Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Efficient Dual Branch Multi-Scale CNN With Reconstruction Learning and Explainable AI for Dental Caries Classification

Shashikala K.S.[1*], Umamaheswaran S.[2]

[1]Research Scholar, Department of Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India.
[2]Research Supervisor, Department of Computer Science and Engineering, New Horizon College of Engineering, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka, India.

E-mail: [1*]shashikalaks.nhce@newhorizonindia.edu, [2]dr.umamaheswaran.nhce@newhorizonindia.edu

Orcid ID: [1*]0009-0003-2644-1381, [2]0000-0003-1941-9605

Abstract

Preventive oral healthcare depends on the early detection of dental caries. Delicate visual differences between healthy tooth and early-stage lesions make automated disease diagnosis difficult. This work introduces a comparative analysis between a Single-Branch Convolutional Neural Network (SBCNN) and a proposed Dual-Branch Multi-Scale Convolutional Neural Network (DBMSCNN) for multiple class dental caries classification. It integrates Sharpness-Aware Minimization (SAM), multi-scale feature extraction, and self-supervised reconstruction learning to enhance feature robustness and generalization. We utilized a three-class dental caries dataset consisting of No Enamel Caries, Advanced Caries and Early Stage Caries. A total of 447 samples were used for testing in each fold. The proposed DBMSCNN framework achieved a mean classification accuracy of 92.43% across five-fold stratified cross-validation. The model demonstrated strong discriminative capability and a low false-positive rate, with a mean specificity of 96.14% and a mean precision of 92.59%. Reconstruction-based self-supervised pretraining improved anatomical feature learning, resulting in mean Dice and IoU scores of 0.8530 and 0.7577 respectively. Feature-level analysis using t-SNE is done for qualitative visualization of latent feature distributions for the proposed model, while Grad-CAM and saliency maps show meaningful and moderate localization of carious regions. The comparative results validate the effectiveness of multi-scale learning, reconstruction-based regularization, and SAM optimization in improving diagnostic reliability for dental caries classification.

Keywords: Dental Disease Diagnosis, Reconstruction Learning, Sharpness Aware Minimization, Multi Scale Learning.

1. Introduction

Teeth play an essential role in daily human life. Tooth care is a lifelong responsibility essential for maintaining oral and overall health. Oral health deteriorates due to poor eating and

*** Corresponding Author**

DOI: https://doi.org/10.36548/jiip.2026.3.014
Received: 07.05.2026, received in revised form: 12.06.2026, accepted: 03.07.2026, published: 16.07.2026

lifestyle habits. As a result, it becomes very essential to regularly check oral health and carry out proper oral care and cleaning [1]. Dental caries is an infectious disease produced by acids from the metabolism of bacteria. They penetrate into the enamel and dentin and dissolve the mineral. Organic acids are produced as a by-product of the metabolism of fermentable carbohydrates by the bacteria. The caries process results from multiple demineralization and remineralization cycles. At the atomic level, demineralization begins at the solid surface inside the enamel. It can extend, unless stopped, with the final point being cavities [2].

Traditional methods of diagnosis rely heavily on visual inspection by dentists. It can be prone to inter-observer variability, especially in early-stage lesions. In practice, detecting dental decay can be the toughest task for dentists, due to the altering lesion characteristics at various stages of development. Initial stages can be handled using non-invasive treatment methods such as changes in diet and by improving oral hygiene. In the later stages, dental cavities can damage the tooth structure and require restorative management. These stages must be clearly distinguished to identify appropriate dental caries management strategies [3].

Deep learning has shown promising accuracy in image classification tasks. However, typical CNN architectures often cope with:

- Multi-scale feature representation, especially when lesions vary in size.[4]
- Overfitting in imbalanced datasets.[5]
- Lack of explainability, which restricts clinical trust and integration.[5]

This study utilizes a multi-scale dual-branch CNN architecture [6] with reconstruction based self-supervised learning (SSUL), explainable AI, and sharpness-aware minimization to overcome these challenges.

Feature extraction can be regularized using reconstruction-based self-supervised learning [7]. P.Huang et al [8]., integrated an SSUL method using an undersampling strategy to improve reconstruction performance [8]. SSUL tends to reduce the dependency on large labeled datasets[9]. Self-supervised models learn a better input representation using supervisory signals of the partial input. They predict the unobserved input part by harnessing the underlying structure in the data [10]. Spyros Gidaris et al [11] developed a self-supervised learning scheme that captures unique visual features useful for image classification.

The sharpness of the loss surface was examined using many techniques. Through SAM, small perturbations were applied to the model parameters before updating them to achieve a flatter loss landscape. First, SAM identifies a set of weights where the loss is maximized around the current weights. Next, SAM updates the model parameters to minimize the model loss based on the gradient of the loss [12].

In the study named GA-SAM, the relationship between local minima and generalization ability to learn algorithms that find flat minima that generalize better was examined [13]. The additional training cost associated with SAM could be reduced by periodically performing inner gradient ascent [14]. In FSAM, SAM's perturbation was broken down into full gradient and stochastic gradient noise components to reduce the adverse impact of the full gradient component [15]. Z-Score Filtered Sharpness-Aware Minimization, was proposed by Vincent-Daniel Yun [16]. Here, Z-score based filtering was applied to each layer gradient to improve test accuracy. A hybrid classifier was developed to capture and classify fine-grained textures and variations in [17].

2. Related Work

Dental caries remains a major public health concern worldwide. Early detection of enamel caries is critical for avoiding invasive treatment processes. Traditional diagnosis based on visual inspection and radiographic interpretation is highly dependent on dentist expertise and is often subjective, particularly in distinguishing early-stage enamel caries from healthy enamel. Mild texture variations, inter-patient variability, and illumination differences complicate reliable diagnosis. Though existing literature shows significant progress in automated dental caries detection, a few limitations exist, as listed below:

2.1 Limited Multi-Scale Feature Modeling

Methods based on single-branch CNN architectures inadequately capture global structural patterns and fine-grained local textures required for differentiating early enamel demineralization from advanced caries.

2.2 Insufficient Use of Self-Supervised Learning

Current approaches, based on fully supervised learning, require large amounts of annotated data. Self-supervised reconstruction learning has the ability to improve feature generalization and reduce overfitting in dental imaging. Gruber et al. [18] proposed a SSUL framework called Noisier2Inverse, that performs image reconstruction without ground-truth data or assuming uncorrelated noise. This work was extended by Charles Millard et al[19] by proposing a loss weighting that compensates for the variable sampling and partitioning densities. Burhaneddin Yaman et al [20] proposed SSUL through data undersampling for deep learning reconstruction.

2.3 Lack of Explainability in Clinical Contexts

Existing studies based on deep learning methods do not provide a visual explanation of their predictions. This includes identifying the features examined by the models and understanding the reasons for incorrect predictions and overfitting [21]. This decreases clinical trust and restricts usage in medical settings.

2.4 Poor Generalization and Optimization Stability

Standard optimization methods often converge to sharp minima, leading to poor generalization on unseen data. Fewer studies have utilised Sharpness-Aware Minimization (SAM) for dental image classification in the literature.

2.5 Inadequate Feature Space Analysis

Fewer research has focused on learned feature distributions to validate class separability, particularly for challenging cases such as early-stage caries [22]. A considerable gap was reported between validation set performance and test set performance for all data splitting methods. These gaps underscore the need for an integrated framework that simultaneously addresses generalization, feature representation, and interpretability in dental caries classification. XAI techniques synthesize feature heatmap images to identify the features considered by a model. Within the test image, These features are represented with different colors, based on their impact on the final decision-making process [23]. Existing convolutional neural network (CNN) based approaches suffer from limitations like poor robustness to image variability, and a lack of interpretability. Therefore, there is a necessity for a robust and explainable deep learning framework capable of classifying dental caries into clinically meaningful stages while providing meaningful decision-making support to clinicians.

3. Methodology

The methodology used in this paper employs the single-branch CNN baseline and dual-branch CNN architectures.

3.1 Single-Branch CNN Baseline

The single-branch CNN baseline model accepts RGB dental panoramic images, resized to 224 × 224 pixels, as input. The architecture consists of three sequential convolutional feature extraction blocks. In the first block, a 3 × 3 convolutional layer with 32 filters and padding of 1 is applied to preserve spatial resolution. This is accompanied by batch normalization to stabilize training, a ReLU activation function for non-linearity, and a 2 × 2 Max Pooling layer. The number of filters is increased to 64 in the second block, and the channel depth is increased to 128 in the third block. The same kernel size, normalization, activation, and pooling strategy are maintained in the SBCNN.

An Adaptive Average Pooling layer is applied to compress the final feature maps into a fixed 1 × 1 spatial representation, after feature extraction, producing a 128-dimensional feature vector. The pooled feature vector is then passed to a fully connected linear classification layer that maps the 128-dimensional representation to the three output classes.

To improve generalization, the model is trained employing cross-entropy loss with label smoothing of 0.1. The AdamW optimizer is employed with a weight decay of 1×10^{-4} and a learning rate of 1×10^{-4} to provide stable optimization. The model is trained for 25 epochs using an 80:20 train-test split. The batch size is set to 16. Data augmentation strategies, namely, image resizing to 224 × 224, random horizontal flipping, and random rotation of ±10 degrees, are applied to improve balanced learning.

3.2 Proposed Dual-Branch Multi-Scale CNN Architecture and Workflow

The proposed DBMSCNN framework is designed as an end-to-end system for automated dental caries classification with improved feature representation, generalization, and explainability. Figure 1 shows the proposed architecture. The proposed Dual-Branch Multi-Scale Convolutional Neural Network (DBMSCNN) framework is designed for automated dental disease classification from panoramic radiographs. The input to the system is an RGB dental panoramic image resized to 224 × 224 pixels. Prior to training, data augmentation techniques including random horizontal flipping and random rotation are applied to improve model generalization and reduce overfitting.

The augmented image is simultaneously supplied to a dual-branch encoder consisting of two parallel feature extraction pathways; the Local Branch and the Global Branch. Both branches follow a similar architectural design but employ different convolution kernel sizes, enabling the extraction of complementary local and global information from dental radiographs.

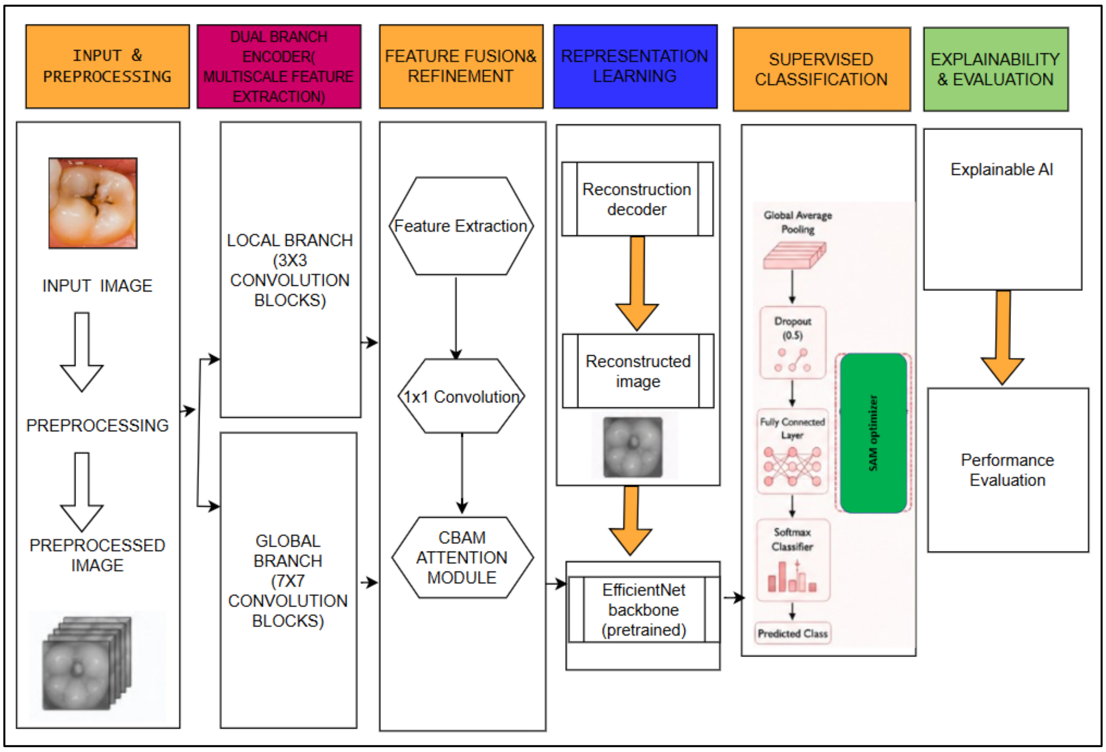

Figure 1. Proposed DBMSCNN architecture

The Local Branch is responsible for capturing fine-grained lesion characteristics such as early enamel demineralization, small cavity boundaries and subtle intensity variations. Since these abnormalities occupy relatively small spatial regions, the branch employs 3 × 3 convolution kernels. The branch consists of two sequential convolutional blocks. Each block comprises a convolutional layer, Batch Normalization, ReLU activation, and 2 × 2 Max Pooling. The first convolution layer transforms the input from 3 channels to 32 channels, while the second increases the feature depth from 32 to 64 channels. Through hierarchical feature extraction, the Local Branch learns discriminative representations associated with early-stage pathological changes.

Let the input image be represented as $X \in R^{224x224x3}$ where X denotes an RGB dental panoramic image resized to 224×224 pixels. The local branch captures fine-grained lesion characteristics using small convolution kernels. (Eqn.1)

$$F_L = \mathcal{B}_{3\times3}(X) \tag{1}$$

where:

- F_L = local feature map
- $\mathcal{B}_{3\times3}$ = two stacked convolution blocks using 3×3 kernels.

The Global Branch is designed to capture broader anatomical and contextual information, including tooth arrangement, lesion spread, structural relationships, and large-scale radiographic patterns associated with disease progression. To achieve a larger receptive field, this branch utilizes 7 × 7 convolution kernels. Similar to the Local Branch, it consists of two convolutional blocks incorporating Batch Normalization, ReLU activation, and Max Pooling. The first convolutional layer generates 32 feature channels, while the second expands the representation to 64 channels. The larger kernels enable effective modeling of contextual dependencies that may not be captured by smaller filters.

Following parallel feature extraction, the outputs of the Local and Global branches are fused through channel-wise concatenation. The resulting fused representation is passed through a dedicated fusion block comprising a 3 × 3 convolutional layer with 128 output channels, batch normalization, ReLU activation, and 2 × 2 max pooling. This fusion mechanism integrates complementary local and global information while suppressing redundant activations. The Global Branch captures broader contextual information using larger receptive fields (Eqn.4-6)

$$F_G = \mathcal{B}_{7\times7}(X) \tag{2}$$

where:

- FG denotes global feature maps
- $\mathcal{B}_{7\times7}$ represents two stacked convolution blocks using 7×7 kernels.

Outputs of both branches are concatenated channel-wise:

$$\mathrm{F_{fusion}} = \mathrm{Concat}(\mathrm{F_L}, \mathrm{F_G}) \tag{3}$$

The fused representation is processed through a fusion layer:

$$F_F = \mathcal{B}_{3\times3}\left(F_{fusion}\right) \tag{4}$$

$$F_f = \phi\left(W_f * F_{fusion} + b_f\right) \tag{5}$$

$$\phi(\cdot) = MaxPool\left(ReLU\left(BN(\cdot)\right)\right) \tag{6}$$

The fused feature representation is projected into a three-channel tensor using a 1×1 convolution (Eqn.7)

$$\mathrm{FR=Conv1\times1(FF)} \tag{7}$$

where FR is the reduced 3-channel representation supplied to EfficientNet-B0.

The fused feature representation is projected into a three-channel feature map using a 1×1 convolution and subsequently forwarded to the EfficientNet-B0 encoder for hierarchical feature extraction: (Eqn.8)

$$FE = EfficientNet(Ff) \tag{8}$$

where FE denotes deep semantic features extracted by EfficientNet-B0.

Feature representation is enhanced by incorporating a Convolutional Block Attention Module (CBAM) after the EfficientNet-B0 feature extraction stage. It is a lightweight attention mechanism that sequentially applies channel attention and spatial attention to enhance diagnostically relevant features while suppressing irrelevant information. By adaptively refining both feature channels and spatial locations, CBAM enhances the discriminative capability of the learned representations before classification (Eqns.9-12)

The Convolutional Block Attention Module (CBAM) is applied to refine EfficientNet features.

$$M_c(F_E) = \sigma\left(MLP\left(AvgPool(F_E)\right) + MLP\left(MaxPool(F_E)\right)\right) \quad (9)$$

$$F_c = M_c(F_E) \odot F_E \quad (10)$$

$$M_s(F_c) = \sigma\left(f^{7\times7}([Avg(F_c), Max(F_c)])\right) \quad (11)$$

$$F_A = M_s(F_c) \odot F_c \quad (12)$$

Mc denotes the channel attention map, which adaptively emphasizes informative feature channels and suppresses less relevant channels. Ms denotes the spatial attention map, which highlights diagnostically important spatial regions within the feature map. Together, these attention mechanisms enable the CBAM module to enhance feature discriminability prior to classification.

To enhance representation learning, an auxiliary reconstruction decoder is incorporated during the pretraining stage. The decoder receives the CBAM-refined EfficientNet feature representation and reconstructs the input image through a sequence of transposed convolution layers. The proposed framework incorporates an auxiliary decoder during self-supervised pretraining. The decoder reconstructs the original image from the learned feature representation (Eqn.13)

$$\hat{X} = D(F_A) \quad (13)$$

where

- D(·) denotes the reconstruction decoder
- $\hat{X}$ is the reconstructed image.

The reconstruction objective is defined using Mean Squared Error (MSE)(Eqn.14)

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \hat{X_i}\right)^2 \quad (14)$$

where N denotes the number of pixels.

The reconstruction stage enables the encoder to learn robust latent representations before supervised classification.

3.2.1 Two-Stage Training Strategy

The proposed framework adopts a two-stage optimization procedure consisting of self-supervised reconstruction pretraining followed by supervised fine-tuning.

In the first stage, the entire DBMSCNN framework, including the dual-branch feature extractor, feature fusion module, EfficientNet encoder, CBAM attention module, and decoder, is trained using a reconstruction objective. Given an input dental radiograph, the decoder reconstructs the original image from the learned latent representations, and the network parameters are optimized using Mean Squared Error (MSE) loss. This self-supervised learning stage enables the model to capture meaningful anatomical and pathological structures without requiring class labels.

In the second stage, the pretrained network is fine-tuned for dental disease classification. Cross-Entropy Loss with label smoothing (0.1), is used to optimize the classification network. Parameter updates are performed using the Sharpness-Aware Minimization (SAM) optimizer with AdamW as the base optimizer.

Furthermore, Sharpness-Aware Minimization (SAM) is employed during supervised training to improve model generalization by seeking flatter minima in the optimization landscape, thereby enhancing classification robustness across different data folds.

After attention refinement, Adaptive Average Pooling generates a compact feature vector (Eqn.15)

$$f = [f_1, f_2, \dots, f_d]^T \tag{15}$$

where d=1280 corresponds to the EfficientNet feature dimension.

The fully connected classifier transforms the feature vector into logits: (Eqn.16)

$$Z=Wf+b \tag{16}$$

$$z = [z_1, z_2, \dots, z_K]^T$$

Where K denotes the number of disease classes.

The Softmax layer converts logits into class probabilities. (Eqn.17)

$$P(y = i|f) = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)} \tag{17}$$

The final prediction is obtained using Eqn.18.

$$\hat{y} = \arg \max_{i \in \{1,\dots,K\}} P\,(\,y = i \mid f\,) \tag{18}$$

Cross entrophy loss is depicted in Eqn.19.

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{i=1}^{K} y_i \log(P_i) \tag{19}$$

where $p_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}$

Sharpness-Aware Minimization (SAM) is incorporated during supervised classification training to reduce overfitting. SAM searches for flatter minima in the loss landscape rather than sharp local minima, resulting in better robustness and improved performance on unseen test samples. In the proposed framework, SAM is applied only during the supervised classification stage and not during reconstruction learning (Eqn.20)

$$w_{SAM} = \arg \min_{w} \max_{|\epsilon|_2 \le \rho} \mathcal{L}(w + \epsilon) \tag{20}$$

where:

- w denotes model parameters.

- ϵ represents an adversarial perturbation applied to the weights.
- ρ controls the perturbation radius.
- $L(\cdot)$ denotes the training loss.

Finally, Explainable Artificial Intelligence (XAI) techniques are integrated into the system to improve clinical interpretability. Grad-CAM and saliency maps are generated from the trained model. They highlight the image regions that contribute most strongly to the final classification decision. These explanation maps validate whether the model prioritizes clinically relevant caries regions rather than irrelevant background structures. Quantitative metrics such as Dice coefficient, Intersection over Union (IoU) and Structural Similarity Index Measure (SSIM) are computed to ensure systematic evaluation of performance, making the proposed DBMSCNN framework suited for automated dental diagnosis.

Confusion matrix analysis, ROC-AUC, and t-SNE feature visualization were used to evaluate the proposed framework. This demonstrated the classification performance and class separability, confirming that and the dental images are effectively classified by the model.

3.2.2 Dataset and Preprocessing

The Caries Spectra dataset [24] was used in this study. The curated version of the dataset consisted of 2234 dental radiographic images belonging to three classes: Advance Enamel Caries (800 images), Early Stage Enamel Caries (800 images), and No Enamel Caries (634 images). An 80:20 hold-out validation strategy was employed, resulting in 1787 images for training. For each fold of the five-fold stratified cross-validation protocol, approximately 446–447 images were used for testing. Resizing and augmentation of the images were done using random horizontal flipping, random rotation (±10°) and tensor normalization.

Unbiased evaluation was ensured using an 80:20 train–test split. A baseline implementation used a single-branch convolutional neural network (SBCNN). The model consists of a single sequential convolutional pathway with progressively increasing channel depth, followed by a fully connected classification layer and global average pooling.

An SBCNN trained using supervised learning without reconstruction pretraining or SAM optimization was implemented as a baseline. It did not employ parallel feature extraction, multi-scale fusion, reconstruction pretraining, or SAM. The model used identical data splits, preprocessing, training epochs, and evaluation metrics to ensure fair comparison. The dataset was divided into five stratified folds to preserve class distribution across all subsets. The proposed DBMSCNN model was trained and evaluated across all folds, and fold-wise accuracy, mean accuracy, and standard deviation were reported. Sharpness-Aware Minimization (SAM) is integrated with multi-scale feature extraction and self-supervised reconstruction learning to enhance generalization and discriminative capability.

3.2.3 Explainable AI (XAI)

Two XAI techniques are applied to ensure clinical interpretability:

1. Grad-CAM: Highlights class-discriminative regions in the final convolutional feature map.

2. Saliency Maps: Depicts Pixel-level sensitivity for class predictions.

Gradients are controlled to avoid interference with evaluation metrics, ensuring correct attribution maps.

The procedures involved in reconstruction learning, SAM-based optimization and explainability generation are described in algorithm 1, algorithm 2 and algorithm 3.

Algorithm 1: Encoder–Decoder reconstruction

Input:
Training images X
Output:
Reconstructed images, $\hat{X}$
Optimized encoder-decoder parameters θ_e, θ_d
1. Resize input images to 224×224.
2. Apply augmentation and tensor conversion.
3. Pass X through local branch:
$F_l = B_{local}(X)$
4. Pass X through global branch:
$F_g = B_{global}(X)$
5. The feature maps are fused as:
$F = [F_l \oplus F_g]$
6. A 1×1 convolution is used to reduce the fused feature channels
7. Extract deep feature maps using the EfficientNet-B0 encoder
E=EfficientNet(F)
8.Refine encoder features using CBAM
Ecbam=CBAM(E)
9.Generate the classification embedding
Z=GAP(Ecbam)
10.Reconstruct the input image from the CBAM-refined feature maps
$\hat{X}$=Decoder(Ecbam)
11.Upsample the reconstructed image to 224×224 using bilinear interpolation.
12.Compute the reconstruction loss

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N} |X_i - \widehat{X_i}|_2^2$$

where

- N = batch size,
- X_i = original image,
- $\hat{X}_i$= reconstructed image,
- $|\cdot|_2^2$ denotes the squared Euclidean norm over all pixels and channels.

13.Backpropagate the reconstruction loss through the decoder, CBAM, EfficientNet encoder, channel-reduction layer, feature-fusion module, and both CNN branches.
14.Update the network parameters using the AdamW optimizer.
15.Repeat until the reconstruction loss converges.

Algorithm 2: SAM based classification

Input:
Training images X
Ground-truth labels Y
Model parameters w
SAM radius ρ
Output:
Optimized DBMSCNN parameters
1. Extract encoder features:
   z = Encoder(X)
2. Apply attention refinement:
   z_att = CBAM(z)
3. Predict class probabilities:
   ŷ = Classifier(z_att)
4. Compute classification loss:

$$\mathcal{L}_{c\ell s} = -\sum_{c=1}^{C} y_c \log(\widehat{y_c})$$

5. Compute gradients:
   g = ∇_w L_cls(w)
6. Compute SAM perturbation:

$$\epsilon(w) = \rho \frac{\nabla_w \mathcal{L}(w)}{|\nabla_w \mathcal{L}(w)|_2}$$

7. Update perturbed weights:

$$w' = w + \varepsilon(w)$$

8. Recompute perturbed predictions:

$$\hat{y}' = \mathrm{Classifier}\big(\mathrm{Encoder}(X; w')\big)$$

9. Compute perturbed SAM loss:

$$L\_sam = L\_cls(w')$$

10. Perform second forward-backward propagation
11. Update model parameters using AdamW:

$$w \leftarrow w - \eta \nabla_w L_{sam}$$

12. Update cosine annealing scheduler
13. Repeat Steps 1–12 for all mini-batches and epochs
14. Return optimized parameters w

Algorithm 3: Grad-CAM generation

Input:
Test image X
Target class c
Trained DBMSCNN model
Output:
Grad-CAM heatmap M_{GC}
Explainability visualization
1. Pass input image through trained DBMSCNN model:
   ŷ = Model(X)
2. Predict target class:
   c = argmax(ŷ)
3. Extract encoder feature representation:
   z = Encoder(X)
4. Extract final convolutional feature maps:

$$A^k = ConvFeatureMaps(z)$$

5. Compute score for target class c:

$$y^c = Score(\hat{y})$$

6. Perform backpropagation for target class:

$$\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

7. Calculate channel importance weights:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

8. Compute Grad-CAM activation map:

$$M_{GC} = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

9. Normalize Grad-CAM heatmap:

$$M_{norm} = \frac{M_{GC} - min(M_{GC})}{max(M_{GC}) - min(M_{GC})}$$

10. Resize heatmap to input image dimensions
11. Overlay heatmap on original image X
12. Generate final explainability visualization
13. Save Grad-CAM visualization results

4. Experimental Results

4.1 Performance of the Proposed Framework

Table 1 shows the fold wise classification performance analysis. Table 2 shows the hyperparameter details. The proposed DBMSCNN framework achieved an average classification accuracy of 92.43% across five-fold stratified cross-validation, demonstrating significant generalization capability and multiclass enamel caries classification performance. Table 3 shows the reconstruction pretraining and supervised training convergence loss across all the folds. The reconstruction learning stage depicted consistent convergence across all five folds. The mean reconstruction loss decreased from 0.0476 at the first epoch to 0.0245 at the 10th epoch, depicting that the decoder progressively improved its ability to reconstruct input dental images from the learned latent representations. During supervised fine-tuning with Sharpness-Aware Minimization (SAM), the classification loss decreased from an average of 1.0616 at the first epoch to 0.5003 at the final epoch, suggesting stable optimization and effective learning of discriminative features for dental caries classification. The convergence behaviour observed in both training stages affirmed the effectiveness of the proposed reconstruction-guided feature learning strategy and subsequent supervised optimization framework.

Table 1. Analysis of fold wise classification performance

Fold #	**Accuracy%**	**Precision %**	**Recall %**	**F1-Score %**	**Specificity %**
1	93.29	93.57	93.29	93.33	96.54
2	92.39	92.79	92.39	92.43	96.10
3	93.51	93.54	93.51	93.51	96.69
4	93.06	93.06	93.06	93.06	96.47
5	89.91	90.01	89.91	89.93	94.90
Mean value	92.43	92.59	92.43	92.45	96.14

Table 2. Hyperparameter details

Image size	**224**
Batch size	16
Reconstruction Epochs	10
Supervised Epochs	25
Train split	0.8
Reconstruction LR 1	1e-3
Classification LR	1e-4
SAM – **ρ**	0.05
Reconstruction loss	MSE loss
Weight decay	1e-4
Loss function	Cross entrophy+ Label smoothing

Table 3. Reconstruction pretraining and supervised training convergence across 5 folds

Fold	**Reconstruction Pretraining Loss (Epoch 1)**	**Reconstruction Pretraining Loss(Epoch 10)**	**Supervised Loss (Epoch 1)**	**Supervised Loss (Epoch 25)**
1	0.0452	0.0240	1.0388	0.4836
2	0.0447	0.0246	1.0601	0.5140
3	0.0596	0.0249	1.0687	0.4963
4	0.0444	0.0239	1.0722	0.4730
5	0.0442	0.0252	1.0681	0.5345
Mean	0.0476	0.0245	1.0616	0.5003

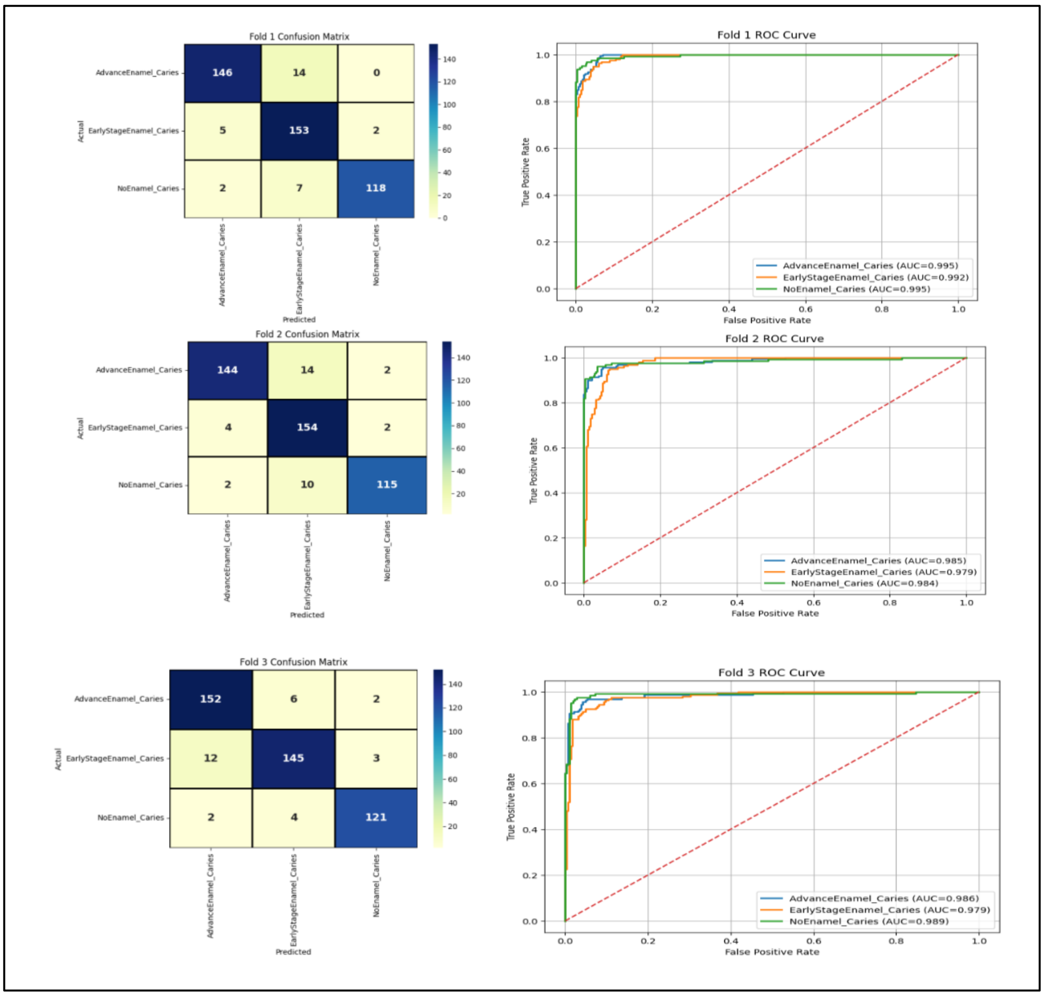

Figure 2. Confusion matrices and ROC curve obtained across the first 3 folds of stratified cross-validation for the proposed DBMSCNN model

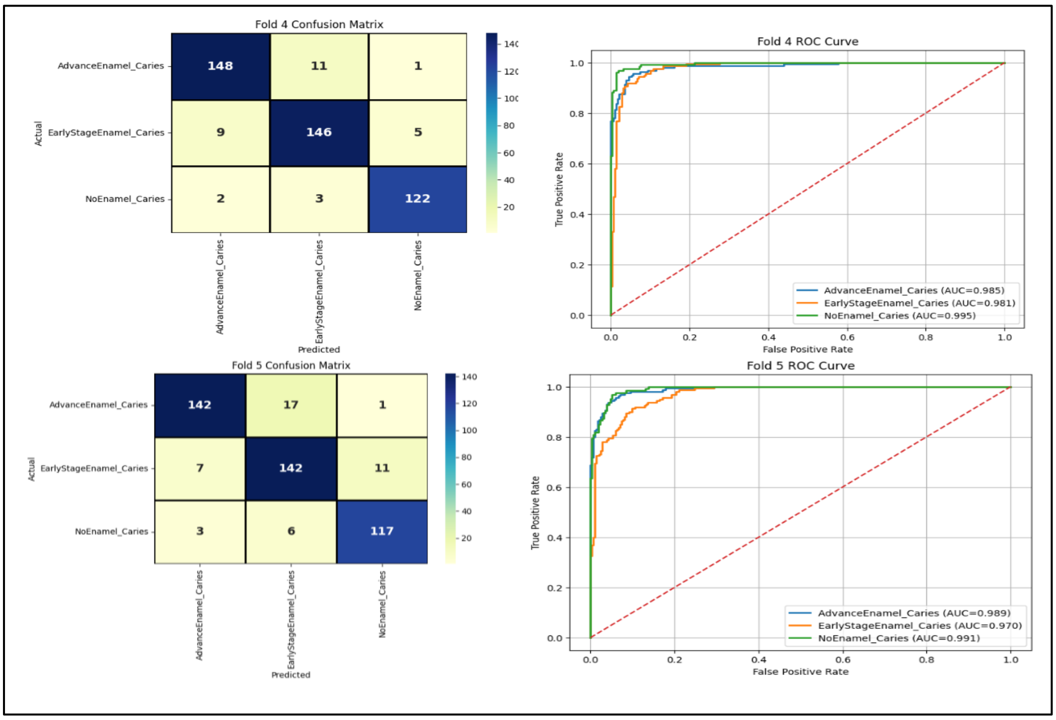

Figure 3. Confusion matrices and ROC curve obtained across the 5 folds of stratified cross-validation for the proposed DBMSCNN model

Figure 3 shows the confusion matrices, ROC curves, and Figure 4 shows the Grad-CAM visualizations. Occlusal pits, enamel discoloration, and structural irregularities are consistently highlighted by the Grad-CAM visualizations. Pixel-level sensitivity in clinically relevant regions is depicted by the saliency maps for improving the interpretability. Systematic quantitative evaluation of explainability was done by comparing the Grad-CAM attention regions with lesion-focused pseudo segmentation masks generated using image processing techniques such as CLAHE enhancement, Gaussian smoothing, Otsu thresholding, and morphological refinement.

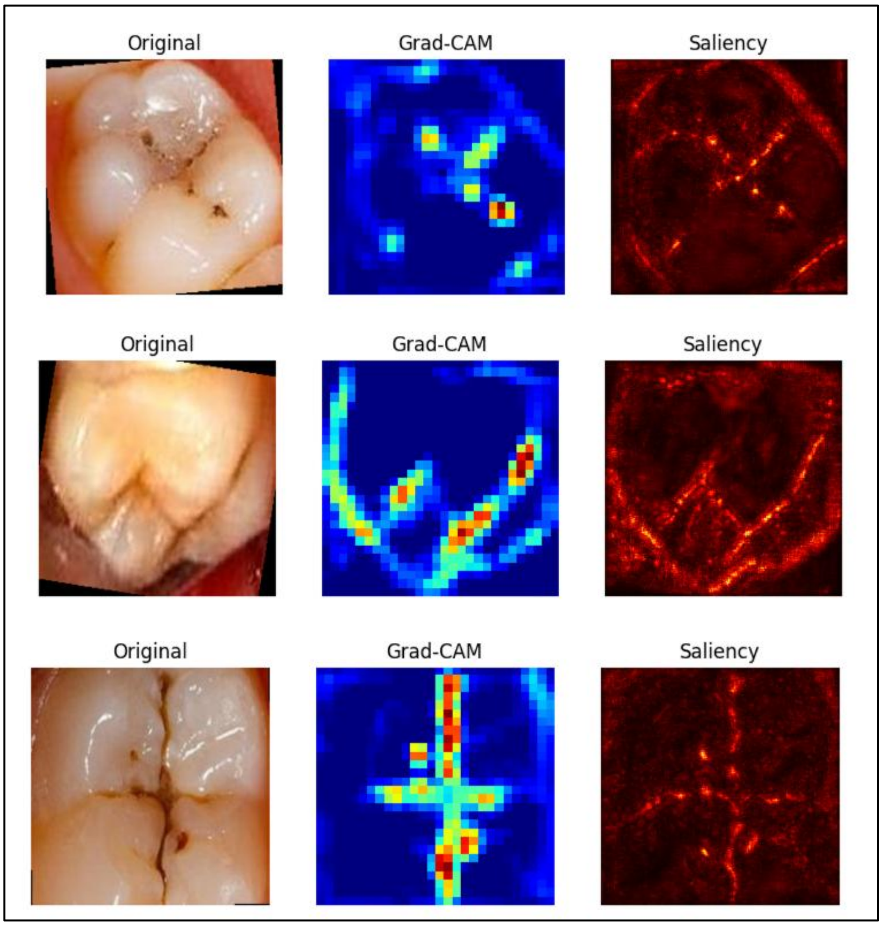

Figure 4. Grad-CAM visualizations

Three quantitative metrics were used for evaluation: IoU, Dice coefficient, and SSIM. Reconstruction-based self-supervised pretraining improved anatomical feature learning. The reconstruction branch achieved an average SSIM of 0.6549, indicating moderate structural similarity between reconstructed and original images. The average Dice coefficient of 0.8530 and IoU of 0.7577 suggest substantial overlap between reconstructed and original image regions after thresholding.

Table 4. Reconstruction performance metrics values of the proposed framework.

Fold	SSIM	Dice	IoU
1	0.6443	0.8483	0.7482
2	0.6586	0.8622	0.7689
3	0.6529	0.8521	0.7542
4	0.6591	0.868	0.7777
5	0.6599	0.8346	0.7395
Mean value	0.6549	0.853	0.7577

The explainability evaluation values improved the model transparency of the decision-making process. The robustness analysis is improved by showing the misclassified samples along with their corresponding predictions in Figure 5. Due to subtle radiographic similarities and overlapping lesion characteristics, most incorrect predictions occurred between Early Stage Enamel Caries and Advanced Enamel Caries. These visualizations provide additional insight into the limitations and decision making behavior of the proposed CNN framework.

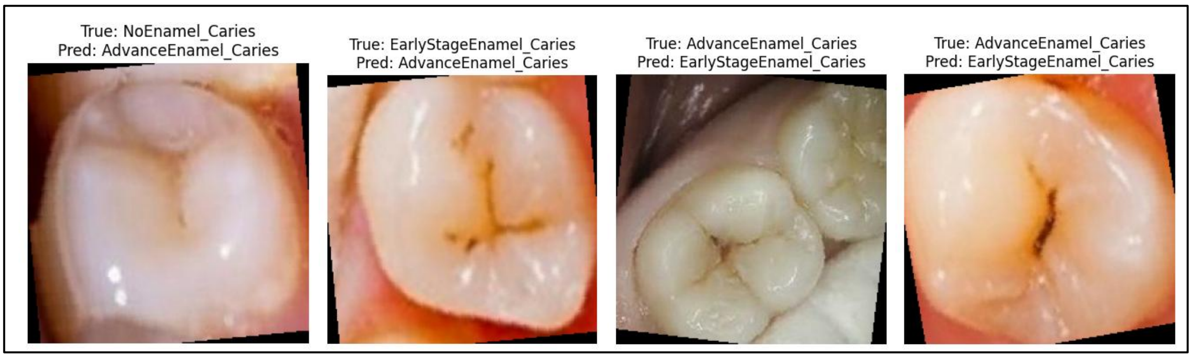

Figure 5. Section of the misclassified samples

5. Results and Discussion

Self-supervised reconstruction learning is a training strategy where a neural network learns useful feature representations without using class labels, by learning to reconstruct its own input. The model is asked to recreate the image from a compressed internal representation. If the model can reconstruct the image well, it must have learned meaningful structure, textures, and spatial relationships present in the data, as shown in Equations (16- 18).

Formally, given an input image x, an encoder E(·) maps it to a latent representation z:

$$z=E(x) \tag{16}$$

The original image is reconstructed using the decoder D(.) as below:

$$\hat{x} = D(z) \tag{17}$$

The proposed DB-MSCNN framework follows a two-stage training strategy. Each training batch involves two forward-backward passes, consisting of (1) reconstruction-based pretraining and (2) supervised classification with SAM optimization. In the first stage, the dual-branch encoder and decoder were trained using self-supervised reconstruction learning. The objective was to learn robust multi-scale feature representations from dental panoramic X-ray images by reconstructing the input image from the fused latent feature map.

The reconstruction loss is defined using Mean Squared Error (MSE) in equation 3:

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N}|X_i - \widehat{X}_i|_2^2 \tag{18}$$

During this stage, the optimizer used was standard AdamW. SAM was not used during reconstruction because the goal of this stage was stable feature representation learning rather than decision boundary optimization. It is self supervised since no ground-truth labels are required. Reconstruction learning forces the encoder to preserve fine-grained details (edges, textures), learn global structure (shape, layout) and avoid learning only class-specific shortcuts.

This leads to better generalization, reduced overfitting and more stable features for downstream tasks. In medical images, labels are expensive and subjective and visual cues are subtle. Self-supervised reconstruction uses all images effectively and encourages learning of anatomically meaningful features and improves performance on small datasets.

Stage 2: Supervised Classification with SAM

In the second stage, the pretrained encoder was fine-tuned jointly with the classifier using labeled supervision. Here, Sharpness-Aware Minimization (SAM) was introduced to improve generalization and reduce overfitting by encouraging convergence toward flatter minima.

The SAM objective is defined as in Equation 19.

$$\min_{w} \max_{|\epsilon|_2 \leq \rho} L_{cls}(w + \epsilon) \tag{19}$$

where:

- w represents model parameters
- ϵ is the adversarial weight perturbation
- ρ=0.05 is the perturbation radius

The classification loss is cross-entropy with label smoothing, as shown in equation 20.

$$L_{cls} = CrossEntropy(y, \hat{y}) \tag{20}$$

y = Ground truth label (actual class label)

$\hat{y}$= Predicted output of the model (predicted class probabilities or logits)

SAM was implemented using AdamW as the base optimizer. SAM was applied only during supervised classification training (Stage 2) and was not used during reconstruction pretraining (Stage 1). This design ensures stable feature learning during reconstruction while improving classifier robustness during final prediction.

In our framework, the dual-branch encoder is first trained with a decoder. It learns to reconstruct dental images. The decoder is then discarded, and the pretrained encoder is reused for classification. This means that the classifier starts from a medically informed feature space, not random weights.

SAM is an optimization technique that minimizes training loss and ensures that the model finds flat minima in the loss landscape (Equation 21). Flat minima correspond to models that generalize better to unseen data. Standard optimizers like SGD and Adam may find sharp minima, very low training loss, and poor generalization. This is especially problematic in medical datasets with class imbalance and high intra-class variability.

SAM solves the following optimization:

$$\min_{f} \max_{|\epsilon| \leq \rho} \mathcal{L}_{c\ell s}(w + \epsilon) \tag{21}$$

where:

- f = model parameters
- ϵ = worst-case perturbation
- ρ = Maximum value of allowable perturbation radius in the parameter space.

If the value of ρ is large, the optimizer is forced to search for flatter regions more aggressively. A small value of ρ makes SAM behave similarly to standard optimization methods such as AdamW. In this work, the value of ρ=0.05 was selected based on empirical validation and existing literature [25]. For image classification tasks, SAM commonly performed well within the range of 0.02 to 0.1. Initially, multiple candidate values for ρ like 0.05,0.02 and 0.1 were considered. When ρ=0.02, the generalization improvement over standard AdamW was limited. When ρ=0.1, convergence slowed due to excessive perturbation and training became less stable. The intermediate value of ρ=0.05 provided the balance between stability and generalization, yielding improved validation accuracy, a better F1-score, and more stable convergence behavior. For each batch, we compute gradients, reset the weights toward higher loss, recompute loss and gradients, and update the weights to minimize the reset loss. Robustness of the model to small weight variations is thus ensured. The selected configuration of SAM with ρ=0.05, AdamW base optimization, and controlled weight decay contributed significantly to the improved generalization performance.

SAM penalizes overconfident solutions, encourages smooth decision boundaries, and reduces sensitivity to noise and spurious features. Empirically, SAM often improves F1-score, ROC-AUC, and robustness on unseen data. Fig.6 shows the reconstruction decoder training epoch and the loss. Stable convergence and effective feature learning were confirmed through this performance. The encoder successfully captures sufficient information to reconstruct input images.

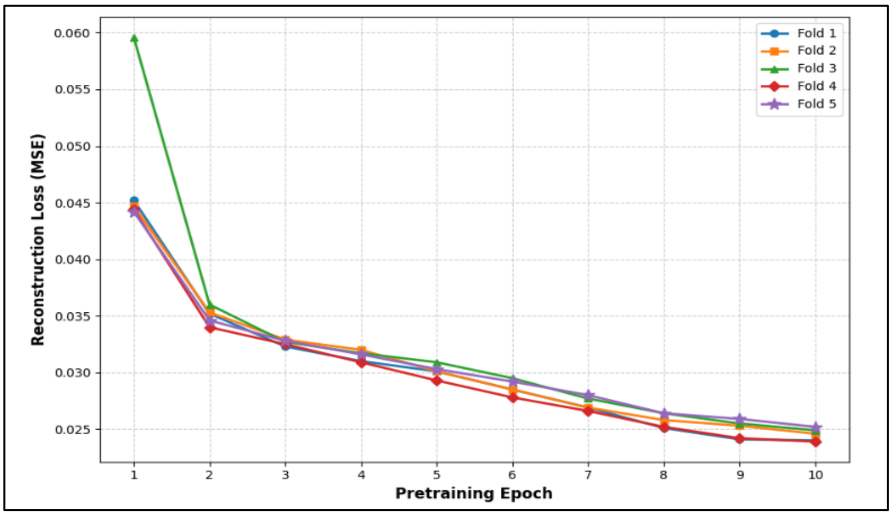

Figure 6. Reconstruction decoder training loss across 5 folds

In our supervised training, SAM is applied to the encoder and classifier and used with label smoothing. It prevents overfitting to texture artifacts, So the recall value for advanced caries is high and the overall ROC-AUC is strong. The single-branch CNN struggles to separate early pathological changes from normal enamel, due to reliance on a single receptive field and lack of feature regularization (Figure 7). The DBMSCNN achieves much cleaner diagonal dominance, reflecting better class separation. The t-SNE clusters (Figure 8) show moderate clustering. In the proposed DBMSCNN framework, the reconstruction branch is not intended merely to reduce reconstruction loss. It is used as a self-supervised pretraining mechanism to improve the quality of feature representations used later for classification. The reconstruction branch was optimized using Mean Squared Error loss. The reconstruction loss decreased progressively across epochs, indicating effective learning of structural and contextual image representations. This improved the encoder's feature quality and contributed to better downstream classification performance. This reduction indicated that the model was able to reconstruct the dental images with high fidelity, thereby improving the quality of the latent feature space. These pretrained encoder features were then transferred to the supervised classification stage, resulting in improved classification performance compared to the single-branch baseline model.

Therefore, reconstruction does not merely reduce loss; it acts as a self-supervised representation learning mechanism that enhances downstream classification accuracy and generalization.

6. Ablation Study

This section presents a detailed evaluation of the different variants of the proposed DBMSCNN framework and a Single-Branch CNN baseline. The experiments were conducted using identical dataset splits, evaluation metrics and preprocessing steps to ensure a fair comparison analysis. Table 5 shows the results of the ablation study across the different model variants.

Variant 2, termed DBMSCNN-EfficientNet, comprises a dual-branch multi-scale convolutional architecture integrated with an EfficientNet-B0 backbone. The model employs local and global feature extraction branches followed by feature fusion to capture both fine-grained and contextual dental caries characteristics. This variant excludes reconstruction learning, CBAM attention, and SAM optimization to isolate the contribution of the core multi-scale feature extraction framework.

Table 5. Results of the ablation study

Variant #	Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
1	SBCNN Baseline	82.27	82.18	82.43	82.27	94.32
2	DBMSCNN+EfficientNet-B0 backbone	82.27	82.72	82.27	82.36	94.53
3	DBMSCNN+EfficientNet-B0 + Reconstruction	60.83	61.59	60.83	60.94	77.36
4	DBMSCNN+EfficientNet-B0 backbone+SAM	80.48	81.12	80.48	80.53	93.87
5	DBMSCNN+ EfficientNet-B0 + Reconstruction+ SAM Optimization	92.43	92.59	92.43	92.45	98.64

Table 6. Branch-level study of accuracy using 5-fold cross-validation

Model	Architecture	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean accuracy
SBCNN	Single-Branch CNN	0.8277	0.8322	0.8568	0.7875	0.8094	0.8227
DBMSCNN	Dual-Branch Multi-Scale CNN	0.9329	0.9239	0.9351	0.9306	0.8991	0.9243

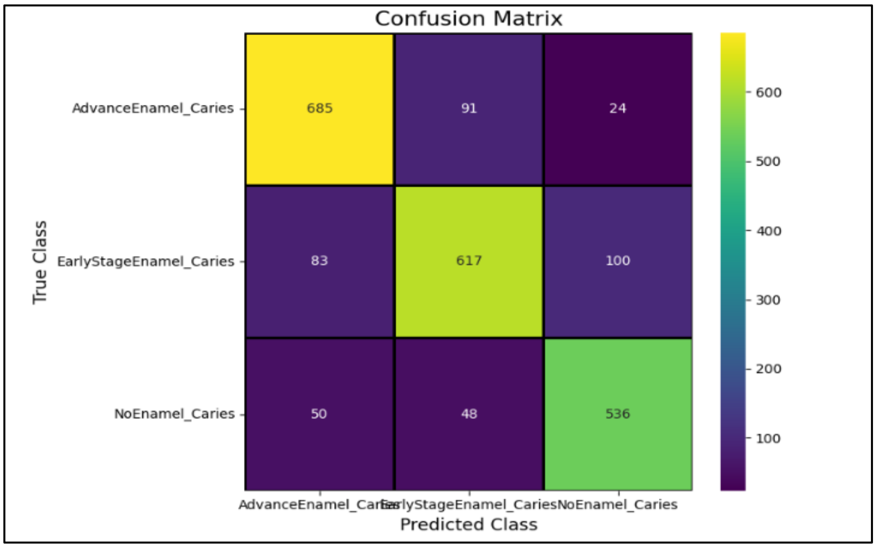

Figure 7. Confusion matrix -Single branch CNN

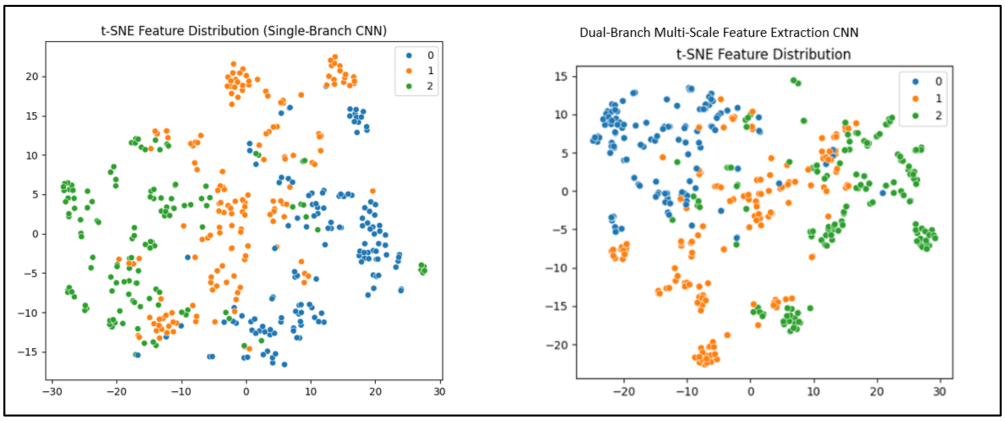

Figure 8. t-SNE feature distribution comparison

Figure 8 presents the t-SNE embedding of features learned by the proposed DBMSCNN with reconstruction regularization and SAM optimization. From the t-SNE Feature Distribution Comparison, Advanced Enamel Caries samples aggregate into dense, well-defined regions. The No Enamel Caries class exhibits spatial separation with moderate overlap. Early Stage Enamel Caries lies between the two extremes, reflecting its transitional pathological nature. To support the qualitative t-SNE feature distribution analysis, two quantitative cluster validation metrics namely Silhouette Score and Davies–Bouldin Index (DBI) were incorporated. The t-SNE visualization demonstrates partial grouping tendencies among the extracted latent features. Quantitative validation (Table 7) using Silhouette Score suggests that the learned feature space possesses reasonable discriminative capability, although some overlap between classes remains due to the challenging nature of dental caries patterns.

Table 7. Quantitative assessment of latent feature separability using t-SNE clustering metrics

Metric	Value	Interpretation
Silhouette Score	0.108	Indicates weak-to-moderate cluster separation. Positive values suggest that samples are generally closer to their own class clusters than to neighboring clusters.
Davies–Bouldin Index (DBI)	2.06	Reflects moderate overlap among clusters. Lower values indicate better cluster compactness and separation.

Statistical significance analysis was conducted to compare the proposed DBMSCNN with the baseline Single-Branch CNN using 5-fold cross-validation accuracies. The paired t-test produced a t-statistic of 14.3513 with a p-value of 0.0001, confirming that the performance improvement is statistically significant ($p < 0.05$). Although the Wilcoxon signed-rank test yielded a p-value of 0.0625, which is slightly above the 0.05 threshold, this is acceptable due to the small sample size and the conservative nature of the non-parametric test. Furthermore, bootstrap analysis produced a 95% confidence interval of [0.1415, 0.1820], indicating that the proposed model consistently improves classification accuracy by approximately 14–18% over the baseline model. These results in Table 8 strongly validate the superiority and reliability of the proposed DBMSCNN framework.

Table 8. Statistical significance of the DBMSCNN framework

Statistical Test	Result
Paired t-test Statistic	14.3513
Paired t-test p-value	0.0001
Wilcoxon Signed-Rank Statistic	0.0
Wilcoxon p-value	0.0625
Bootstrap 95% CI Lower Bound	0.1415
Bootstrap 95% CI Upper Bound	0.1820

The multi-scale dual-branch architecture captures both fine enamel texture changes and global structural cues. Self-supervised reconstruction learning enforces anatomically consistent representations. Feature generalization was improved, and flatter minima were promoted using Sharpness Aware Minimization. During the reconstruction-based self-supervised pretraining stage, the reconstruction loss decreased substantially within the first three epochs. Across the five folds, the loss reduction ranged from 25.79% to 45.13%, with an average decrease of 30.53%. Thus, self-supervised reconstruction worked as an implicit regularizer and enabled faster optimization convergence. Each architectural component contributed incrementally to high classification performance. Thus, the dual-branch multi-scale design enhances feature representation, reconstruction learning improves robustness, and SAM optimization improves generalization. The proposed DBMSCNN framework achieves superior diagnostic reliability and balanced performance across all classes.

7. Conclusion and Future Work

Across the five folds, the DBMSCNN framework achieved an average classification accuracy of 92.43% and an average F1-score of 92.45%. The reconstruction-learning component produced average SSIM, Dice, and IoU values of 0.6549, 0.8530, and 0.7577, respectively. The results demonstrate that the framework was able to learn discriminative representations for three-class dental caries classification while simultaneously maintaining reconstruction capability through the auxiliary reconstruction branch. Furthermore, the reconstruction and explainability framework demonstrated the ability of the proposed model to generate meaningful visual explanations. Rapid convergence was observed during reconstruction pretraining. It showcased the stabilizing effect of self-supervised learning, facilitating more reliable optimization compared to supervised training. The results portrayed improved internal generalization and robustness rather than complete external generalization. The dual-branch architecture generated well-separated class clusters, compared to the overlapping distributions noted in the single-branch model. DBMSCNN focused on clinically relevant enamel regions associated with carious lesions and improved model interpretability. Integrating multi-scale feature extraction with reconstruction-based regularization and SAM optimization led to explainable and robust dental caries classification. Future study will examine external validation on independent multi-center datasets, lesion severity regression, and integration with clinical decision-support systems to further enhance real-world applicability. The system can be extended to validate the proposed framework using externally curated datasets with explicit patient identifiers to enable strict patient-wise train–test separation. This will further augment the clinical applicability of the proposed framework.

References

[1] Li, Jiaxing. "DA-Net: A Classification-Guided Network for Dental Anomaly Detection from Dental and Maxillofacial Images." Journal of King Saud University-Computer and Information Sciences 36, no. 9 (2024): 102229.

[2] Featherstone, John DB. "Dental Caries: A Dynamic Disease Process." Australian dental journal 53, no. 3 (2008): 286-291.

[3] Jones, Bree, Mathias Lambach, Tong Chen, Stavroula Michou, Nicky Kilpatrick, Nigel Curtis, David P. Burgner, Christoph Vannahme, and Mihiri Silva. "Dental Caries Detection in Children Using Intraoral Scans and Deep Learning." Journal of Dentistry 160 (2025): 105906.

[4] PR, Bipin, and Anoop BK. "Automated Biomedical Image Classification Using Multi-Scale Dense Dilated Semi-Supervised U-Net with Cnn Architecture." Multimedia Tools and Applications 83, no. 10 (2024): 30641-30673.

[5] Barragan-Montero, Ana, Adrien Bibal, Margerie Huet Dastarac, Camille Draguet, Gilmer Valdes, Dan Nguyen, Siri Willems et al. "Towards a Safe and Efficient Clinical Implementation of Machine Learning in Radiation Oncology by Exploring Model Interpretability, Explainability and Data-Model Dependency." Physics in Medicine & Biology 67, no. 11 (2022): 11TR01.

[6] Lin, Jianwu, Xin Zhang, Yongbin Qin, Shengxian Yang, Xingtian Wen, Tomislav Cernava, Quirico Migheli, and Xiaoyulong Chen. "Local and Global Feature-Aware Dual-Branch Networks for Plant Disease Recognition." Plant Phenomics 6 (2024): 0208.

[7] Rupesh Devidas Sushir “Exploring Novel Self-Supervised Learning Techniques for Image Reconstruction Tasks”. International Journal of Advanced Research in Science, Communication and Technology, (2024). 48–55.

[8] Huang, Peizhou, Chaoyi Zhang, Xiaoliang Zhang, Xiaojuan Li, Liang Dong, and Leslie Ying. "Self-Supervised Deep Unrolled Reconstruction Using Regularization by Denoising." IEEE transactions on medical imaging 43, no. 3 (2023): 1203-1213.

[9] Rani, Veenu, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. "Self-Supervised Learning: A Succinct Review." Archives of Computational Methods in Engineering 30, no. 4 (2023): 2761-2775.

[10] Ohri, Kriti, and Mukesh Kumar. "Review on Self-Supervised Image Recognition Using Deep Neural Networks." Knowledge-Based Systems 224 (2021): 107090.

[11] Gidaris, Spyros, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. "Learning Representations by Predicting Bags of Visual Words." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6928-6938. 2020.

[12] T. Ratchatorn and M. Tanaka, "Improving Sharpness-Aware Minimization Using Label Smoothing and Adaptive Adversarial Cross-Entropy Loss," in IEEE Access, vol. 13, 2025, 100326-100337.

[13] Zhang, Zhiyuan, Ruixuan Luo, Qi Su, and Xu Sun. "GA-SAM: Gradient-Strength Based Adaptive Sharpness-Aware Minimization for Improved Generalization." In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, 3888-3903.

[14] Liu, Yong, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. "Towards Efficient and Scalable Sharpness-Aware Minimization." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 12360-12370.

[15] Li, Tao, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. "Friendly Sharpness-Aware Minimization." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, 5631-5640.

[16] Yun, Vincent-Daniel. "Sharpness-Aware Minimization with Z-Score Gradient Filtering." In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2026, 2961-2965.

[17] J., Priya, Kavya S., Gunavathie M.A., and Jacophine Susmi S. 2026. "A Novel Optimized Hybrid ResNextTransform for Dental Diseases Classification With Crested Porcupine Optimizer". Journal of Innovative Image Processing 8 (1): 233-49. https://doi.org/10.36548/jiip.2026.1.013.

[18] Gruber, Nadja & Schwab, Johannes & Haltmeier, Markus & Biguri, Ander & Dlaska, Clemens & Hwang, Gyeongha. (2025). Noisier2Inverse: Self-Supervised Learning for Image Reconstruction with Correlated Noise. 10.48550/arXiv.2503.19468.

[19] Millard, Charles, and Mark Chiew. "A Theoretical Framework for Self-Supervised MR Image Reconstruction Using Sub-Sampling via Variable Density Noisier2Noise." IEEE transactions on computational imaging 9 (2023): 707-720.

[20] Yaman, Burhaneddin & Hosseini, Seyed Amir Hossein & Moeller, Steen & Ellermann, Jutta & Uğurbil, Kâmil & Akçakaya, Mehmet. (2020). Self-Supervised Learning of Physics-Guided Reconstruction Neural Networks Without Fully Sampled Reference Data. Magnetic Resonance in Medicine. 84. 10.1002/mrm.28378.

[21] Islam, Sheikh Rabiul, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. "Explainable Artificial Intelligence Approaches: A Survey." arXiv preprint arXiv:2101.09429 (2021).

[22] Srinivasu, Parvathaneni Naga, N. Sandhya, Rutvij H. Jhaveri, and Roshani Raut. "From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies." Mobile Information Systems 2022, no. 1 (2022): 8167821.

[23] Kisten, Melvin, Absalom El-Shamir Ezugwu, and Micheal O. Olusanya. "Explainable Artificial Intelligence Model for Predictive Maintenance in Smart Agricultural Facilities." IEEE Access 12 (2024): 24348-24367.

[24] Himel, Galib Muhammad Shahriar; Islam, Md Masudul; Hannan, Ummae Hamida (2023), “Caries-Spectra: A dataset of Enamel Caries”, Mendeley Data, V2, doi: 10.17632/9jnf2jvghy.2

[25] Foret, Pierre, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. "Sharpness-Aware Minimization for Efficiently Improving Generalization." arXiv preprint arXiv:2010.01412 (2020).