SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I9P117

Volume 12 Issue 9, 194-202, September 2025
© 2025 Seventh Sense Research Group®

Original Article

A Residual Cross-Attention Fusion Network with Adaptive Focal-Margin Loss for Robust Dental Caries Detection from Intraoral Radiographs

Sheetal Kulkarni[1], N. Rama Rao[2]

[1,2]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India*

[1]*Corresponding Author : sheetalkulkarni529@gmail.com*

Received: 17 July 2025 Revised: 19 August 2025 Accepted: 18 September 2025 Published: 29 September 2025

***Abstract* - *In recent years, Automated dental caries detection using intraoral radiographs has attained increasing potential to assist dentists in early diagnosis and treatment planning. Traditional research primarily relies on general-purpose Convolutional Neural Networks (CNNs) such as VGG16 or ResNet, which have demonstrated reasonable performance in medical imaging tasks. The major challenge remains in accurately identifying carious lesions in the occurrence of anatomical noise, restorations and poor image contrast, which often leads to high false positive rates. This research proposes a novel hybrid architecture called ResXformer to address these limitations. Initially, the data is collected from diverse clinical sources and pre-processed to enhance the poor contrast. A depth-wise separable CNN backbone extracts fine-grained features with reduced computational cost. Later, a residual cross-attention fusion mechanism allows bidirectional information flow between the CNN and Transformer branches, enhancing spatial and contextual learning. Further, an adaptive focal-margin loss is introduced that penalizes ambiguous predictions based on per-sample logit variance, reducing false positives near restorations. Together, these steps create a robust, lightweight model tailored for accurate and interpretable dental caries detection in clinical practice.*

***Keywords* - *Adaptive Focal-Margin Loss, Convolutional Neural Networks, Dental Caries Detection, Intraoral Radiographs, Residual Cross-Attention Fusion Network.*

1. Introduction

Tooth decay or dental caries is among the most common chronic diseases affecting people of all age groups around the world [1]. Detection of caries early and accurately is paramount for obtaining timely treatment to prevent serious oral health issues [2, 3]. Intraoral radiographs, particularly bitewing radiographs, are widely accepted in clinical application for identifying and detecting carious lesions on the interproximal surfaces of teeth that would otherwise not be visible through visual examination [4]. Artificial Intelligence (AI) based tools, specifically those using deep learning techniques, have been utilized in dental translational research and clinical practice for the automated detection of caries in dentistry in order to minimize the error potential of human-based queuing and observations and to boost clinician workflow or throughput potentially [5, 6]. While the development of automated detection of caries has made significant improvements over the last few years, it remains challenging [7]. Several inherent characteristics of dental X-ray images are responsible for posing significant obstacles to the automated detection of caries. Low-contrast structures, overlapping anatomical structures, radiographic artifacts, restorations, and dental hardware cause difficult-to-detect artifacts [8, 9]. These subtle differences in noise and artifacts can create difficulties for generalized deep learning models to correctly classify carious tooth structure from non-carious tooth structure [10].

Traditional CNN-based models such as VGG16 and ResNet have been successfully applied to general medical imaging but have not been optimized for dental radiographs' heterogeneous texture and spatial structural characteristics [11]. Moreover, the existing models are not fine-tuned for different textures and fine-grained spatial features with respect to dental radiographs. Increased false-positive rates, particularly within or around metallic restorations or compromised radiographic contrast, are caused by inferring structural information of the teeth based on an acceptable pattern within the neighboring, but potentially different structure [12]. The major research gap is that the existing models are not complex and optimized for dental radiographs, capable of both local and global relationships in noisy images, which increases the false negatives in classification. The motivation for this research is to create a model architecture specific to its intended domain that overcomes the limitations of traditional CNNs by incorporating local and global context [13]. The primary goal is to create a model to improve classification performance while maintaining trust and interpretability in a clinical setting [14]. A reliable architecture that balances lightweight computation with fine-grained feature representation would enable scaling up for real-time, real-world applications in dental diagnoses [15]. The major contributions of the research are listed as below:

- To introduce ResXformer, a hybrid architecture that combines depth-wise separable CNN layers with a Transformer attention mechanism with residual cross-attention [16]. This enables bi-directional flow of information, strengthening spatial accuracy and global comprehension.
- A contextual encoding approach by Transformer blocks is used for the extraction of CNN features, and the representation of the data is improved by modelling non-local interactions [17].
- To introduce a novel adaptive focal-margin loss that dynamically penalizes penalties based on the characteristics of the hard-to-classify sample and minimizes false positives [18].

The current research manuscript is organized as follows: Section 2 discusses the literature review, Section 3 explains the proposed methodology, Section 4 illustrates the experimental results, and Section 5 gives the research conclusion.

2. Literature Review

Some of the existing research works that are applied for the dental caries diagnosis are discussed: Haihua Zhu et al. [19] suggested a new U-shape DL framework with a complete-scale axial attention mechanism, called CariesNet, aimed at the segmentation of shallow, moderate and deep dental caries in panoramic X-ray images. This included skip connection and encoder-decoder designs in its architecture, increasing the Sensitivity of segmenting tiny lesions. However, the model showed issues with under-represented lesions, and it is only tested on panoramic radiographs, which cannot be generalized.

Tareq, A et al. [20] presented the fused YOLO model incorporating transfer learning, such as VGG16 and DenseNet, to detect dental cavitations from unstandardized smartphone images. An experiment was performed on an augmented dataset of 1,703 images with an effective diagnostic accuracy. The approach has flexibility in handling non-standardized images, which is favorable for remote diagnosis; however, it requires high-quality images with similar performance in different areas of clinical practice.

Yi Liu et al. [21] presented the Oral-Mamba system, a DL framework that is constituted of Mamba blocks integrated into a modified U-Net framework, to detect dental caries, calculus and gingivitis using intraoral photographic images. The model had improved accuracy in caries detection, which was much faster than the standard U-Net models. The drawback comprises a lack of data diversity and limitations of generalization to the unknown clinical setting.

Geetha Chandrashekar et al. [22] suggested a collaborative learning framework that integrates self-trained models (Mask R-CNN and Faster R-CNN) in dental image segmentation and identification tasks on panoramic dental radiographs. The collaborative model increased the reliability and accuracy, particularly in scenarios where orthodontic tools or the absence of teeth are involved. However, the system was effective only in the segmentation of overlapping teeth and dental implants.

Mehmet Boztuna et al. [23] developed an automated periapical lesion detection system through DL based on U^2-Net architecture. Research focused on diagnosing periapical periodontitis through panoramic radiographs due to the variable lesion manifestations that made manual evaluation subjective. The model enhanced the segmentation accuracy and detected periapical lesions. However, the system requires additional panoramic radiographs for better generalization.

Ahmed, W.M. et al. [24] employed CNN to identify the different caries types from dental images of multiple datasets. Bitewings radiographs with a resolution of 1876×1402 pixels were gathered, segmented, and anonymized using a dental caries analysis software application. The strategy used supervised learning algorithms trained for semantic segmentation tasks. This approach achieved a superior classification result, establishing itself as a sensitive and precise approach that excels in all evaluated classes. However, the model requires huge computation, making it very hard to apply in real-time and widely if not supported by adequate hardware.

Ragodos, R. et al. [25] trained the Deep Neural Network (DNN) to recognize 10 dental abnormalities, hypoplasia, and microdontia on the large intraoral images dataset and impacted teeth, particularly in children with Orofacial Clefting (OFC). This multi-class classification technique got an efficient F1 score nearly similar to that of experienced dentists in detecting seven of the anomalies. The study emphasized that the model was particularly rapid and accurate in detecting anomalies, but low effective for particular anomalies and requires extensive data annotation. Tareq, A et al. [26] presented the fused YOLO model incorporating transfer learning, such as VGG16 and DenseNet, to detect dental cavitations from unstandardized smartphone images [27]. An experiment was performed on an augmented dataset of 1,703 images with an effective diagnostic accuracy. The approach has flexibility in handling non-standardized images, which is favorable for remote diagnosis; however, it requires high-quality images with similar performance in different areas of clinical practice.

3. Proposed Work

This section explains the proposed methodology for precise dental caries diagnosis. Initially, the data is collected from diverse clinical sources and pre-processed to enhance the poor contrast. A depth-wise separable CNN backbone extracts fine-grained features with reduced computational cost. Later, a ResXformer allows bidirectional information flow between the CNN and Transformer branches, enhancing spatial and contextual learning. Further, an adaptive focal-margin loss is introduced that penalizes ambiguous predictions based on per-sample logit variance, reducing false positives near restorations.

3.1. Dataset and Pre-Processing

A total of 4,875 dental radiographic images were collected and used in this research. The images are separated into a training set with 2,242 images, a validation set with 600 images and a testing set with 600 images. The training set images consist of 28,014 labels, the validation set consists of 3,567 labels and the test set consists of 3,463 labels. The input image that shows original caries and augmented caries is illustrated in Figure 1.

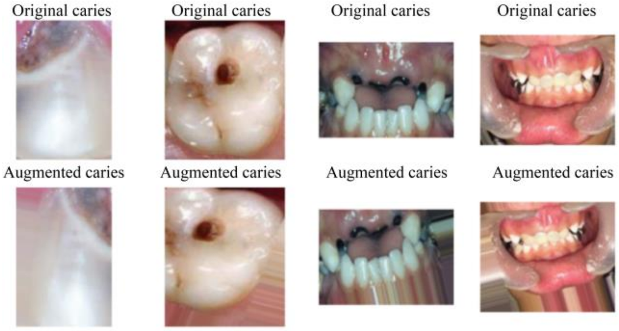

Fig. 1 Original dental caries input image

The Frost filter is an effective non-linear filter that is effectively used in removing speckle noise that is present in medical images, such as ultrasound and X-ray radiographs. Multi-scaled analysis is done using pixel-based local mean and standard deviation data. When a pixel with an intensity value significantly varies from the mean value of the region it is located in, it will be replaced by an intensity value that will be calculated by adding up the weight values of the neighboring pixels. The filter reduces noise by operating them without destroying edges or detailed information required in a medical and dental diagnostic application. As a noise filter, the Frost filter application is effective with a digital periapical image, thus enhancing its performance in recognition and description of the lesions. This type of image enhancement has essential advantages in the bone repositioning process following periapical surgeries, as it allows clinicians to find out where it happens.

3.2. Feature Extraction

Depth-wise separable convolution is a low-cost, efficient alternative to conventional convolutional layers, saving more computation and memory at the cost of no performance loss. This is done by separating the convolution into two parts, depth-wise and point-wise; the former applies one spatial filter to each of the input channels and later a 1×1 convolution to project the depth-wise-filtered outputs across channels. Such separation reduces the number of parameters and floating-point operations. If the standard convolution cost is expressed mathematically as in Equation (1), then the cost of depth-wise separable convolution becomes as in Equation (2)

$$Cost_{std}=k^2\times M\times N\times F^2 \tag{1}$$

$$Cost_{DWSC}=k^2\times M\times F^2+M\times N\times F^2 \tag{2}$$

Where, k^2 is the kernel size, M and N are the input and output channels, and F^2 is the spatial resolution of the feature map. This computation cost reduction is suitable for dental caries detection, where the inference speed and model efficiency are crucial. The depth-wise separable CNN architecture is particularly effective for intraoral dental radiographs. The poor quality of dental imaging is a factor that gives it low levels of contrast, insignificant variations in texture and superimposition of anatomical structures. The capacity of depth-wise convolution to focus on and maintain specific spatial elements in every channel assures the conservation of the fine lesion textures. The second point-wise convolution allows the network to merge such local patterns in the channels, which will allow more reliable classifications of caries areas. Each separable convolution block usually comprises a Batch Normalization and Gaussian Error Linear Unit (GELU) activation to enhance learning stability and feature refinement more effectively by permitting non-linear transformation to be learned. Then, selective pooling is used to downsample the feature maps and maintain spatial saliency.

Following the utilization of the CNN backbone, the extracted lightweight feature maps, such as a tensor of shape $128\times H/4\times W/4$ are passed to a transformer. This is used to learn longer-range dependencies that the CNN alone cannot and generalizes its local features over the full dental arch. These features informed at a global level are then combined with the original CNN outputs by way of cross-attention modules or residual links. Lastly, a prediction head that has been trained using an adaptive focal-margin loss is used as a way of tightening decision boundaries and minimizing false positive errors, particularly in the area of ambiguity, such as restorations and shadows [28]. Therefore, the depth-wise separable CNN will not only be the basis for efficient and accurate extraction of features but also ensure the subsequent use of a hybrid CNN architecture for precise dental diagnosis.

3.3. Global Context Modelling via Transformer

After the semantic and special features are extracted by the depth-wise separable CNN, the second step is the contextual encoding by Transformer blocks. As much as CNNs are perfectly suited to learn local patterns using convolution, they are naturally incapable of modelling long-range relationships and global contextual relationships in an image. Such a limitation is particularly important in the case of any dental radiographs, where pathological elements of interest, such as caries, can cover many regions of the image, or some less-obvious context can play a role in informing a diagnosis. To overcome this, a Transformer encoder is introduced to the extracted CNN features and improves the data's representation by modelling such non-local interactions. A transformer works by first encoding the input tensor into a series of patch embeddings, which are further enriched with multi-head self-attention operations. For a given input feature map $X \in R^{C\times H\times W}$The sequence is reshaped into the set of N flattened patches, each of which is linearly projected into a fixed-dimensional embedding. The self-attention mechanism computes attention scores as in Equation (3)

$$\text{Attention }(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) V \qquad (3)$$

Where Q, K, V are the query, key and value matrices derived from the input embeddings and d_K is the dimensionality of the key vectors. The formulation then enables every patch to consider the rest of the patches within the image, incorporating spatially distant but semantically correlated information. This is particularly effective in dental imaging. As an illustration, lesions at an occlusal surface can be more conspicuous in the light of the general orientation of teeth, distribution of density, or proximity of other restorations.

Transformers make the model recognize these larger-scale spatial relations, resulting in a more informative and accurate representation of features. The application of positional encodings also ensures that the spatial structure does not disintegrate during the attention mechanism, which is necessary to preserve the anatomical fidelity [29]. After the global context has been incorporated in the feature representations, they are combined back with the CNN features by means of residual or cross-attention, or sent to segmentation or classification heads. The Transformer blocks integration ensures that the network will no longer be confined to local analysis, but the image as a whole can be understood, which is essential to robust and clinically sound diagnosis of dental caries.

3.4. Residual Cross-Modality Attention Fusion Network (ResXformer)

The next inevitable step is featuring integration with ResXformer, which is necessary to successfully combine local and global features after contextual encoding is done using Transformer blocks. Whereas CNNs would be useful regarding local textures in high-resolution, and Transformers would be more advantageous in providing global contextual relationships, they both may excel in different ways. It is possible to simply concatenate their outputs and lose some information due to information redundancy. Hence, the mechanism of fusion needs to enable both modalities to go back and forth, enrich one another's features and pose spatial integrity, which is paramount in the precise detection of carious lesions in complex dental radiographs.

In the residual cross-attention fusion module, the CNN stream and Transformer stream feature maps, as cross-attention directions, are considered as sets of query-key-values. The CNN features act as the queries. Q_{CNN} In the first path, querying the transformers key and values K_{TR}, V_{TR}, which enables the integration of global semantics into localized spatial descriptors as in Equation (4)

$$Z_{CNN} = softmax\left(\frac{Q_{CNN}K_{TR}^T}{\sqrt{d_k}}\right) V_{TR} \qquad (4)$$

On the opposite pathway, the Transformer stream takes its features as queries to perfect world knowledge with the high-res spatial features of the CNN, as in Equation (5)

$$Z_{TR} = softmax\left(\frac{Q_{TR}K_{CNN}^T}{\sqrt{d_k}}\right) V_{CNN} \qquad (5)$$

This reciprocal querying only gets the features that are not just comprehensive but are also contextually filtered. To stabilize this fusion process and reduce destabilization through layers, residual connections are used in both streams regularly. Such skip connections allow the original activations to be reused and make it easy to have the gradient flow throughout training. It is important to get detailed features of dental areas susceptible to restorations, fillings, or anatomical noise. A mechanism of bidirectional fusion is beneficial in clinical imaging, particularly when used to perform tasks such as caries segmentation or detection. Once the fusion is done, the output is directed to a prediction module in which fine-grained boundaries are drawn with increased conviction [30]. In this way, the remaining cross-attention merge works as a connection between the two representational spaces and indeed balances the advantages of both building blocks of architecture.

3.5. Adaptive Focal-Margin Loss

The model will enter the prediction phase when local and global features are aligned using ResXformer, generating the final output. In order to ensure that this process is maximized, an optimization of the training stage is done using an advanced loss known as adaptive focal-margin loss that eliminates the use of conventional loss functions such as cross-entropy. The designed standard focal loss solves the class imbalance problem by down-weighting easy examples and committing learning to more difficult, misclassified cases. Adaptive focal-margin loss takes this further to add a running margin per sample and allows it to move as confidence and variance of batch predictions vary. The main one is the formulation that mixes logit distance penalties and a focal variant of scaling as in Equation (6)

$$L_i\text{=-}\alpha_i\left(1\text{-}p_{t,i}\right)^{\gamma_i}\log\left(\sigma^{-1}\left(p_{t,i}\right)\text{-}m_i\right) \qquad (6)$$

Where, $p_{t,i}$ is the predicted probability for the target class, γ_i It is a focusing parameter that increases with the prediction error δ_i=|y_i-p_i|and m_i is a sample-specific margin derived from the logit variance across the mini batch. This loss penalizes predictions that are both wrong and uncertain, particularly in the dental radiograph, because some of these regions are deceptively similar to the carious tissue [31]. The obtained final output images are shown in Figure 2. The confusion matrix for the true and predicted caries is illustrated in Figure 3.

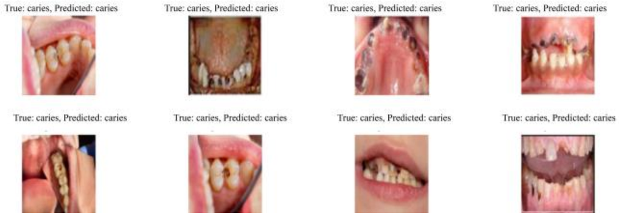
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries
True: caries, Predicted: caries

Fig. 2 Various sample output images if caries are predicted

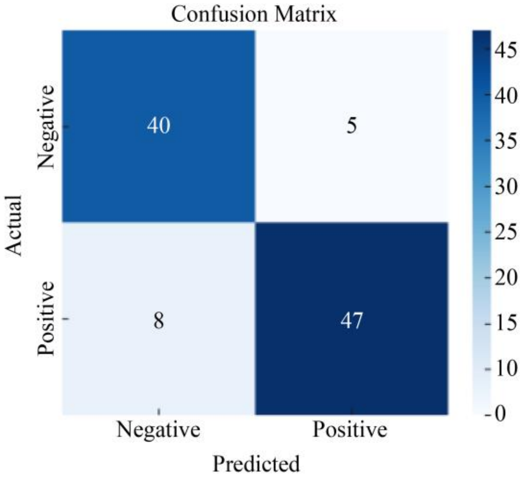

Fig. 3 Confusion matrix for true and predicted caries lesions

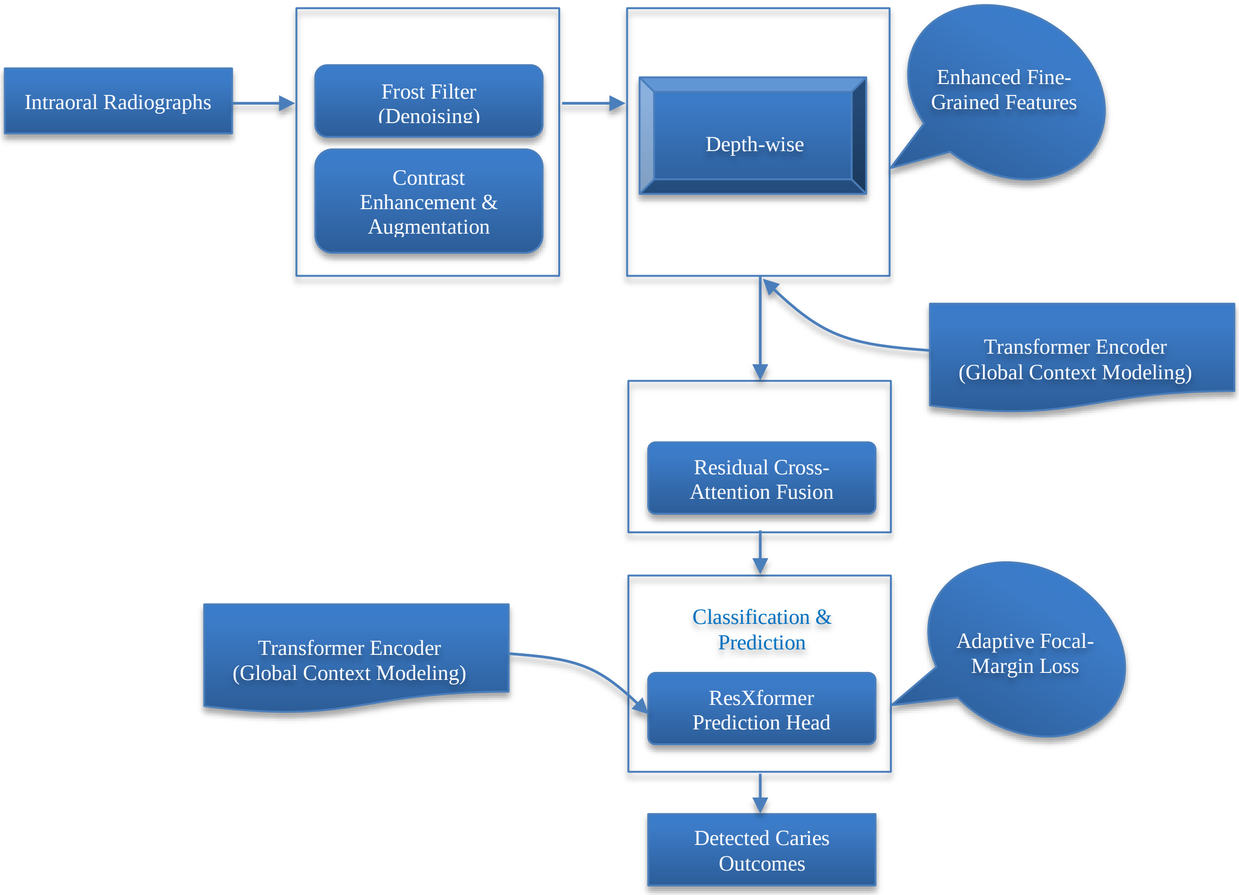
Intraoral Radiographs
Frost Filter (Denoising)
Contrast Enhancement & Augmentation
Depth-wise
Enhanced Fine-Grained Features
Transformer Encoder (Global Context Modeling)
Residual Cross-Attention Fusion
Classification & Prediction
Transformer Encoder (Global Context Modeling)
ResXformer Prediction Head
Adaptive Focal-Margin Loss
Detected Caries Outcomes

Fig. 4 Workflow of the proposed ResXformer model for dental caries detection

The model uses this adaptive loss in such a way that the number of false positives is decreased significantly, particularly in areas of the image where the noise levels are high, which can refer to fillings, shadows, or artifacts. It also adapts the Sensitivity of the model to the complexity of the presented training examples, which will be necessary in medical imaging, where insignificant distinctions may now hold diagnostic significance. Adaptive focal-margin loss directs the prediction phase, which can be regarded as a precision engine of the whole pipeline, as this ensures clinical applicability and trustworthiness of a decision.

In Figure 4, the proposed ResXformer system incorporates pre-processing, depth-wise separable CNN, Transformer encoder, residual cross-attention fusion, and adaptive focal-margin loss to produce effective, efficient, and effective dental caries detection.

4. Experiment Results

This section examines the suggested methods for the most recent DL based segmentation approaches. The study experiments are executed using Python 3.8 on Windows 10 on an Intel i5 having 16GB RAM, 6GB GPU, and 1TB SSD. The comparison of the proposed ResXformer model with two popular baselines, VGG16 and ViT-Base, is given in Table 1.

The ResXformer performs much better than the two and has the highest overall accuracy of 0.91, precision of 0.89, recall of 0.8 and F1-score of 0.88. It is also important to note that the model gives a much higher PR-AUC of 0.93, which indicates a better ability to deal with class imbalance and handle uncertain predictions. The graphical representation of accuracy scores and loss values is shown in Figure 5.

Table 1. Comparison with baseline models

Model	Accuracy	Precision	Recall	F1-Score	PR-AUC	FPR (@90% Recall)	Params (M)	Inference (ms)
VGG16	0.84	0.79	0.81	0.8	0.78	0.25	138	120
ViT-Base	0.86	0.82	0.83	0.82	0.81	0.18	86	95
ResXformer	0.91	0.89	0.87	0.88	0.93	0.07	4.5	45

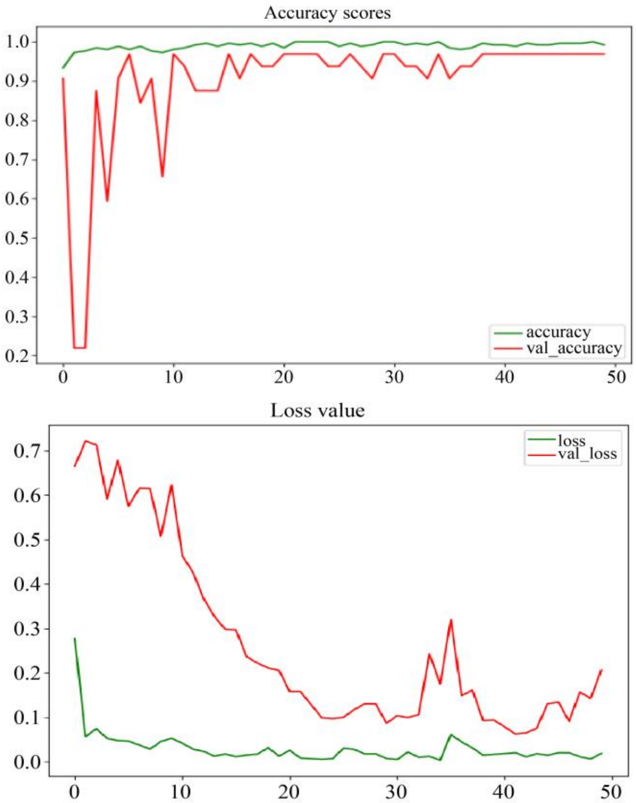

Fig. 5 Graphical representation of accuracy scores and loss values

Moreover, it has a low False Positive Rate (FPR), indicating its strength in the clinical scenario where high recall is necessary. Notably, ResXformer produces these improvements when only 4.5M parameters are used and achieves the best inference time of 45 ms, implying it is far more efficient than VGG16 (138M, 120ms) and ViT-Base (86M, 95ms). These findings not only point out its overall excellence in accuracy, but also its applicability in application when it comes to dental diagnostics in real time scenarios. The tabular form of the results is given in Table 1, and the graphical representation of the True Positive Rate (TPR) and False Positive Rate (FPR) is shown in Figure 6.

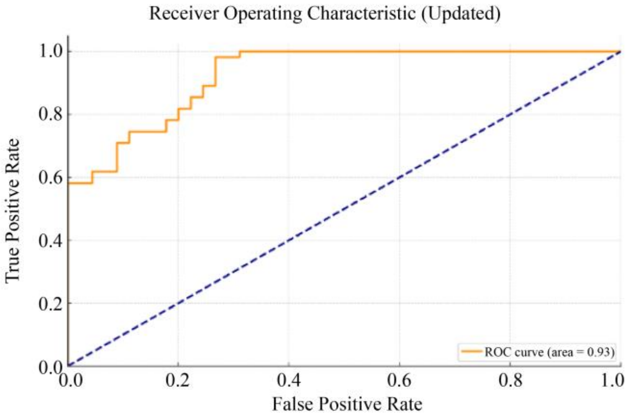

Fig. 6 TPR Vs FPR

Table 2. Ablation study

Variant	PR-AUC	FPR (@90% Rec)	Δ PR-AUC	Δ FPR
Full ResXformer	0.93	0.07	-	-
Without residual cross-attention fusion	0.88	0.13	-0.05	+0.06
Replacing Adaptive Focal-Margin with standard FL	0.9	0.12	-0.03	+0.05
Using standard convolutions vs. depth-wise separable	0.91	0.1	-	+0.03
Without curriculum-guided hard-negative mining	0.92	0.09	-0.01	+0.02

Table 2 shows the results of an ablation study carried out to understand the role of each architectural component in ResXformer. The substitution of adaptive focal-margin loss with regular focal loss is also deteriorating, and its PR-AUC is 0.90 with an elevated FPR (0.12), which means that the loss formula produces less FPR around ambiguous edges. Replacing depth-wise separable convolutions with standard convolutions does not change PR-AUC and doesn’t show increased FPR, indicating that parameter efficiency is maintained at the cost of spatial filtering quality. Finally, removing curriculum-guided hard-negative mining also leads to a slight drop in performance, evidence that it can be helpful in modifying the focus of learning in the model. The study generally affirms that every innovation is significant to the final model and helps it to be effective.

Table 3 shows the results of hyperparameter sensitivity analysis, demonstrating that the performance of the proposed model depends on the configurations of the parameters. By using a large variety of values of the base parameter, concentrating parameter γ_0, margin multiplier and the number of Transformer stages, the PR-AUC exhibits high scores between 0.90 and 0.94and the FPR has relatively small ranges between 0.06 and 0.11. This proves that the model is effective, stable, and robust to changes in hyperparameters. These findings have justified the reliability of the model during other training conditions and that the model can be optimized with flexibility without any performance loss.

Table 3. Hyperparameter sensitivity

Hyperparameter	Values Tested	PR-AUC Range	FPR Range
Base focusing γ_0	{1.0, 2.0, 4.0}	0.91–0.93	0.06–0.09
Margin scaling factor	{0.5, 1.0, 1.5}	0.90–0.93	0.07–0.11
Transformer layers	{2, 4, 6}	0.91–0.93	0.07–0.10
Hard-negative ratio	{10%, 20%, 30%}	0.92–0.94	0.06–0.08

Table 4 is the last structure, and this evaluates the explainability of the model using the various attribution methods. Grad-CAM obtains an Intersection over Union (IoU) of 0.72 and a standard deviation of 0.05, whereas SHAP alone gets a rather lower average IoU of 0.68 and a standard deviation of 0.10. However, the compositional explanation that integrates Grad-CAM and SHAP created a much better understanding, with a mean IoU of 0.85 and reduced variance to 0.04. It means that the more visualisation tools are used, the more consistent and correct interpretations of the model predictions are obtained. The finding underlines the importance of explainability alignment since it has a crucial role in obtaining the final output in practical applications.

Table 4. Explainability alignment

Method	Mean IoU	Std IoU
Grad-CAM alone	0.72	0.05
SHAP alone	0.68	0.06
Grad-CAM + SHAP (composite)	0.85	0.04

Table 5 shows the comparative analysis of the research, where the existing models, such as MobileNet-v3 + U-Net [16], nnU-Net [17], and Ensemble Inception-ResNet-v2 [18], are implemented in the proposed simulation. The obtained results are then compared with the results of the proposed model for a validated state-of-the-art comparison. The results are evaluated by means of performance metrics like accuracy, recall, and specificity, as shown in Table 5.

The proposed model was found to be 97.08% accurate, 95.04% recall, and 99.04% specific in comparison with the existing models. Although the accuracy and specificity of nnU-Net [17] were very high, it had a relatively lower recall, which constrains the Sensitivity of detecting more subtle caries lesions. The balance between accuracy and specificity of MobileNet-v3 + U-Net [16] was good; however, it was still not as effective in recall as the proposed algorithm. Equally, Ensemble Inception-ResNet-v2 [18] had competitive recall values at the expense of accuracy and specificity. The proposed model outperformed these alternatives in all three measures, indicating that the model is robust and can generalize.

Table 5. State-of-the-art comparison to the proposed model

DL models	Accuracy (%)	Recall / Sensitivity (%)	Specificity (%)
MobileNet-v3 + U-Net [16]	93.40	81.31	95.65
nnU-Net [17]	98.6	82.1	100.0
Ensemble Inception-ResNet-v2 [18]	87.1	85.8	89.3
Proposed Model	97.08	95.04	99.04

5. Discussion

The experimental testing performed on the proposed model showed better performance than the traditional deep learning approaches. A ResXformer architecture, which added depth-wise separable CNN and Transformer encoders and adaptive focal-margin loss to enhance the accuracy of caries detection. There were great improvements in the new system framework in terms of accuracy, recall, specificity and Dice scores. Several comparative studies demonstrated higher-quality balanced and reliable performance of the proposed method over traditional state-of-the-art models, including MobileNet-v3 + U-Net [16], U-Net [17], and Ensemble Inception-ResNet-v2 [18]. Its practical applicability to large-scale deployment systems is more robust due to reduced computational wastage as well as increased diagnostic reliability in clinical practice.

6. Conclusion

The proposed research introduced a new DL architecture, ResXformer, which efficiently helps diagnose dental caries using intraoral radiographs. The model incorporates fine-grained textures and global context, pairing with global contextual features through a residual cross-attention fusion mechanism.

The model integrated a lightweight depth-wise separable CNN with transformer encoders. It also added an adaptive focal-margin loss that is more uncertain and adaptively down-weights classifications. This limits the number of false positives, especially instances of restorations or anatomical noise. A large number of experiments prove that ResXformer has the highest accuracy, precision, recall and computational efficiency compared with traditional architectures, such as VGG16 and ViT-Base.

The entire research gives a very clear, efficient and interpretable solution that is applicable to real-time implementations of dental diagnostics, which holds great potential for increasing early caries detection and lowering the subjectivity of dental diagnostics in clinical practice

Acknowledgements

I would like to express my sincere gratitude to all those who contributed to completing this research paper.

References

[1] Bree Jones et al., “Dental Caries Detection in Children using Intraoral Scans and Deep Learning,” *Journal of Dentistry*, vol. 160, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[2] Viktor Szabó et al., “Validation of Artificial Intelligence Application for Dental Caries Diagnosis on Intraoral Bitewing and Periapical Radiographs,” *Journal of Dentistry*, vol. 147, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Javier Pérez De Frutos et al., “AI-Dentify: Deep Learning for Proximal Caries Detection on Bitewing X-Ray - Hunt4 Oral Health Study,” *BMC Oral Health*, vol. 24, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Betül Ayhan, Enes Ayan, and Saadet Atsü, “Detection of Dental Caries Under Fixed Dental Prostheses by Analyzing Digital Panoramic Radiographs with Artificial Intelligence Algorithms based on Deep Learning Methods,” *BMC Oral Health*, vol. 25, pp. 1-10, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] S. Siji Rani et al., “Deep Learning-based Cavity Detection in Diverse Intraoral Images: A Web-based Tool for Accessible Dental Care,” *Procedia Computer Science*, vol. 233, pp. 882-891, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yutong Liang et al., “AI-Driven Dental Caries Management Strategies: From Clinical Practice to Professional Education and Public Self Care,” *International Dental Journal*, vol. 75, no. 4, pp. 1-11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mashail Alsolamy et al., “Automated Detection and Labeling of Posterior Teeth in Dental Bitewing X-Rays Using Deep Learning,” *Computers in Biology and Medicine*, vol. 183, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Leena Rohan Mehta et al., “Identifying Suitable Deep Learning Approaches for Dental Caries Detection Using Smartphone Imaging,” *International Journal of Computational Methods and Experimental Measurements*, vol. 12, no. 3, pp. 251-267, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Xiaoyan Zhang, “Integrating U-Net and Multi-Model Approaches for Accurate Dental Occlusal Surface and Caries Detection,” *2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC)*, Mianyang, China, pp. 33-40, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lubaina T. Arsiwala-Scheppach et al., “Impact of Artificial Intelligence on Dentists’ Gaze during Caries Detection: A Randomized Controlled Trial,” *Journal of Dentistry*, vol. 140, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Gelareh Haghi Ashtiani et al., "Diagnostic Accuracy of Tele-Dentistry in Screening Children for Dental Caries by Community Health Workers in a Lower-Middle-Income Country," *International Journal of Paediatric Dentistry*, vol. 34, no. 5, pp. 567-575, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Bipin Kumar Rai et al., "Leveraging 3D Faster R-CNN for 3D Dental X-ray Restoration and Treatment Identification," *Proceedings of Computing and Machine Learning*, Sikkim, India, pp. 241-260, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Kathrin Becker et al., "Eligibility of a Novel BW + Technology and Comparison of Sensitivity and Specificity of Different Imaging Methods for Radiological Caries Detection," *Oral Radiology*, vol. 40, pp. 424-435, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Reinhilde Jacobs et al., "Radiographic Diagnosis of Periodontal Diseases – Current Evidence Versus Innovations," *Periodontology 2000*, vol. 95, no. 1, pp. 51-69, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Kyubaek Yoon et al., "AI-Based Dental Caries and Tooth Number Detection in Intraoral Photos: Model Development and Performance Evaluation," *Journal of Dentistry*, vol. 141, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jing-Wen Zhang et al., "Diagnostic Accuracy of Artificial Intelligence-Assisted Caries Detection: A Clinical Evaluation," *BMC Oral Health*, vol. 24, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Luya Lian et al., "Deep Learning for Caries Detection and Classification," *Diagnostics*, vol. 11, no. 9, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sohee Kang et al., "Diagnostic Accuracy of Dental Caries Detection using Ensemble Techniques in Deep Learning with Intraoral Camera Images," *Plos One*, vol. 19, no. 9, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Haihua Zhu et al., "CariesNet: A Deep Learning Approach for Segmentation of Multi-Stage Caries Lesion from Oral Panoramic X-ray Image," *Neural Computing and Applications*, vol. 35, pp. 16051-16059, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Abu Tareq et al., "Visual Diagnostics of Dental Caries through Deep Learning of Non-Standardised Photographs Using a Hybrid YOLO Ensemble and Transfer Learning Model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 7, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yi Liu et al., "Oral Screening of Dental Calculus, Gingivitis and Dental Caries through Segmentation on Intraoral Photographic Images Using Deep Learning," *BMC Oral Health*, vol. 24, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Geetha Chandrashekar et al., "Collaborative Deep Learning Model for Tooth Segmentation and Identification using Panoramic Radiographs," *Computers in Biology and Medicine*, vol. 148, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mehmet Boztuna et al., "Segmentation of Periapical Lesions with Automatic Deep Learning on Panoramic Radiographs: An Artificial Intelligence Study," *BMC Oral Health*, vol. 24, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Walaa Magdy Ahmed et al., "Artificial Intelligence in the Detection and Classification of Dental Caries," *The Journal of Prosthetic Dentistry*, vol. 133, no. 5, pp.1326-1332, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ronilo Ragodos et al., "Dental Anomaly Detection using Intraoral Photos via Deep Learning," *Scientific Reports*, vol. 12, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Taseef Hasan Farook MScDent et al., "A Virtual Analysis of the Precision and Accuracy of 3-Dimensional Ear Casts Generated from Smartphone Camera Images," *The Journal of Prosthetic Dentistry*, vol. 128, no. 4, pp. 830-836, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Esra Asci et al., "A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition," *Children*, vol. 11, no. 6, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Md Arid Hasan, and Krishno Dey, "Depthwise Separable Convolutions with Deep Residual Convolutions," *arXiv Preprint*, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Omid Nejati Manzari et al., "BEFUnet: A Hybrid CNN-Transformer Architecture for Precise Medical Image Segmentation," *arXiv preprint*, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Jiaxuan Li et al., "CFFormer: Cross CNN-Transformer Channel Attention and Spatial Feature Fusion for Improved Segmentation of Heterogeneous Medical Images," *Expert Systems with Applications*, vol. 295, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[31] Md Rakibul Islam et al., "Enhancing Semantic Segmentation with Adaptive Focal Loss: A Novel Approach," *arXiv preprint*, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]