9.1.1 Dataset Bias

The datasets employed in this research are characterized by several intrinsic sources of bias that could potentially impact the accuracy of the outcome.

- The OPG dataset is deemed small as it has been composed of only 231 images spread over six classes, and its multi-label structure makes it even more complicated since several conditions can occur together in one radiograph. The small size of the dataset limits the variety of visual patterns that can be used for the training of the model and augments the likelihood of overfitting, especially when the subtle radiographic abnormalities are being dealt with.
- The RGB Teeth dataset is of medium size and gives a lesser degree of imbalance than the OPG dataset; however, it still does not have enough variation in terms of lighting, angles, and demographics. These restrictions weaken the extracted features and limit CNN models to learning less general representations.
- Although the RGB dataset outperformed the others in general, the lack of intra-class variability and the uneven photographic quality were still factors that limited the model’s performance, thus revealing a certain extent of dataset-driven bias in both classification pipelines.

9.1.2 Impact of Preprocessing

The preprocessing choices have also influenced the internal validity of the different datasets which include various factors.

- In the case of the OPG object detection task, we depended on the augmented dataset that accompanied the benchmark, which contained various transformations like rotations, flips, shearing, exposure adjustments, and mosaic transformations. Still, we were not privy to the complete augmentation pipeline or its validation criteria, so it is possible that some augmented samples do not correspond to real diagnostic variations, thereby introducing a potential bias to the model’s learning.
- In the OPG classification pipeline, padding was applied to maintain the aspect ratio before resizing images to 224×224, thus ensuring that structural proportions were preserved without distortion.
- In contrast, the RGB Teeth dataset was not padded or standardized in the same manner since the images varied greatly in size, aspect ratio, and resolution, which

might have caused inconsistencies in the spatial representations that were input into different CNN models.

- The preprocessing inconsistencies that existed among the datasets opened up variability in feature learning and thus may have played a part in the performance outcomes of the comparison being influenced.

9.1.3 Data Splitting and Leakage

- The data splitting process was a major methodological problem for the OPG multi-label classification task because of the very irregular nature of the label combinations. The same patient might have had inconsistent patterns of co-occurring labels in different images, while a few samples might even have included the Healthy label along with disease labels, which is not logically invalid in real world. In order to eliminate such inconsistencies from happening between training and testing partitions, the whole dataset was subjected to a deterministic splitting procedure.
- The contradictory labels were corrected by removing the Healthy labels wherever a disease was present. Next, images were assembled by a signature, which was defined as the sorted list of their cleaned labels, thus assuring that identical combinations remained united. These signature groups were assigned deterministically among training, validation, and testing, with extremely rare signatures (the ones appearing only once) being allotted to a special 'test-rare' group to prevent data leakage. This tactic was essential as a normal random split would have distributed unique or contradictory combinations across partitions, causing leakage and thus wrong performance estimates.
- The data merging in the RGB Teeth dataset for dental caries class, in further investigation, might have introduced bias for dental caries class, leading to better results later on, conducted in the NASNet-Mobile architecture.
- For OPG object detection, the dataset had 604 images after augmentation; however, there is a limitation since the test set does not have any examples of the fractured teeth class, thus the overall mAP@50 and mAP@95 scores may likely underestimate model performance for this specific category.
- The initial labels in the RGB Intraoral Cavity dataset consisted of Permanent and Temporary Tooth Decay classes, but the distribution was highly imbalanced and therefore not appropriate for training that would yield reliable results. To ensure that the analysis was consistent with the goals of the research detection of cavities the two categories were combined to form one class labeled Cavity. On the contrary, this decision improves stability, it removes granularity, and may limit downstream interpretability of decay types.

9.2 External Validity

9.2.1 Generalization to Real-World Settings

- The promising results from the experiments made on both classification and detection tasks do not guarantee that these results could be transferred to real clinical or real-world environments for all cases.

- The variability in different aspects of real-world dental imaging is huge; it goes to different patient positioning along with the anatomical differences, occlusions, and then it combines with staining, soft-tissue interference, motion blur, device heterogeneity, and practitioner skill.
- The workflow of clinical conditions will require the model to manage unseen pathologies, mixed conditions, and non-standard cases, which the datasets used in this study do not fully represent.
- Consequently, the models might be working very well in controlled experimental conditions but still need domain adaptation, more diverse training data, and further clinical validation before they can be used in operational dental settings.

9.2.2 Device and Lighting Variations

Both the classification and detection datasets have constraints that are specific to devices, which lead to limitations in generalization. All the dental images in RGB format were taken with smartphones, and the lighting conditions of the samples differed greatly because of the varying use of flash, the quality of the ambient light, and the cameras. Such effects may drive models to recognize not anatomical features but lighting-dependent ones instead. The OPG radiographs were first displayed on computer monitors and then captured by phones, which made them come in artifacts like glare, shadowing, and contrast inconsistencies that do not represent the actually obtained raw DICOM radiographs. These differences imply that the models developed might not yield the same results in dental clinics with standardized imaging systems or even in those with advanced radiographic sensors; therefore, the external validity is reduced.

9.2.3 Population Diversity

Population diversity is a factor that may also negatively impact the validity of the study results.

- The OPG dataset, was obtained solely from Bangladesh, a country with a very uniform population regarding dental morphology, bone density patterns, and disease prevalence.
- The RGB Teeth dataset did not check the population and could be from an unknown area.
- The RGB Intraoral Cavity dataset was collected in Pakistan.
- The dental anatomy, oral hygiene practices, diet, and socio-economic factors differ a lot between regions, it is probable that the models created with these datasets will not be able to cover the whole range of populations with different dental characteristics.

Therefore, the limited diversity present in the datasets not only restricts the trained models generalization but also highlights the necessity for more inclusive and geographically diverse datasets.

9.3 Theoretical Validity

9.3.1 Appropriateness of Metrics

The evaluation metrics applied in this research study were selected to mirror the precise goals and features of each task; nevertheless, a critical analysis of their suitability is still required.

In the case of the classification experiments, macro-precision, macro-recall, macro-F1, and accuracy were selected due to the datasets' major class imbalances and uneven sample distributions. The macro-averaged metrics were able to equalize the significance given to the minority classes like Caries and Mouth Ulcer, and thus made sure that the model's performance was not overly influenced by the majority classes.

For the multilabel OPG classification, the datasets contradictory label combinations validated the application of macro-averaged scores and multilabel-compatible evaluation logic; however, the unavailability of clinically verified multilabel distributions limits the absolute values of these metrics in terms of interpretation.

The object detection aspect of the study employed Precision, Recall, mAP@50, and mAP@50-95, which are recognized metrics in the YOLO community and demonstrate both localization and classification performance.

However, the omission of particular classes from the test set may falsely inflate or deflate the overall mAP values. The metrics selected, in general, are normative and methodologically sound, but certain restrictions imposed by the datasets limit the extent to which the numerical results can be generalized or interpreted with strength.

9.3.2 Label Definitions and Class Boundaries

The study's reliability is also tied to how closely the labels and class definitions reflect the clinical phenomena.

- In the OPG multilabel dataset, the class distribution were sometimes hard to determine since there were overlapping diseases and diverse annotation regulations. Some samples contained logically impossible combinations, for instance, the label "Healthy Teeth" was used together with disease labels. Inconsistencies like these were resolved by applying preprocessing steps, yet they still pointed to the problems with the dataset's labels.
- Elements like Tooth Discoloration and Calculus in the RGB Teeth dataset were visually defined, but oftentimes, visual overlap between stains, lighting artifacts, and noise made it difficult to clearly differentiate these classes.
- Likewise, the RGB Intraoral Cavity dataset, at first, had categories for Permanent and Temporary Tooth Decay, but because of the extreme class imbalance and vague boundary definitions, these two categories were combined into one Cavity class. Although this makes experiments more stable, it also neglects real-world differences in dental pathology. Consequently, certain model outputs might be based on dataset-driven label definitions instead of on clinically rigorous categories.

9.3.3 Clinical Alignment of Outputs

Even though the models performed very well within the datasets' ranges, the extent of their outputs being able to cater to the real clinical diagnostic needs is still the main issue.

Only the appearances of the surfaces are the basis of the CNN predictions for the RGB images, and clinical factors such as patient history, symptoms, or co-occurring conditions, which dentists always consider, are not included. Besides, missing the healthy class and multilabels in the dataset limits it from real word deployment.

Models in OPG analysis detect radiographic abnormalities, but the resolution of the diagnostic categories is not as fine as in real clinical practice for instance, several radiographic abnormalities may arise in radiographic findings can come from the same anatomical area. The detection tasks create bounding boxes that show the general area of the findings, yet they do not indicate the levels of severity or the treatment implications.

Therefore, the models are still, to a certain extent, aligned with the clinical setting through their detection of visually identifiable clues; however, this alignment is curtailed by the use of simplified labels, the lack of integration of domain knowledge, and the absence of multi-level diagnostic granularity. Before concluding the application of the models in real-world clinical settings, one has to acknowledge these limitations.

9.4 Statistical Conclusion Validity

9.4.1 Statistical Robustness

The performance metrics lead to the conclusion that one should consider the statistical robustness of the situation. Several things can justify the variance in statistics; one of them is the small size of the datasets, especially the OPG dataset, which consists of only 231 samples. Model predictions are likely to fluctuate, and the macro-averaged scores would be meaningfully affected by those small fluctuations. Moreover, the lack of several independent test sets, cross-validation folds, or bootstrapping diminishes the ability to quantify the confidence intervals or statistical significance across the models.

Although the performance differences between the deep models (e.g., ResNet101 beating ResNet50) seem stable and large, the lack of formal statistical testing like paired t-tests or non-parametric comparisons implies that the conclusions are drawn from trends that were observed rather than differences that were verified statistically. The conclusions are therefore based on evidence that is strong but also requires caution in terms of the inferential claims that are made.

9.4.2 Model Comparisons

The inclusion of suitable ground signifies one of the major factors of conclusion validity. The lightweight architectures MobileNetV2 and NASNet-Mobile in this work were used as a ground for comparing the deeper and more expressive models of ResNet101, DenseNet121, and InceptionResNetV2. The same was done with the two variants of YOLO Nano (N) and Small (S) that acted as ground for the newer generations of YOLO architectures across v8 to v12. These ground function as points of reference that are meaningful and thus allow for easy comparison between the advantages of having more depth, the availability of richer feature representations, and the use of improved detection heads. On the other hand, the lack of non-deep-learning grounds, such as classical machine learning or handcrafted feature approaches, hampers the assessment of deep models' superiority over traditional pipelines in terms of the added value they bring. However, the ground chosen within the scope of modern CNN and YOLO architectures is sufficient contrast to uphold the study's conclusions.

9.4.3 Stability Across Runs

The models were trained just once for each of the different settings because of the limitations related to the computational power; therefore, the findings would not be able to fully indicate the stability of the training process and the variance across the runs. Stochastic models in deep learning are generally affected by weight initialization, shuffling of data, and tiny stochastic variations in the optimizer, especially when they are fine-tuning small datasets. The stability cannot be specified exactly, nor can the effect of different performance across the trials be measured if no repeated training runs or variance reporting are done.

The YOLO object detection models tend to be more stable due to the larger size of their training set instances and the use of standard pipelines. However, this could not be claimed for the OPG multilabel classification, where the small dataset size and irregular label combinations cause instability. Nevertheless, the trends seen across models were consistent, led by architecture expectations, and thus, the overall conclusions were credible, although more reliable future work should involve repeated runs.

9.5 System Limitations

The RGB dental classification pipeline could not be integrated into the web system due to the dataset lacking a "healthy" class, and it was not suitable for multilabel scenarios. As a result, real-world inference was affected, particularly for scenarios in which users upload images of healthy teeth. The same issue was observed in the multi-label OPG classification pipeline, which suffered from severe class imbalance, contained very few healthy samples with fully visible and properly aligned teeth, and exhibited inconsistencies in dataset design factors that collectively prevented stable, clinical-grade deployment.

Thus, only the pipelines of YOLO12s detection were implemented: (1) an intraoral caries detection system for public health monitoring and (2) an OPG-based detection interface to support dentists.

These detection systems, however, also cover only a few dental conditions, do not have a diverse population, and have not been tested in the real world for validation or usability. Both systems, though, are currently running in a local server through two different web applications, yet large-scale deployment is still a challenge. The future work will primarily be on the generation of large, balanced, and multimodal datasets along with integration of these pipelines into one comprehensive m-Health application for oral-health monitoring and clinical decision support, which will then be subjected to thorough testing and the planning of deployment.

10 Conclusion and Future Work

Our framework developed and evaluated a YOLO-based detection model and multiple CNN-based classification models, supported by systematic ablation studies and Grad-CAM explanations. Through these investigations, the work successfully identified the best-performing models for each domain and proposed two practical solutions: a clinical decision-support system based on OPG analysis, and a consumer-oriented screening tool based on RGB images. In terms of object detection, the YOLO12 variants constantly gave the highest performance, however for RGB and OPG image classification, ResNet models and NASNet-Mobile remained the competitors; yet, no definitive statements about the results could be granted because of the dissimilarities between single vs multi-label and the exclusion of the Healthy class in a wrong manner from datasets.

The results demonstrate that DL has strong potential to support both professional diagnostic settings and general public self-assessment. The OPG detection and classification pipeline achieved reliable localization and disease recognition, indicating its usefulness in clinical workflows in real life. The RGB system worked well for spotting common tooth problems, so it can help people who don't often visit the dentist initially. Grad-CAM also showed that the model was looking at the right parts of the teeth, which makes its predictions more trustworthy.

There are still some limits. The models were trained on small datasets and have not been checked by dental experts or tested with everyday users. They need this testing before they can be used in real clinics or public settings. Also, the data did not include all image types or diverse groups, which may affect how well the models work for everyone. To address these limits, our future goal is to build a much larger and better dental image dataset that shows real clinical and everyday conditions. The full framework will be tested on real clinic images and real user inputs to check how strong, easy to use, and practical it is. Dental experts will help give professional feedback, and user studies will see how regular people use the RGB screening tool.

This study is very important for South Asian developing countries like Bangladesh, where access to dental care is limited due to economic constraints, high population density, and shortages of dental specialists. In such environments, AI-based solutions can play a meaningful role in early detection, reducing the burden on healthcare systems and offering low-cost preliminary assessment tools for underserved communities. The frameworks in this work are an important step toward creating easy, scalable, and tech-based tools to support dental health.

Although more testing and a larger dataset are needed, this study's results suggest that the framework can help both dental professionals and the general public when used responsibly. The proposed solutions offer a strong base for future real-world use and can play an important role in improving dental health, mainly in areas with limited resources. With continued work and development, these tools can help create better access to care and support healthier communities.

Bibliography

- [1] A. Garg, J. Lu, and A. Maji, "Towards earlier detection of oral diseases on smart-phones using oral and dental rgb images," *arXiv preprint arXiv:2308.15705*, 2023.
- [2] D. Ali and H. Sadeeq, "An interpretable deep learning framework for multi-class dental disease classification from intraoral rgb images," *Statistics, Optimization & Information Computing*, 2025.
- [3] S. A. Sony, F. Haseen, S. S. Islam, and I. Jahan, "Prevalence and assessment of experience of dental caries among school going adolescents in a rural area of sylhet, bangladesh," *Journal Name*, 2021.
- [4] M. N. Islam, M. S. I. Pial, M. A. A. Noman, F. K. Madilo, M. M. Hasan, S. Munmun, T. Hassan, C. S. N. A. Mensah, and N. Roy, "Oral health knowledge, attitudes, and practices among university students in bangladesh: a cross-sectional study," *Discover Public Health*, vol. 22, no. 1, p. 242, 2025.
- [5] R. Archana and P. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, 2024.
- [6] R. Bouali, O. Mahboub, and M. Lazaar, "Review of dental diagnosis by deep learning models: trends, applications and challenges," *Procedia Computer Science*, vol. 231, pp. 221–228, 2024.
- [7] M. Saraei, M. Lalinia, and E.-J. Lee, "Deep learning-based medical object detection: A survey," *IEEE Access*, 2025.
- [8] S. Park, H. Erkinov, M. A. M. Hasan, S.-H. Nam, Y.-R. Kim, J. Shin, and W.-D. Chang, "Periodontal disease classification with color teeth images using convolutional neural networks," *Electronics*, vol. 12, no. 7, p. 1518, 2023.
- [9] 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, p. 25193, 2024.
- [10] Y. Li and J. Zhang, "multi-label dental image classification via vision transformer for orthopantomography x-ray images," *Computer-Aided Design and Applications*, pp. 198–207, 2024.
- [11] S. Araidy, G. Batshon, and R. Mirochnik, "Artificial intelligence applications in dentistry: A systematic review," *Oral*, vol. 5, no. 4, 2025.
- [12] S. Minoo, F. Ghasemi, *et al.*, "Automated teeth disease classification using deep learning models," *International Journal of Applied Data Science in Engineering and Health*, vol. 1, no. 2, pp. 23–31, 2024.

- [13] M. Kamran, M. Waseem, N. Shafi, and A. Zahid, "Detection of caries and hypodontia using deep learning and explainable ai," in *2024 18th International Conference on Open Source Systems and Technologies (ICOSST)*, pp. 1–6, IEEE, 2024.
- [14] J. Krois, T. Ekert, L. Meinhold, T. Golla, B. Kharbot, A. Wittemeier, C. Dörfer, and F. Schwendicke, "Deep learning for the radiographic detection of periodontal bone loss," *Scientific reports*, vol. 9, no. 1, p. 8495, 2019.
- [15] L. Lian, T. Zhu, F. Zhu, and H. Zhu, "Deep learning for caries detection and classification," *Diagnostics*, vol. 11, no. 9, p. 1672, 2021.
- [16] L. Wang, Y. Xu, W. Wang, and Y. Lu, "Application of machine learning in dentistry: insights, prospects and challenges," *Acta Odontologica Scandinavica*, vol. 84, p. 43345, 2025.
- [17] S. M. Faizan Ahmed, M. H. Ghorri, A. Khalid, A. Nooruddin, N. Adnan, A. Lal, and F. Umer, "Annotated intraoral image dataset for dental caries detection," *Scientific data*, vol. 12, no. 1, p. 1297, 2025.
- [18] A. Adnan, M. T. Rizwan, H. M. Attaullah, S. Basheer, and M. T. Quasim, "Deep architectural classification of dental pathologies using orthopantomogram imaging," *Computers, Materials and Continua*, vol. 85, no. 3, pp. 5073–5091, 2025.
- [19] A. HaghaniFar *et al.*, *Automated Teeth Extraction and Dental Caries Detection in Panoramic X-ray*. PhD thesis, University of Saskatchewan, 2022.
- [20] F. Campbell and H. Rogers, "Through the looking glass: a review of the literature surrounding reflective practice in dentistry," *British dental journal*, vol. 232, no. 10, pp. 729–734, 2022.
- [21] E. Efitli, A. A. Karcioglu, A. Ozdogan, F. Karatas, and T. Senocak, "Tooth color prediction in intraoral images under different clinical lights using ml algorithms and clahe technique: an in-vivo study," *Lasers in Medical Science*, vol. 40, no. 1, pp. 1–10, 2025.
- [22] M. Najeeb and S. Islam, "Artificial intelligence (ai) in restorative dentistry: current trends and future prospects," *BMC Oral Health*, vol. 25, no. 1, p. 592, 2025.
- [23] A. Fatima, I. Shafi, H. Afzal, I. Díez, D. Lourdes, J. Breñosa, J. Espinosa, and I. Ashraf, "Advancements in dentistry with artificial intelligence: current clinical applications and future perspectives. healthcare (basel) 2022; 10: 2188."
- [24] M. Jaiswal, M. Sharma, P. Khandnor, A. Goyal, R. Belokar, S. Harit, T. Sood, K. Goyal, P. Dua, D. P. Khandnor, *et al.*, "Deep learning models for classification of deciduous and permanent teeth from digital panoramic images," *Cureus*, vol. 15, no. 12, 2023.
- [25] Y. E. Almalki, A. I. Din, M. Ramzan, M. Irfan, K. M. Aamir, A. Almalki, S. Alotaibi, G. Alaglan, H. A. Alshamrani, and S. Rahman, "Deep learning models for classification of dental diseases using orthopantomography x-ray opg images," *Sensors*, vol. 22, no. 19, p. 7370, 2022.
- [26] I. D. S. Chen, C.-M. Yang, M.-J. Chen, M.-C. Chen, R.-M. Weng, and C.-H. Yeh, "Deep learning-based recognition of periodontitis and dental caries in dental x-ray images," *Bioengineering*, vol. 10, no. 8, p. 911, 2023.

- [27] A. AlSayyed, A. M. Taqateq, R. Al-Sayyed, D. Suleiman, S. Shukri, E. Alhenawi, and A. M. Albsheish, "Employing cnn ensemble models in classifying dental caries using oral photographs.," *International Journal of Data & Network Science*, vol. 7, no. 4, 2023.
- [28] Z. Hua, K. Aranganadin, C.-C. Yeh, X. Hai, C.-Y. Huang, T.-C. Leung, H.-Y. Hsu, Y.-C. Lan, and M.-C. Lin, "A benchmark review of yolo algorithm developments for object detection," *IEEE Access*, 2025.
- [29] X. Wei, L. Chanjuan, J. Ke, Y. Linyun, G. Jinxing, and W. Quanbing, "Convolutional neural network for oral cancer detection combined with improved tunicate swarm algorithm to detect oral cancer," *Scientific Reports*, vol. 14, no. 1, p. 28675, 2024.
- [30] Z. Can, S. Isik, and Y. Anagun, "Cvapool: using null-space of cnn weights for the tooth disease classification," *Neural Computing and Applications*, vol. 36, no. 26, pp. 16567–16579, 2024.
- [31] U. Sirisha, S. P. Praveen, P. N. Srinivasu, P. Barsocchi, and A. K. Bhoi, "Statistical analysis of design aspects of various yolo-based deep learning models for object detection," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 126, 2023.
- [32] A. Vijayakumar and S. Vairavasundaram, "Yolo-based object detection models: A review and its applications," *Multimedia Tools and Applications*, vol. 83, no. 35, pp. 83535–83574, 2024.
- [33] D. Palaniappan, R. Jain, T. Premavathi, K. Parmar, W. Ghribi, A. M. Ahmed, and N. Ahmad, "Yolo in healthcare: A comprehensive review of detection architectures, domain applications, and future innovations," *IEEE Access*, 2025.
- [34] S. B. Singh, A. Laishram, K. Thongam, and K. M. Singh, "Automatic detection and classification of dental anomalies and tooth types using transformer-based yolo with ga optimization," *IEEE Access*, 2025.
- [35] J. G. de Almeida, C. Messiou, S. J. Withey, C. Matos, D.-M. Koh, and N. Papanikolaou, "Medical machine learning operations: a framework to facilitate clinical ai development and deployment in radiology," *European Radiology*, pp. 1–14, 2025.
- [36] A. Ramírez-Pedraza, S. Salazar-Colores, C. Cardenas-Valle, J. Terven, J.-J. González-Barbosa, F.-J. Ornelas-Rodriguez, J.-B. Hurtado-Ramos, R. Ramirez-Pedraza, D.-M. Córdova-Esparza, and J.-A. Romero-González, "Deep learning in oral hygiene: Automated dental plaque detection via yolo frameworks and quantification using the oleary index," *Diagnostics*, vol. 15, no. 2, p. 231, 2025.
- [37] M. Razaghi, H. E. Komleh, F. Dehghani, and Z. Shahidi, "Innovative diagnosis of dental diseases using yolo v8 deep learning model," in *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, pp. 1–5, IEEE, 2024.
- [38] U. Şevik and O. Mutlu, "Detection of dental anomalies in digital panoramic images using yolo: A next generation approach based on single stage detection models," *Diagnostics*, vol. 15, no. 15, p. 1961, 2025.

- [39] Y. Ma, M. A. Al-Arooomi, Y. Zheng, W. Ren, P. Liu, Q. Wu, Y. Liang, and C. Jiang, "Application of mask r-cnn for automatic recognition of teeth and caries in cone-beam computerized tomography," *BMC Oral Health*, vol. 25, no. 1, pp. 1–14, 2025.
- [40] M. L. Ali and Z. Zhang, "The yolo framework: A comprehensive review of evolution, applications, and benchmarks in object detection," *Computers*, vol. 13, no. 12, p. 336, 2024.
- [41] Q. T. Lam, M. H. N. Le, I.-T. Lee, and N. Q. K. Le, "Evaluating yolo for dental caries diagnosis: a systematic review and meta-analysis," *Evidence-Based Dentistry*, pp. 1–14, 2025.
- [42] A. Davut and M. Burukanli, "Deep learning for dentistry: Yolo variants in tooth abnormality detection," *Computer Science Engineering*, 2025.
- [43] Y. Zhang, E. Hao, and X. Ding, "Development of ai-powered oral disease detection using yolo deep learning model," in *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, pp. 1–6, IEEE, 2025.
- [44] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is grad-cam explainable in medical images?," in *International Conference on Computer Vision and Image Processing*, pp. 124–135, Springer, 2023.
- [45] C. Chauhan and R. R. Gullapalli, "Ethics of ai in pathology: current paradigms and emerging issues," *Artificial Intelligence in Pathology*, pp. 159–180, 2025.
- [46] A. Guerra-Manzanares, L. J. L. Lopez, M. Maniatakos, and F. E. Shamout, "Privacy-preserving machine learning for healthcare: open challenges and future perspectives," in *International Workshop on Trustworthy Machine Learning for Healthcare*, pp. 25–40, Springer, 2023.
- [47] R. B. Rahman, S. A. Tanim, N. Alfaz, T. E. Shrestha, M. S. U. Miah, and M. Mridha, "A comprehensive dental dataset of six classes for deep learning based object detection study," *Data in Brief*, vol. 57, p. 110970, 2024.
- [48] P. Raj, "Teeth dataset." <https://www.kaggle.com/datasets/rajapriyanshu/teeth-dataset>, 2022. Accessed: 2025-07-20. [Online].
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [50] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- [51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
- [52] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

- [53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” *Proceedings of the AAAI conference on artificial intelligence*, pp. 4278–4284, 2017.
- [55] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
- [56] G. Jocher, A. Chaurasia, T. Qiu, and P. Kaur, “Yolov8: New vision ai models from ultralytics,” *Ultralytics*, 2023. Accessed: Nov. 2025.
- [57] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov9: Learning what you want to learn using programmable gradient information,” *arXiv preprint arXiv:2402.13616*, 2024.
- [58] T. Chen, H. Zhou, G. Wang, X. Zhang, and D. Lin, “Yolov10: Real-time end-to-end object detection without nms,” *arXiv preprint arXiv:2405.03060*, 2024.
- [59] Ultralytics, “Yolov11 release announcement,” 2024. Accessed: Nov. 2025.
- [60] Z. Li, Y. Xu, H. Wang, and M. Zhang, “Yolov12: Efficient and accurate object detection with attention-based architecture,” *arXiv preprint arXiv:2509.00012*, 2025.
- [61] Ultralytics, “Yolov8 documentation,” 2025. Accessed: Nov. 2025.
- [62] M. T. Islam, M. Alam, J. M. S.-U. I. Smaron, J. H. Setu, S. T. Bhuiyan, R. Rahman, A. Mahmud, S. B. Alam, and A. Islam, “Comparative evaluation of deep learning models for multi-disease classification using reflective dental imaging,” in *2025 IEEE 23rd Student Conference on Research and Development (SCoReD)*, (Kuala Lumpur, Malaysia), Nov. 2025. Presented at the conference; not yet available online.
- [63] L. K. Meng, L. J. Xin, H. H. Yi, Z. A. A. Salam, and N. B. Wei, “A machine learning approach for face mask detection system with adamw optimizer,” *J Appl Technol Innov*, vol. 7, no. 3, p. 25, 2023.
- [64] R. Wali, “Xtreme margin: A tunable loss function for binary classification problems,” *arXiv preprint arXiv:2211.00176*, 2022.
- [65] M. Hossin and M. N. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.
- [66] M. T. Islam, M. T. Hossain, S. A. Dip, J. H. Setu, N. Halder, A. Islam, and M. A. Amin, “Predictive modeling of infant birth weight with kolmogorov-arnold networks,” in *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pp. 2974–2979, IEEE, 2024.
- [67] R. Padilla, W. L. Passos, T. L. Dias, S. L. Netto, and E. A. Da Silva, “A comparative analysis of object detection metrics with a companion open-source toolkit,” *Electronics*, vol. 10, no. 3, p. 279, 2021.