JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 03, Issue No. 02, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0302-0048

OralVision: A Deep Learning Framework for Multi-Class Oral Disease Classification

Moiz Ud Din[1], Ibtisam Asghar[1], Ejlal Hameed[1], Saad Irfan Khan[1,*]

[1] Department of Robotics & Artificial Intelligence, SZABIST University, Islamabad, Pakistan.

*Corresponding author: saad.khan@szabist-isb.edu.pk

Contributing authors: 22108183@szabist-isb.pk, 22108178@szabist-isb.pk, 22108171@szabist-isb.pk, saad.khan@szabist-isb.edu.pk

Received: September 01, 2026 **Accepted:** September 28, 2026

Abstract- Oral diseases impact a large percentage of the world population but are typically diagnosed by manual visual examination by the dental health care provider, which may be scarce. This work is a deep learning framework, OralVision, that classifies oral disease into six classes (Calculus, Caries, Gingivitis, Hypodontia, Mouth Ulcer, and Tooth Discoloration) using an EfficientNet-B0 backbone network, factorized Coordinate Attention, and a linear classification head with a single layer. The model was trained on a corrected set of 8,430 source images collected from the Holy Family Hospital, Rawalpindi, Pakistan with 5900, 1265 and 1265 images used for training, validation and testing, respectively. Duplicate detection and label conflict auditing were performed to evaluate the quality of the data sets. The best test accuracy of 99.53% and macro-F1 of 98.41% were obtained at the primary checkpoint. Across five seeds, mean accuracy was 99.24% ± 0.21%, and mean macro-F1 was 97.68% ± 0.51%. It has 4,171,474 parameters, 16.20MB size and around 0.831 GFLOPs per forward pass. The ablation experiments demonstrated that the removal of Coordinate Attention, curriculum learning, or the hard-example cascade resulted in a higher macro-F1 compared to the Full Model, but did not reveal any empirical evidence of an improvement in the overall classification performance. After fine-tuning, macro-F1 scores were 0.93-0.97, and the Tooth Discoloration recall was 0.23-0.32, when evaluated on an external dataset of 1,057 images, which were not included in the training set.

Keywords- Oral disease classification, dental image analysis, attention mechanism, deep learning, explainable AI

1. Introduction

Oral diseases continue to be one of the most common non-communicable diseases of the world. In recent years, a Global Burden of Disease (GBD) analysis estimated that lip and oral cavity cancer represents a substantial and growing proportion of the total burden of disease from cancer, with differences across regions, sexes, and exposures to risk factors [1]. In addition to malignant diseases, many other common diseases occur, affecting large numbers of people and, when identified early, are known to have downstream treatment cost and disease progression implications such as dental caries, gingivitis, calculus deposition, hypodontia, oral ulceration, and tooth discoloration. Conventional diagnosis of these conditions has relied on manual, visual examination by a dental professional, which is open to inter-observer variation and limited by the geographical distribution of specialists available [2] and can result in an important diagnostic bottleneck in low-resource and remote areas.

To overcome this bottleneck, artificial intelligence, specifically convolutional neural networks (CNNs), has been suggested, achieving classification accuracy rates often surpassing 90% in many oral and dental imaging tasks and as part of the increasing interest in the use of generative AI in dentistry [3]. However, automated, image-based screening

is potentially useful because it can be carried out without the presence of a specialist at the time of image capture, and before formal referral, which could be made if needed.

Solutions to this challenge have been developed sequentially: from handcrafted-feature machine-learning processes to deep convolutional architectures, and, more recently, to transfer learning from large pretrained backbones, which are discussed in detail in Section 2. This has tended to increase accuracy in the reported results and decrease the amount of feature engineering, but hasn't addressed some common methodological drawbacks that have been observed in the literature.

The present study is motivated by four of these limitations.

- First, models are trained and tested almost always on a single dataset and a single train/test split, and cross-dataset or testing on externally sourced data is rarely reported as standard practice [4, 5].
- Second, data leakage between the training and evaluation partition may cause an overly optimistic evaluation of the model's internal performance, when the images are not the same or when they have finite variation in their attributes [6, 7].
- Third, oral disease taxonomies are often imbalanced, that is, the conditions that occur less often (e.g., tooth discoloration) are systematically underrepresented compared with the more common conditions (e.g., tooth decay), and mechanisms designed to correct imbalances are often presumed to operate as intended but not actually tested [8].
- Fourth, interpretability tools such as Grad-CAM are increasingly used to support trust in these models, but their outputs are qualitative and are frequently presented without the caveats needed to prevent over-interpretation [9, 10].

Only limitations directly supported by the literature reviewed in Section 2 are listed here.

Overall, this literature suggests that while the accuracy on an internal test set of an oral disease classifier might be near its maximum, the real-world value is not established: the behavior of a model on data not provided by the model developer is not known. The model's sensitivity to architectural and training decisions is unknown and the model's interpretability is not known. We do not know of any paper reporting a classification head that is minimal and statistically justified, as well as being factorized and position-preserving for the task of oral disease classification.

We address this gap with OralVision, which combines an EfficientNet-B0 backbone, factorized Coordinate Attention, and a single-layer linear classification head. EfficientNet-B0 was selected as a comparatively compact pretrained backbone, while Coordinate Attention was included to preserve height- and width-specific information in the feature map before global average pooling. These choices provide the architectural motivation for the model and their empirical contribution is assessed through the ablation study in Section 4.5 rather than assumed in advance. The classification head is kept deliberately simple, and its design is examined separately from the classifier-family comparison in Section 4.5.

This paper has the following contributions:

- An exact-duplicate and label-conflict-corrected dataset acquired in a real-world hospital setting
- A complete ablation study on the individual role of Coordinate Attention, curriculum learning, hard-example cascading, and classifier-head depth, with a full training budget.
- A post hoc comparison of families of classifiers, based on a statistically sound approach (bootstrap confidence intervals and paired McNemar's tests)
- A strict evaluation of the test set, using an external dataset different from the one used for training, and without any fine-tuning or parameter updates and
- A qualitative Grad-CAM interpretability analysis plus an explicit computational-complexity evaluation.

The rest of the paper describes related work (Section 2), the proposed model and methodology (Section 3), the experimental data, process, and results (Section 4), a discussion of the results and a comparison with the state of the art (Section 5), and concludes the research (Section 6).

2. Literature Review

2.1. Conventional Machine-Learning Approaches for Oral/Dental Classification

Machine-learning approaches predating widespread CNN adoption in this domain typically combined handcrafted or hybrid feature representations with classical classifiers. Malik et al. [11] combined a vanilla generative adversarial network (GAN) for data augmentation with a quantum machine-learning classifier for binary dental caries detection from bitewing radiographs, reporting 98.95% accuracy on the GAN-augmented dataset. The study demonstrates that hybrid classical/quantum processes are technically viable for dental imaging and that GAN-based augmentation can substantially expand a limited radiographic dataset. However, the task is binary rather than multi-class, the imaging modality is radiographic rather than photographic, and no external test set was used, limiting direct comparability to the six-class photographic taxonomy addressed in the present study. Salmanpour et al. [12] introduced PySERA, an open-source library combining handcrafted and deep radiomic feature extraction with an explicit emphasis on reproducibility. As a tooling contribution rather than a benchmarked classifier, it reports no accuracy figure directly comparable to a classification

task, but it shows that hybrid handcrafted/learned feature processes remain an active methodological direction in medical image analysis. And reproducibility-focused engineering is itself treated as a publishable contribution in this literature which this study adopts directly in its own reproducibility reporting (Section 3.3).

2.2. Deep Learning for Oral/Dental Classification

Deep convolutional networks now dominate oral and dental image classification. Eight recent studies address the same or a closely related six-class taxonomy or examine methodological choices relevant to the present work. Their datasets, tasks, and reported metrics differ, so their results are not all directly comparable. Ahmed [13] used an optimized, layer-pruned ResNet-18 (selected ReLU, batch-normalization, and convolutional layers removed from the 128-filter residual blocks) for the identical six-class oral disease taxonomy adopted here, reporting 90.94% accuracy with post-training quantization reducing model size by roughly 75% at no measurable accuracy cost. This shows that deployment-oriented compression is feasible for this task but leaves further evaluation of generalization and interpretability uninvestigated, as the authors did not test their model on an external test set and did not include a mechanism to attend. dos Santos et al. [14] fine-tuned a ResNet-50 for this task through a two-phase transfer learning protocol on a leakage-controlled, MD5-hash-deduplicated subset of 3,507 images of the same taxonomy, yielding 96.18% accuracy and 90.78% macro-F1 on a held-out set of 524 images. The authors' leakage-control protocol is a direct methodological precursor for the deduplication and group-aware splitting used in Section 3.2, but they did not validate accuracy on an external test set and attention was not mentioned. MultiDentNet [15] fuses four backbones (DenseNet-121, EfficientNetV2-S, ResNet-50, and Inception-V3) with the Squeeze-and-Excitation attention module, and used a graph-convolutional inter-class modeling, evaluated on 10,235 images across a five-condition taxonomy, with an accuracy of 99.70% and, importantly, a simulated domain-shift robustness test showing 8-12% accuracy drop with simulated brightness/contrast perturbation, though the perturbation is applied to the same set of source images and not tested on a genuinely independent test set. Oktiana et al. [16] found that changing the backbone has a significant effect on the performance of the model even if the classifier head and training protocol are kept the same, but analyzed only one dataset without cross-source testing and did not consider attention mechanisms and the complexity of the classifier head as separate design parameters. Marie et al. [17] proposed a latent-diffusion framework for augmenting pediatric dental images. The reported gains are relative to the baseline described in that study and do not provide an absolute accuracy result directly comparable with the six-class classification metrics considered here. The work is therefore discussed as an augmentation contribution rather than included in the quantitative benchmark.

Three additional studies employ the same six-class taxonomy (Calculus, Caries, Gingivitis, Hypodontia, Mouth Ulcer, Tooth Discoloration) as the present work. Inspired by this, Yang et al. [18] introduced Dynamic Compositional Prototype Learning (DyCoP) that decomposes oral pathologies into reusable category-level prototypes and localized visual feature prototypes (e.g., enamel demineralization), and dynamically combined them for sparse feature activation to achieve 93.30% accuracy and 91.90% macro-F1 on this exact taxonomy. DyCoP's prototype-matching approach is conceptually related to this study's emphasis on localized visual features, but its evaluation is internal and does not include external or cross-dataset testing. Ali and Sadeeq [19] report an EfficientNet-based model and Grad-CAM visualization to achieve the highest accuracy of 95.40%, and a Cohen's Kappa of 0.989, but they also use internal evaluation with only a single dataset, and class-imbalance correction is done only at the loss level and not at the combined sampler-and-loss level as explored in Section 4.5. George et al. [20] benchmarked modern Vision Transformer variants (ViT, DeiT, BEiT, Swin) against ConvNeXt on 12,653 images spanning the same six classes under real-world class imbalance, finding that ConvNeXt (81.06% accuracy) outperformed the transformer variants (79.73-80.00% accuracy) and that all architectures, transformer and convolutional alike, showed pronounced performance drop-offs on minority classes such as Caries and Tooth Discoloration. This result is directly relevant evidence that heavier, attention-heavy transformer backbones do not automatically resolve the class-imbalance problem this study addresses through its sampler-and-loss design (Section 3.1.5), and that substantially lower accuracy than EfficientNet-based approaches on this taxonomy is not confined to lightweight architectures specifically.

Beyond these eight directly comparable studies, related work addresses adjacent oral- and dental-imaging tasks. Asiri [21] evaluated a lightweight custom CNN (ExplainableDentalNet) and a Squeeze-and-Excitation-augmented ResNet-50 (Interpretable ResNet50-SE), both paired with Grad-CAM, for a narrower three-class enamel caries grading task, reporting up to 98.30% accuracy. The shared emphasis on lightweight backbones, attention, and Grad-CAM makes this study methodologically close to the present one, but its three-class caries-grading scope is narrower than the six-condition taxonomy addressed here. Michelin et al. [22] applied YOLOv8s object detection to 7,963 intraoral video frames extracted from handheld toothbrush-camera recordings, detecting findings including decay, plaque, and staining (conceptually overlapping this study's Caries, Calculus, and Tooth Discoloration classes) with an overall mAP@0.5 of 83.0%. This video-frame object-detection framing, and its explicit real-world handheld acquisition noise, is a useful point of contrast to the single-photograph classification protocol used here, illustrating that acquisition-condition variability is a concern documented elsewhere in this literature and not unique to the external-testing gap this study investigates (Section 4.7). Ren et al. [23] introduced Illumination-Reflection Aware Attention (IRAA) within a dual-branch network (DBAttNet) for caries classification across both intraoral photographs and bitewing radiographs, reporting 97.90% and 95.72% F1-score respectively and outperforming ResNet-50, Inception-V3, and other CNN baselines. IRAA's specific

design target specular illumination artifacts in intraoral photographs and is a complementary attention objective to Coordinate Attention's positional-preservation goal (Section 2.4). Its dual-modality (photographic and radiographic) evaluation is methodologically distinct from, though thematically related to, this study's single-modality, cross-dataset generalization focus.

Desai et al. [24], Liu and Bagi [25], and Mohapatra and Dash [26] each target oral cancer or oral potentially malignant disorders specifically, using deep-learning models on photographic or lip/tongue imagery together with Vissamsetti and Shivakanth's [27] general-purpose oral lesion classification framework, these illustrate that cancer-specific screening and general-purpose multi-condition classification remain largely separate research threads despite relying on similar CNN backbones. On the radiographic side, Pham and Al-Hebshi [28] and Ayhan et al. [29] classify pediatric dental disease and caries under fixed prostheses, respectively, from panoramic radiographs, reinforcing that radiographic and photographic oral imaging remain distinct sub-literatures with different preprocessing and evaluation conventions. A distinction the present study maintains by scoping itself to photographic classification only (Section 5.8). Both of these methods focused on data privacy from cross-institutional data, with Qureshi et al. [30] applying it to dental arch classification on panoramic radiographs, and Xu et al. [31] proposing the KC-UNet approach of integrating a Kolmogorov-Arnold Network with the CBAM attention for medical image segmentation (which is also close to the method used here, but in a different task, imaging domain, and with cross-institutional data). MedNet [32] is a lightweight attention-augmented CNN designed explicitly to be parameter-efficient, a design goal that is directly shared with OralVision, but tested on generic medical imaging benchmarks instead of a set of images from a specific oral or dental dataset.

Putting it all together: CNN-based approaches have significantly reduced reliance on handcrafted feature engineering, compared to the conventional approaches in Section 2.1, and transfer learning (Section 2.3) has given good results on relatively small datasets. Recent evidence from George et al. [20] further indicates that this pattern extends to Vision Transformer architectures, which do not resolve class imbalance simply by virtue of using self-attention throughout the network. However, near-ceiling internal accuracy recurs across almost every study surveyed, while cross-dataset or externally sourced testing, the more demanding evidence of generalization remains the exception rather than the rule even among the most recent, closest-taxonomy studies [18-20], and model complexity varies considerably without a corresponding standardization of evaluation protocol [13],[15].

2.3. Transfer Learning and Efficient CNN Backbones

Transfer learning from ImageNet-pretrained backbones is the de facto standard in this literature, used by dos Santos et al. [14], Oktiana et al. [16], and the present study alike. Tan and Le [33] introduced EfficientNet's compound-scaling principle, jointly scaling network depth, width, and resolution under a fixed computational budget. EfficientNet-B0, the smallest member of this family, achieves competitive ImageNet accuracy at a fraction of the parameter count of ResNet-family backbones (5.3M native parameters versus approximately 25.6M for ResNet-50), making it scientifically relevant to a study emphasizing parameter and FLOPs efficiency (Section 4.8). In the context of oral disease classification, Ahmadam's cross-dataset, federated, and explainable transfer-learning framework [34] makes an explicit argument that generalization across data sources, rather than single-dataset accuracy, should be the primary reported outcome of a classification study of transfer-learning, and the external testing protocol in Section 4.1.2 follows that argument. Ahmadam's own reported cross-dataset accuracy applies to a different clinical domain. Musa et al. [35] sought to tackle cross-population domain shift problems in chest radiograph classification, which this study empirically documents on a different cross-population domain shift, in a different imaging modality without the oral-imaging equivalent task. The transfer-learning-plus-lightweight-backbone approach outlined here was also widely used in other medical imaging applications with similar efficiency-accuracy dilemmas, as demonstrated by Bhuria et al. [36] in automated bone fracture classification using X-ray images.

2.4. Attention Mechanisms in Medical and Dental Image Classification

Attention mechanisms are now common additions to CNN classifiers on the premise that they help the network focus on diagnostically relevant regions rather than incidental background content. Hou et al. [37] introduced Coordinate Attention, which factorizes 2D global pooling into two 1D pooling operations along the height and width axes, preserving positional information that channel-only mechanisms such as Squeeze-and-Excitation collapse into a single descriptor. This factorization achieves the positional sensitivity of full 2D spatial attention at substantially lower computational cost, motivating its selection for the present, parameter-constrained model (Section 3.1.2). Wei et al. [38] applied Coordinate Attention combined with multi-scale feature fusion to PCB surface-defect detection, an industrial inspection task sharing with oral lesion classification the underlying need to localize small, spatially-precise regions of interest, and reported measurable detection improvements over non-attention baselines evidence that the mechanism's benefit is not domain-specific to natural images. Abdian et al. [15] embedded Squeeze-and-Excitation attention within a four-backbone ensemble (Section 2.2), improving on non-attention baselines but at a substantially larger parameter budget than the present single-backbone design. Ozdemir and Pacal [39] combined a ConvNeXtV2 backbone with focal self-attention for skin cancer detection, using focal-loss-style reweighting as an alternative class-imbalance correction mechanism to the sampler-based approach used here (Section 3.1.5). The study reports a strong accuracy figure on a dermatological

dataset but does not test whether the imbalance-correction mechanism specifically benefits the most severely imbalanced class, the check performed directly in Section 4.5 of the present study. Xu et al.'s [31] CBAM-augmented KC-UNet (Section 2.2) and Ferdous et al.'s [32] attention-augmented MedNet further illustrate that channel-and-spatial attention variants are an active design space across medical imaging tasks. Synthesizing across these studies: attention mechanisms consistently report a measurable, though architecture- and task-dependent, improvement over non-attention baselines, and Coordinate Attention's specific factorization preserving rather than discarding positional information which is a scientifically motivated choice for oral lesion classification, where lesion location (near the gumline, along the tooth surface, at the inter-dental gap) is itself diagnostically informative. However, no prior study identified in this review evaluates Coordinate Attention specifically for oral disease classification, a gap this study addresses directly.

2.5. Data Reliability, Leakage, and Medical-AI Evaluation

Trusting any accuracy value quoted in this literature is dependent upon rigorous experimental design. Near-duplicate or derived images can be strategically placed in the training and evaluation partitions and "fool " the machine learning system into claiming high accuracy, potentially without being noticed, as pointed out by Apicella et al. [6] in their comprehensive study on data-leakage risks in machine learning and transfer learning. In the oral-disease literature, the closest precedent to data encoding for healthcare data democratization was provided by dos Santos et al. [14] (Section 2.2), which directly inspired the deduplication and group-aware splitting protocol in Section 3.2 of the present study for the prevention of information leakage. This approach of filtering low-confidence or mislabeled training examples using conformal prediction is not widely explored. It is noted as a direction to be explored in the future in the Conclusion by Zhan et al. [40]. Although highly relevant, this literature is not the topic of the present paper but rather is considered as supporting evidence for methodological rigor.

2.6. Domain Shift and Cross-Dataset Generalization

A large body of empirical evidence shows that what works in a development sample's distribution often does not generalize to data collected in different settings. Musa et al. [4] conducted a systematic review of the cross-population shift in medical imaging analysis when using deep learning, and Van Calster et al. [5] offered more general guidance on the evaluation of predictive AI models for medical decision support, warning that acquisition devices, patient populations, and preprocessing processes each have the potential to introduce their own generalization error. Within the oral-disease literature, one quantified example is Abdian et al.'s [15] brightness/contrast perturbation (Section 2.2), but disturbing the images in the same way as what is described in the paper does not reflect differences in acquisition device or annotation protocol that a truly independent set of images could bring, which is the motivation for the external, zero-shot evaluation protocol in Section 4.1.2. Musa et al. [35] (Section 2.3) demonstrate, in a different imaging modality, that domain shift can be actively mitigated through adversarial domain adaptation, a direction identified as future work in the Conclusion rather than implemented in the present study.

2.7. External Validation in Medical Imaging

External validation evaluate an already-trained model on data it has never encountered, with no fine-tuning or parameter updates. This is a stronger test of generalization than internal validation on a held-out split from the same source. Udomluck et al. [41] demonstrated a multi-center external validation design for automated lumbar spinal stenosis grading on MRI, evaluating a single trained model across multiple independent centers. This two-or-more-source design is mirrored, at a smaller scale, by the two-condition (raw and Ben Graham-normalized) external protocol adopted in Section 4.1.2. Ahammad [34] (Section 2.3) argues explicitly that cross-dataset generalization, rather than single-dataset accuracy, should be the primary reported outcome of a transfer-learning study. None of the eight directly comparable oral-disease studies reviewed in Section 2.2 report evaluation on a separately sourced dataset. Abdian et al.'s [15] simulated domain-shift test is the closest precedent, but it is not external validation in the strict sense used here, since it perturbs rather than replaces the evaluation images.

2.8. Class Imbalance and Evaluation Metrics

Multi-class medical image classification is frequently affected by substantial class imbalance, and accuracy alone can be misleading when minority classes are clinically important despite low prevalence in the available data. Zhou et al. [8] reviewed class-imbalanced learning strategies across multi-sensor medical imaging, directly motivating the class-weighted, difficulty-cascaded sampling design examined in Section 3.1.5, and as this study's own ablation evidence later shows (Section 4.5) motivation why such a design must be verified empirically rather than assumed to work as intended. Ozdemir and Pacal [39] (Section 2.4) apply focal-loss-style reweighting as an alternative imbalance-correction mechanism, illustrating that sampler-based and loss-based reweighting are not the only available strategies. Accordingly, this study reports macro-precision, macro-recall, and macro-F1 alongside accuracy. Macro-averaging gives each class equal weight regardless of its support, making performance on minority classes such as Tooth Discoloration more visible than accuracy alone (Section 4.4).

Table 1: *Literature review summary of 20 primary benchmarked studies spanning the themes reviewed in Sections 2.1-2.9.*

Ref.	Year	Approach	Methodology / Results	Strengths	Weaknesses
Ahmed [13]	2026	DL (similar work)	Optimized/quantized ResNet-18; 90.94% accuracy.	Deployment-efficient; 75% size reduction.	No external validation; no attention.
dos Santos et al. [14]	2026	Transfer learning	ResNet-50; two-phase fine-tuning; leakage-controlled split; 96.18% accuracy and 90.78% macro-F1.	Explicit leakage control.	Small test set; caries recall 64.3%.
Abdian et al. [15]	2026	DL ensemble	Four-backbone SE+GCN ensemble; 99.70% accuracy.	Simulated domain-shift evaluation.	Simulated rather than genuinely external validation.
Oktiana et al. [16]	2026	Transfer learning	ResNet-50, EfficientNet-B4, and MobileNetV2 comparative evaluation.	Direct backbone comparison.	Single-dataset evaluation.
Marie et al. [17]	2025	DL augmentation	Latent-diffusion augmentation for six classes; approximately 6-7 point improvement in mAP/F1.	Same disease taxonomy.	Augmentation study rather than standalone classifier.
Malik et al. [11]	2026	ML hybrid	V-GAN with quantum-ML classifier; 98.95% accuracy.	Novel quantum-ML formulation.	Binary classification only.
Ozdemir & Pacal [39]	2025	ML/DL hybrid	ConvNeXtV2 with focal self-attention for skin-cancer classification.	Alternative imbalance-handling mechanism.	Different clinical domain.
Qureshi et al. [30]	2026	DL + federated	Federated-learning framework for dental-arch classification.	Privacy-preserving design.	Radiographic modality and different task.
Xu et al. [31]	2025	DL + attention	KC-UNet with CBAM and KAN; evaluated using segmentation-specific metrics.	Attention-enhanced U-Net.	Segmentation rather than classification.
Ferdous et al. [32]	2025	Lightweight DL	MedNet attention-augmented CNN with efficient inference.	Efficiency-oriented design.	Not oral/dental specific.
Ahammad [34]	2026	Transfer learning	Cross-dataset, federated, explainable transfer learning with reported cross-dataset accuracy.	Explicit generalization focus.	Parkinson's disease rather than oral imaging.
Musa et al. [35]	2025	Transfer learning	Supervised adversarial domain adaptation with cross-population evaluation.	Directly addresses domain shift.	Chest X-ray modality.
Alqhatni et al. [42]	2025	Transfer learning + XAI	EfficientNet with Grad-CAM for lung-cancer classification.	Same backbone family and Grad-CAM.	Different clinical organ/domain.
Akram et al. [43]	2025	XAI + uncertainty	Bayesian uncertainty-aware deep learning with accuracy and calibration evaluation.	Trustworthiness-oriented analysis.	Retinopathy rather than oral imaging.
Yang et al. [18]	2025	DL (same taxonomy)	Dynamic Compositional Prototype Learning (DyCoP); 93.30% accuracy and 91.90% macro-F1.	Localized prototype matching.	Internal evaluation only.
Ali & Sadeeq [19]	2025	Transfer learning	Five-backbone benchmark with Grad-CAM; EfficientNet-B3 achieved 95.40% accuracy and $\kappa = 0.989$.	Same backbone family with Grad-CAM.	Internal evaluation; loss-level imbalance correction only.
George et al. [20]	2025	DL (same taxonomy)	ViT, DeiT, BEiT, Swin, and ConvNeXt benchmark; best accuracy 81.06%.	Direct transformer-versus-CNN comparison.	Pronounced minority-class degradation.
Asiri [21]	2025	DL + attention + XAI	ExplainableDentalNet using ResNet50-SE and Grad-CAM; 98.30% accuracy.	Lightweight attention and explainability.	Limited to three-class caries grading.
Michelin et al. [22]	2026	Object detection	YOLOv8s on toothbrush-camera video frames; 83.0% mAP@0.5.	Real-world handheld acquisition.	Detection rather than classification.
Ren et al. [23]	2026	DL + attention + XAI	Dual-branch IRAA with XGrad-CAM using photographs and radiographs; 97.90% F1 on photographs.	Illumination-aware attention.	Caries only; not multi-condition.

2.9. Explainability and Reproducibility

The qualitative interpretability in this study is done using Grad-CAM, and thus, the explainable-AI literature is discussed separately (Section 4.6). The closest methodological predecessor to this study's backbone-and-interpretability union is Alqhatni et al. [42], who combined an EfficientNet backbone with Grad-CAM for lung cancer stage classification without including a minority-class or a cross-dataset assessment of the robustness of the model behind the interpretability claim. Besides Grad-CAM, there are complementary trustworthiness mechanisms, such as Bayesian uncertainty quantification to mark the certainty of a prediction instead of where it looked, as proposed by Akram et al. [43], which are not combined with Grad-CAM in the same study and will be addressed in the future (Conclusion). As a more general overview of medical imaging, Chaddad et al. [9] argued that interpretability and generalization are complementary issues that should be considered together rather than separately, which is picked up here directly in the pairing of Grad-CAM (Section 4.6) and external testing (Section 4.7). In this paper, in the section dedicated to interpretability (Section 4.6), the authors make it clear that activation overlays are qualitative only and do not provide evidence of clinical reasoning, which is what Demircioğlu [10] pointed out is often overlooked in medical imaging papers. The authors saw that, as in the aforementioned pattern, cross-dataset evaluation was uncommon in the reviewed studies [44] as well. Mirfendereski et al. [45] summarized recent artificial intelligence (AI) methods for the diagnosis of oral cancer from clinical photographs and clearly mentioned the clinical requirement for explainable (rather than accurate) AI model outputs. A.Tariq et al. [46] also paired EfficientNetV2 with Grad-CAM for multi-class brain-tumor MRI classification. This cross-domain example is relevant to Grad-CAM use in medical imaging, but it does not establish that activation maps are validated lesion segmentations. The present study therefore treats its own Grad-CAM overlays as qualitative saliency visualizations only.In an adjacent computer-vision domain, Mujtaba et al. [47] proposed a spatio-temporal deep-learning framework. Although video forensics is distinct from oral-disease classification and its results are not directly comparable to the present task, this work provides a relevant example of deep-learning methods, specifically CNNs, being applied to visual-integrity analysis.

2.10. Literature Summary Table

Table 1 summarizes the 20 primary, benchmarked studies discussed in Sections 2.1-2.9, in the same format used to synthesize this literature: approach, methodology, principal results, strengths, and weaknesses. Studies contributing tooling or review-only findings with no benchmarked classification result of their own are discussed in the text above but omitted here, since they contribute no result to compare against.

2.11. Research Gap

The literature reviewed above supports a consistent, multi-part synthesis. Conventional machine-learning approaches (Section 2.1) depended heavily on handcrafted or hybrid feature representations and were typically demonstrated on narrow, binary, or radiographic tasks. Deep-learning techniques (Section 2.2) greatly alleviated this feature-engineering

burden and demonstrated high internal accuracy in a variety of oral and dental taxonomies, but by far the most common form of evidence in nearly all of the studies reviewed is internal accuracy. Transfer learning of efficient backbones (Section 2.3) brought better data efficiency without losing the reliance on the training distribution, as well as attention mechanisms (Section 2.4), which brought in more selective, in Coordinate Attention's case positionally-aware, feature representation. However, a study could not be found in this review that used Coordinate Attention specifically for the classification of oral diseases. The literature on data-reliability (Section 2.5) indicates that the accuracy figures listed above can only be trusted if the leakage-aware splitting is performed as specified in the protocol. Only one of the eight studies listed above that are directly comparable with oral disease, dos Santos et al. [14], reports an explicit leakage-control protocol. The evidence from domain-shift and external-validation literature (Sections 2.6-2.7) suggests that the kind of generalization gaps reported in medical imaging in general are feasible in this domain in particular, but are virtually nonexistent in the eight directly comparable studies: see Abdian et al.'s [15] domain-shift test for simulated rather than independently-sourced data. The literature on class imbalance (Section 2.8) suggests the need for sampler- and loss-based correction mechanisms, but warns that their performance should be evaluated on a class-specific basis. Literature on explainability (Section 2.9) supports Grad-CAM as a qualitative tool and reminds readers that over-interpretation should not be considered clinical reasoning.

Despite the success of existing work for each of the evaluation procedures, the literature is rather limited in the number of papers that address all four of these aspects, leakage-audited dataset construction, a positionally-aware attention mechanism specifically tested for this taxonomy, statistically-supported (rather than point-estimate) selection of classifiers, and independent external testing within a single study. OralVision tackles these two challenges head-on: it is trained on an explicitly audited and corrected dataset for exact duplicates and label-conflict-contaminated data (Section 4.1.1), evaluates Coordinate Attention for oral disease classification for what is to our knowledge the first time (Section 3.1.2), and provides a strictly zero-shot evaluation on an independent, publicly documented external dataset (Section 4.7) rather than relying on a simple point estimate comparison of classifiers.

3. Proposed Model and Methodology

3.1. Proposed OralVision Architecture

OralVision consists of three stages: ImageNet-pretrained EfficientNet-B0 is used to extract the spatial feature map from an input oral photo, a factorized Coordinate Attention module processes the feature map independently along its height and width, and a simple linear layer (equal to multinomial logistic regression) maps the feature map into six class scores. The design is focused on the two questions that are frequently mixed in previous work (Section 2.4): Does the quality of the features improve when using a lightweight, attention-preserving attention mechanism, and is that sufficient to make a linear decision boundary adequate so that there is no need for a deeper classification head with additional parameters and risk of overfitting? The two questions are tested directly and separately in Section 4.5, and in the same section the question about the classifier family is tested statistically.

3.1.1. EfficientNet-B0 Backbone

The feature-extraction backbone is EfficientNet-B0 [33], trained with ImageNet weights and was chosen about the accuracy-to-parameter ratio of the backbone network. As its depth is limited by its compound-scaling design, it is able to achieve competitive ImageNet accuracy at approximately 1/25 of the number of parameters of ResNet-50 (25.6M), and the backbone has been designed for computational efficiency in comparison with the deeper backbones, namely MBConv. Given input image $x \in \mathbb{R}^{224\times224\times3}$, the backbone produces a feature map $F = f_\theta(x) \in \mathbb{R}^{7\times7\times1280}$, passed directly into the Coordinate Attention module rather than through the backbone's native ImageNet classification head, which is discarded.

3.1.2. Coordinate Attention

Coordinate Attention [37] factorizes the 2D global pooling operation used by channel-only mechanisms such as Squeeze-and-Excitation into two 1D pooling operations along the height and width axes, preserving positional information that a single global descriptor would discard (Section 2.4). Writing the channel index c as a subscript and pooling direction (h or w) as a superscript:

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i), \qquad z_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w) \tag{1}$$

These direction-aware descriptors are concatenated, passed through a shared 1×1 convolution with reduction ratio $r = 32$ (bottleneck width $\max(8, C/r)$), batch normalization, and a Hardswish nonlinearity, then split and passed through

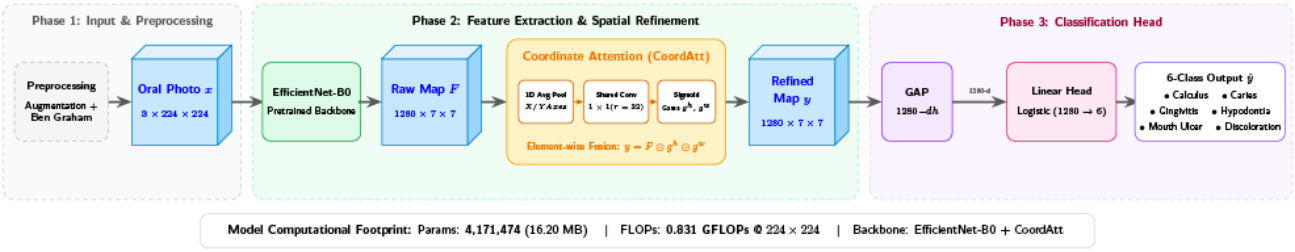


Figure 1: *OralVision architecture.*

separate 1×1 convolutions and sigmoid activations to produce attention weights g_c^h and g_c^w, applied multiplicatively to the input feature map:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{2}$$

This factorization achieves positional sensitivity along both spatial axes at a fraction of the computational cost of full 2D spatial attention; its measured contribution is quantified in Section 4.5.

3.1.3. Feature Aggregation

The refined feature map $y \in \mathbb{R}^{7\times7\times1280}$ is collapsed to a single 1280-dimensional embedding h via global average pooling: $h = \text{GAP}(y)$, $h \in \mathbb{R}^{1280}$. No additional dimensionality reduction is applied before the classification head.

3.1.4. Classification Head

The classification head is deliberately minimal: a single linear layer with no dropout or hidden layers, producing class logits $z \in \mathbb{R}^6$ from the pooled embedding $h \in \mathbb{R}^{1280}$ via $z = Wh + b$, with $W \in \mathbb{R}^{6\times1280}$ and $b \in \mathbb{R}^6$. Class probabilities follow the softmax:

$$p(y = k \mid h) = \frac{\exp(z_k)}{\sum_{j=1}^{6} \exp(z_j)}, \qquad k \in \{1, \ldots, 6\} \tag{3}$$

This single affine linear layer followed by softmax has the multinomial logistic-regression form. The primary checkpoint uses `Linear(1280 -> num_classes)` with no dropout or hidden layer. Table 8 separately compares a Logistic Regression classifier fitted on frozen embeddings with the joint end-to-end head. It presents a classifier-family comparison rather than the primary checkpoint's training implementation.

3.1.5. Training Objective

The model is trained end-to-end with class-weighted softmax cross-entropy: $\mathcal{L} = -\omega_{y_{\text{true}}} \log p(y = y_{\text{true}} \mid h)$, where ω_c denotes the inverse-frequency class weight for class c (sklearn "balanced" weighting). Training additionally uses a two-phase curriculum implemented via a class- and difficulty-weighted sampler: an 8-epoch warm-up phase sampling in proportion to ω_c alone, followed by a 32-epoch hard-example cascade phase in which each sample i's weight is further multiplied by d_i^α, with $d_i \leftarrow 0.5\, d_i + 0.5\, \ell_i$ updated every epoch (d_i initialized to 1.0, ℓ_i the sample's current training loss, $\alpha = 1.5$). Class imbalance is therefore corrected through both the loss function and the sampling distribution simultaneously.

3.1.6. Complete Forward Pass

The full inference path is: $x \to f_\theta(x) = F \to \text{CoordAtt}(F) = y \to \text{GAP}(y) = h \to Wh + b = z \to \text{softmax}(z) = p(y \mid x)$, with the predicted class $\hat{y} = \arg\max_k p(y = k \mid x)$.

The architecture in Figure 1 makes explicit that Coordinate Attention operates on the backbone's output feature map, before pooling so its positional refinement acts on spatially-resolved features rather than an already-collapsed embedding, which is what allows it to preserve height- and width-specific information that a post-pooling attention mechanism could

not recover. The total trainable parameter count is dominated by the EfficientNet-B0 backbone, with Coordinate Attention and the classification head each contributing a comparatively small fraction, consistent with the framework's lightweight design objective.

3.2. Data Preparation

To prepare the datasets, a leakage-aware protocol was used, directly inspired by Section 2.5. Exact-duplicate images were identified via cryptographic hashing. Duplicate groups in which the identical image file carried conflicting class labels across copies were excluded entirely rather than resolved by arbitrary tie-breaking, since no single label in such a group is defensible as ground truth. The remaining clean, deduplicated corpus was partitioned at the source-image level into training, validation, and test splits using a stratified split with a fixed random seed (seed = 42). Validation and test images are original photographs only. No augmentation and no Ben Graham illumination normalization is applied to either split while the training split receives on-the-fly geometric and photometric augmentation (random horizontal flip, rotation, and color jitter) plus a Ben Graham illumination-normalized copy, merged with the augmented originals into the final training corpus. A post-hoc leakage audit, re-run after splitting, confirmed zero exact-duplicate-hash overlap between the training, validation, and test partitions. Exact dataset sizes, class distributions, and the private/external dataset descriptions are reported in Section 4.1, consistent with reporting datasets as part of the experimental evidence rather than the modeling methodology.

3.3. Training Configuration

Optimization used Adam with a learning rate of 1×10^{-4} and weight decay 1×10^{-4}, a ReduceLROnPlateau schedule on validation loss (patience 3, factor 0.5, minimum LR 1×10^{-7}), and early stopping with patience 10. Training used mixed-precision (AMP), a real batch size of 16 with gradient accumulation over 2 steps (effective batch size 32), and 224×224 input resolution, with images normalized using standard ImageNet statistics. The main reported checkpoint was trained using a fixed seed of 42 over the entire budget of 40 epochs (including 8 warmup epochs and 32 cascade epochs). Five runs of the same setup with different seeds 42, 123, 2024, 3407, and 7777 were also performed for each end-to-end training configuration (Section 4.4) to provide variance due to random initialization. The hyperparameters and environmental conditions are mentioned in Table 2.

Table 2: *Training hyperparameters and software/hardware environment.*

Hyperparameter	Value
Input resolution	224×224
Real batch size/grad. accumulation / effective	16 / 2 / 32
Mixed precision (AMP)	Enabled
Warm-up / cascade / total epochs	8 / 32 / 40
Learning rate / weight decay	1×10^{-4} / 1×10^{-4}
LR schedule	ReduceLROnPlateau (patience 3, factor 0.5, min 1×10^{-7})
Early-stop patience	10
CoordAtt reduction ratio r	32
Difficulty exponent α	1.5
Seeds (primary / variance study)	42 / {42, 123, 2024, 3407, 7777}
Software	Python 3.10.9, PyTorch 2.5.1+cu121, torchvision 0.20.1+cu121, CUDA 12.1, cuDNN 90100
Hardware	NVIDIA GeForce RTX 2050 (4.29 GB)

3.4. Evaluation Protocol

Rather than reporting metrics as single point estimates from one train/test split, this study reports two complementary forms of statistical uncertainty. First, five independent training runs across different random seeds yield a mean ± standard deviation for accuracy and macro-F1, directly quantifying variance due to random initialization (Section 4.4). Second, for a single trained checkpoint evaluated once, non-parametric bootstrap resampling (10,000 resamples with replacement, seed = 42) constructs 95% confidence intervals for accuracy, macro-precision, macro-recall, macro-F1, macro-averaged one-vs-rest ROC-AUC, and per-class recall. Where two classifiers are compared on identical test images (Section 4.5), a paired bootstrap resamples the same indices for both models simultaneously, yielding a 95% confidence interval on the metric difference, and McNemar's test (continuity-corrected) is reported as a paired significance test on the two classifiers' per-image correct/incorrect decisions. Quadratic Weighted Kappa (QWK) and a chi-square test of independence were computed for continuity with earlier analysis of this project but are reported only as secondary, supplementary statistics: the six disease categories are clinically nominal, not ordinal, so QWK's quadratic penalty reflects an arbitrary label ordering rather than a clinically meaningful severity gradient, and a chi-square test on a confusion matrix built from over a thousand test images is expected to be significant ($p < 0.001$) regardless of clinical adequacy. Model complexity is reported via parameter count, on-disk size, and FLOPs at 224×224 resolution (Section 4.8). Grad-CAM visualizations

Table 3: *Dataset composition after corrected, leakage-audited splitting.*

Split	Unique images	Samples used in training
Train	5,900	17,700 (post-augmentation/Ben Graham)
Validation	1,265	1,265 (unaugmented)
Test	1,265	1,265 (unaugmented)

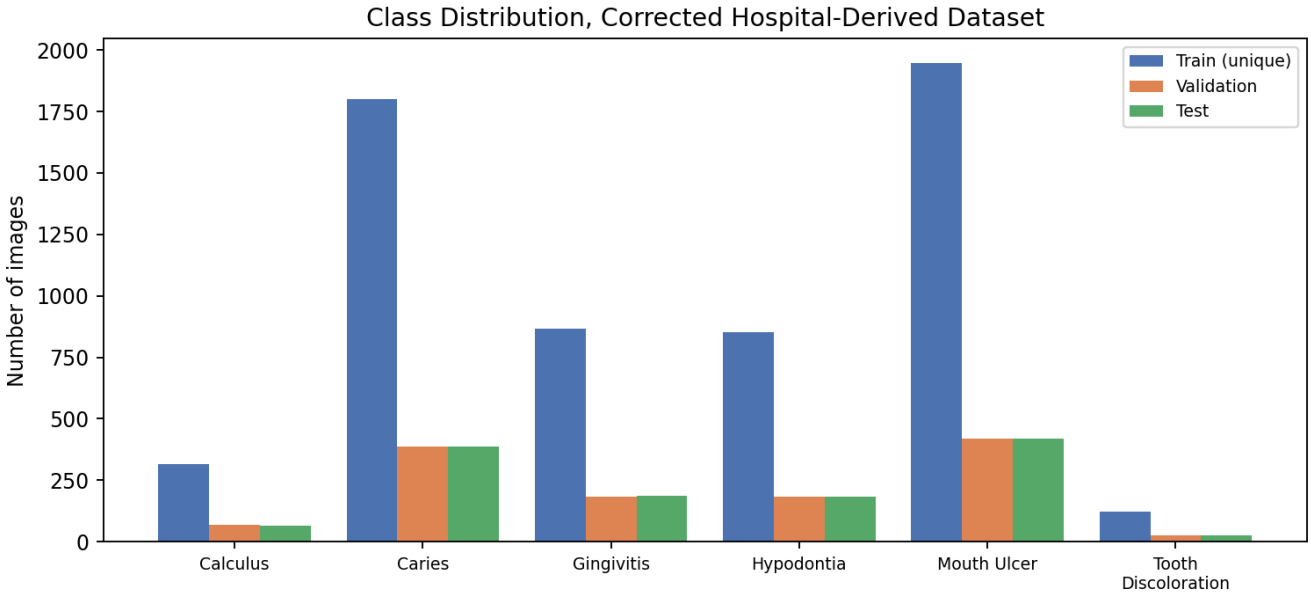


Figure 2: *Class distribution across train, validation, and test splits, corrected hospital-derived dataset.*

on the post-CoordAtt feature map are used strictly as a qualitative saliency tool, per the caveats established in Section 2.9, and are not treated as evidence of clinical interpretability (Section 4.6).

4. Experimentation and Results

4.1. Experimental Datasets

The training and internal-evaluation dataset is a private, hospital-derived collection of clinical oral photographs obtained from Holy Family Hospital, a government hospital in Rawalpindi, Pakistan. This dataset is not publicly available. Images were organized into six disease-class folders (Calculus, Caries, Gingivitis, Hypodontia, Mouth Ulcer, Tooth Discoloration) and were acquired under real-world clinical conditions. Each image was labeled by a dentist, who served as the sole reference-standard rater, with no automated or crowd-sourced labeling used at any stage. The dataset provides one folder label per image, so the task is single-label rather than multi-label classification. When more than one oral condition may have been visible, no co-occurrence labels or adjudication rule are available in the source records. This study therefore uses only the recorded folder label and does not claim to identify every condition present. The unit of analysis throughout this study is the unique source image rather than the patient, since no confirmed patient, case, or visit identifier was available in the source records to establish a patient-level count (Section 5.8).

An audit of this collection, conducted before the analyses reported here, identified exact-duplicate images, the identical file, byte-for-byte filed under two different class labels, most often between visually similar categories. Of 15,866 source images, cryptographic-hash deduplication identified 8,822 unique image hashes, 3,340 duplicate groups shared a consistent label and were collapsed to a single unique instance, while 392 duplicate groups carried conflicting labels and were excluded entirely (3,016 images), yielding a final corpus of 8,430 unique, label-consistent images. This corpus was partitioned at the source-image level into 5,900 training / 1,265 validation / 1,265 test images (Table 3). Only the training split is expanded, via augmentation and Ben Graham merging, into the 17,700 samples actually presented to the model during training.

As can be seen in the per-class distribution (Figure 2), there is a large disparity in the number of images in each class, with Tooth Discoloration being significantly underrepresented (26 test images) compared to the other five classes (67-418 test images).

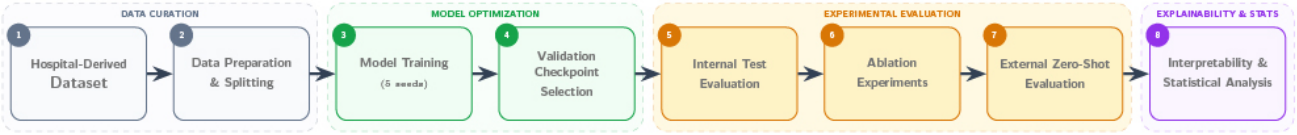


Figure 3: *Experimental process, from dataset preparation through final statistical and interpretability analysis.*

A second, entirely independent set of 1,057 images was sourced from a publicly available, open-source Roboflow collection, the *Dental Diseases Classification Dataset* [48], covering the same six-class label space but disjoint from the hospital-derived process in both origin and acquisition conditions. This dataset is not a validation subset of the hospital dataset. It is an independent external test dataset, used exclusively for a strict zero-shot evaluation of the already-trained checkpoint, with no fine-tuning, no retraining, and no parameter updates of any kind. The same 1,057 images are evaluated under two fixed preprocessing conditions (raw and Ben Graham illumination-normalized) not as two disjoint sets, so $N = 1{,}057$ for each condition throughout this study (Section 4.7).

4.2. Preprocessing

All images are resized to 224×224 and normalized using standard ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). Training images additionally receive on-the-fly random horizontal flip, random rotation ($\pm 15°$), and color jitter (brightness, contrast, saturation each varied by up to ± 0.2), applied at load time on top of the offline Ben Graham illumination-normalized copy merged into the training corpus (Section 3.2). Validation, internal test, and external test images receive resizing and normalization only. No augmentation and no Ben Graham normalization is applied to validation or internal test images at any stage. The external dataset is evaluated separately under both raw and Ben Graham conditions specifically to isolate the effect of this single preprocessing step at inference time (Section 4.7).

4.3. Experimental Process

Figure 3 summarizes the full experimental process, from the hospital-derived dataset through to final statistical analysis. The corrected dataset is split and prepared (Section 3.2), the model is trained across five random seeds (Section 3.3), the best-validation-macro-F1 checkpoint is selected per run, internal test performance is evaluated with bootstrap confidence intervals (Section 4.4), ablation variants are trained and compared at the identical full training budget (Section 4.5), the resulting checkpoint is evaluated zero-shot on the external dataset (Section 4.7), and interpretability and remaining statistical comparisons (Grad-CAM, McNemar's tests) are computed last, since they depend on a finalized checkpoint.

4.4. Main Model Results

Table 4 reports macro-averaged performance for the primary OralVision checkpoint. The checkpoint uses EfficientNet-B0 with Coordinate Attention, a class-weighted loss, a weighted sampler, and curriculum training (8 warm-up epochs plus a 32-epoch cascade budget, maximum 40 epochs). It achieved a validation macro-F1 of 0.986220 at epoch 30. The checkpoint summary gives test accuracy 0.995257 and macro-F1 0.984145 (Table 4 rounds these to 0.9953 and 0.9841). Table 6, Variant 6, the Naive Training ablation (no Coordinate Attention, curriculum, or class weighting), reports accuracy 0.9952569 and macro-F1 0.9840564 (rounded in Table 6 to 0.9953 and 0.9841). The full-precision values are close but not identical. The reported macro-precision also differs (0.9906 in Table 4, 0.9940 in Table 6). Table 4 and Variant 6 therefore report distinct model and training configurations. The Full Model ablation baseline reports test accuracy/F1 of 0.9905/0.9695. Performance is closely matched between validation (accuracy 0.9897, macro-F1 0.9862) and test. The train-split metrics (accuracy 0.9999) are higher than validation and test.

Table 4: *Macro-averaged performance by data split, primary OralVision checkpoint.*

Split	N	Accuracy	Precision	Recall	F1
Train	17,700	0.9999	0.9999	0.9999	0.9999
Validation	1,265	0.9897	0.9896	0.9831	0.9862
Test	1,265	0.9953	0.9906	0.9790	0.9841

To quantify variance across random initialization, five independent training runs (seeds 42, 123, 2024, 3407, 7777) achieved test accuracy of 99.24% ± 0.21% and test macro-F1 of 97.68% ± 0.51% (mean ± SD), with balanced accuracy of 96.95% ± 0.66%. Bootstrap 95% confidence intervals (10,000 resamples) for the deployed checkpoint on the internal test set are: accuracy 0.9953 [0.9913, 0.9984], macro-precision 0.9907 [0.9813, 0.9981], macro-recall 0.9791 [0.9561, 0.9986], macro-F1 0.9840 [0.9667, 0.9974], macro-averaged one-vs-rest ROC-AUC 0.9999 [0.9997, 1.0000]. Only reported here as a secondary, supplemental statistic, as described in Section 3.4. Expected Calibration Error (ECE) = 0.0019 from predicted-probability calibration on the internal test set, indicating close agreement between softmax

confidence and empirical accuracy on this test set specifically. Calibration was not separately assessed on the external dataset (Section 5.8).

Table 5 presents the per-class precision, recall and F1 scores. The results show that although the recall for Calculus is perfect (1.0000), the precision is not as high (0.9571). Tooth Discoloration is at its best with perfect precision (1.0000), but the lowest recall (0.8846, 23 of 26 test images). The results for the ablation indicate that recall of Tooth Discoloration varies as a function of configuration, but ablation is not a sufficient evidence to conclude that the difficulty-weighted sampler or the curriculum caused the under-recall.

Table 5: *Per-class internal test-set performance, deployed checkpoint.*

Class	Precision	Recall	F1	Support
Calculus	0.9571	1.0000	0.9781	67
Caries	0.9974	0.9948	0.9961	386
Gingivitis	0.9946	0.9946	0.9946	186
Hypodontia	0.9945	1.0000	0.9973	182
Mouth Ulcer	1.0000	1.0000	1.0000	418
Tooth Discoloration	1.0000	0.8846	0.9388	26
Macro avg	**0.9906**	**0.9790**	**0.9841**	**1,265**

Figure 4 shows the corresponding confusion matrix, making explicit which off-diagonal errors underlie the precision and recall figures reported above.

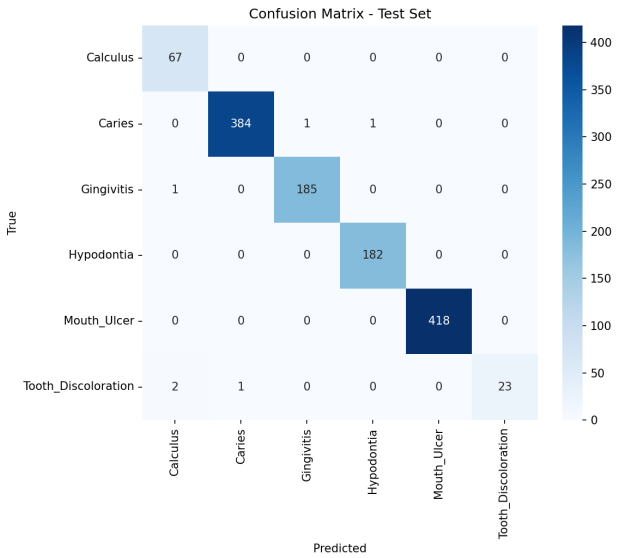


Figure 4: *Confusion matrix on the internal test set (1,265 images, 6 classes).*

4.5. Ablation Study

Table 6 reports an ablation study conducted at the full training budget used by the main model (8 warm-up + 32 cascade epochs, or 40 single-phase epochs), rather than a reduced budget, so that no observed effect can be attributed to a training-budget mismatch. ΔF1 is defined as $F1_{\text{baseline}} - F1_{\text{variant}}$. A positive value indicates the removed component helped, a negative value indicates the variant outperformed the baseline. Variants are numbered sequentially in the order evaluated.

Table 6: *Full-training-budget ablation results, corrected dataset.*

Variant	Acc.	Prec.	Rec.	F1	ΔF1	Best Epoch	Runtime (s)
1. Baseline (Full Model)	0.9905	0.9808	0.9604	0.9695	0.0000	19	5844.0
2. No Attention	0.9905	0.9908	0.9668	0.9776	−0.0081	8	3527.6
3. No Curriculum (single-phase)	0.9929	0.9876	0.9660	0.9758	−0.0062	30	7969.4
4. No Hard-Example Cascade	0.9929	0.9876	0.9660	0.9758	−0.0062	30	8016.3
5. No Class Weighting	0.9913	0.9834	0.9633	0.9715	−0.0019	6	3208.6
6. Naive Training (no CoordAtt, no curriculum, no weighting)	0.9953	0.9940	0.9754	0.9841	−0.0145	24	6750.2
7. Linear+Dropout Head (instead of primary linear head)	0.9960	0.9944	0.9838	0.9888	−0.0193	26	7332.1

Coordinate Attention. Removing CoordAtt (Variant 2) increased macro-F1 from 0.9695 to 0.9776 (ΔF1 = −0.0081 under the table's baseline-minus-variant definition). This single-run result does not support a positive contribution from Coordinate Attention. The no-attention variant performed better. **Curriculum and cascade.** The no-curriculum and no-hard-example-cascade variants (Variants 3-4) each achieved 0.9758 macro-F1, also exceeding the Full Model baseline of 0.9695 (ΔF1 = −0.0062). Thus, neither removal variant supports the claim that these training components improve aggregate F1.

Classifier-head depth. Replacing the primary single-layer linear head with a deeper Linear+Dropout head (Variant 7) produced the highest Table 6 macro-F1 (0.9888, ΔF1 = −0.0193), favoring the deeper head in this single-run ablation. This is a distinct question from classifier-family choice, addressed statistically below, and the two should not be conflated.

Class-weighting and sampler comparison. Variants 5 and 6 show negative ΔF1 and outperform the Full Model baseline. Per-class analysis (Table 7) shows that class-weighted loss provides no measurable Tooth Discoloration recall benefit (21/26 correct with or without weighting), while removing the curriculum/difficulty-weighted sampler together with class weighting (Variant 6) improves this recall (21/26 → 23/26). These observations are consistent with a possible interaction among training choices, but this single-run ablation does not isolate a causal mechanism. A previously reported sub-finding shows that class weighting visibly shifts Calculus/Gingivitis confusion at moderate imbalance ratios that does not replicate on this corrected, deduplicated dataset (0-2 errors in either direction across all three configurations). The cause of the discrepancy remains uncertain.

Table 7: *Per-class breakdown for the three full-budget checkpoints most affected by class weighting.*

Class / Metric	Baseline	No Class Weighting	Naive Training
Tooth Discoloration-correct/support	21/26	21/26	23/26
Tooth Discoloration-recall	0.8077	0.8077	0.8846
Calculus↔Gingivitis errors (total)	2	1	2

The primary OralVision checkpoint retains the CoordAtt, curriculum, and class-weighted configuration. However, the ablation results do not independently verify a performance contribution from CoordAtt or the training-strategy components: the no-attention and no-curriculum/no-cascade variants outperform the Full Model baseline in macro-F1. Table 7 suggests no Tooth Discoloration recall change from removing class weighting alone and a higher recall for Naive Training, but this single-run comparison does not isolate a causal mechanism for the aggregate differences.

Classifier-family comparison. To compare logistic regression against three alternative families of classifiers, trained on frozen CoordAtt embeddings with the same checkpoint and tested on the same 1,265 test images, we perform a separate, post-hoc analysis (Table 8) that includes bootstrap 95% confidence intervals and McNemar's tests for each pair of classifiers.

Table 8: *Classifier-family comparison on frozen embeddings, with paired statistical testing against the joint head.*

Classifier	Accuracy [95% CI]	Macro-F1 [95% CI]	McNemar's p	Δ Macro-F1 vs. joint head
Joint end-to-end head	0.9953 [0.9913, 0.9984]	0.9840 [0.9667, 0.9974]	-	-
Logistic Regression	0.9937 [0.9889, 0.9976]	0.9768 [0.9566, 0.9925]	0.617	+0.0071 [−0.0034, +0.0214]
SVM (RBF)	0.9905 [0.9850, 0.9953]	0.9658 [0.9425, 0.9848]	0.077	+0.0181 [+0.0023, +0.0381]
Random Forest	0.9921 [0.9874, 0.9961]	0.9716 [0.9500, 0.9887]	0.221	+0.0124 [−0.0012, +0.0297]
XGBoost	0.9889 [0.9826, 0.9945]	0.9557 [0.9277, 0.9780]	0.027	+0.0283 [+0.0087, +0.0530]

In the paired McNemar tests of per-image correctness, the differences between the joint head and Logistic Regression (p=0.617), Random Forest (p=0.221), and SVM (p=0.077) were not statistically significant at the 0.05 level. The difference between the joint head and XGBoost was statistically significant (p=0.027). These tests assess paired correctness decisions, not whether one classifier is universally superior or whether its macro-F1 is statistically higher. The bootstrap confidence intervals for macro-F1 differences should therefore be considered alongside the McNemar results. No conclusion about comparative runtime is drawn because a directly comparable runtime analysis is not reported here. Table 8 reports a post-hoc classifier-family comparison on frozen embeddings, whereas the primary checkpoint uses the end-to-end linear head described in Section 3.1.4.

4.6. Explainability Analysis Using Grad-CAM

For two representative test images per class (comprising 12 images), the Grad-CAM activation map over the post-CoordAtt feature map for each image is shown in Figure 5. The predicted class, the softmax confidence for that class, the accuracy (True or False) of the prediction, and the original photograph are all labeled on each panel. The heatmap of the original image is shown, as is the raw heatmap extracted using Grad-CAM, and the heatmap overlaid on the original image. Table 9 summarizes the qualitative region of concentrated activation and a brief interpretive note, per class.

Table 9: *Qualitative summary of Grad-CAM activation regions by class.*

Class	Region of Concentrated Activation	Interpretive Note
Calculus	Near the gumline, along visible deposit surfaces	Consistent with calculus forming as a mineralized deposit at the gum margin
Caries	Lesion surface and surrounding enamel	Activation follows the visible dark lesion rather than adjacent healthy enamel
Gingivitis	Gingival tissue margins	Consistent with inflammation localized to the soft-tissue margin rather than the tooth surface
Hypodontia	Inter-dental gap	Activation localizes to the missing-tooth space rather than adjacent teeth
Mouth Ulcer	Lesion boundary	Activation follows the ulcer perimeter rather than the surrounding unaffected mucosa
Tooth Discoloration	Tooth surface	Activation is diffuse across the discolored tooth surface rather than a single focal point

From Figure 5 and Table 9, it is observed that: First, the activation pattern is directionally consistent with each class's expected clinical presentation. Attention concentrates at the gumline for Calculus and the gingival margin for Gingivitis rather than at a shared, generic "tooth region," suggesting the model has not collapsed these two visually similar classes into an identical spatial attention pattern despite their frequent visual overlap. Second, the Tooth Discoloration panels show a more diffuse activation pattern than the other five classes, without a single, tightly localized focal region. This is consistent with the class's own diagnostic character. Discoloration is a surface-wide property rather than a spatially bounded lesion but also plausibly reflects the class's limited training support (Section 4.1.1), since a sparsely-represented class offers the network fewer examples from which to learn a more spatially selective representation. Neither observation is offered as evidence of correct model reasoning: no pixel-level ground-truth lesion annotation was available to compute a quantitative overlap score between activation regions and clinically relevant regions, so both observations remain qualitative and descriptive rather than measured, consistent with the caveats established in Section 2.9.

4.7. External Dataset Generalization

Table 10 compares internal test performance against the independent external dataset (Section 4.1.2), evaluated with the identical, frozen checkpoint under both raw and Ben Graham illumination-normalized conditions.

Table 10: *Internal test versus external raw and Ben Graham generalization performance.*

Metric	Internal Test	External (Raw)	External (Ben Graham)
N	1,265	1,057	1,057
Accuracy	0.9953	0.8335	0.8127
Macro Precision	0.9906	0.8918	0.8761
Macro Recall	0.9790	0.8376	0.8207
Macro F1	0.9841	0.8211	0.7892
Tooth Discoloration recall	0.8846	0.2786	0.1592
QWK (secondary)	0.9849	0.5807	0.5229

All primary metrics are lower on the external dataset than the internal test set: accuracy drops by 16.2-18.3 percentage points and macro-F1 drops by 16.3-19.5 percentage points. External macro-F1 is between 0.7892 and 0.8211, which suggests that the performance is still higher than the uniform guessing model, but there is a significant gap in generalization, which is not a criterion for clinical adequacy. Tooth Discoloration has the greatest class specific recall reduction (-0.8846 internally, -0.2786 raw external, and -0.1592 after Ben Graham normalization). Low precision of Caries and perfect recall in external per-class results (Table 11) are consistent with the high frequency of over-prediction of Caries, and high precision of Tooth Discoloration with low recall. These patterns suggest that the external error profile cannot be completely summarized by aggregate measures.

Table 11: *Per-class external generalization performance.*

Class	Prec. (Raw)	Rec. (Raw)	F1 (Raw)	Prec. (BG)	Rec. (BG)	F1 (BG)
Calculus	0.9278	0.8182	0.8696	0.9200	0.8364	0.8762
Caries	0.5882	1.0000	0.7407	0.5682	1.0000	0.7246
Gingivitis	0.8961	0.9388	0.9169	0.8903	0.9388	0.9139
Hypodontia	0.9949	0.9899	0.9924	0.9949	0.9899	0.9924
Mouth Ulcer	0.9434	1.0000	0.9709	0.9132	1.0000	0.9547
Tooth Discoloration	1.0000	0.2786	0.4358	0.9697	0.1592	0.2735

The external confusion matrices for the two preprocessing conditions are presented in Figure 6.

Ben Graham normalization reduced external accuracy by 2.08 percentage points and macro-F1 by 3.19 percentage points. Tooth Discoloration recall also declined, from 0.2786 on raw images to 0.1592 after normalization, a decrease of

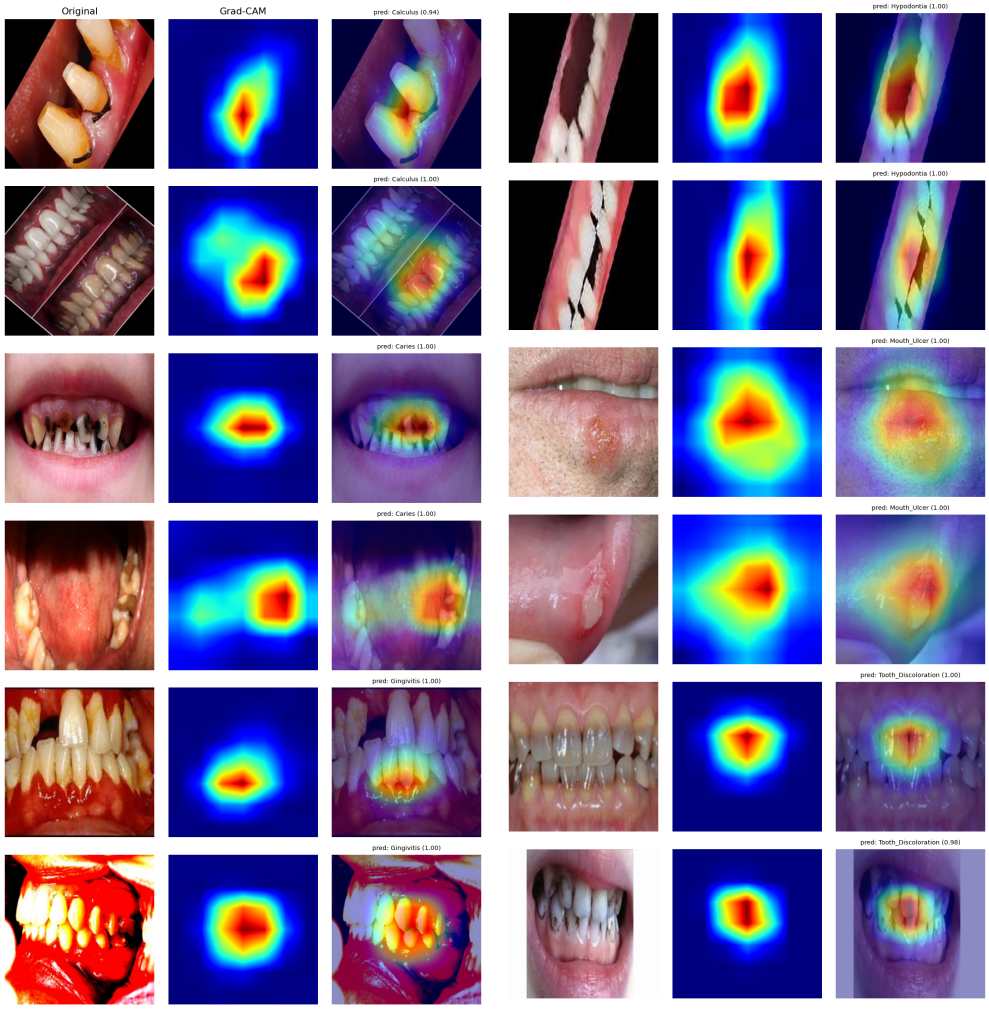


Figure 5: *Grad-CAM activation overlays, two representative test images per class (twelve overlays total). Each panel is labeled with the predicted class, prediction confidence, and correctness.*

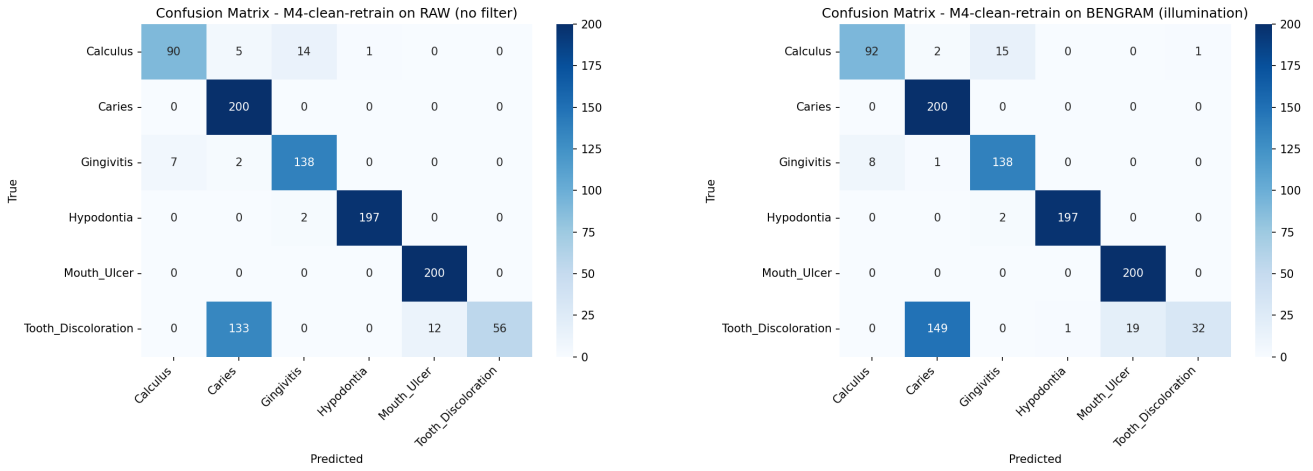
11.94 percentage points. Thus, this preprocessing step did not improve generalization under the evaluated conditions, and its effect was particularly pronounced for Tooth Discoloration recall.

4.8. Model Complexity

The computational footprint is summarized in Table 12: 4,171,474 parameters (all trainable), 16.20 MB on disk, and $\approx$ 0.831 GFLOPs per 224×224 forward pass, which is about one-sixth of the number of parameters in a standard ResNet-50 backbone before any classification head. The figures shown below only reflect the performance of the trained network and do not reflect any latency, memory, or battery measurement taken at the device level. No claims of any real-time or mobile deployment performance have been measured (Section 5.7).

Table 12: *Model complexity summary.*

Metric	Value
Total/trainable parameters	4,171,474
Model size on disk	16.20 MB
FLOPs (224×224 input)	830,903,808 ($\approx$ 0.831 GFLOPs)



(a) Raw external images

(b) Ben Graham illumination-normalized images

Figure 6: *External confusion matrices: (a) raw external images and (b) Ben Graham illumination-normalized external images.*

5. Discussion

5.1. Principal Findings

This study tested OralVision on a leakage-audited, deduplicated hospital dataset. The primary checkpoint's test accuracy was 0.995257 and macro-F1 was 0.984145. The accuracy is 0.9952569 and macro-F1 is 0.9840564 in Table 6, Variant 6 (Naive Training), where Coordinate Attention, curriculum, and class weighting are not included. They are not exactly equal, but round the same way to four decimal places. Five-seed results refer to multiple training runs, the full-budget ablation is a single-run comparison, and it does not support positive contributions from Coordinate Attention or the training-strategy components. Without any fine-tuning, the primary checkpoint's macro-F1 dropped to 0.79-0.82, with a significant and class-specific decline in Tooth Discoloration recall. It is not a single architectural outcome that is the main result of this work but the internal-versus-external contrast.

5.2. Architectural Rationale and Ablation Evidence

Coordinate Attention was inspired by its ability to maintain positional information (Section 3.1.2), which could be useful for oral-image classification. However, the full-budget ablation does not reveal an empirical performance gain: macro-F1 with Coordinate Attention ablation remains at 0.9695, while the macro-F1 with the ablation of the curriculum or hard-example cascade is 0.9758. The results reported here are from a single run and do not provide evidence towards better classification performance being due to these components. The proposed spatial rationale is still a motivation rather than an evidence of better classification performance.

5.3. Internal Performance Versus External Generalization

The gap between the internal macro-F1 score of 0.9841 and the external macro-F1 scores of 0.8211 (raw images) and 0.7892 (Ben Graham-normalized images) shows that there is a significant generalization gap. The internal results show good performance on the development distribution, while the external results show lower performance with external images that are sourced separately. In particular, Tooth Discoloration recall is decreased to 0.2786 on raw images and 0.1592 after Ben Graham normalization, which shows that the model is unable to recognize a significant percentage of cases from this class under the analyzed external conditions. These results illustrate the need to rely on external tests to determine generalizability across independent data distributions. OralVision shows good internal classification performance, but needs to be validated externally on a variety of datasets and acquisition conditions before it can be used for clinical screening or clinical deployment. No clinical readiness or deployment suitability claims are made.

5.4. Comparison with Existing State of the Art

We compare OralVision with nine studies from Section 2.2: six studies utilizing the same or a similar 6-class taxonomy, and three additional studies investigating a complementary set of oral-imaging tasks (narrower caries grading, video-frame object detection and dual-modality caries classification) for architectural and domain context. The two studies-Oktiana et al. [16] (2026) and Marie et al. [17] (2025)-are discussed in the qualitative synthesis and Table 1 but are excluded from the quantitative comparison because their reported results are not directly comparable with the selected classification

metrics. The values are reported as directly quoted by each source, and the metric type is identified in line as such, with the exception of the object-detection metric mAP, which is not an accuracy figure. “NR" indicates a metric that the source paper did not report.

Table 13: *Comparison of OralVision with related oral-disease classification studies.*

Study	Year	Dataset	Model	Metric 1	Metric 2	External Validation
Ahmed [13]	2026	Private	Pruned ResNet-18	90.94% accuracy	NR	No
dos Santos et al. [14]	2026	OdontoAI (3,507 images)	ResNet-50 (25.6M parameters)	96.18% accuracy	90.78% macro-F1	No
Abdian et al. [15]	2026	Private (10,235 images)	4-backbone SE+GCN ensemble	99.70% accuracy	NR	Simulated only
Yang et al. [18]	2025	Private	DyCoP (prototype learning)	93.30% accuracy	91.90% macro-F1	No
Ali & Sadeeq [19]	2025	Intraoral RGB	EfficientNet-B3 + Grad-CAM	95.40% accuracy	$\kappa = 0.989$	No
George et al. [20]	2025	12,653 images	ConvNeXt (best of 5 backbones)	81.06% accuracy	NR	No
Asiri [21]	2025	2,000 images (3-class)	ResNet50-SE + Grad-CAM	98.30% accuracy	MCC = 0.975	No
Michelin et al. [22]	2026	7,963 video frames	YOLOv8s	83.0% mAP@0.5	77.9% F1	No
Ren et al. [23]	2026	639 photos + 456 radiographs	DBAttNet (IRAA)	97.90% F1 (photo)	95.72% F1 (radiograph)	No
OralVision	**2026**	**Holy Family Hospital**	**EfficientNet-B0 + CoordAtt**	**99.53% accuracy**	**98.41% macro-F1**	**Yes*, N=1,057, zero-shot**

Note: NR indicates not reported. *External validation refers to zero-shot evaluation on an independent dataset of 1,057 images without retraining or fine-tuning.

The OralVision row reports the internal test results of the primary checkpoint summarized in Tables 4 and 10. External evaluation was performed on the independent Roboflow dataset (N = 1,057) without fine-tuning. The Full Model ablation baseline is a separate run reported in Table 6 and is not used for the internal metrics in this comparison.

There are four observations to follow and the differences in the datasets and protocols are explained. Firstly, the primary OralVision checkpoint has an internal accuracy of 99.53% while Abdian et al. [15] reported 99.70% accuracy. The numbers are not directly comparable as the studies are based on different datasets, splits, taxonomies and model configurations and the difference between the numbers is not meant to be interpreted as a superiority or inferiority. Results for the baseline Full Model ablation were not the primary-checkpoint results found in Table 13, and the accuracy from this baseline was 99.05%. Secondly, no one of the nine comparator studies reviewed in Table 13 provides evaluation on an independent, separately sourced external data set. A simulated domain-shift test on perturbed source images was reported by Abdian et al. [15] but this is not the same as test on independent source data. In the context of the studies analyzed for this comparison, OralVision states that this observation is based on an external evaluation of the studies that was conducted separately from this study, and must not be generalized to the whole literature. Thirdly, George et al. [20] claim an accuracy range of 79.73%-81.06% for the variants of the transformer and ConvNeXt. The fact that this is the case indicates that scale alone is not sufficient for high accuracy on this taxonomy, though this is not necessarily due to a difference in architecture, as it may also be due to different datasets and protocols. Fourth, OralVision has 4.17 million parameters which is less than what ResNet-50 of dos Santos et al. [14] reported as 25.6 million parameters. Here, no quantitative comparisons with the four-backbone ensemble of Abdian et al. [15] and with the models of George et al. [20] are made since they do not directly compare the number of parameters.

Overall, the primary OralVision checkpoint reports high internal accuracy and macro-F1, although the comparisons in Table 13 are descriptive due to differences in the datasets, splits, and evaluation protocols across the different studies. The number of parameters of the model is fairly low as compared to ResNet-50 reported by dos Santos et al. [14] but cannot be unconditionally compared across all the architectures. The external evaluation, which is conducted separately, provides a valuable additional source of evidence in this comparison set, but has some limitations, such as the use of only one public data set, which are discussed in Section 5.8.

5.5. Ablation Interpretation

The full-budget ablation (Section 4.5) shows that the Full Model baseline is outperformed by both the no-attention variant and the no-curriculum/no-cascade variants. The Naive Training variant also has higher aggregate F1. These results do not support the earlier claim that Coordinate Attention, curriculum learning, or hard-example cascading improve macro-F1. Because each ablation configuration was run once, the observed differences should be treated as descriptive and require multi-seed confirmation. The per-class results show no measurable Tooth Discoloration recall benefit from class-weighted loss in the compared runs, but do not by themselves establish a causal mechanism for the aggregate differences.

5.6. Explainability

The class-level activation patterns in Section 4.6 show visually different Grad-CAM responses for Calculus and Gingivitis, with activation concentrated around the gumline and gingival margin, respectively. These visualizations are descriptive only and do not establish that the model relies on clinically valid features or explain its classification decisions. No

quantitative interpretability metric, such as overlap with expert-annotated regions or deletion/insertion analysis, was computed. Therefore, no causal explanation for the observed classification performance is inferred from these maps.

5.7. Practical and Deployment Implications

At 4.17M parameters, 16.20 MB, and 0.831 GFLOPs, OralVision is substantially smaller than ResNet-50-based comparators (Section 4.8). This characterization concerns measured model size and computational cost only. No latency, memory, or battery benchmarking was performed on any target device, and no claim of real-time or mobile-deployment performance is made.

5.8. Limitations

Four limitations are consequential to interpreting these findings.

- First, the external test set is a single publicly available Roboflow collection (N = 1,057, [48]). Its collection and annotation processes are not independently verified here, and its provenance does not establish that it represents the same acquisition process or population as the hospital-derived training data. Performance should therefore not be assumed to generalize to other acquisition settings or populations. Evaluation on additional independently collected, preferably multi-center datasets is an important next step.
- Second, the Tooth Discoloration class is severely under-represented in the corrected internal test set, with only 26 images. Therefore, conclusions about its generalization performance remain provisional.
- Third, the primary checkpoint reports accuracy 0.995257 and macro-F1 0.984145, whereas Table 6, Variant 6 (Naive Training), reports accuracy 0.9952569 and macro-F1 0.9840564. These results are close but not identical at full precision and arise from different experimental configurations. The primary checkpoint uses EfficientNet-B0 with Coordinate Attention and the specified class-weighted curriculum strategy. The Naive Training variant omits Coordinate Attention, curriculum, and class weighting. The two results are reported separately and should not be treated as repeated measurements of the same trained model.
- Fourth, the full-budget ablation reports seven configurations, with each reported variant reflecting one fixed-seed run. Consequently, the observed performance differences may partly reflect run-to-run variation and require confirmation across multiple random seeds.

6. Conclusion and Future Work

Although there are compelling reasons for moving toward computer-based methods for analyzing photographs to diagnose oral disease, the kind of data used to train and validate the algorithms is almost always from the development set, which is limited to a single source, and the question remains whether reported accuracy is a true representation of the algorithms' diagnostic capacity or an artifact of limited, single-source evaluation. This paper introduces a new method, named OralVision, based on combining an EfficientNet-B0 backbone with a factorized Coordinate Attention module and a single-layer linear classification head, which is trained on a leakage-audited, deduplicated dataset from Holy Family Hospital in Rawalpindi, Pakistan, and tested using a full training-budget ablation study, a statistically grounded comparison of classifiers in the family, and a strict, zero-shot evaluation on an independent, publicly documented external dataset. The primary internal test checkpoint achieved accuracy 0.995257 and macro-F1 0.984145. Table 6, Variant 6 (Naive Training), which omits Coordinate Attention, curriculum, and class weighting, reports accuracy 0.9952569 and macro-F1 0.9840564. These values are close but not identical at full precision, although they round alike at four decimal places. They represent different configurations and are reported separately. Five-run results were 99.24%±0.21% accuracy and 97.68%±0.51% macro-F1. The full-budget ablation does not provide evidence that Coordinate Attention improves performance: the no-attention variant scored higher than the Full Model baseline, as did the no-curriculum and no-hard-example-cascade variants. The classifier-family comparison found no significant difference between the separately evaluated Logistic Regression classifier and the joint head (McNemar's $p = 0.617$). Table 8 reports that comparison on frozen embeddings, while the primary checkpoint uses the end-to-end linear head described in Section 3.1.4. Without optimization on the independent external test set, macro-F1 was 0.79-0.82, with Tooth Discoloration recall of 0.16-0.28. The internal-versus-external gap demonstrates that high internal performance does not guarantee generalization and should be assessed directly.

Building on these findings, the most immediately valuable next steps are: extending external testing to a genuinely multi-center design beyond the single external dataset used here. A dedicated four-way (sampler × loss) decomposition experiment to test the possible class-weighting/sampler interaction suggested by the present ablation. Re-running the six remaining ablation configurations and a multi-seed ablation study at full training budget. Targeted data collection to increase Tooth Discoloration representation, quantitative interpretability evaluation against expert-annotated lesion boundaries and conformal-prediction-based data cleaning as a complementary safeguard against residual label noise.

Author Declarations

Author Contributions (CRediT): Moiz Ud Din: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Validation, Investigation, and Writing-Original Draft; Ibtisam Asghar: Data Curation and Resources; Ejlal Hameed: Software, Visualization, Application Development, and Writing-Review and Editing; Saad Irfan Khan: Conceptualization, Methodology, Supervision, Project Administration, and Writing-Review and Editing. All authors reviewed and approved the final version of the manuscript.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Statement: The internal training dataset consists of private clinical oral photographs obtained from Holy Family Hospital and is not publicly available. The independent external test dataset is the publicly available Roboflow Dental Diseases Classification Dataset [48].

Acknowledgments: The authors (Moiz, Ibtisam, Ejlal) would like to thank their supervisor (Saad Irfan Khan) and institution (SZABIST University) for supporting this research.

Conflicts of Interest: The authors declare no known competing financial interests or personal relationships that could have appeared to influence this work.

Ethics Statement: Not applicable.

Declaration of Generative AI: The authors declare that LLM model was used solely for improving grammar and enhancing the structure of the paper.

References

[1] M. Hu, H. Chen, R. Wang, R. Zhang, J. Yao, X. Qiu, and J. Guo, "Global, regional, and national burden of lip and oral cavity cancer and its attributable risk factors from 1990 to 2021: An analysis of the Global Burden of Disease Study 2021," *International Journal of Clinical Pharmacy*, vol. 48, no. 1, pp. 93-106, 2026, doi: 10.1007/s11096-025-01961-9.

[2] H. Li et al., "Application of artificial intelligence in oral health management: Challenges and opportunities," *Frontiers in Medicine*, vol. 13, Art. no. 1700529, 2026, doi: 10.3389/fmed.2026.1700529.

[3] Z. Ma, C. Du, Q. Lao, and X. Xie, "Generative artificial intelligence: Applications and future prospects in dentistry," *International Dental Journal*, vol. 76, no. 1, Art. no. 108276, 2026, doi: 10.1016/j.identj.2025.108276.

[4] A. Musa, R. Prasad, P. Onwualu, and M. Hernandez, "A systematic review of cross-population shifts in medical imaging analysis with deep learning," *Big Data and Cognitive Computing*, vol. 10, no. 3, Art. no. 76, 2026, doi: 10.3390/bdcc10030076.

[5] B. Van Calster et al., "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance," *The Lancet Digital Health*, vol. 7, no. 12, Art. no. 100916, 2025, doi: 10.1016/j.landig.2025.100916.

[6] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! Exploring data leakage risks in machine learning and transfer learning," *Artificial Intelligence Review*, vol. 58, no. 11, Art. no. 339, 2025, doi: 10.1007/s10462-025-11326-3.

[7] A. Thakur, T. Zhu, V. Abrol, J. Armstrong, Y. Wang, and D. A. Clifton, "Data encoding for healthcare data democratization and information leakage prevention," *Nature Communications*, vol. 15, Art. no. 1582, 2024, doi: 10.1038/s41467-024-45777-z.

[8] D. Zhou, S. Gao, and X. Huang, "Strategies for class-imbalanced learning in multi-sensor medical imaging," *Sensors*, vol. 26, no. 6, Art. no. 1998, 2026, doi: 10.3390/s26061998.

[9] A. Chaddad, Y. Hu, Y. Wu, B. Wen, and R. Kateb, "Generalizable and explainable deep learning for medical image computing: An overview," *Current Opinion in Biomedical Engineering*, vol. 33, Art. no. 100567, 2025, doi: 10.1016/j.cobme.2024.100567.

[10] A. Demircioğlu, "Reproducibility and interpretability in radiomics: A critical assessment," *Diagnostic and Interventional Radiology*, vol. 31, no. 4, pp. 321-328, 2025, doi: 10.4274/dir.2024.242719.

[11] H. Malik, A. Naeem, D. Jeong, R. A. Naqvi, and S.-W. Lee, "QML: Quantum machine learning and vanilla GAN-based classification of dental caries using bitewing radiographs," *Cluster Computing*, vol. 29, Art. no. 355, 2026, doi: 10.1007/s10586-026-06116-7.

[12] M. R. Salmanpour, A. H. Pouria, S. Barichin, Y. Salehi, S. Falahati, I. Shiri, M. Oveisi, and A. Rahmim, "PySERA: Open-source standardized Python library for automated, scalable, and reproducible handcrafted and deep radiomics," *Computer Methods and Programs in Biomedicine*, vol. 284, Art. no. 109463, 2026, doi: 10.1016/j.cmpb.2026.109463.

[13] A. Ahmed, "Optimized ResNet-18 architecture for multi-class oral diseases classification," *Karbala International Journal of Modern Science*, vol. 12, no. 1, Art. no. 8, 2026, doi: 10.33640/2405-609X.3446.

[14] R. M. dos Santos, K. F. Gameiro, V. M. da Silva Ferreira, S. C. Brum, A. S. Carlos, A. M. Moreno, and S. T. A. de Mello, "Development and evaluation of a deep learning model for dental image classification," *Research, Society and Development*, vol. 15, no. 7, Art. no. e6615751427, 2026, doi: 10.33448/rsd-v15i7.51427.

[15] A. Z. Abdian, M. M. Javidi, N. Mansouri, F. Mehranfar, and S. Cheperli, "MultiDentNet: A unified deep learning framework for multi-class dental condition screening and preliminary oral lesion triage," *Scientific Reports*, vol. 16, Art. no. 21815, 2026, doi: 10.1038/s41598-026-52264-6.

[16] M. Oktiana, M. Fitria, H. Habibie, K. Saddami, Y. Elma, H. Rahayu, S. Janura, R. Putri, and R. Novita, "Performance analysis of ResNet-50, EfficientNet B4, and MobileNet V2 for dental caries classification," *JOIV: International Journal on Informatics Visualization*, vol. 10, no. 1, pp. 10-18, 2026, doi: 10.62527/joiv.10.1.3713.

[17] H. S. Marie, M. Elbaz, R. S. Soliman, and A. A. Elkhatib, "DentoMorph-LDMs: Diffusion models based on novel adaptive 8-connected gum tissue and deciduous teeth loss for dental image augmentation," *Scientific Reports*, vol. 15, Art. no. 27268, 2025, doi: 10.1038/s41598-025-11955-2.

[18] W. Wang, R. Yang, J. Xiao, Y. Wang, and Z. Chen, "Mining hierarchical oral disease prototypes: Oral disease diagnosis using a novel dynamic compositional prototype learning framework," *Preprints*, version 2, 2025, doi: 10.20944/preprints202504.0003.v2.

[19] D. A. Ali and H. T. Sadeeq, "An interpretable deep learning framework for multi-class dental disease classification from intraoral RGB images," *Statistics, Optimization & Information Computing*, vol. 14, no. 6, pp. 3380-3397, 2025, doi: 10.19139/soic-2310-5070-2880.

[20] A. B. George, S. Bathini, and N. S. R., "Deep learning for oral health: Benchmarking ViT, DeiT, BEiT, ConvNeXt, and Swin Transformer," *arXiv preprint arXiv:2509.23100*, 2025.

[21] F. Y. I. Asiri, "Automated classification of enamel caries from intraoral images using deep learning models: A diagnostic study," *Journal of Clinical Medicine*, vol. 14, no. 24, Art. no. 8959, 2025, doi: 10.3390/jcm14248959.

[22] J. Michelin, A. Nooruddin, A. Khalid, S. M. F. Ahmed, B. King, M. M. Shaikh, A. J. T. Neto, and F. Umer, "Deep learning models for the detection of dental-findings and tooth-types using video data," *BMC Oral Health*, vol. 26, Art. no. 1502, 2026, doi: 10.1186/s12903-026-08861-y.

[23] L. Ren and J. Chen, "A dual-branch deep learning framework with explainability for dental caries classification using intra-oral photographs and radiographs," *Journal of Imaging*, vol. 12, no. 5, Art. no. 207, 2026, doi: 10.3390/jimaging12050207.

[24] K. M. Desai et al., "Screening of oral potentially malignant disorders and oral cancer using deep learning models," *Scientific Reports*, vol. 15, Art. no. 17949, 2025, doi: 10.1038/s41598-025-02802-5.

[25] P. Liu and K. Bagi, "A tailored deep learning approach for early detection of oral cancer using a 19-layer CNN on clinical lip and tongue images," *Scientific Reports*, vol. 15, Art. no. 23851, 2025, doi: 10.1038/s41598-025-07957-9.

[26] A. Mohapatra and R. Dash, "A unified framework for oral squamous cell carcinoma detection using machine learning and deep learning techniques," *Discover Computing*, vol. 29, Art. no. 169, 2026, doi: 10.1007/s10791-026-10054-8.

[27] N. V. S. S. Vissamsetti and G. Shivakanth, "A deep learning-based image processing framework for oral lesion classification," *Engineering, Technology & Applied Science Research*, vol. 16, no. 2, pp. 33266-33271, 2026, doi: 10.48084/etasr.15583.

[28] T. D. Pham and S. Al-Hebshi, "Classification of pediatric dental diseases from panoramic radiographs using natural language transformer and deep learning models," *Frontiers in Artificial Intelligence*, vol. 9, Art. no. 1754498, 2026, doi: 10.3389/frai.2026.1754498.

[29] B. Ayhan, E. Ayan, and S. Atsü, "Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence algorithms based on deep learning methods," *BMC Oral Health*, vol. 25, Art. no. 216, 2025, doi: 10.1186/s12903-025-05577-3.

[30] A. Qureshi, Q. U. A. Aini, E. Solomin, N. Ahmad, Z. Uddin, and A. H. Milyani, "Dental arch classification using deep learning and federated learning on DenPAR radiographs," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-62329-1.

[31] J. Xu, H. Gao, and Z. Wang, "KC-UNet: Enhancing U-Net with KAN and CBAM for medical image segmentation," *IEEE Access*, vol. 13, pp. 153788-153797, 2025, doi: 10.1109/ACCESS.2025.3605148.

[32] M. Ferdous, S. Mahmud, and M. E. Z. Shimul, "MedNet: A lightweight attention-augmented CNN for medical image classification," *Scientific Reports*, vol. 15, Art. no. 41936, 2025, doi: 10.1038/s41598-025-25857-w.

[33] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, pp. 6105-6114, 2019.

[34] I. Ahammad, "Generalization over accuracy: A cross-dataset, explainable, and federated learning framework for Parkinson's disease detection," *Digital Health*, vol. 12, Art. no. 20552076261467837, 2026, doi: 10.1177/20552076261467837.

[35] A. Musa, R. Prasad, and M. Hernandez, "Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation," *Scientific Reports*, vol. 15, Art. no. 11383, 2025, doi: 10.1038/s41598-025-95390-3.

[36] R. Bhuria, S. Gupta, R. M. Ghoniem, J. Singh, S. Rani, B. M. Taye, and S. Bharany, "A transfer learning-based approach for automated bone fracture classification in X-ray imaging," *Therapeutic Advances in Musculoskeletal Disease*, vol. 18, Art. no. 1759720X251405099, 2026, doi: 10.1177/1759720X251405099.

[37] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13713-13722, doi: 10.1109/CVPR46437.2021.01350.

[38] Z. Wei, F. Yang, K. Zhong, and L. Yao, "PCB-YOLO: Enhancing PCB surface defect detection with coordinate attention and multi-scale feature fusion," *PLOS ONE*, vol. 20, no. 6, Art. no. e0323684, 2025, doi: 10.1371/journal.pone.0323684.

[39] B. Ozdemir and I. Pacal, "An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms," *Results in Engineering*, vol. 25, Art. no. 103692, 2025, doi: 10.1016/j.rineng.2024.103692.

[40] X. Zhan, Q. Xu, Y. Zheng, G. Lu, and O. Gevaert, "Reliability-enhanced data cleaning in biomedical machine learning using inductive conformal prediction," *PLOS Computational Biology*, vol. 21, no. 2, Art. no. e1012803, 2025, doi: 10.1371/journal.pcbi.1012803.

[41] P. Udomluck, W. Cholamjiak, J. Inpun, and W. Waratamrongpatai, "Multi-task deep learning model for automated detection and severity grading of lumbar spinal stenosis on MRI: Multi-center external validation," *Diseases*, vol. 14, no. 1, Art. no. 32, 2026, doi: 10.3390/diseases14010032.

[42] A. Alqhatani, T. K. S. R. Babu, T. R. Mahesh, S. B. Khan, O. Saidani, and M. T. Quasim, "Automated classification and explainable AI analysis of lung cancer stages using EfficientNet and gradient-weighted class activation mapping," *Frontiers in Medicine*, vol. 12, Art. no. 1625183, 2025, doi: 10.3389/fmed.2025.1625183.

[43] M. Akram, M. Adnan, S. F. Ali, J. Ahmad, A. Yousef, T. A. N. Alshalali, and Z. A. Shaikh, "Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches," *Scientific Reports*, vol. 15, Art. no. 1342, 2025, doi: 10.1038/s41598-024-84478-x.

[44] M. Moharrami, E. Vahab, M. Bagherianlemraski, G. Hemmati, S. Singhal, C. Quinonez, F. Schwendicke, and M. Glogauer, "Deep learning for detecting dental plaque and gingivitis from oral photographs: A systematic review," *Community Dentistry and Oral Epidemiology*, vol. 53, no. 6, pp. 617-632, 2025, doi: 10.1111/cdoe.70001.

[45] P. Mirfendereski, G. Y. Li, A. T. Pearson, and A. R. Kerr, "Artificial intelligence and the diagnosis of oral cavity cancer and oral potentially malignant disorders from clinical photographs: A narrative review," *Frontiers in Oral Health*, vol. 6, Art. no. 1569567, 2025, doi: 10.3389/froh.2025.1569567.

[46] A. Tariq, S. I. Khan, M. Imran, and H. Dawood, "Enhanced Grad-CAM segmentation for multi-class brain tumor classification," in *Proc. 4th Int. Conf. Communication, Computing and Digital Systems (C-CODE)*, 2025, doi: 10.1109/C-CODE67372.2025.11204150.

[47] M. H. Mujtaba, M. Imran, S. Irfan, U. Muhammad, and H. Dawood, "TamperNet: A spatio-temporal deep learning framework for video manipulation detection," *Neural Processing Letters*, 2026, doi: 10.1007/s11063-026-11885-8.

[48] Final Year Project, "Dental Diseases Classification Dataset," *Roboflow Universe*, Roboflow, 2025. [Online]. Available: https://universe.roboflow.com/final-year-project-gifyj/dental-diseases-classification. [Accessed: Sep. 6, 2026].