Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 8** : 1–16.
ISSN-0037-8712
doi:10.5269/bspm.83893

DQ-SAM: Data-Quality-Aware Optimization for Robust G. V. Black Class I–III Caries Classification on Clinical Intraoral Periapical Radiographs

Aneetta Joy Parathanath and A. Manimaran

ABSTRACT: Early and reliable classification of dental caries in intraoral periapical radiographs is challenged by heterogeneous image quality in routine clinical workflows. This study develops a data-quality-aware Swin Transformer pipeline that is robust to such variability and evaluates it on a curated single-centre dataset from the SIBAR Institute of Dental Sciences, Guntur. We employ a lightweight Swin-Tiny backbone with a dental-biased transformer block trained using Data-Quality–gated Sharpness-Aware Minimization (DQ-SAM). Each training image is assigned a scalar quality score $q \in [0, 1]$, fused from six fast, interpretable cues (Laplacian variance, Tenengrad, tiled RMS contrast, dynamic range, entropy, and edge density) that are robustly normalized within each G.V. Black class. The resulting q modulates both the SAM perturbation radius and the effective AdamW update, emphasizing reliable gradients from higher-quality batches while tempering updates from lower-quality ones. On a three-class periapical test set of 360 radiographs (G.V. Black Classes I–III), the proposed DQ-SAM Swin-Tiny classifier attains 91.9% accuracy with a macro-F1 score of 0.92 and class-wise ROC–AUC values above 0.97. Quality-stratified analysis shows stable performance across low-, medium-, and high-quality bands ($\approx$ 89–93% accuracy), indicating robustness to blur and contrast variation. Compared with DenseNet-121 and ResNet-50 baselines trained under the same protocol, the proposed model delivers consistently higher accuracy, particularly in challenging quality regimes. The method supports reliable artificial intelligence for clinical dental imaging.

Keywords: DQ-SAM, sharpness-aware minimization, swin-tiny transformer, dental caries, intraoral periapical radiographs.

Contents

1. Introduction

Dental caries remains one of the most common dental diseases worldwide and continues to be a major public health problem. A dysbiotic dental biofilm interacts with dietary sugars and host variables to cause it. Cariogenic bacteria turn sugars into organic acids, which lower the pH of plaque and cause repeated demineralization of enamel and dentin during acid exposure [1]. When the acid attack exceeds the buffering capacity of saliva and the redeposition of minerals, white spots can form and progress to

2020 *Mathematics Subject Classification:* 68T07, 68U10.
Submitted May 01, 2026. Published June 05, 2026.

cavities over time [2]. It affects a large population, and the burden is unevenly distributed across regions and populations. The disease is highly prevalent worldwide [3]. Hundreds of millions of children alone have untreated carious lesions, which shows how important it is to have diagnostic methods that can be used on a large scale [4]. Dental caries causes pain, infection, and tooth loss, and it also commonly leads to more complicated restorative or endodontic procedures. This costs patients and health systems significant amounts of money worldwide [5].

In clinical dentistry, the G.V. Black classification has long served as a standard scheme for describing carious lesions according to their anatomical location and involved surfaces. Originally developed by Greene Vardiman Black in the late 19th and early 20th centuries, this system groups lesions into classes depending on whether they occur in pits and fissures, proximal surfaces, incisal edges, or cervical regions, and it is still widely taught in operative dentistry. Although more recent systems, such as minimal intervention–oriented classifications and ICDAS, emphasize lesion activity and early enamel changes, the traditional G.V. Black framework continues to guide everyday restorative planning and documentation in many clinical settings [6]. Assigning the correct G.V. Black class from radiographic images is, however, inherently challenging because intraoral radiographs vary in exposure, contrast, sharpness, and anatomical superimposition, which can obscure lesion margins and distort apparent morphology. Moreover, interpretation is heavily dependent on clinician experience and is vulnerable to inter- and intra-observer variability, particularly for subtle or borderline proximal lesions. In this context, automated classification systems based on deep learning offer a promising complement to human diagnosis. Recent work in dental AI has shown that convolutional and related models can improve diagnostic accuracy for caries detection and help standardize decision-making across operators and settings, supporting more consistent, data-driven care [7].

Over the past few years, convolutional neural networks (CNNs) have been widely applied across dental radiographic analysis tasks, demonstrating strong promise in caries detection and segmentation. For instance, in a study of bitewing radiographs, a U-shaped CNN architecture was developed to detect early carious lesions; the authors reported modest precision ($\sim$63%) and recall ($\sim$65%), yet showed that clinicians' sensitivity improved when assisted by the model outputs (i.e., the model acted as a second reader) [8]. A systematic review of CNNs applied to periapical radiographs for early-stage caries detection further confirmed that deep convolutional architectures can match or exceed conventional radiographic assessment in many cases, though performance often degrades when image quality is poor or when datasets are small [9]. In another clinical-scale investigation, a CNN-based object detection method (Faster R-CNN) was used for proximal caries in 978 bitewing radiographs: the network differentiated lesions into five depth categories (E1, E2, D1, D2, and D3), indicating that detailed staging rather than binary classification is feasible in the dental imaging domain [10].

Beyond pure classification, integrated detection-classification pipelines have also been explored. For example, in a review of methods for caries detection, segmentation, and classification, it was observed that many works lean toward segmentation or detection first, followed by classification or rule-based diagnosis, rather than designing end-to-end multiclass classification architectures [11]. Likewise, a survey of deep learning in dental radiographs notes that the majority of studies focus on either binary classification (caries vs no caries) or lesion localization/segmentation, with fewer targeting richer lesion taxonomy such as lesion classes or stages [12]. More recently, transformer and hybrid architectures have begun to appear in dental imaging, aiming to leverage their capacity for modeling long-range dependencies and contextual relationships. In a 2025 comparative study, Schneider et al. [13] assessed CNNs, transformers, and hybrid models on dental imaging tasks such as caries and hypomineralization detection, observing that transformer-based backbones sometimes outperformed CNNs in generalization, especially under variable imaging conditions. Another work in dental disease diagnosis proposes a hybrid of MobileNetV2 and Swin Transformer. The model fuses local feature extraction and global contextual reasoning to classify multiple dental conditions (fillings, implants, cavities, and impacted teeth), showing that transformer modules can complement lightweight CNNs in dental X-ray classification [14].

Despite this progress, substantial gaps remain when it comes to multiclass G.V. Black classification under heterogeneous image conditions. Most studies reduce the problem to binary lesion detection or coarse staging rather than classification. Even when multiple lesion stages are used, they rarely cover the complete anatomical classification. In addition, existing methods generally treat all training images

equally, without acknowledging that certain radiographs may be degraded (e.g., blur, low contrast, noise) and thus less informative. The use of sharpness-aware optimizers in deep vision tasks has improved generalization in some medical imaging studies, but their adoption in dental radiograph classification remains minimal. To the best of our knowledge, no prior work adaptively modulates the strength of sharpness perturbation on a per-batch basis as a function of image quality. Finally, while aggregate performance (accuracy, sensitivity, F1) is commonly reported, few works stratify results according to image quality, which is crucial for understanding how robustness degrades under real-world imaging variability.

To address these gaps, we propose a unified framework that embeds data-quality awareness into sharpness-aware optimization and couples this with a Swin Transformer + AdamW architecture tailored for the G.V. Black classification of dental caries. We term this optimization scheme *Data-Quality–gated Sharpness-Aware Minimization (DQ-SAM)*. During training, each image is assigned a no-reference quality score q (derived from blur and contrast-related metrics), and for each mini-batch the average $\bar{q}$ is computed and used to (i) gate the SAM perturbation magnitude so that lower-quality batches induce smaller adversarial perturbations, and (ii) scale the AdamW update step so that "messy" batches contribute more gently to parameter updates. The Swin backbone is further regularized via dropout, domain-appropriate augmentations, and a cosine-annealed learning-rate schedule.

2. Existing Works

Deep learning has been increasingly adopted for radiographic caries diagnosis, and a small but growing subset of this literature treats G.V. Black classes as explicit targets rather than collapsing lesions to a binary label. Singh and Sehgal [15] proposed one of the earliest deep learning systems mapping periapical radiographs directly to G.V. Black classes I–VI using an optimal CNN–LSTM classifier, reporting high accuracy and demonstrating that fine-grained, anatomy-based labelling is feasible on routinely acquired intraoral images. More recently, Parathanath et al. [16] introduced CBMNet, a dual-attention ConvNeXt-Tiny model for automatic classification of G.V. Black Classes I–III on intraoral periapical radiographs, combining CBAM and multi-scale attention with GAN-based augmentation and PSO-driven hyperparameter tuning to achieve test accuracy above 90% and consistent gains over ResNet-50, EfficientNet-B0, and DenseNet-121 baselines. Together, these studies indicate that G.V. Black-oriented classification is clinically meaningful and that modern architectures can exploit the localized but structured patterns of periapical caries to deliver high diagnostic performance.

A larger body of work has focused on caries detection and staging on bitewing or periapical radiographs without always encoding G.V. Black explicitly. Lee et al. [17] evaluated deep CNN algorithms for the detection and diagnosis of dental caries on periapical radiographs, demonstrating that convolutional models can match or surpass human performance on interproximal lesions when trained and tested under controlled single-center conditions. ForouzeshFar et al. [18] presented a CNN-based classifier for the diagnosis of dental caries utilizing bitewing radiographs from 713 individuals, exhibiting commendable accuracy across four network modifications. Chaves et al. [19] advanced this research by training a CNN to identify both primary and secondary caries surrounding restorations on bitewing radiographs, demonstrating that deep learning models can detect subtle radiolucencies adjacent to restorative materials that are frequently misinterpreted in clinical practice. These task-specific studies are contextualized by systematic reviews, such as the work of Prados-Privado et al. [20], which synthesized 13 neural-network-based studies on caries detection and diagnosis using periapical, bitewing, and near-infrared images, highlighting the heterogeneity in definitions, labels, and outcome metrics as a significant impediment to comparability and reproducibility. More recently, transformer-based and hybrid architectures have been introduced into dental imaging. Zhou et al. [21] proposed a tooth-type–enhanced Swin Transformer for caries diagnosis in children on panoramic radiographs, modelling shared and tooth-type–specific representations to capture differences among incisors, canines, and molars, achieving an AUC of 0.92 and outperforming several CNN baselines. A broader review by Gao et al. [22] on the application of transformers in stomatological imaging similarly concluded that Swin-style architectures are under-exploited but promising for tasks such as caries detection, periodontal bone loss assessment, and lesion segmentation, especially when combined with domain-specific adaptations. Collectively, these studies motivate the choice of a Swin backbone for G.V. Black classification from periapical radiographs, as it naturally

captures both local texture and long-range anatomical context.

Finally, because the proposed method is centered on optimization under variable image quality, it is essential to situate it within recent work on sharpness-aware training. Hassan et al. [23] systematically evaluated a suite of sharpness-based optimizers, including SAM and several variants, on CNN and vision transformer models for medical image analysis and found that vanilla SAM was the only method that consistently improved generalization and produced flatter loss landscapes relative to standard Adam, particularly on classification tasks. In follow-up work, the same group introduced GCSAM, a gradient-centralized variant of SAM, and demonstrated further gains in generalization across both natural and medical imaging benchmarks by explicitly regularizing gradient directions [24]. These findings set a precedent for using SAM-style optimizers in medical imaging, but they do not incorporate data-quality awareness. All images and batches are treated equally, regardless of radiographic noise or blur. To our knowledge, no prior study in dental radiograph analysis combines Swin Transformers with a data-quality–gated sharpness-aware optimizer for multiclass G.V. Black classification on intraoral periapical images, which defines the methodological gap of the present work.

3. Methodology

3.1. Dataset Description and Preparation

We used a retrospective collection of 1,103 anonymized intra-oral periapical radiographs from the SIBAR Institute of Dental Sciences, Guntur (India), acquired on a digital radiography system. All intraoral periapical radiographs were obtained at a single institution using a standardized imaging setup comprising an X-Mind DC dental X-ray unit (Acteon, France) and a VistaScan Mini Easy PSP scanner (Dürr Dental, Germany). Size 2 PSP plates were used, with exposure times ranging from 0.12 to 0.32 seconds. Plates were scanned at 20 lp/mm, processed using VistaScan imaging software, and exported in PNG format. Because the dataset was collected using the same imaging system and workflow, scanner-related variability was limited. Nevertheless, as the data were obtained retrospectively during routine clinical practice, some minor variability related to operator handling, positioning, and exposure adjustments may remain. For this study, we focused on radiographically driven G.V. Black categories and restricted labels to Classes I–III, as these are primarily diagnosed from radiographs, while excluding Classes IV–VI that are typically determined by direct clinical and tactile examination. All images were screened for gross technical artifacts, and objective image quality was enforced using the BRISQUE score, retaining only radiographs with BRISQUE ≤ 40 to support stable feature learning. At source, the class distribution was imbalanced (Class I = 408, Class II = 490, Class III = 205).

To mitigate this imbalance while preserving realistic morphology, we first increased the number of Class III samples through mild, radiograph-appropriate augmentations (small rotations, horizontal flips, and modest brightness changes). This expanded Class III from 205 to 453 authentic images (+248) without introducing unrealistic structures. We then balanced all three classes to a common target of 600 images per class by adding synthetic samples, ensuring that authentic images remained the majority in each category. The final composition was Class I (408 real / 192 synthetic), Class II (490 / 110), and Class III (453 / 147), yielding a total of 1,800 images with matched class counts while avoiding over-representation of synthetic samples.

We employed StyleGAN2-ADA to generate synthetic intraoral periapical radiographs under limited-data conditions. Real periapical radiographs were converted to 8-bit RGB and upsampled to 512×512 before GAN training. StyleGAN2-ADA was trained on a cloud GPU (A100) with a batch size of 32 and adaptive augmentation for approximately 2,500 kimg, initialized from an FFHQ-512 checkpoint. During training, the Fréchet Inception Distance (FID) steadily improved from 311 to 35, after which snapshot samples were drawn. Following generation, synthetic radiographs were manually reviewed and sorted into the three diagnostic classes. Visibly implausible images were discarded, and all retained synthetic samples were subjected to the same BRISQUE ≤ 40 thresholding and preprocessing pipeline as the real data. Perceptual realism was further evaluated via a blinded human review, in which 11 dentists each classified 100 images (50% real, 50% synthetic); the mean real-vs-synthetic accuracy was 54.45% (range: 41–72%), indicating that many synthetic periapical radiographs were difficult to distinguish from real radiographs. The complete data acquisition and preprocessing workflow is summarized in Figure 1.

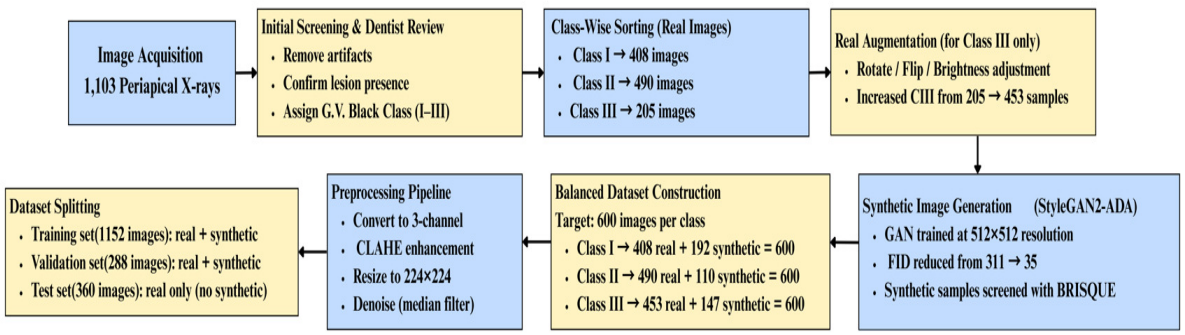

Figure 1: Data acquisition and preprocessing pipeline for periapical radiographs

3.2. Data Preprocessing

All real and synthetic radiographs across the training and test sets were passed through a unified preprocessing pipeline to ensure consistent appearance and stable input to the Swin Transformer. Each image was first loaded in its native format and converted to a standard three-channel BGR representation, allowing uniform handling of grayscale, RGB, and RGBA images. The radiographs were then resized to the target Swin input resolution. To enhance diagnostically relevant structures, particularly enamel–dentin boundaries and fissural regions, we applied Contrast-Limited Adaptive Histogram Equalization (CLAHE) to the grayscale component, followed by reconversion to three channels. A 3×3 median blur was applied to suppress impulse noise and minor acquisition artifacts without compromising edge integrity. The resulting images were normalized to the $[0, 1]$ floating-point range and stored as standardized PNG files. This preprocessing was applied identically to both real and StyleGAN2-ADA–synthesized radiographs to maintain distributional consistency. Representative examples of real and StyleGAN2-ADA–generated periapical radiographs for each G.V. Black class are shown in Figure 2.

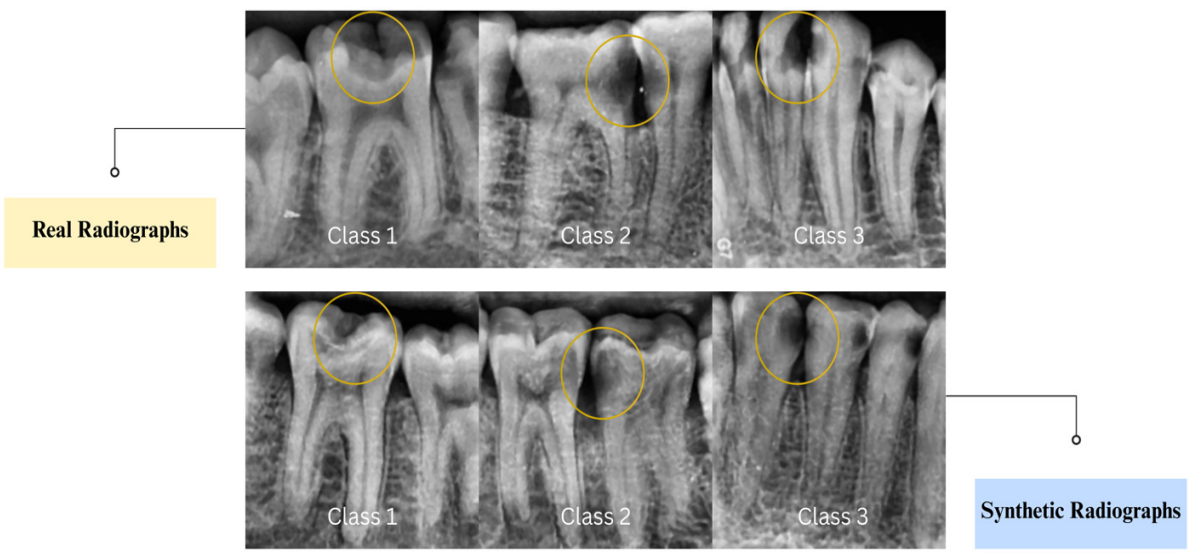

Figure 2: Representative real and synthetic periapical radiographs

3.3. Quality-score computation (q-score)

Since the proposed optimization framework incorporates data-quality awareness, we computed a per-image quality score q for all training images after preprocessing. For each image $I \in \mathbb{R}^{H\times W}$, we extracted a six-dimensional feature vector capturing sharpness, structural visibility, contrast, and texture richness:

$$f(I) = [f_{\text{lap}}, f_{\text{ten}}, f_{\text{rms}}, f_{\text{dyn}}, f_{\text{ent}}, f_{\text{edge}}]\,. \tag{3.1}$$

The individual components are defined as follows:

Laplacian variance (sharpness):

$$f_{\text{lap}} = \text{Var}(\nabla^2 I). \tag{3.2}$$

Tenengrad gradient magnitude (edge acuity):

$$f_{\text{ten}} = \mathbb{E}\left[G_x^2 + G_y^2\right], \tag{3.3}$$

where (G_x, G_y) are Sobel gradients in horizontal and vertical directions.

Local RMS contrast:

$$f_{\text{rms}} = \frac{1}{K}\sum_{k=1}^{K} \sigma(I_k), \tag{3.4}$$

where I_k denotes the k-th local tile and σ is the standard deviation.

Dynamic range (exposure index):

$$f_{\text{dyn}} = \frac{P_{98} - P_2}{255}, \tag{3.5}$$

using the 2nd and 98th grayscale percentiles.

Entropy (texture richness):

$$f_{\text{ent}} = -\sum_{i=0}^{255} p_i \log_2(p_i), \tag{3.6}$$

where p_i is the normalized histogram of I.

Edge density (structural visibility):

$$f_{\text{edge}} = \frac{1}{HW}\sum_{x,y} \mathbb{1}\left[\text{Canny}(I)_{x,y} > 0\right]. \tag{3.7}$$

To account for class-dependent differences in radiographic appearance, quality normalization was performed separately within each G.V. Black class, not globally across the full dataset. Specifically, for each class c and feature k, the class-specific median and interquartile range (IQR) were estimated from the training set only:

$$m_{c,k} = \text{median}(f_{c,k}), \tag{3.8}$$

$$\text{IQR}_{c,k} = Q_{75}(f_{c,k}) - Q_{25}(f_{c,k}). \tag{3.9}$$

Using these class-wise robust statistics, each raw feature value was standardized as

$$z_{c,k} = \frac{f_{c,k} - m_{c,k}}{\text{IQR}_{c,k}}, \quad z_{c,k} \in [-3, 3], \tag{3.10}$$

and then linearly mapped to $[0, 1]$:

$$\vartheta_{c,k} = \frac{z_{c,k} + 3}{6}. \tag{3.11}$$

A weighted fusion step combined these normalized features:

$$s = \sum_{k=1}^{6} w_k\, \vartheta_{c,k}, \tag{3.12}$$

with weights $w = \{0.18, 0.12, 0.15, 0.10, 0.15, 0.30\}$ assigned to Laplacian variance, Tenengrad energy, tile-based RMS contrast, dynamic range, entropy, and edge density, respectively. These weights emphasize structural visibility and sharpness while still incorporating contrast and texture information.

To stabilize the distribution, the fused score was passed through a logistic squeeze:

$$q = \frac{1}{1 + e^{-4(s-0.5)}}, \tag{3.13}$$

followed by bounding $q \in [0.20, 0.98]$ to avoid extreme values. Thus, a higher q_i indicates better estimated radiographic quality.

Across classes, the final q-distributions were well centred (Class I: 0.496 ± 0.099; Class II: 0.503 ± 0.098; Class III: 0.504 ± 0.092), indicating consistent quality characteristics in the curated dataset. These per-image q-scores serve as the core signal used in the subsequent DQ-SAM + AdamW training stage to modulate perturbation magnitude and learning-rate scaling on a batch-wise basis. From a computational standpoint, the proposed quality-scoring module is lightweight because it relies on six deterministic handcrafted cues computed directly from the preprocessed grayscale image. These features are extracted once offline and stored in the metadata file, after which the training stage uses only the resulting scalar q-score. Consequently, the added runtime cost during optimization is minimal compared with the forward/backward cost of the classifier itself. The class-wise distribution of the resulting q-scores is illustrated in Figure 3. Representative examples from the low-, medium-, and high-quality bands are shown in Figure 4. The low-quality examples exhibit weaker edge definition, reduced contrast, and lower structural clarity, whereas the high-quality examples show sharper tooth boundaries, improved contrast, and clearer anatomical visibility. The medium-quality band lies between these extremes, supporting the interpretability of the proposed q-score stratification.

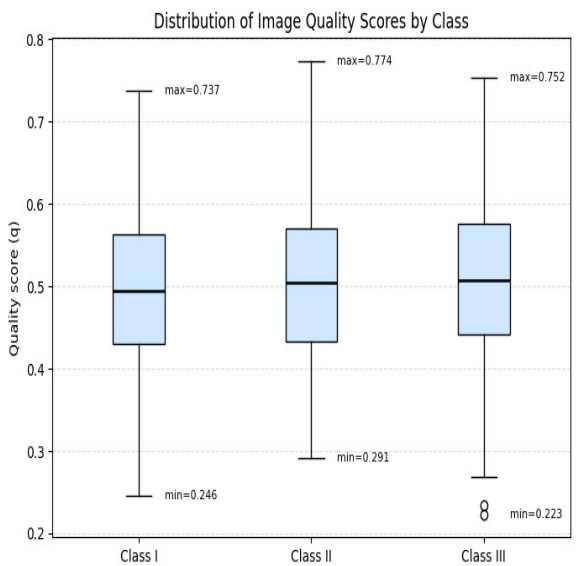

Figure 3: Distribution of image quality scores by class

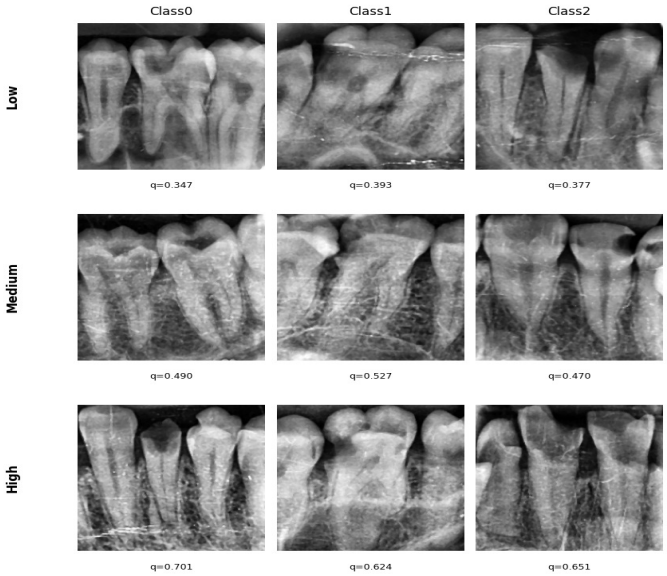

Figure 4: Representative low-, medium-, and high-quality intraoral periapical radiographs according to the proposed q-score.The examples illustrate the gradual change in radiographic sharpness, contrast, and structural visibility across the three quality bands, with higher q values generally corresponding to clearer anatomical detail.

3.4. Architecture Design

The proposed model is built on a Swin-Tiny Transformer backbone, chosen for its ability to capture both local texture patterns and broader spatial context in dental radiographs. We use the features-only variant of Swin, which provides four hierarchical feature maps. The final stage (Stage 4) contains the most semantically rich representation and serves as the input to our classification head. The output of this stage is first standardized to a consistent layout and then flattened into a sequence of spatial tokens. These tokens are subsequently processed by a dental-biased transformer block, a lightweight attention module tailored to radiographic caries patterns. Within this block, we apply: (i) global multi-head self-attention over all tokens (feasible at the final 7×7 resolution), enabling long-range interactions between fissure regions, proximal contact areas, and dentin–enamel junctions; and (ii) a depthwise convolutional MLP (DWConv-MLP), which injects local 3×3 spatial context into the channel-wise mixing step. Both the attention output and the DWConv-MLP output are incorporated via quality-gated residual connections. Specifically, an image-level quality score $q \in [0, 1]$ is passed through a lightweight gating function to produce scaling factors that modulate the residual pathways. High-quality radiographs, therefore, receive stronger refinement (larger gate values), whereas lower-quality images are updated more conservatively, reducing the risk of amplifying noise or acquisition artifacts. This design makes the refinement stage explicitly quality-aware and consistent with the DQ-SAM optimization strategy employed during training. After attention refinement, the token sequence is reshaped back into a spatial feature map and passed through adaptive average pooling, which aggregates spatial information into a compact feature vector. A final fully connected layer maps this representation to three output logits corresponding to the radiographically diagnosable G.V. Black classes (I–III). Overall, the proposed architecture preserves the hierarchical representation strengths of Swin Transformers while introducing a simple, domain-guided attention mechanism that integrates image-quality cues directly into the feature learning process. This combination yields a stable and interpretable backbone tailored for intraoral periapical caries classification. The overall architecture of the proposed model is illustrated in Figure 5.

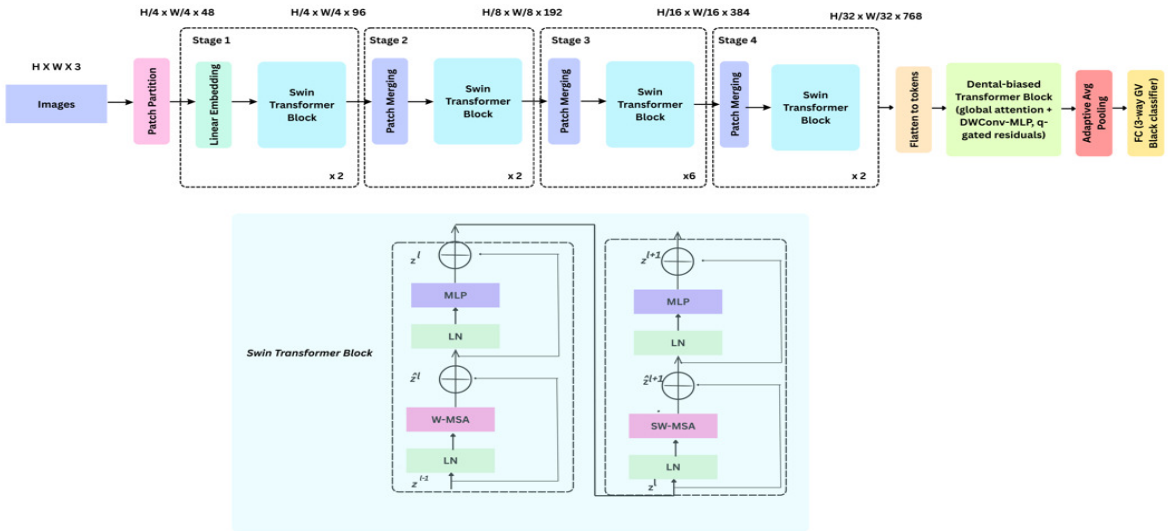

Figure 5: Overall architecture of the proposed DQ-SAM Swin-Tiny model

3.5. Model Compilation and Training Process

We trained the proposed Swin-Tiny–based classifier using a simple but stable pipeline designed for small medical datasets. As no separate validation folder was predefined, model validation was performed using an internal 80:20 split of the development pool with class-stratified sampling and a fixed random seed. This preserved the proportionality of the three diagnostic classes across training and validation subsets. However, the split was not explicitly stratified or logged according to image origin (real versus synthetic). Therefore, although the overall real/synthetic composition of the dataset is known, the exact number of synthetic images included in the validation subset cannot be retrospectively specified. Images were loaded in batches of 16 and resized to (224×224) pixels. For data augmentation, we used light, dental-appropriate transforms only on the training set: horizontal flips, small affine scaling (about $\pm 5\%$), and mild brightness/contrast changes. These augmentations help the model generalize to small variations in positioning and exposure without distorting lesion morphology. The validation set was only resized and normalized. All images were normalized using ImageNet mean and standard deviation, matching the Swin pretraining setup. The network was optimized with AdamW, starting from a base learning rate of 3×10^{-4} and weight decay of 0.05. The loss function was standard cross-entropy (optionally with class weights and label smoothing, both set to zero in our main experiments). To schedule the learning rate, we used a cosine decay with linear warm-up. The learning rate ramps up over the first few epochs and then smoothly decays to a small floor value, which improves stability compared to a constant schedule. Training was run for up to 300 epochs with mixed-precision (AMP) on the GPU, gradient clipping (max norm 5), and checks to skip steps with non-finite gradients. On top of AdamW, we wrapped a DQ-SAM optimizer, a data-quality–gated variant of Sharpness-Aware Minimization (SAM). SAM encourages flat, well-generalizing minima by updating the model to perform well under small worst-case perturbations of the weights. At each step it first finds an adversarial perturbation within a norm ball of radius ρ, then updates the parameters to reduce the loss at the perturbed point. In DQ-SAM, for each mini-batch, the mean quality score $\bar{q}$ is converted into a gate ($g \in [g_{\min}, 1]$). This gate controls both the size of the SAM sharpness perturbation and the practical learning step for that batch. High-quality batches use stronger SAM (larger perturbations) and larger steps; lower-quality batches are updated more gently. Each iteration thus consists of two forward–backward passes (SAM style) followed by a quality-gated AdamW update. Finally, we maintained an exponential moving average (EMA) of the model weights. Validation accuracy and loss were continuously computed using this EMA model, and the best EMA checkpoint on validation accuracy was saved as the final model. An early-stopping rule (after a minimum

of 40 epochs and 100 epochs without improvement) prevented overfitting and unnecessary training once the validation performance had converged.

4. Results and Discussion

4.1. Quantitative analysis

The proposed DQ-SAM Swin-Tiny classifier demonstrated strong performance on the held-out test set of 360 periapical radiographs. Using test-time augmentation (TTA), the model achieved an overall accuracy of 91.94%, with balanced performance across all three G.V. Black classes. The overall test-set performance of the proposed classifier, including per-class and macro-averaged metrics, is summarized in Table 1. Per-class accuracy was 91.67% for Class I, 94.17% for Class II, and 90.00% for Class III, indicating that the model is consistently able to distinguish occlusal, proximal, and anterior lesions despite subtle morphological differences on radiographs. The classification report shows high precision, recall, and F1-scores across classes (macro-F1 = 0.9202). Class III lesions, which are often the most visually subtle on radiographs, achieved an F1-score of 0.9391, reflecting the model's ability to capture fine structural cues. Class II caries, which can be clinically missed due to overlapping contacts, achieved a recall rate of 0.9417, suggesting the model's potential to reduce false negatives in routine diagnosis. ROC–AUC analysis further confirms strong discriminative ability, with class-wise AUC values of 0.9872, 0.9732, and 0.9943 for Classes I–III. Both micro- and macro-averaged AUCs (> 0.98) demonstrate excellent separability between caries categories. Clinically, this level of discrimination implies that the model responds reliably to early radiographic changes and maintains low false-positive and false-negative tendencies across lesion types.

Table 1: Summary of classification performance

Metric	Class I	Class II	Class III	Overall/Average
Per-class Accuracy	91.67%	94.17%	90.00%	91.94% (overall)
Precision	0.9402	0.8496	0.9818	Macro: 0.9239
Recall	0.9167	0.9417	0.9000	Macro: 0.9194
F1-score	0.9283	0.8933	0.9391	Macro: 0.9202
ROC–AUC	0.9872	0.9732	0.9943	Micro: 0.9839, Macro: 0.9849

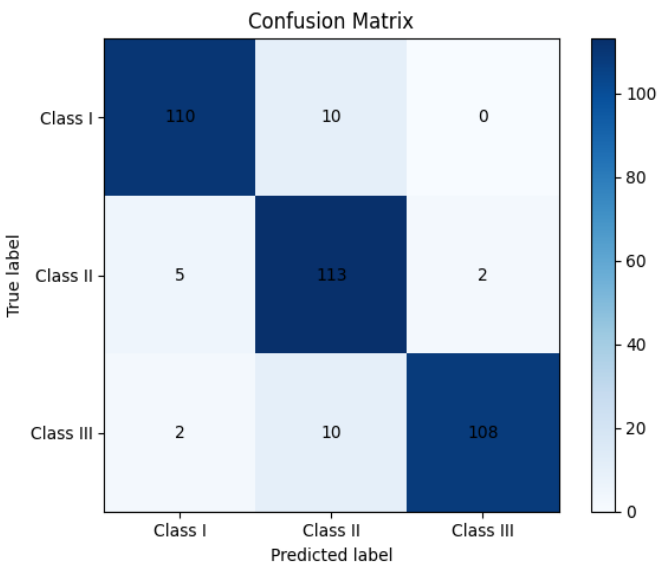

Figure 6: Confusion matrix for G.V. Black Class I–III predictions

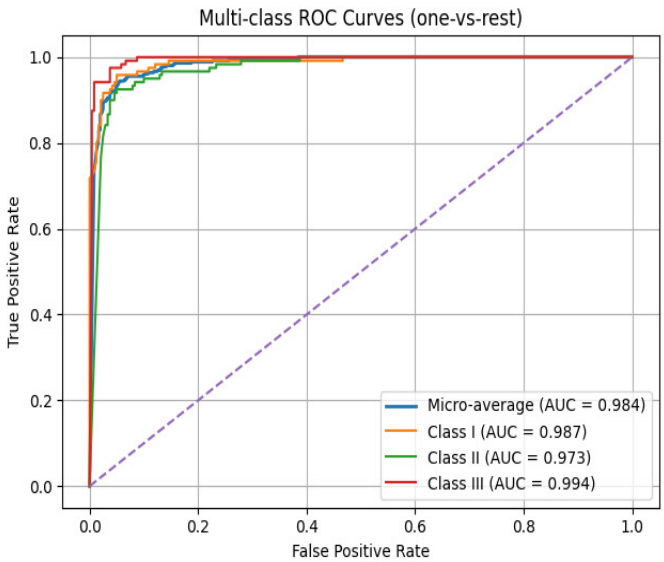

Figure 7: One-vs-rest ROC curves for the three G.V. Black classes

The confusion matrix (Figure 6) provides a detailed view of class-wise errors and shows that the model maintains highly reliable predictions across all three G.V. Black categories. Class I achieved 110 correct classifications out of 120, with only 10 instances misidentified as Class II and none misclassified as Class III. Class II exhibited the strongest diagonal dominance (113/120), with minimal confusion toward neighbouring categories (5 $\rightarrow$ Class I, 2 $\rightarrow$ Class III). Class III also showed a robust pattern of correct predictions (108/120), with only modest spill-over into Class II (10 cases) and very few mislabels to Class I (2 cases). This distribution indicates that the model captures the characteristic radiographic patterns specific to each lesion type and maintains low cross-class ambiguity, which is an essential property for clinical reliability. Receiver operating characteristic (ROC) curves (Figure 7) further highlight the model's discriminative strength. Class-wise AUC scores were 0.9872 for Class I, 0.9732 for Class II, and 0.9943 for Class III, reflecting excellent separability between lesion categories. Both micro-averaged (0.9839) and macro-averaged (0.9849) AUCs approached 1.0, demonstrating consistent performance across classes and confirming that the model produces well-calibrated decision boundaries. Clinically, such high AUC values imply that even subtle radiographic manifestations of caries can be confidently distinguished, supporting the model's potential as a reliable decision-support tool for early caries detection and treatment planning.

4.2. Quality-Stratified Performance Analysis

To assess robustness under varying radiographic conditions, the test images were grouped into low-, medium-, and high-quality bands using the learned q-scores. The model maintained consistently strong performance across all groups, indicating that DQ-SAM training improved resilience to noise, blur, and contrast variations. Low-quality images ($n = 56$) still achieved 91.07% accuracy, with stable macro-F1 (0.9110). Notably, Class II recall reached 1.000, suggesting that the model avoids missing proximal caries even when visibility is compromised. Medium-quality radiographs, which are the most common in clinical practice, yielded the highest accuracy (92.55%) and strongest overall balance of metrics (macro-F1 = 0.9271), reflecting close alignment between training and real-world quality distribution. High-quality images ($n = 49$) showed slightly lower accuracy (89.80%), likely due to their smaller representation in the dataset and the presence of sharper fine-grained details that introduce greater intra-class variation. Nevertheless, macro-F1 remained high (0.8991), indicating stable discriminative behaviour. Overall, this analysis demonstrates that the model generalizes well across heterogeneous imaging conditions, supporting its suitability for deployment in routine dental workflows where radiograph quality is often inconsistent. These results are summarized in Table 2, and the class-wise F1-scores, precision, and recall across

quality bands are visualized in Figure 8.

Table 2: Quality-stratified metrics across low-, medium-, and high-quality bands

Quality Band	n	Accuracy	Macro Precision	Macro Recall	Macro F1
Low	56	91.07%	0.9090	0.9258	0.9110
Medium	255	92.55%	0.9294	0.9264	0.9271
High	49	89.80%	0.9041	0.8991	0.8991

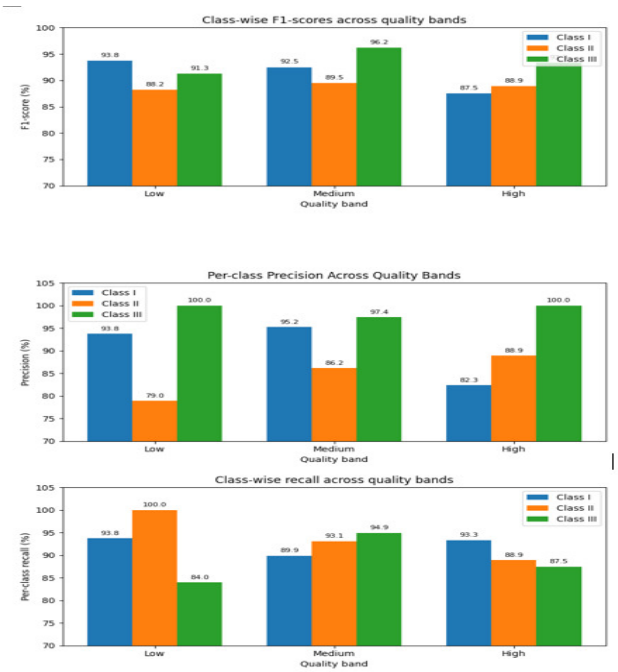

Figure 8: Quality-stratified performance by G.V. Black class

The proposed Dental Swin-Tiny DQ model demonstrates stable and clinically aligned behaviour across diverse radiographic conditions. Its quality-gated transformer block enables the network to adaptively modulate attention and refine features based on image clarity, resulting in consistently high accuracy ($\sim$ 91%) even on low-quality scans. This robustness is reinforced by substantial class-wise precision and recall, minimal clinically severe misclassifications, and near-perfect ROC–AUC values (> 0.97), indicating excellent separation between G.V. Black Class I–III lesions. The confusion patterns show that most errors occur between adjacent severity levels, mirroring fundamental diagnostic ambiguity, while large-gap confusions are rare, reducing risks of overtreatment or missed advanced lesions. Notably, the architecture remains lightweight and operates directly on standard periapical radiographs without requiring segmentation masks, making it feasible for real-time chairside use. By integrating image-quality awareness, reliable probability outputs, and clinically intuitive error modes, the model enhances its suitability as a decision-support tool that assists dentists with early detection, triage, and treatment planning, especially in settings where radiograph quality and operator skill vary.

4.3. Ablation Study: Standard SAM vs. DQ-SAM

To isolate the effect of the proposed quality-gated sharpness-aware optimization, we replaced DQ-SAM with standard SAM while keeping the same Swin-Tiny backbone, q-aware residual architecture, preprocessing, data split, and training schedule unchanged. The standard SAM model achieved 90.56%

accuracy, 90.95% macro precision, 90.56% macro recall, and 90.65% macro F1-score, whereas DQ-SAM achieved 91.94%, 92.29%, 91.94%, and 92.02%, respectively. These gains show that the proposed quality-aware gating provides a measurable benefit beyond standard SAM.

4.4. Comparison with Baseline Models

To contextualize the performance of the proposed quality-aware Swin-Tiny model, we compared it against two strong convolutional baselines: ResNet-50 and DenseNet-121, all trained on the same dataset and evaluated on the identical test split. On the overall test set, DQ-SAM achieved the strongest performance, with 91.94% accuracy and a macro F1-score of 92.02%, compared with 88.06% and 88.25% for ResNet-50, and 91.39% and 91.49% for DenseNet-121, respectively. Bootstrap analysis yielded a 95% confidence interval of 89.17–94.44% for DQ-SAM accuracy and 89.13–94.57% for DQ-SAM macro F1-score. The paired comparison against ResNet-50 was statistically significant. McNemar's exact test showed 25 discordant cases favouring DQ-SAM versus 11 favouring ResNet-50 ($p = 0.0288$). Paired bootstrap analysis also showed significant improvements for DQ-SAM over ResNet-50 in all principal metrics, including +3.84% accuracy (95% CI: +0.56% to +7.22%, bootstrap $p = 0.0233$) and +3.73% macro F1-score (95% CI: +0.53% to +7.03%, bootstrap $p = 0.0207$). In contrast, although DQ-SAM achieved slightly higher point estimates than DenseNet-121, the difference was not statistically significant. McNemar's exact test yielded 13 discordant cases favouring DQ-SAM and 11 favouring DenseNet-121 ($p = 0.8388$). Similarly, paired bootstrap analysis showed only small differences, such as +0.51% accuracy (95% CI: −2.22% to +3.06%, bootstrap $p = 0.794$) and +0.48% macro F1-score (95% CI: −2.18% to +3.03%, bootstrap $p = 0.743$), with all confidence intervals including zero. Quality-stratified analysis shows that the proposed model is consistently more robust in the challenging low- and high-quality regimes: in the low-quality bin, it reached 91.07% accuracy (macro F1 91.10%), compared with 89.29% (DenseNet-121) and 85.71% (ResNet-50); in the high-quality bin, it achieved 89.80% accuracy versus 85.71% and 83.67%, respectively. DenseNet-121 attained a marginally higher accuracy than the proposed model in the medium-quality bin (92.94% vs. 92.55%), but its performance dropped more notably at the extremes of the quality spectrum. However, these subgroup-level comparisons were not emphasized as the primary inferential analysis because the subgroup sample sizes were substantially smaller than the full test set, particularly in the low- and high-quality bins. This reduced the statistical power of McNemar and bootstrap subgroup analyses and led to wide confidence intervals with bin-level differences that did not reach statistical significance. Therefore, the bin-wise findings are presented as supportive descriptive evidence of robustness across image-quality levels rather than as definitive subgroup-level superiority claims. The DQ-SAM Swin-Tiny maintains a smoother performance profile across low, medium, and high q-bands, suggesting that the combination of transformer-based features and quality-gated optimization better stabilizes learning under heterogeneous radiographic conditions. From a clinical perspective, this behaviour is desirable. The model maintains high accuracy even when images are noisy or suboptimal, a condition where traditional CNN baselines degrade more sharply, making it a more reliable candidate for real-world deployment in routine dental practice. Figure 9 compares the proposed model with DenseNet-121 and ResNet-50 across the three quality bands in terms of accuracy, macro precision, macro recall, and macro F1-score.

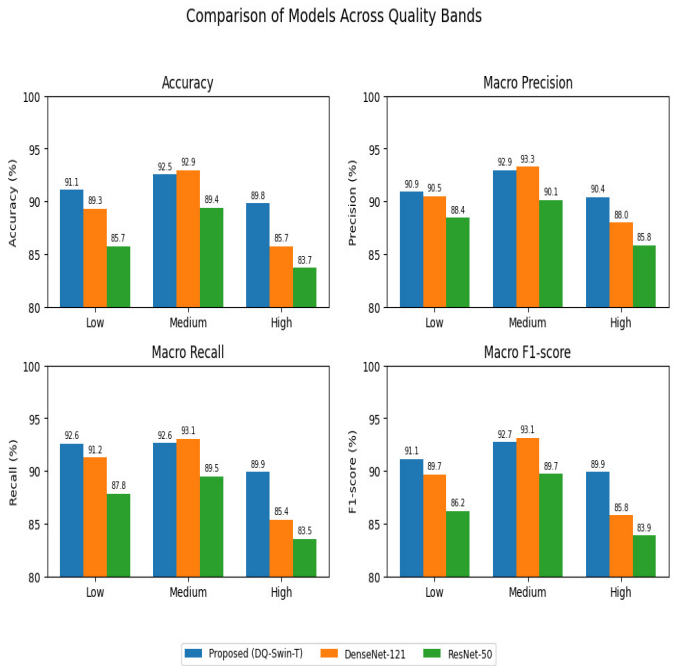

Figure 9: Comparison of models across image-quality bands

5. Conclusion

This study presented a quality-aware Swin Transformer classifier for automated G.V. Black Class I–III caries identification on periapical radiographs, combining a lightweight backbone with a dental-biased attention block and a quality-gated optimization strategy. The proposed automated caries detection framework may have practical value as a decision-support tool in routine dental radiographic assessment. By assisting in the earlier identification of subtle proximal or occlusal lesions on intraoral periapical radiographs, the system may support timely preventive or minimally invasive intervention before lesion progression. From a clinical safety perspective, false negatives are particularly important because missed carious lesions may remain untreated and advance to deeper structural involvement, potentially increasing the need for restorative or endodontic procedures. Therefore, in real-world deployment, the model should be used as an adjunct to, rather than a replacement for, clinician judgment, with emphasis on maintaining high sensitivity for screening purposes. In practice, such a model could be integrated into digital dental imaging workflows as a chairside triage or second-reader tool, including possible incorporation into radiology viewing software or PACS-like dental image management systems, where it may help flag suspicious regions for closer review and improve reporting consistency. The model achieved strong overall performance, stable class-wise behaviour, and high ROC–AUC values, showing that integrating image-quality cues helps maintain accuracy even when radiographs vary in sharpness or exposure. However, the work has a few limitations, such as the dataset, which, although carefully curated, is single-centre and limited to three radiographic classes. Accordingly, the generalizability of the model to multi-centre or multi-scanner settings remains to be established. In addition, the current system performs only global classification without explicit lesion localization, which may restrict its interpretability in borderline cases. Future work will focus on expanding the dataset across institutions, incorporating full G.V. Black Classes I–VI, computing quality estimates directly on test images, and extending the framework to joint detection–classification models with saliency-based explanations. These steps may support broader clinical deployment and improve the reliability of AI-assisted diagnosis in routine dental workflows.

Acknowledgments

The authors would like to express their sincere gratitude to Dr. Sagi Sai from SIBAR Institute of Dental Sciences, Guntur, for valuable support in the data collection process and for providing expert guidance on dental anatomy and caries classification.

References

1. Marsh, P. D., Zaura, E., *Dental biofilm: ecological interactions in health and disease*, Journal of Clinical Periodontology, 44, S12–S22, (2017).
2. Yu, O. Y., Lam, W. Y.-H., Wong, A. W.-Y., Duangthip, D., Chu, C.-H., *Nonrestorative management of dental caries*, Dentistry Journal, 9(10), 121, (2021).
3. Pitts, N. B., Twetman, S., Fisher, J., Marsh, P. D., *Understanding dental caries as a non-communicable disease*, British Dental Journal, 231(12), 749–753, (2021).
4. Chen, J., Chen, W., Lin, L., Ma, H., Huang, F., *The prevalence of dental caries and its associated factors among preschool children in Huizhou, China: a cross-sectional study*, Frontiers in Oral Health, 5, 1461959, (2024).
5. Dunleavy, G. et al., *Inequalities in oral health: estimating the longitudinal economic burden of dental caries by deprivation status in six countries*, BMC Public Health, 24(1), 3239, (2024).
6. Tyas, M. J., Anusavice, K. J., Frencken, J. E., Mount, G. J., *Minimal intervention dentistry—a review (FDI Commission Project 1–97)*, International Dental Journal, 50(1), 1–12, (2000).
7. Schwendicke, F., Cejudo Grano de Oro, J., Garcia Cantu, A., Meyer-Lueckel, H., Chaurasia, A., Krois, J., *Artificial intelligence for caries detection: value of data and information*, Journal of Dental Research, 101(11), 1350–1356, (2022).
8. Lee, S., Oh, S.-il, Jo, J., Kang, S., Shin, Y., Park, J.-won, *Deep learning for early dental caries detection in bitewing radiographs*, Scientific Reports, 11(1), 16807, (2021).
9. Musri, N., Christie, B., Ichwan, S. J. A., Cahyanto, A., *Deep learning convolutional neural network algorithms for the early detection and diagnosis of dental caries on periapical radiographs: A systematic review*, Imaging Science in Dentistry, 51(3), 237, (2021).
10. Chen, X., Guo, J., Ye, J., Zhang, M., Liang, Y., *Detection of proximal caries lesions on bitewing radiographs using deep learning method*, Caries Research, 56(5–6), 455–463, (2022).
11. Zanini, L. G. K., Rubira-Bullen, I. R. F., Nunes, F. de L. dos S., *A systematic review on caries detection, classification, and segmentation from X-ray images: methods, datasets, evaluation, and open opportunities*, Journal of Imaging Informatics in Medicine, 37(4), 1824–1845, (2024).
12. Bhat, S., Birajdar, G. K., Patil, M. D., *A comprehensive survey of deep learning algorithms and applications in dental radiograph analysis*, Healthcare Analytics, 4, 100282, (2023).
13. Schneider, L., Krasowski, A., Pitchika, V., Bombeck, L., Schwendicke, F., Buettner, M., *Assessment of CNNs, transformers, and hybrid architectures in dental image segmentation*, Journal of Dentistry, 156, 105668, (2025).
14. Alsakar, Y. M., Elazab, N., Nader, N., Mohamed, W., Ezzat, M., Elmogy, M., *Multi-label dental disorder diagnosis based on MobileNetV2 and Swin Transformer using bagging ensemble classifier*, Scientific Reports, 14(1), 25193, (2024).
15. Singh, P., Sehgal, P., *GV Black dental caries classification and preparation technique using optimal CNN-LSTM classifier*, Multimedia Tools and Applications, 80(4), 5255–5272, (2021).
16. Parathanath, A. J., Manimaran, A., *CBMNet: a dual-attention enhanced ConvNeXt model for accurate GV Black type I–III classification in intraoral periapical radiographs*, Scientific Reports, 15(1), 39287, (2025).
17. Lee, J.-H., Kim, D.-H., Jeong, S.-Nyum, Choi, S.-H., *Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm*, Journal of Dentistry, 77, 106–111, (2018).
18. ForouzeshFar, P., Safaei, A. A., Ghaderi, F., Hashemikamangar, S. S., *Dental caries diagnosis from bitewing images using convolutional neural networks*, BMC Oral Health, 24(1), 211, (2024).
19. Chaves, E. T. et al., *Detection of caries around restorations on bitewings using deep learning*, Journal of Dentistry, 143, 104886, (2024).
20. Prados-Privado, M., García Villalón, J., Martínez-Martínez, C. H., Ivorra, C., Prados-Frutos, J. C., *Dental caries diagnosis and detection using neural networks: a systematic review*, Journal of Clinical Medicine, 9(11), 3579, (2020).
21. Zhou, X., Yu, G., Yin, Q., Yang, J., Sun, J., Lv, S., Shi, Q., *Tooth type enhanced transformer for children caries diagnosis on dental panoramic radiographs*, Diagnostics, 13(4), 689, (2023).
22. Gao, Y., Zhang, P., Xie, Y., Han, J., Zeng, L., Ning, N., Zheng, Q., Li, H., Chen, X., Chen, Z., *Application of transformers in stomatological imaging: A review*, Digital Medicine, 10(3), e24–00001, (2024).
23. Hassan, M., Vakanski, A., Zhang, B., Xian, M., *Do Sharpness-Based Optimizers Improve Generalization in Medical Image Analysis?*, IEEE Access, 13, 82972–82985, (2025).

24. Hassan, M., Vakanski, A., Zhang, B., Xian, M., *GCSAM: Gradient Centralized Sharpness Aware Minimization*, arXiv:2501.11584, (2025).

Aneetta Joy Parathanath,
Department of Mathematics,
School of Advanced Sciences,
VIT-AP University, Beside AP Secretariat, Amaravati, 522241, Andhra Pradesh
India.
E-mail address: aneetta.23phd7154@vitap.ac.in

and

A. Manimaran,
Department of Mathematics,
School of Advanced Sciences,
VIT-AP University, Beside AP Secretariat, Amaravati, 522241, Andhra Pradesh,
India.
E-mail address: manimaran.a@vitap.ac.in