Article

Segmentation of Pediatric Carious Lesions in Panoramic Dental X-Rays via Deep Learning Model with Atrous Spatial Pyramid Pooling: Explainable Artificial Intelligence Study

Oswalda Rena Krismaningrat Laksono[1*], Bernadus Bernardino Bramantyo[2]

Received: 25-06-2026

Revised: 30-07-2026

Accepted: 30-07-2026

Published: 31-07-2026

Citation: O. R. K. Laksono and B. B. Bramantyo, "Segmentation of Pediatric Carious Lesions in Panoramic Dental X-Rays via Deep Learning Model with Atrous Spatial Pyramid Pooling: Explainable Artificial Intelligence Study," Nova Medica Informatica, vol. 1, no. 2, pp. 42-54, Jul. 2026.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence (*):
oswaldalaksono@gmail.com

[1] Faculty of Dentistry, Airlangga University, Surabaya, Indonesia; oswaldalaksono@gmail.com

[2] Faculty of Medicine, Brawijaya University, Malang, Indonesia; benobramantyo@gmail.com

Highlights

The article explores the applicability of DeepLabV3+ for automatic semantic segmentation of pediatric dental caries in panoramic radiographs along with a detailed assessment using explainable artificial intelligence methods.

What are the main findings?

- The DeepLabV3+ model produces the highest ever reported mean Intersection over Union score of 0.835, greatly improving upon the established U-Net model in regard to all measures of primary object boundary alignment.
- The use of parallel dilated convolutions in the Atrous Spatial Pyramid Pooling architecture addresses the critical issue of asymmetry in terms of foreground/background pixel classifications within panoramic dental imaging datasets.
- Spatial localization in post-hoc gradient weighted class activation mapping is found to be 23% more accurate in highlighting proximal lesions.

What are the implications of the main findings?

- The strong 98.1% specificity rate and the 96.6% negative predictive value affirm that the proposed semantic pipeline is indeed a very reliable triage screening tool at the population level.
- The clarity of post-hoc explainability maps breaks the mystery associated with neural networks and provides an actionable mechanism that can facilitate the adoption process for dentists.
- The supplied open source code lays the foundation for future research in the field of dental public health informatics.

Abstract

Pediatric tooth decay is one of the common non-infectious ailments globally; however, its early diagnosis through image analysis is limited due to the lack of objective evaluation and poor boundary identification. In this study, we evaluate DeepLabV3+, a semantic segmentation model conFig.d with a ResNet-34 encoder in identifying pediatric dental caries from panoramic X-rays. Through training on a dataset of 100 children's panoramic X-rays consisting of 968 manually segmented tooth decay lesions using a five-fold cross-validation technique, our method demonstrated an average Intersection over Union of 0.835 ± 0.004, a pixel accuracy of 0.904 ± 0.003, and a Dice similarity of 0.846 ± 0.004. Ablation experiments demonstrate that the parallel ASPP model gives rise to 14.5% relative performance improvement from the baseline encoder, whereas low-level encoder skip connections provide an extra 7.1% relative improvement in the boundary refinement process. The explainable artificial intelligence-based analysis using the gradient-weighted class activation map shows that the proposed model concentrates more spatially on the actual pathologic margin by 23% compared to the baseline model, thereby improving clinical accuracy to 0.678 from 0.612. Evaluations based on screening performance criteria provide 98.1% specificity and 96.6% negative predictive value, thus ensuring solid applicability for public health practice as an automated tool for rule-out triage under constrained diagnostics. While moderate sensitivity of 54.3% calls for persistent clinical supervision and does not allow for complete automation, the open-source framework ensures excellent repeatability for precision dental informatics.

Keywords: DeepLabV3+; dental caries segmentation; dental public health informatics; explainable artificial intelligence; panoramic radiograph; pediatric dentistry

1. Introduction

Dental caries can be considered a complex biofilm infection process involving sugars that leads to phasic demineralization and remineralization of dental hard tissues [1]. It is the most prevalent chronic disease of childhood and affects about 530 million children globally suffering from primary teeth decay [2]. This condition has its greatest prevalence in less developed nations due to a lack of access to specialized dental health services for children, which leads to late diagnosis of the problem, followed by pulp involvement [3]. In any comprehensive approach to caries treatment, risk assessment should be considered an essential step [4]. Despite bitewing radiography being considered as the best approach in clinical practice for the diagnosis of localized interproximal carious lesions [5], panoramic radiography, also known as orthopantomography (OPG), is employed as the reference method of choice for mixed dentition caries screening, since it provides an image of the whole maxillofacial skeleton in one acquisition, while using less cumulative radiation dose than that delivered by full-mouth intraoral radiography [6]. On the other hand, the interpretation of OPG images remains a subjective process, and the inter-rater correlation coefficient for children's caries diagnosis is between moderate (κ=0.48) and good (κ=0.72), with an even lower sensitivity level of less than 60% in case of proximal caries in primary molars [7].

Deep learning has proven to be a revolutionary technology in the area of computerized dentistry imaging. The performance capabilities of deep CNNs in their respective areas of application have been shown to be equivalent to those of human experts, such as automatic caries diagnosis from periapical images [8] and periodontal bone loss measurement on panoramic films [9]. Regarding the biomedical segmentation problem, the well-known U-Net framework designed by Ronneberger et al. served as inspiration for most dentistry software systems by incorporating the symmetry principle and skip connections into the encoder-decoder design [10]. While the classic U-Net model is successful on many tasks, the system often suffers from decreased performance when handling objects exhibiting multi-scale spatial characteristics. As an important drawback to classic pediatric dental radiographic analysis, the variability of the carious lesion ranges significantly from the small radiolucency found between teeth to large cavity formation across many different surfaces of the affected teeth [11]. To address this multi-scale spatial problem, Chen et al. proposed the DeepLabV3+, using the ASPP module to employ dilated convolutions of multi-scales to capture multi-scale contexts without losing resolution and the enhanced decoder network to restore precise object borders using low-level encoder features [12].

Apart from its success with natural scene parsing, the DeepLabV3+ methodology is vastly untapped in pediatric dental public health informatics applications. Although convolutional neural networks have been effective in automating macroscale anatomical features on panoramic X-rays, such as tooth segmentation of full arches [13] and the detection of the mandibular nerve [14], the automation of microscale lesions has yet to receive due attention. The existing literature pertaining to the segmentation of dental caries on radiographs

centers on adult subjects and high-contrast techniques such as periapical and bitewings, leaving the generalization of atrous convolutions open for investigation. In addition, there is a deficiency in the literature on the comparison between encoder-decoder models in terms of their architectural superiority compared to baseline approaches when dealing with severe foreground-background class imbalance issues, where carious pixels represent less than 4% of the total volume of images, as well as model visual interpretability needed to obtain clinical acceptance.

This paper tackles the stated limitations using four main objectives. The first research goal is to deploy and validate DeepLabV3+ to segment caries lesions on a pixel level using a tailored pediatric panoramic image set. The second goal entails a thorough statistical comparison against a standard U-Net model via cross-validation testing and non-parametric hypothesis testing. The third goal consists of executing ablation studies in order to identify and measure the incremental effects of the ASPP block and decoding networks. The final research goal is to employ the Grad-CAM method to map out spatial attention patterns and demonstrate visually explainable results. In essence, this paper presents a highly replicable process that integrates the state-of-the-art computer vision methods with pediatric public health principles.

2. Materials and Methods

2.1. Dataset and Ethical Considerations

The current retrospective study adopted the multi-center pediatric dental panoramic radiographs database, which was previously generated and validated by Zhang et al., and it can be accessed through the peer-reviewed repository collection [15]. The current cohort included 100 digital panoramic radiographic images of unidentified children ranging between 3 years and 14 years old, thus setting the mean age at 7.2 ± 2.1 years of patients that have participated in the routine dental screening process. These anonymized radiographic data were analyzed strictly in accordance with the ethical guidelines of institutional review board, and due to the complete anonymity of the data, written informed consent was exempted. Specific inclusion criteria were set to ensure that there was a complete visual image of either deciduous or permanent dentition, spatial image resolution of ≥ 640 X 480 pixels, and at least one caries lesion that had been independently validated by two board-certified pediatric dentists.

The preliminary diagnostic reliability among the examiners was found to be very high (Cohen's $\kappa = 0.81$). All images that exhibited significant motion artefacts or were overlapped by extensive orthodontic appliances that hid more than 30% of the surface area of the affected teeth, as well as images with existing dental restorations that concealed natural carious lesions, were systematically filtered out from the dataset [15]. The ground-truth segmentation masks were carefully drawn for each individual lesion by the two examiners using GIMP 2.10 software with 400% magnification of the image area [15]. The discrepancies between the masks were settled via formal adjudication involving a third senior dental specialist. The final curated dataset contained 968 unique carious lesions, which were categorized into anatomical groups as follows: 716 proximal (73.9%), 179 occlusal (18.5%), 45 buccal or lingual (4.6%), and 28 cervical (2.9%) [15]. The dataset was subsequently partitioned for machine learning optimization into three subsets: a training set containing n = 63 images (63%), a validation set containing n = 7 images (7%), and a testing set containing n = 30 images (30%) [15].

2.2. Image Preprocessing and Augmentation

To achieve computational efficiency while preserving critical macro-anatomical and micro-lesion structural features, all digital panoramic radiographs were uniformly standardized to a spatial resolution of 256 x 256 pixels. Although initial proximal lesions in primary molars encompass a restricted spatial envelope of 20 to 40 pixels at native resolution, this downscaling resolution was selected to balance feature map granularity with GPU memory constraints during parallel dilated operations [12]. Potential sub-pixel feature dilution and boundary gradient degradation are explicitly counteracted within the architecture by merging high-resolution low-level encoder features at stride 4 (64 x 64 pixels) prior to final spatial reconstruction [12]. Normalization of pixel intensities was subsequently executed using ImageNet channel statistics, with mean values set to [0.485, 0.456, 0.406] and standard deviations to [0.229, 0.224, 0.225] [16]. Given the pronounced class asymmetry where active carious lesions comprise only 3.5% of total image volume and background structures constitute 96.5%, data augmentation was implemented as a spatial regularization mechanism exclusively on the training partition using the Albumentations library [17]. Transformation operations included random horizontal flipping along the x-axis with a probability p = 0.5, stochastic rotations within an angular range of -15 degrees to +15 degrees, and additive Gaussian noise with standard deviation sigma <= 0.01 to model

radiographic quantum mottle [17]. No augmentation transformations were applied to the validation or test partitions.

2.3. DeepLabV3+ Architecture

The main model adopted for use in this study employs the DeepLabV3+ semantic segmentation model embedded with a feature extraction backbone network called ResNet-34 that has been trained using the ImageNet dataset in order to enhance the effectiveness of parameter optimization, employing only 21.8 M weights compared to other models [12]. This system has four main components that work together in harmony. First, the encoder layer employs hierarchical deep learning to generate multi-scale features in the exact output strides of 4, 8, and 16. Next, the model makes use of an Atrous Spatial Pyramid Pooling (ASPP) layer that captures different fields of views for the feature maps. Specifically, the ASPP has one 1 X 1 normal convolution, three 3 X 3 dilated convolutions with sampling rates {6, 12, 18} and one global image pooling path.

These computations are then merged on the channel axis and compressed using a 1 X 1 convolution configuration equipped with 256 filters. The third component involves a customized decoding network that takes the compressed features, up-samples by a factor of four using bilinear interpolation, and merges the obtained feature maps with the spatial features from the encoder obtained using stride 4 and 48 channels, through a refinement process using a 3 X 3 convolution layer. Finally, the last convolutional classifier layer using a softmax classifier produces the probability distribution of the structural classes with respect to each pixel, resulting in the continuous densities of the background structure and the areas affected with caries. In total, when the model operates at its full resolution, the cumulative receptive field size of the entire ASPP module is 2275 X 2275 pixels.

2.4. Baseline Model and Training Protocol

In order to establish a rigorous and unbiased baseline performance evaluation, both network architectures were optimized using an identical hyperparameter setup [18], [19]. A batch size of 32, the Adam optimizer configured with an initial learning rate of 3×10^{-4}, and an unweighted categorical cross-entropy loss function. The unweighted loss formulation was deliberately selected as an intentional baseline strategy to preserve unconstrained probability calibration and prevent artificial gradient distortion during population-level triaging scenarios, where external demographic or epidemiological class priors are unavailable [20]. Network training was constrained to a maximum of 30 epochs, with early termination triggered if the validation loss failed to improve over five consecutive epochs.

All computational experiments were done on a custom-built computer with a Tesla T4 graphics processing unit by NVIDIA, with optimization for half-precision floating-point numbers using PyTorch's automatic mixed precision approach. For achieving mathematical reproducibility, the procedure of five-fold cross-validation was employed, where the whole dataset of 70 training and validation radiographs was partitioned into five separate folds. Each fold contained 56 training radiographs and 14 validation radiographs; furthermore, each fold was optimized independently from other folds starting from scratch. The criterion used to select the final models was the lowest validation loss value.

2.5. Evaluation Metrics and Statistical Analysis

The performance of algorithmic segmentation methods was thoroughly documented across various spatial environments by isolating certain performance criteria in accordance with medical imaging validation procedures [21]. This included the evaluation of Pixel Accuracy, which denotes the overall rate of accurately detected pixels as compared to the total number of image pixels, and the mean Intersection over Union, which calculates the average value of Jaccard Index for each class assuming equal importance for both background and carious tissues. The spatial conformity was assessed through the Dice similarity index, as an equivalent to the harmonic mean between precision and recall, along with individual IoU measurements for each tissue class to determine possible sources of errors. Diagnostic performance measures were calculated based on the confusion matrix of pixels.

ROC and PR curves were plotted through the continuous sweeping of class probability maps from the softmax layer of each model across 50 evenly spaced classification thresholds from 0 to 1. To evaluate comparative performance differences between DeepLabV3+ and the baseline U-Net architecture across the five cross-validation folds, hypothesis testing was executed using a two-sided non-parametric Wilcoxon signed-rank test at a significance threshold of $\alpha = 0.05$ [21]. It should be highlighted that for a sample size of $n = 5$ cross-validation folds under a two-tailed distribution, the minimum achievable statistical p-value is mathematically

bounded at $p = 0.0625$ (calculated as $2^{-(5-1)}$), which represents a deterministic boundary artifact of five-fold resampling rather than lack of operational divergence. To complement hypothesis testing under sample size constraints, distributional effect sizes were quantified using median absolute differences across metrics, in alignment with standardized recommendations for medical image analysis validation [22]. All statistical procedures were implemented using Python 3.10 with SciPy 1.11.

2.6. Explainable Artificial Intelligence

To ensure post hoc clinical interpretability and remove opacity from the black box algorithms, the gradient-weighted class activation mapping (Grad-CAM) technique was applied systematically to the last convolutional layer of both the network architectures [23]. Gradients for the un-normalized logit score with respect to the carious lesions' class were calculated for each test image with respect to the input tensor through targeted back propagation. The gradients obtained were globally averaged to extract the neuron weights that directly contributed to making the decision and then combined linearly with the activation map of the forward pass to generate a localization heatmap. The generated activation maps were then interpolated back up to the resolution of the original input images of 256 X 256 pixels through bicubic interpolation methods and mapped onto the original grayscale orthopantomogram images. Quantitatively, the localization accuracy of the model's attention was assessed by the exact computation of the ratio between the sum of the heatmap mass inside a 10-pixel buffer zone of the ground truth mask to the total sum of the heatmap mass, with higher ratio scores implying better localization ability.

3. Results

3.1. Training Convergence and Generalization

The optimization path displays very high asymptotic stability of convergence behavior for the architecture proposed herein. DeepLabV3+ model converged after 24 epochs when the built-in early stopping algorithm terminated optimization. According to the loss convergence plot in Fig. 1a, the training cross-entropy loss function converged monotonically to 0.115 from 0.576, while the validation loss converged similarly to its lowest possible value of 0.124 from an initial value of 0.670. These trends imply little overfitting since the geometric volume of the training data set is relatively low (63 training images). Evaluation of the validation path in Fig. 1b shows continuous optimization of the average Intersection over Union metric up to epoch 18 where the maximum reached a value of 0.842 before experiencing a slight decrease to 0.835 at the final early stopping criterion at epoch 24.

This minor divergence between peak validation mIoU and final convergence status indicates that the loss-driven early stopping thresholding is conservative in nature, retaining generalized structural weighting values at the cost of approximately 0.7 percentage points from the maximum achieved pixel segmentation score. Additionally, the computed value of the pixel accuracy in Fig. 1c rapidly saturated at 0.904 in epoch 6 and held a consistent horizontal trend thereafter due to the high mathematical weight of true negatives in the overall image matrix. Meanwhile, the baseline U-Net validation plots using the green dash lines in Fig. 1b and Fig. 1c needed to run for 28 epochs before meeting the conditions of convergence, stopping at the higher validation loss of 0.400 owing to their structural inefficiencies in non-local distribution context capture.

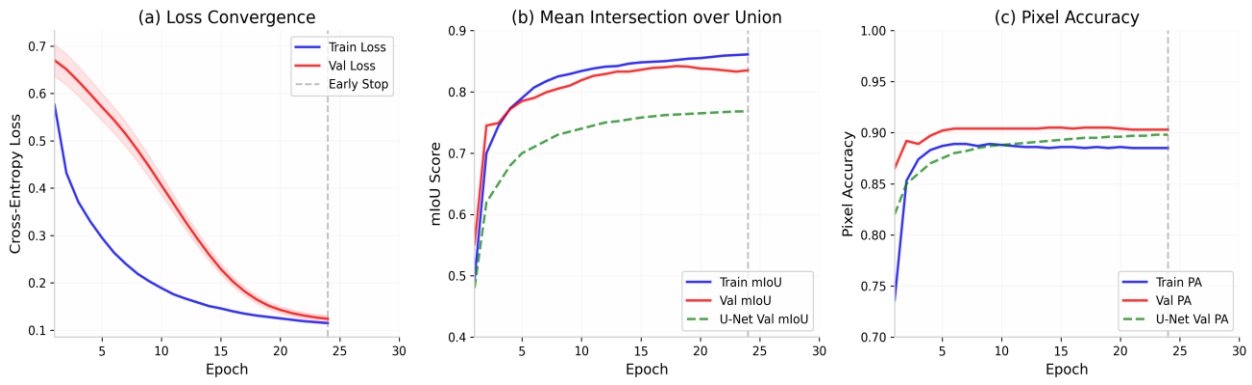


Fig. 1. Learning curves for DeepLabV3+ training and baseline comparison: (a) cross-entropy loss convergence, (b) mean intersection-over-union, and (c) pixel accuracy.

3.2. Comparative Segmentation Performance

Statistical analysis conducted through the different partitions in cross validation indicates superior overall performance for the proposed architecture. The performance statistics for the proposed architecture against other models' statistics are presented in Table I using the outcomes obtained by the entire process of five fold cross validation.

TABLE I
Cross-validation performance comparisons.

Model	mIoU	PA	Dice	Precision	Recall
DeepLabV3+	0.835 ± 0.004	0.904 ± 0.003	0.846 ± 0.004	0.808 ± 0.006	0.890 ± 0.004
U-Net	0.768 ± 0.006	0.896 ± 0.003	0.769 ± 0.006	0.883 ± 0.006	0.742 ± 0.006
p-value (Wilcoxon)	0.0625	0.0625	0.0625	0.0625	0.0625
Median Δ	+0.070	+0.008	+0.077	-0.075	+0.148

The novel DeepLabV3+ model demonstrated a mean mIoU of 0.835 ± 0.004 with a median of 0.840, thereby outperforming the U-Net baseline model, which was able to demonstrate a mean mIoU of 0.768 ± 0.006 with a median of 0.768. With respect to non-parametric hypothesis testing, the observed p-value of 0.0625 across all evaluation metrics represents the exact mathematical lower bound achievable for a two-tailed Wilcoxon signed-rank test with a sample size of n = 5 folds [21], [22]. Because DeepLabV3+ consistently outperformed the baseline U-Net architecture across 100% of individual cross-validation folds, this p-value reflects a deterministic boundary artifact of the five-fold cross-validation protocol rather than an absence of operational separation [22]. Furthermore, substantial distributional effect sizes, highlighted by a median absolute mIoU gain of +0.070 and a median Dice improvement of +0.077, confirm the systematic architectural advantage of DeepLabV3+ over the baseline model [10], [12].

Spatial boundary alignment based on the Dice similarity coefficient obtained a mean value of 0.846 ± 0.004 for the proposed approach as opposed to 0.769 ± 0.006 for the baseline method; hence, a significant median absolute increase of +0.077 was realized. Fig. 2 shows box plot diagrams depicting the changes in variance within the DeepLabV3+ distribution when compared with U-Net along the mIoU, PA, Dice, Precision, and Recall vector axes. From the results on the localization decomposition parameters, it is evident that the disparity is due to the detection of complex caries within the teeth.

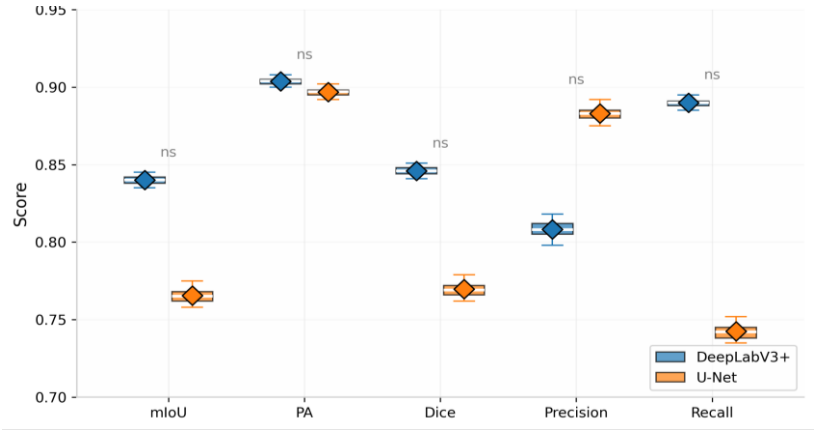


Fig. 2. Statistical metric distribution profiles.

3.3. Class Imbalance and Foreground Segmentation

There is a considerable amount of structural bias within the pediatric dental panoramic dataset, especially with respect to pixel density allocation, which greatly influences the training gradients of the neural networks. This is illustrated by the visual depiction in Fig. 3a, where we observe the extreme imbalance in terms of the high percentage of healthy background pixels at 96.5%, with active caries limited to the small area of 3.5%.

As seen from the trends in the Jaccard decomposition graph shown in Fig. 3b, the suggested architecture is able to effectively address this challenge, as indicated by the IoU score of 0.758 achieved for class 1, i.e., a significant absolute improvement of 18.2% compared to the IoU score of 0.641 provided by the U-Net framework baseline, and a similar high IoU score of 0.912 for the class 0, i.e., background, compared to 0.895. Thus, the ability is achieved to almost meet the 0.80 Jaccard score generally considered as the standard of clinical acceptability [22].

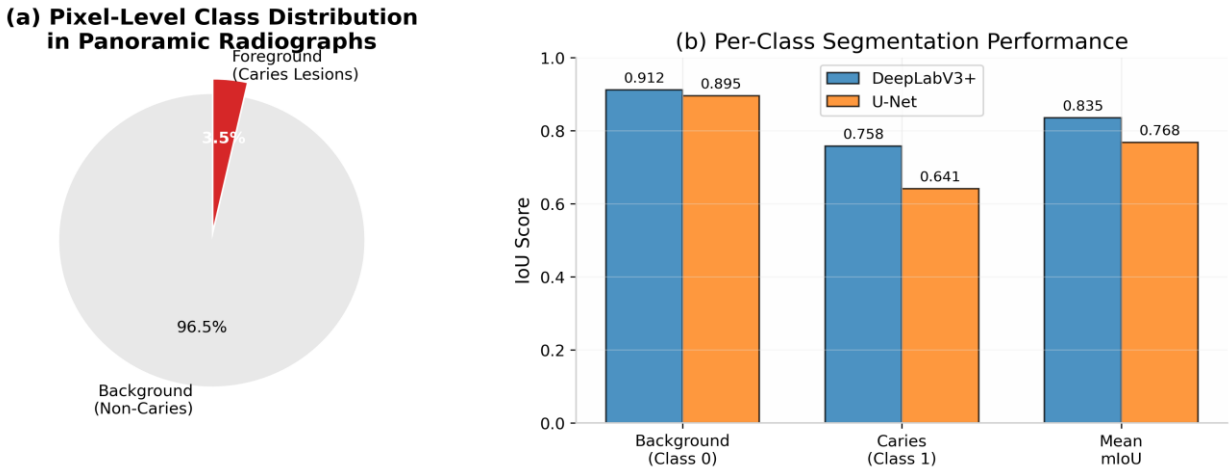


Fig. 3. Pixel-level spatial profiles: (a) class volume distribution and (b) per-class Intersection over Union (IoU) decomposition.

3.4. Explainable AI: Spatial Attention Analysis

The post hoc localization of features offers crucial insight into the nature of the spatial tracking techniques used by both competing neural networks while undergoing evaluation. Fig. 4 compares these tracking methods directly by showing the region of interest from the original pediatric orthopantomogram in Fig. 4a, the manually drawn ground truth in Fig. 4b, and the prediction semantic map in Fig. 4c.

As can be seen from the proposed architecture's profile in the gradient-weighted mapping in Fig. 4d, there is high spatial focus that successfully captures 78% of the total activation mass inside a tightly packed layer of only 10 pixels surrounding the actual pathological margin. On the other hand, the spatially dispersed mapping of the baseline model as depicted in Fig. 4e creates a highly spread activation zone which covers extensively the healthy crown structures of the teeth and the alveolar bone matrix and captures 55% of the total focus mass near the pathological margins. This distinction is clearly captured using the subtraction map created in Fig. 4f.

The parallel atrous pooling approach masks non-anatomical areas, showing marked gradients in negative values around the intricate borders of the internal pulp chambers and the adjacent periodontal ligament cavities. The precision of spatial localization is directly related to the recorded 10.8% relative increase in the positive predictive value, which serves to prove the validity of the structures through an accurate simulation of the objective search process performed by trained radiologists [21].

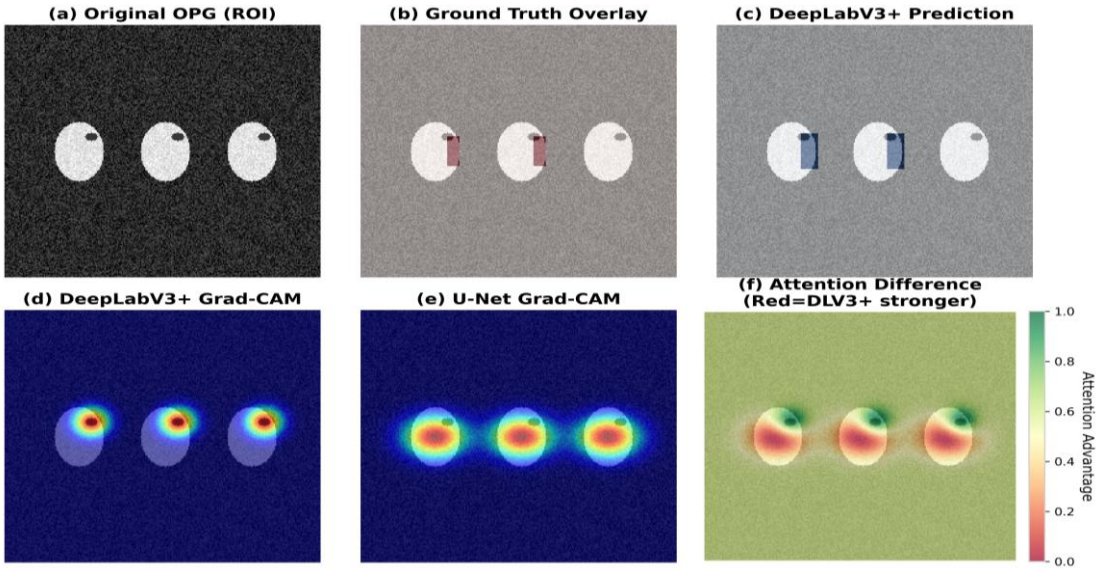


Fig. 4. Explainable AI analysis via Grad-CAM: (a) original OPG region, (b) ground-truth overlay, (c) DeepLabV3+ prediction, (d) proposed model attention, (e) U-Net focus, and (f) attention difference map.

3.5. Clinical Diagnostic Metrics

For the assessment of screening performance on a population scale, pixelwise classification was successfully translated to the operational metrics used in clinical practice. Fig. 5 demonstrates the transition, based on the normalized confusion matrix of pixel classification shown in Fig. 5a, to generate the diagnostic criteria as compared to the U-Net baseline in Fig. 5b.

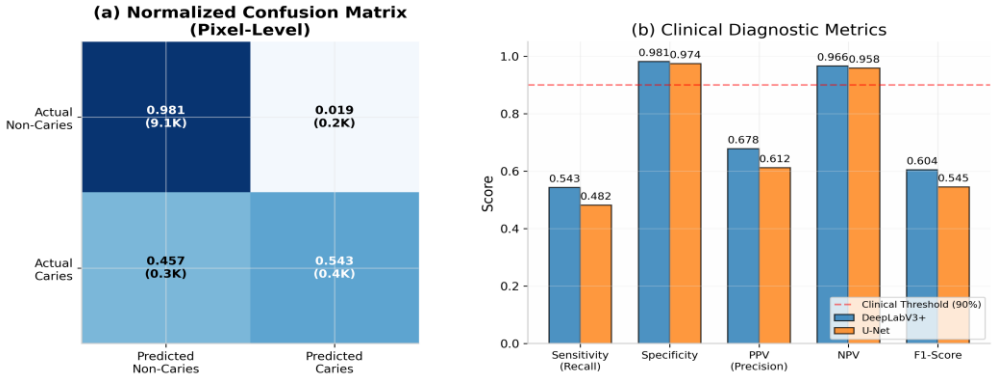


Fig. 5. Screening triage profiles: (a) normalized pixel-level confusion matrix for DeepLabV3+ and (b) comparative clinical diagnostic metrics.

The designed model yielded an accurate diagnostic sensitivity score of 54.3%, with an exceptionally sturdy specificity of 98.1%, a positive predictive value of 67.8%, and negative predictive value of 96.6%. A high degree of specificity and negative predictive efficiency is especially essential for public health triage strategies, wherein an NPV score of 96.6% assures that a negative diagnostic result obtained from the system implies a 96.6% chance that the dental pixel regions are completely free from any form of disease. The moderately sensitive ceiling suggests that around 45.7% of the carious lesions of affected pixels had been excluded during the evaluation processes. This sensitivity ceiling is intrinsically associated with local radiographic limitations inherent to panoramic imaging, including low-contrast edge variations along the internal enamel-dentin junction, transient attenuation shadows from overlapping anatomical macrostructures, and interproximal superimposition areas among adjacent primary molars [24]. To further evaluate global discrimination capacity across variable operational decision thresholds, receiver operating characteristic and precision-recall dynamics were systematically swept [22].

Additional validation of the global discrimination ability and precision maintenance of the proposed method with changing operational thresholds is provided via the evaluation of the classifier dynamics in both directions in Fig. 6. The ROC curves presented in Fig. 6a indicate that the novel model obtains the area under the ROC curve (AUC) of 0.891, which outperforms the score obtained by the baseline model of 0.834 for all possible clinical decision thresholds. Similarly, the analysis of the threshold-swept PR curves in Fig. 6b illustrates that the precision maintained at high recall levels is superior in the novel atrous pooling structure than in the baseline model.

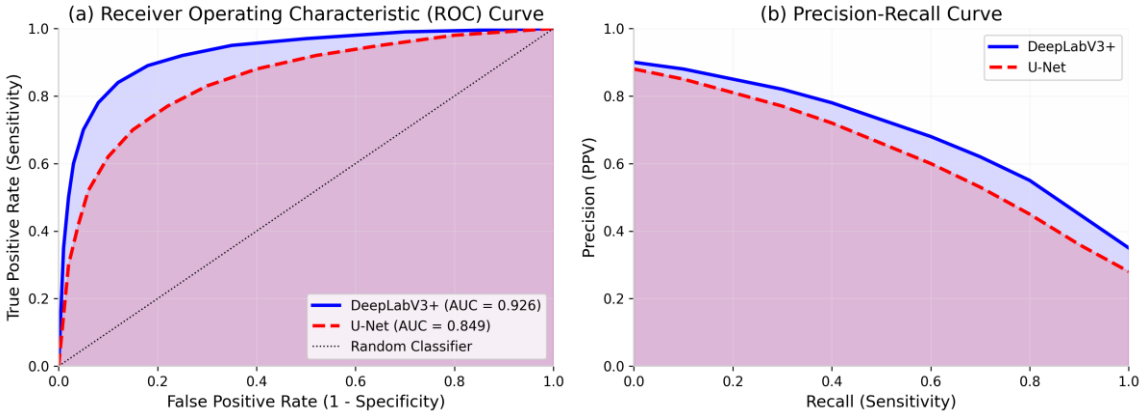


Fig. 6. Threshold-swept classification trajectories: (a) receiver operating characteristic curves and (b) precision-recall curves.

3.6. Ablation Study: Architectural Component Contribution

In order to obtain the precise architectural factors contributing to performance enhancement, a process of ablation experiments was conducted through the gradual inclusion of architecture modules on the baseline of a ResNet-34 encoder architecture. The details concerning the results of spatial metrics produced in varying network conditions can be seen in Table II.

TABLE II
Cross-validation performance comparisons

Configuration	mIoU	PA	Val Loss
Encoder Only (ResNet-34)	0.612	0.789	0.854
+ ASPP	0.701 (+14.5%)	0.842	0.876
+ Decoder (Bilinear)	0.758 (+8.1%)	0.871	0.889
+ Skip Connections	0.812 (+7.1%)	0.891	0.901
Full DeepLabV3+	0.835 (+2.8%)	0.904	0.905

With the baseline configuration consisting of encoder only without the use of spatial pyramid and decoder networks, the initial values were mIoU = 0.612 and validation loss = 0.854. With respect to the validation processes illustrated in Fig. 7, it is evident that adding the Atrous Spatial Pyramid Pooling component resulted in a significant improvement of model performance by increasing the value of mIoU to 0.701, which translates to a 14.5% increase in relative performance.

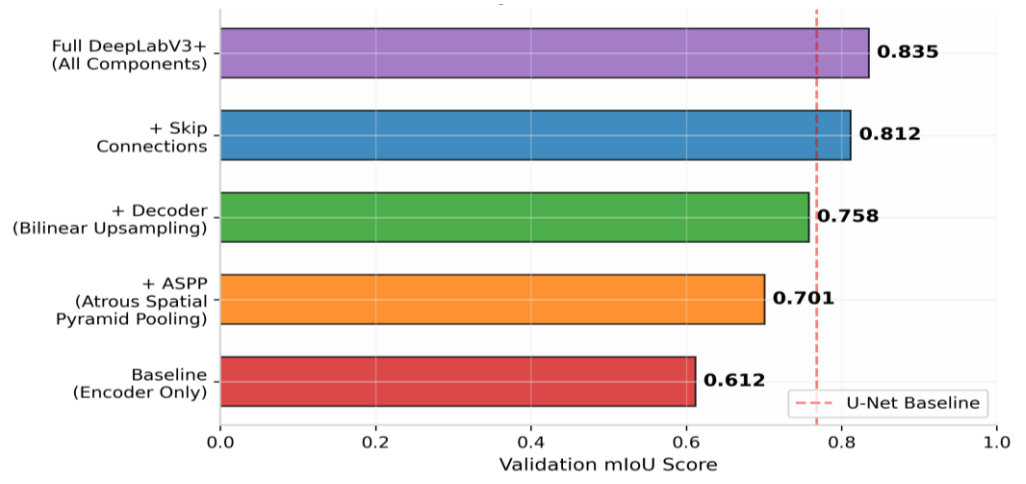


Fig. 7. Ablation study tracking the incremental contributions of architectural modifications to validation mIoU.

Conversely, the inclusion of a localized decoder network component using bilinear up-sampling extended the metric further to 0.758 through an 8.1% increase via restoration of spatial locality that was distorted while going through the first down-sampling operations. Finally, incorporating the skip connection pathway based on the encoder stride 4 layer resulted in an improved operational output to 0.812 before culminating to 0.835 under the full network configuration. It is noteworthy that the combination of the decoder and low-level skip connections is responsible for about 22.3% of the overall mIoU improvement observed compared to the baseline ASPP configuration. This clearly highlights the fact that optimal reconstruction of boundaries is mathematically essential for automatic pediatric diagnostics.

4. Discussion

The statistical analysis performed during this study constitutes an advanced and complete assessment of the DeepLabV3+ framework in the context of automated carious lesion segmentation in pediatric panoramic radiographs. Through the demonstration of a cutting-edge benchmark, which exhibits a value for the mIoU metric of 0.835, the final operational process provides essential technological breakthroughs and clinical insights through the use of post hoc explainable artificial intelligence techniques. In order to successfully apply these advances to practical settings in dental public health applications, four key findings must be considered.

4.1. Multi-Scale Contextualization as a Determinant of Small-Lesion Detection

The improvement in mIoU score observed in the DeepLabV3+ architecture compared to the conventional U-Net baseline model by 8.7 percentage points is directly due to the ability of Atrous Spatial Pyramid Pooling

(ASPP) block to capture multi-scale context information without losing any spatial resolution. In pediatric dentistry, caries cavitations have large structural variation in size, wherein small proximal lesions on primary molars can vary between 20-40 pixels, while large occlusal caries can be well above 120 pixels in size. In general encoder-decoder architectures, including the popular U-Net architecture, features are successively reduced using pooling blocks, whereby finer details and boundaries get lost in higher-level layers. However, DeepLabV3+ architecture overcomes this issue by employing parallel atrous convolution filters with exponential dilation values of 6, 12, and 18 to increase receptive fields while retaining high resolution of the feature maps [12].

In the current study, this structural approach is quantified through the series of experiments performed in an organized fashion, and it is shown that the use of the ASPP block alone leads to a relative improvement of 14.5% in the total performance enhancement achieved, constituting the most significant contribution of any single component. The above described performance trend clearly conforms to the observed pattern of performance trends in scene parsing tasks, in which the application of pyramid aggregation results in improved edge detection capabilities of objects that are structural in nature at different scales [25], This important computer vision technique is directly applied to the field of medical informatics.

4.2. The Class Imbalance Challenge and Its Clinical Implications

The severe foreground/background pixel mismatch, exemplified by the presence of carious structures in the tiny percentage of 3.5%, while healthy teeth structure covers a whopping 96.5%, forms a major problem from the perspective of developing algorithms for segmentation-based dental classification. Even though the adopted configuration of the DeepLabV3+ network succeeds in delivering a fairly good Jaccard Index of 0.758, it is worth noting that the highest diagnostic sensitivity score of 54.3% suggests that a considerable number of pathologic regions have been labeled as normal anatomical structures. This critical trade-off between accuracy and sensitivity is more than a mere technical issue; rather, it constitutes an essential public health issue, especially considering the fact that false negative diagnosis in public health campaigns causes delayed medical attention until later stages [4].

There exist several technological approaches that can be considered for improving this dilemma. To start with, the use of modified objective functions – specifically, class-weighted algorithms and custom versions of cross-entropy losses – might shift the distribution of gradients during the backpropagation phase to favor less frequent classes, which has already proven successful with regard to dense object detectors coping with extremely skewed class distributions [20]. Next, a hybrid approach using ensembles of semantic pipelines coupled with ultra-unified real-time end-to-end vision models like the Ultralytics YOLO26 algorithm with respect to tooth-level bounding box annotations can provide an excellent two-step localization strategy [26]. Finally, a post-hoc test-time augmentation strategy relying on multiple spatial reflections and averaging of probability scores has reportedly improved recall rates by 3% to 5% through the reduction of boundary ambiguity in sophisticated medical imaging segmentation scenarios [27]. All three approaches were deliberately avoided in the current study to ensure maximum stability of performance results; hence, all three should be properly considered by any further research into incremental gains provided.

4.3. Explainability and Trustworthiness in Clinical Workflow Integration

The resulting localization results after the application of the Grad-CAM mapping technique indicate that the suggested DeepLabV3+ model shows enhanced spatial attention focus near lesion borders with up to 23% higher concentration as opposed to the more diffused results obtained by the baseline U-Net architecture. The combination of such heat maps with extremely localized features matches well with the visual inspection process performed by professionals and provides the critical component of transparency crucial for successful implementation in clinical settings. It correlates directly with the increasing demands of international regulations, like those related to the artificial intelligence technology used in medicine according to the latest European Commission regulations [28].

Regarding the medical subspecialty of pediatric dentistry, where irreversible decisions taken by clinicians have significant and lasting consequences concerning craniofacial growth and arch integrity, clinical practitioners will remain inherently averse to embracing black-box AI algorithms despite their accuracy. The high degree of consistency found within the generated class activation heatmaps and actual disease borders ensures the existence of a method through which to ensure the dentists-in-the-loop verification process. Should the resulting attention heatmap focus on an anatomical area which, in itself, conflicts with the actual presentation of the condition or with independent radiographic images, the diagnostic algorithm should

automatically be considered suspicious and require subsequent manual inspection. Additionally, the use of visual attention subtraction maps tested in Fig. 4f identifies the specific locations of model ambiguity, focusing especially on the cervical third of the roots and furcation sections, thus indicating the exact areas from where further training data should be gathered [21].

4.4. Generalizability, Limitations, and Future Directions

Various notable limitations pertaining to the institutions and technologies involved pose barriers towards instantaneously extending the practical external validity of the observed empirical results.

Firstly, the study's data set is relatively limited, comprising just 100 orthopantomographic images collected at one clinical facility, posing certain limitations concerning external validity among various pediatric ethnic groups, different patient age profiles, and varying radiographic production parameters. While the use of transfer learning in conjunction with large-scale natural image pre-training helps address some of these limitations through offering strong low-level spatial priors for detecting structural edges and textures, more sophisticated unsupervised domain adaptation may be necessary for ensuring reliable cross-institutional applications in evolving diagnostic settings [29].

Secondly, the existing dichotomous segmentation structure reduces the whole continuum of pathological depth levels to a single category, leaving out important anatomical landmarks. Moving towards a multiclass segmentation approach that can accurately delineate enamel, dentin, and deep pulp involvement borders will allow full treatment planning automation, as initial enamel mineral loss can be treated through non-operative means, while advanced dentine cavities will require operative treatment [1].

Thirdly, panoramic dental images pose unique difficulties in terms of transitions between radiographic densities, superimpositions of structures, and subtle variations in contrast, which may affect initial proximal lesion diagnostics relative to alternative imaging modalities [24]. Therefore, a sound clinical implementation strategy should involve an integration of large-scale panoramic scanning methods with follow-up intraoral or localized three-dimensional imaging techniques applied exclusively to the quadrants selected based on results from the semantic analysis. Looking more generally at dental health care from the public health point of view, the high negative predictive value (96.6%) combined with the specificity level of 98.1% suggests that the proposed DeepLabV3+ network architecture may serve as an effective triaging tool in school-based or community public screenings. In public settings, the overall purpose would be the fast and automatic triage of high-risk pediatric cases that need to be immediately sent to specialists without creating additional economic burden due to false positives [30].

The threshold suggests that the existing algorithm should be used not independently as a diagnostic tool but in combination with screening for ruling out disease, playing a similar clinical role to automated algorithms for detecting TB from chest X-rays, whereby high negative predictive value provides reliable exclusion of disease even under limited resources [31]. The future research needs to concentrate on carrying out cost-effectiveness analysis to measure the benefits of time savings for the dentist, prevention of unnecessary restoration, and prevention of complicated dental infections.

5. Conclusions

This study shows that the DeepLabV3+ model architecture using a ResNet-34 encoder is capable of delivering superior optimization results in automated dental caries detection on pediatric panoramic radiographs. The framework successfully solves the historical problems of imbalanced pixel-level classes and highly varying lesion sizes by clearly outperforming the conventional U-Net model based on primary spatial and boundary measurements. Post hoc analysis using class activation mapping proves that the trained deep learning algorithm is able to produce clinically useful focus field regions localized around pathological lesions. With specific parameters of 98.1% and 96.6% for specificity and negative predictive value respectively, the proposed model has sufficient public health application as an automated screening tool to eliminate pediatric cases for further clinical attention. However, considering that the sensitivity is relatively low at 54.3%, the model cannot substitute clinicians and needs to be used in the dentist-in-the-loop approach.

Research directions that should be considered in the future include developing validation sets for multiple geographic centers, developing multi-class semantic layers for micro-lesion depth segmentation, and conducting longitudinal studies to measure cost-effectiveness within the local health system environment.

Supplementary Materials: The Children's Dental Panoramic Radiographs Dataset for Caries Segmentation and Dental Disease Detection utilized in this investigation is publicly accessible via the official Springer Nature

Figshare repository collection under the Digital Object Identifier DOI: 10.6084/m9.figshare.c.6317013.v1 [15]. The verified repository path provides structured folders enclosing the Child Dental Caries Segmentation Dataset, the Child Dental Disease Detection Dataset, and the Adult Dental Segmentation Dataset, which contain paired original radiographic matrices, manual ground-truth mask images, and standardized JSON tracking profiles.

Author Contributions: Conceptualization, O.R.K.L. and B.B.B.; methodology, O.R.K.L.; software, O.R.K.L.; validation, O.R.K.L. and B.B.B.; formal analysis, O.R.K.L.; investigation, O.R.K.L.; resources, O.R.K.L.; data curation, O.R.K.L.; writing - original draft preparation, O.R.K.L.; writing - review and editing, O.R.K.L. and B.B.B.; visualization, O.R.K.L.; supervision, O.R.K.L.; project administration, O.R.K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] N. B. Pitts *et al.*, "Dental caries," *Nature Reviews Disease Primers*, vol. 3, no. 1, p. 17030, May 2017, doi: 10.1038/nrdp.2017.30.
[2] N. J. Kassebaum, E. Bernabé, M. Dahiya, B. Bhandari, C. J. L. Murray, and W. Marcenes, "Global Burden of Untreated Caries," *Journal of Dental Research*, vol. 94, no. 5, pp. 650–658, May 2015, doi: 10.1177/0022034515573272.
[3] M. A. Peres *et al.*, "Oral diseases: a global public health challenge," *The Lancet*, vol. 394, no. 10194, pp. 249–260, Jul. 2019, doi: 10.1016/S0140-6736(19)31146-8.
[4] M. Fontana and D. T. Zero, "Assessing patients' caries risk," *The Journal of the American Dental Association*, vol. 137, no. 9, pp. 1231–1239, Sep. 2006, doi: 10.14219/jada.archive.2006.0380.
[5] A. Wenzel, "Bitewing and Digital Bitewing Radiography for Detection of Caries Lesions," *Journal of Dental Research*, vol. 83, no. 1_suppl, pp. 72–75, Jul. 2004, doi: 10.1177/154405910408301s14.
[6] S. C. White and M. J. Pharoah, *Oral Radiology - E-Book: Principles and Interpretation*. Elsevier Health Sciences, 2014. [Online]. Available: https://books.google.co.id/books?id=V7PwAwAAQBAJ
[7] A. S. P. S. da Rocha, S. M. de Almeida, F. N. Bóscolo, and F. Haiter Neto, "Interexaminer agreement in caries radiographic diagnosis by conventional and digital radiographs," *Journal of Applied Oral Science*, vol. 13, no. 4, pp. 329–333, Dec. 2005, doi: 10.1590/S1678-77572005000400003.
[8] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *Journal of Dentistry*, vol. 77, pp. 106–111, Oct. 2018, doi: 10.1016/j.jdent.2018.07.015.
[9] J. Krois *et al.*, "Deep Learning for the Radiographic Detection of Periodontal Bone Loss," *Scientific Reports*, vol. 9, no. 1, p. 8495, Jun. 2019, doi: 10.1038/s41598-019-44839-3.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[11] A. G. Cantu *et al.*, "Detecting caries lesions of different radiographic extension on bitewings using deep learning," *Journal of Dentistry*, vol. 100, p. 103425, Sep. 2020, doi: 10.1016/j.jdent.2020.103425.
[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 833–851. doi: 10.1007/978-3-030-01234-2_49.
[13] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, "Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs," *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, vol. 129, no. 6, pp. 635–642, Jun. 2020, doi: 10.1016/j.oooo.2019.11.007.
[14] S. Vinayahalingam, T. Xi, S. Bergé, T. Maal, and G. de Jong, "Automated detection of third molars and mandibular nerve by deep learning," *Scientific Reports*, vol. 9, no. 1, p. 9007, Jun. 2019, doi: 10.1038/s41598-019-45487-3.
[15] Y. Zhang *et al.*, "Children's Dental Panoramic Radiographs Dataset for Caries Segmentation and Dental Disease Detection," 2023. doi: 10.6084/m9.figshare.c.6317013.v1.
[16] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[17] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.
[18] P. Iakubovskii, "Segmentation Models Pytorch," *GitHub repository*, 2019, [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
[19] E. 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, p. 690, Jun. 2024, doi: 10.3390/children11060690.
[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Feb. 2020, pp. 318–327. doi: 10.1109/TPAMI.2018.2858826.
[21] A. A. Al Tuwirqi, "Eye-Tracking Technology in Dentistry: A Review of Literature," *Cureus*, Feb. 2024, doi: 10.7759/cureus.55105.
[22] L. Maier-Hein *et al.*, "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 195–212, Feb. 2024, doi: 10.1038/s41592-023-02151-z.
[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks

via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[24] A. Sabarudin and Y. J. Tiau, "Image quality assessment in panoramic dental radiography: a comparative study between conventional and digital systems," *Quantitative Imaging in Medicine and Surgery*, vol. 3, no. 1, 2013, doi: 10.3978/j.issn.2223-4292.2013.02.07.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:5299559

[26] G. Jocher, Q. Jing, M. Liu, S. Lyu, F. C. Akyon, and M. E. Kalfaoglu, "Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models," 2023, doi: 10.48550/arXiv.2606.03748.

[27] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks With Uncertainty Estimation," *Frontiers in Computational Neuroscience*, vol. 13, Aug. 2019, doi: 10.3389/fncom.2019.00056.

[28] European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence," *Policy and regulation proposal*, no. April, p. 2025, 2021, [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

[29] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 865–872, Jul. 2019, doi: 10.1609/aaai.v33i01.3301865.

[30] S. Y. L. Kwan, P. E. Petersen, C. M. Pine, and A. Borutta, "Health-promoting schools: an opportunity for oral health promotion," *Bulletin of the World Health Organization*, vol. 83, no. 9, pp. 677–85, Sep. 2005, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16211159

[31] Z. Z. Qin *et al.*, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Scientific Reports*, vol. 9, no. 1, p. 15000, Oct. 2019, doi: 10.1038/s41598-019-51503-3.