*JDHBI - Journal of Digital Health and Biomedical Informatics, 2026, 1:1*
doi: 10.5753/jdhbi.2026.7916


**Research Paper**

# Tooth Detection and Segmentation on Occlusal Surfaces in Early Mixed Dentition Using Deep Learning

**Bruna Cristine Dias** [Universidade Federal do Paraná | *brunacristined@gmail.com* ]
**Mateus Felipe de Cássio Ferreira** [Universidade Federal do Paraná | *mateus.fecassio@gmail.com* ]
**Luan Matheus Trindade Dalmazo** [Universidade Federal do Paraná | *luantrindade@ufpr.br* ]
**Luciana Reichert da Silva Assunção** [Universidade Federal do Paraná | *lurassuncao@yahoo.com.br* ]
**Lucas Ferrari de Oliveira** [Universidade Federal do Paraná | *lferrari@inf.ufpr.br* ]

*Informatic Department, Universidade Federal do Paraná,R. Evaristo F. Ferreira da Costa, 383-391 - Jardim das Américas, Curitiba, PR, 81530-090, Brazil.*

**Abstract.** Artificial intelligence, particularly Convolutional Neural Networks (CNNs), has shown great potential for the detection of oral health conditions, including dental caries. Objectives: To evaluate the performance of CNNs in the detection and segmentation of posterior teeth during the early phase of mixed dentition. A total of 945 images of posterior teeth were collected in a school-based setting. The dataset was divided into training and testing subsets. Three CNN architectures — YOLOv11, U-Net, and DeepLabv3 — were compared based on precision, recall, and their ability to detect partially erupted teeth. YOLOv11 achieved a precision of 0.967 and a recall of 0.938, successfully identifying 96.1% of partially erupted permanent teeth. U-Net demonstrated a precision of 0.953 and a recall of 0.951, detecting 78.4% of partially erupted teeth, while DeepLabv3 achieved a precision of 0.963 and a recall of 0.944, detecting 76.5% of these teeth. All three CNNs demonstrated high accuracy in detecting and segmenting posterior teeth in children. However, YOLOv11 outperformed both U-Net and DeepLabv3 in detecting partially erupted teeth, highlighting its potential for use in studies involving early mixed dentition. Identifying the most suitable CNN architecture for detecting caries in mixed dentition may help standardize diagnosis and reduce clinical time.

**Keywords:** Artificial Intelligence, Deep Learning, Convolutional Neural Networks, Mixed Dentition

**Received:** 23 Mar 2026 • **Accepted:** 14 May 2026 • **Published:** 27 May 2026

## 1 Introduction

Dental caries is a disease dependent on the presence of biofilm and determined by sugar in the diet (Conrads and About (2018), Selwitz *et al*. (2007)). However, some host specificities may increase the risk of the disease, such as the mixed dentition phase (Lynch (2013)). This phase begins around six years of age and is characterized by the concomitant presence of primary and permanent teeth at different stages of development. The initial phase of mixed dentition, which occurs from six to eight years of age, is particularly challenging for the development of the disease due to the immature formation of the enamel of newly erupted permanent teeth and the difficulty in cleaning posterior teeth (Lynch (2013), Shi *et al*. (2016)).

The carious lesion represents the clinical sign of the disease and begins with the demineralization of the enamel, the outermost layer of the tooth, and may progress to deeper tissues such as dentin and, in advanced stages, the pulp (Marsh and Zaura (2017)). Accurately identifying the different stages of carious lesions is essential, since the therapeutic approach depends on the extent and depth of the affected tissues (Nyvad *et al*. (2011)). On the other hand, identifying the different stages can present challenges, especially in their early phases, depending on the examiner's experience (Ismail (2004)). In this sense, the literature shows a growing interest in investigations that use CNNs as an aid in the classification of carious lesions by means of intraoral photographs (Alharbi and Alhasson (2024)).

Deep learning, in particular, Convolutional Neural Networks (CNNs), has been widely explored in recent years as a tool to support the diagnosis of various health conditions, due to its ability to automatically learn hierarchical and abstract representations from images (LeCun *et al*. (2015)). Medical images can present high complexity, resulting from variability in contrast, the presence of artifacts, and anatomical differences between patients, factors that make manual feature extraction a difficult process and subject to biases (Litjens *et al*. (2017)). In this context, CNNs demonstrate significant advantages by modeling multiple levels of representation through successive layers of convolutional filters, enhancing the sensitivity and specificity of tasks such as segmentation, detection, and image classification (Hwang *et al*. (2019)).

Dentistry has stood out as a promising field for the use of Machine Learning (ML), especially CNNs, since diagnosis and treatment planning strongly depend on image analysis, from initial screening to treatment execution. Dental practice integrates the clinical examination with complementary evaluations such as radiographs and cone-beam computed tomography, associated with systemic data obtained in the medical history (Schwendicke *et al*. (2021)). Nevertheless, the clinical examination is considered the gold standard for diagnosing conditions that specifically involve dental structures, particularly dental caries, through visual-tactile inspection (Gomez (2015), Pretty and Ellwood (2013)).

However, for deep learning methods to be able to identify carious lesions, tooth detection and segmentation must occur first, allowing the delimitation of relevant structures and the exclusion of noise arising from soft tissues or clinical artifacts

(Tuzoff *et al*. (2019), Zhang *et al*. (2018), Lian *et al*. (2021)). Nevertheless, few studies have been published addressing dental segmentation using CNNs from intraoral photographs (Park *et al*. (2022), Liu *et al*. (2024), Nguyen *et al*. (2025)). To the best of our knowledge, only one study has evaluated the performance of CNNs using photographs of patients in this developmental phase. However, that investigation also included images from other stages of development, namely, primary and permanent dentition, and did not stratify the results according to each phase Ghorbani *et al*. (2025). This limitation reveals an important gap, since this phase presents varied patterns, which may hinder the training of the models (Hwang *et al*. (2019)).

Given this context, it becomes essential to develop investigations that compare different deep learning architectures and segmentation approaches specifically directed to the mixed dentition phase. This study aims to evaluate the use of neural network–based models for the detection and segmentation of this stage of dental development, examining their ability to accurately delineate regions of interest in the dental arch. This scope constitutes a fundamental preliminary step for future work aimed at identifying patterns in each individual tooth that require rigorous isolation of the areas corresponding to the teeth in relation to other oral structures.

## 2 Related Works

The application of CNNs for the identification of normative patterns, dental anomalies, and carious lesions from intraoral photographs has exhibited substantial and accelerating growth in the recent literature (Moharrami *et al*. (2024)). This burgeoning interest is attributable to several factors, including the increasing accessibility of large-scale clinical imaging datasets, marked advancements in deep learning architectures, and a growing recognition of the potential for automated systems to augment clinical decision-making in restorative and pediatric dentistry. Despite these technological and methodological advances, studies specifically addressing the mixed dentition phase, a critical and highly dynamic stage of oral development characterized by the concurrent presence of primary and permanent dentition, remain remarkably scarce. To date, only two investigations have approached this transitional stage of dental development using panoramic radiographs (Asci *et al*. (2024), Mine *et al*. (2022)). These studies, while valuable, are constrained by the inherent limitations of two-dimensional radiographic imaging, including tissue superimposition, projection distortions, and the absence of colorimetric and textural information essential for the detection of early carious lesions. In contrast, only a single study to date has employed intraoral photographs for the explicit purpose of tooth detection and segmentation in mixed dentition(Ghorbani *et al*. (2025)), underscoring a significant gap in the literature and reinforcing the novelty and timeliness of the present investigation.

A systematic review conducted by Alharbi and Alhasson (2024) highlighted the potential of CNNs in the detection and segmentation of dental structures, emphasizing widely adopted architectures such as Faster R-CNN, U-Net, and YOLO. Faster R-CNN, by integrating a region proposal network with a detector, has demonstrated considerable efficacy in the identification of dental elements. U-Net is widely recognized for its robustness and accuracy in segmentation tasks, whereas models from the YOLO family are distinguished by their capacity for real-time detection combined with high levels of reliability (Alharbi and Alhasson (2024)). However, the majority of these studies have concentrated on permanent dentition, thereby limiting the extrapolation of findings to pediatric clinical contexts.

Within the specific scenario of mixed dentition, the study by Ghorbani *et al*. (2025) reported high accuracy in tooth detection and segmentation using the YOLOv8 architecture. Nevertheless, the authors did not stratify the results considering exclusively images of patients in this developmental phase, which constrains the interpretation of model performance in light of the morphological complexity characteristic of the coexistence of primary and permanent teeth at varying stages of eruption.

From a methodological standpoint, it is observed that most studies employing CNNs for carious lesion detection reported only a single training round (Moutselos *et al*. (2019), Xiong *et al*. (2024), Zhang *et al*. (2022), Zhang *et al*. (2018), Mehdizadeh *et al*. (2024), Kang *et al*. (2024)). This approach may limit the assessment of model stability and robustness. To the best of current knowledge, only the studies by Kühnisch *et al*. (2022) and Park *et al*. (2022) incorporated multiple training iterations into their methodologies, four and five rounds, respectively, with the aim of evaluating result reproducibility.

Another relevant aspect pertains to the eligibility criteria applied to the images comprising the datasets. While some studies adopted highly controlled databases, others opted for strategies more closely aligned with clinical practice, including photographs depicting restorations, developmental enamel defects, the presence of saliva, and other clinical conditions, removing only duplicate or blurred images (Kang *et al*. (2024); Park *et al*. (2022)). This approach tends to favor greater model generalization and enhances applicability in real-world clinical scenarios, particularly within the context of mixed dentition.

## 3 Materials and Methods

### 3.1 Ethical Considerations

This project was reviewed and approved by the Human Research Ethics Committee of the Health Sciences Sector of the Federal University of Paraná (UFPR) on November 20, 2019 (Opinion No.: 3.715.610 / Certificate of Presentation for Ethical Consideration (CAEE) Number: 25001219.5.0000.0102).

### 3.2 Assembly of the Image Database

The study cohort comprised a total of 211 children, aged 8 years, of both sexes. Participants were recruited via convenience sampling from the public school system in the municipality of Curitiba, Paraná, Brazil. The specific age of 8 years was selected as it represents a key stage in mixed dentition, where the first permanent molars are typically in an active phase of eruption or have recently achieved full occlusion.

Inclusion was predicated on the presence of at least one first permanent molar, either partially or completely erupted, at the time of the clinical examination. This criterion ensured the relevance of all subjects to the core objective of analyzing

mixed dentition.

Several exclusion criteria were applied to maintain sample homogeneity and procedural feasibility. Children were excluded from participation if they presented with: (1) the absence of any erupted first permanent molar, (2) phenotypic signs indicative of known syndromes, or (3) any physical condition or behavioral factor that would preclude the safe and effective acquisition of standardized intraoral photographs.

## 3.3 Intraoral Dental Photographs

The photographs were obtained from each child following a standardized protocol, recording the posterior teeth of each quadrant at different angles in order to increase the variability of the images. For the photographic acquisitions, a professional camera (Canon EOS Rebel T6I®) with a ring flash (Canon Macro Ring Lite MR-14ex®) and a macro lens (Ultrasonic®) was used. The standardization of the photographs was obtained through the camera settings, with a shutter speed of 1/160, aperture at f/25, and ISO at 200.

The photographs were taken by indirect vision or with the aid of a crystal mirror for image reflection. The children were seated in school chairs, and retractors and mouth mirrors were used. A dental surgeon, previously trained to perform intraoral photographs, carried out the image acquisition. All necessary personal protective equipment was used.

Regardless of image quality, all photographs were included in the analysis process. For photographs with lower visual quality, the annotation was performed using a clear reference image of the same occlusal surface to ensure accurate identification of the areas of interest.

## 3.4 Convolutional Neural Networks

For the evaluation of the experiments, three distinct models were employed: YOLOv11, U-Net, and DeepLabv3, which will be described in the following sections. Figure 1 presents a visual comparison between the outputs generated by each model. The experiments were carried out on a server comprising an Intel® i9-9900K (3.60 GHz) processor (16 cores), 64GB of RAM, and two NVIDIA® RTX 2080 TI (12 GB) GPUs.

### 3.4.1 YOLOv11

YOLO (You Only Look Once), initially proposed by Redmon *et al*. (2016), consists of a network focused on object detection, standing out for its speed, accuracy, and generalization capability, showing significant performance in several areas, including medical applications (Murat and Kiran (2025)). The architecture of YOLO (Figure 2) is composed of successive convolutional layers, with max pooling stages, ReLU-based activations, and linear activation in the last layer, operating in a single pass over the image (Kang and Kim (2023)). In general, the functioning of the algorithm occurs as follows: the image is subdivided into grid cells, and each cell is responsible for predicting bounding boxes and object classes. Then, there is a box regression stage, in which only the regions that contain objects are highlighted. Based on the intersection between the boxes, the model refines the detections to keep only the truly relevant areas. Finally, since multiple boxes can detect the same object, the Non-Maximum Suppression (NMS) technique is applied, which eliminates redundant overlaps and results in more precise detection.

For this work, version 11 of YOLO, made available by Jocher and Qiu (2024), was employed, in which accuracy and inference speed were improved in comparison with previous generations. In addition, this version includes more compact architectures, such as YOLOv11-M, which has fewer parameters than YOLOv8-M while still maintaining superior performance on certain datasets and providing richer spatial information to the network. In this study, the YOLOv11-X model was used, a larger variant aimed at segmentation, with greater representation capacity and output layers specialized in object detection and delineation, exploiting multiple feature scales and producing more detailed masks of the regions of interest.

### 3.4.2 U-Net

The U-Net architecture (Figure 3), initially pioneered by Ronneberger *et al*. (2015) for biomedical image segmentation, has become a foundational and highly influential model for a wide range of semantic segmentation tasks. Its design directly addresses a core challenge in pixel-wise classification: combining high-level contextual information with precise spatial localization.

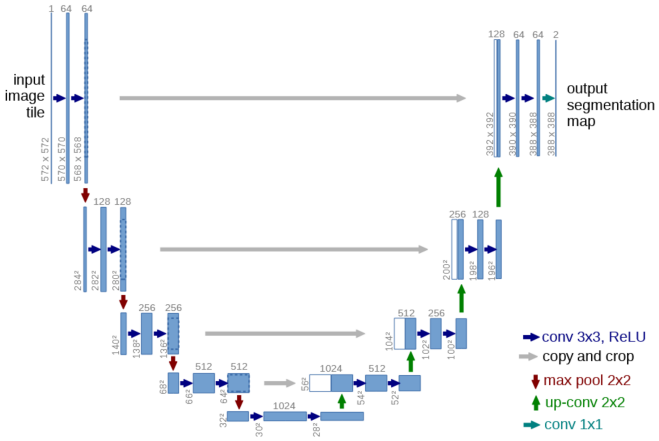


**Figure 3.** U-Net architecture (Extracted from Ronneberger *et al*. (2015)).

The network's symmetric, U-shaped architecture is composed of two distinct yet interconnected pathways. The contracting path (encoder) functions as a feature extractor. It consists of repeated blocks, typically two convolutional layers with ReLU activations followed by a 2x2 max-pooling operation for downsampling. With each pooling step, the spatial dimensions are halved while the number of feature channels is doubled, allowing the network to learn increasingly complex and abstract hierarchical representations of the input image.

Conversely, the expanding path (decoder) is responsible for precise localization and segmentation map generation Azad *et al*. (2024). This path performs upsampling—often via transposed convolution—at each stage, increasing the spatial resolution. Crucially, the defining innovation of U-Net is the use of skip connections. These connections concatenate the high-resolution feature maps from a layer in the contracting path with the corresponding upsampled feature map in the expanding path. This mechanism bridges the semantic gap between the encoder and decoder, allowing the latter to reintegrate fine-grained spatial details that were lost during downsampling, thereby enabling precise boundary delineation.

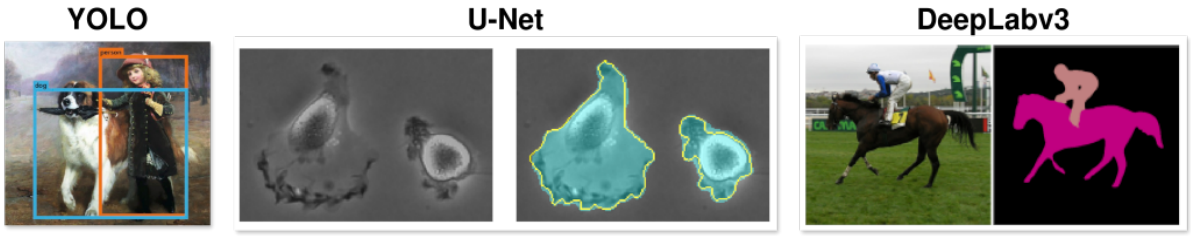


**Figure 1.** From left to right, examples of outputs obtained by the YOLO, U-Net, and DeepLabv3 networks, respectively. The illustrations were extracted from the original works (Redmon *et al*. (2016), Ronneberger *et al*. (2015), Chen *et al*. (2017)).

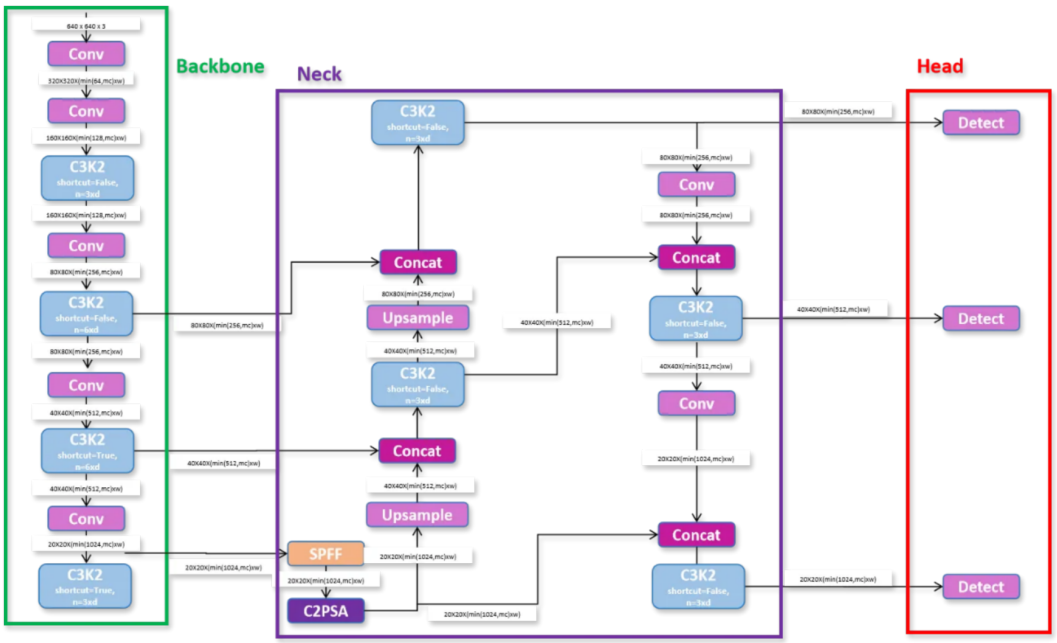


**Figure 2.** YOLOv11 architecture (Extracted from Rao (2024)).

Beyond its original biomedical application, U-Net's success is attributed to several key advantages. First, its ability to produce accurate segmentations with relatively few training images, thanks to effective feature reuse via skip connections. Second, its end-to-end training process and fully convolutional nature allow it to handle input images of variable sizes. These characteristics have made U-Net not just a specific model, but a template for countless variants (such as U-Net++, Attention U-Net, and ResUNet), which have been adapted and optimized for diverse domains, including the analysis of medical radiographs, satellite imagery, and, as in this work, intraoral photographs.

### *3.4.3 DeepLabv3*

DeepLabv3, introduced by Chen *et al*. (2017), represents a significant advancement in semantic segmentation by effectively addressing the fundamental trade-off between spatial resolution and the size of the receptive field. Its core innovation lies in the strategic use of dilated convolutions (atrous convolutions), a technique popularized by Yu and Koltun (2015). Unlike standard convolutions, dilated convolutions insert spaces (or ´´holes") between the kernel weights, allowing the filter to sample from a wider area of the input without increasing the number of parameters or losing resolution through pooling. This enables the network to capture rich multi-scale contextual information while preserving fine-grained spatial details that are often crucial for precise boundary segmentation. Figure 4 shows the architecture of DeepLabv3.

The architectural centerpiece of DeepLabv3 is the Atrolus Spatial Pyramid Pooling (ASPP) module. This module operates on the principle of parallel multi-scale feature extraction. It applies multiple dilated convolutions simultaneously to the same input feature map, each with a different dilation rate (e.g., rates of 1, 6, 12, 18). A 1x1 convolution and global average pooling branch are typically included alongside these parallel paths. While the 1x1 convolution captures local image context, the dilated convolutions with progressively larger rates capture context over increasingly broader regions, effectively simulating the analysis of the image at multiple scales or ´´fields of view". The feature maps from all these parallel branches are then concatenated and fused through a final 1x1 convolution.

This ASPP-based design provides several key advantages. It allows a single forward pass to encode objects and contextual information at vastly different scales within the same image, making it exceptionally robust to size variations of target structures. Furthermore, by avoiding the aggressive spatial downsampling common in encoder-decoder architectures, DeepLabv3 maintains higher-resolution feature maps throughout much of the network, leading to more accurate segmentation masks, particularly along object boundaries. These characteristics have established DeepLabv3 and its variants as a leading architectural paradigm for tasks requiring high precision in complex scenes, such as autonomous

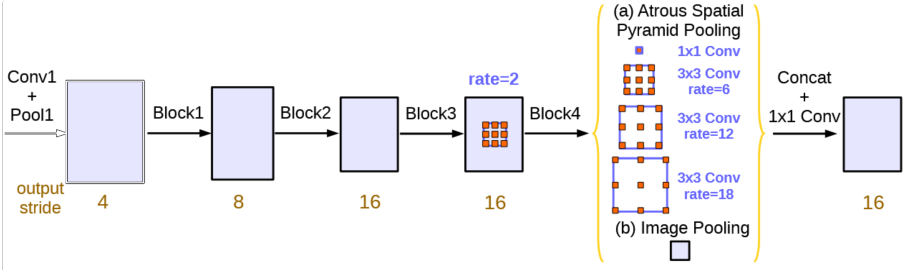


**Figure 4.** DeepLabv3 architecture (Extracted from Chen *et al*. (2017)).

driving, medical imaging, and, as relevant to this study, the detailed segmentation of anatomical structures in intraoral photographs.

### 3.5 Image Annotation and CNN Training

A total of 945 intraoral photographs were analyzed, with the posterior teeth delineated by a researcher specialized in Pediatric Dentistry, using the CVAT software (Figure 5).

**Figure 5.** Tooth annotation using the CVAT software. The annotation was performed for each tooth by marking its entire visible area and a bounding box annotation.

The 945 intraoral images were divided into training (n = 760) and validation (n = 185). Due to the acquisition of multiple images per patient, a patient-based split was adopted, ensuring no overlap between the training set and the other datasets.

Semantic segmentation tests of the teeth were performed on the three aforementioned networks (YOLOv11, U-Net, and DeepLabv3). Network training used transfer learning with the weights provided by each architecture. Each network was trained for at least 100 epochs, with Early Stopping options to prevent overfitting during training. The hyperparameters used were the default ones available in the architectures. Figure 6 presents an overview of our experiment. Figure 6 (A) shows the input image tested on the three selected networks (Figure6 (B)); their resulting annotations (Figure 6 (C) are compared with the Ground Truth (Figure 6 (D)) annotated in the previously described step. Precision, sensitivity, and F1-score metrics were calculated for all teeth (mixed dentition) and for erupting teeth (partially erupted teeth).

## 4 Performance and Accuracy Metrics

A single metric cannot comprehensively evaluate a machine learning model's performance. Consequently, multiple metrics are reported together to provide a more complete and nuanced assessment. In this work, the task is formulated as binary classification, distinguishing "tooth" (positive) from "non-tooth" (negative) instances. The core performance indicators used in this context are defined below.

Precision (Equation 1) measures the reliability of positive predictions. It calculates the fraction of predicted positives that are actually correct, penalizing false positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

A high-precision model is trustworthy when it labels an instance as positive.

Sensitivity (Recall, Equation 2) assesses the model's ability to identify all positive cases. It measures the proportion of actual positives that were correctly identified, penalizing false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

High sensitivity is crucial in applications like disease screening, where missing a positive case is costly.

F-score (F1-score, Equation 3 ) provides a single, balanced measure that harmonizes precision and recall. It is the harmonic mean of the two, offering a more informative metric than accuracy in scenarios with class imbalance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

This metric is most useful when a balance between avoiding false positives and false negatives is required.

## 5 Results

The comparative performance of the three CNNs architectures, YOLOv11, U-Net, and DeepLabv3 is summarized in Table 1, with precision serving as the metric. All three models demonstrated high reliability, achieving precision values exceeding 0.95. This indicates a consistently low rate of false positive predictions across the board. This strong performance was corroborated by similarly high results in the sensitivity (recall) and F1-score metrics, confirming that the networks were not only precise but also comprehensive in their detection, effectively balancing the identification of true positives against false negatives.

A more granular analysis of overall tooth detection within the mixed dentition dataset, which contained a total of 3819 teeth, further illustrates model efficacy. The U-Net architecture achieved the highest detection rate, correctly identifying 751 teeth (0.951), followed closely by DeepLabv3 with 746 teeth (0.944) and YOLOv11 with 741 teeth (0.938). The minor variance in these overall scores suggests that all architectures are well-suited for general tooth segmentation in this context.

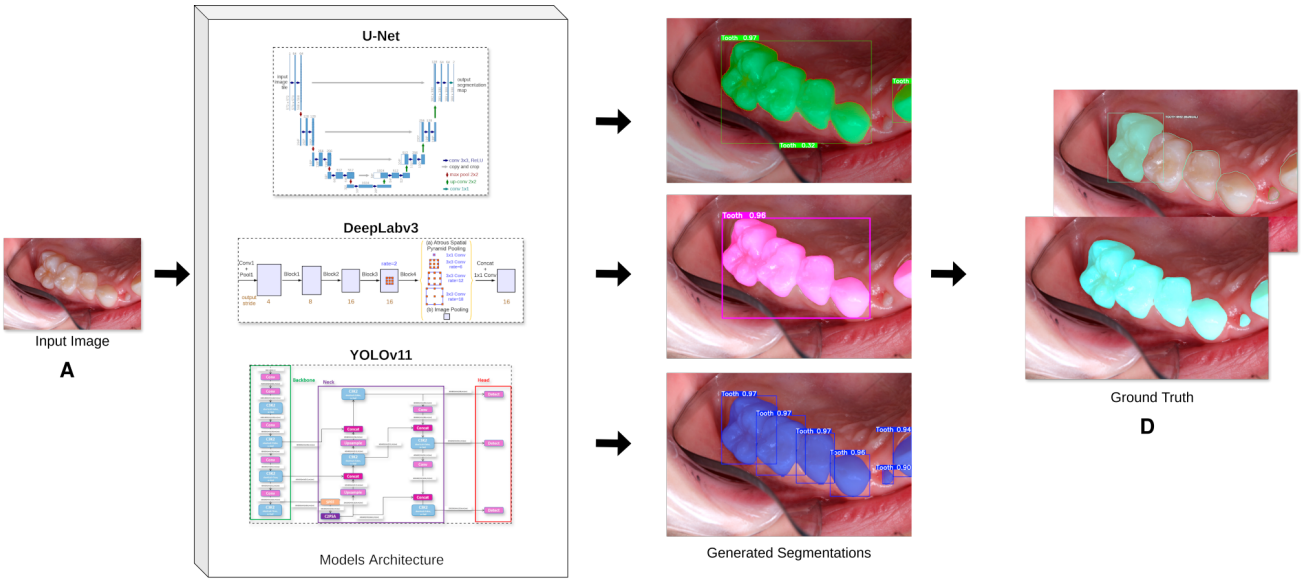


**Figure 6.** Proposed pipeline for comparing segmentation models. The architectural figures were taken from their respective sources: U-Net from Ronneberger *et al*. (2015), DeepLabv3 from Chen *et al*. (2017), and YOLOv11 from Rao (2024).

The most revealing assessment, however, came from the challenging subset of partially erupted and erupting permanent teeth, comprising 51 instances in the dataset. Here, the architectures exhibited more distinct performance characteristics. YOLOv11 demonstrated superior robustness, successfully detecting 49 teeth (96.08%), significantly outperforming U-Net (40 teeth, 78.43%) and DeepLabv3 (39 teeth, 76.47%). A closer examination of partially erupted teeth—a particularly ambiguous category—revealed that U-Net and DeepLabv3 recognized only 8 (15.68%) and 7 (13.72%) of these, respectively, suggesting a potential limitation in handling highly incomplete tooth structures.

Conversely, the failure cases for fully erupting teeth were relatively low but varied. YOLOv11 failed to recognize only 2 teeth (3.92%), whereas U-Net missed 3 (5.88%), and DeepLabv3 missed 5 (9.80%). These results, detailed in Table 2, highlight YOLOv11's particular strength in managing the visual ambiguity and varied presentation of teeth in transitional eruption phases. The performance gap in this specific subset underscores the importance of evaluating models on clinically challenging edge cases, not just on aggregate metrics, as these scenarios are critical for real-world diagnostic reliability in pediatric dentistry.

**Table 1.** Detection of erupting teeth by the YOLOv11, U-Net, and DeepLabv3 networks

| Networks | YOLOv11 | U-Net | DeepLabv3 |
|---|---|---|---|
| Detected | 49 (96.08%) | 40 (78.43%) | 39 (76.47%) |
| Partially detected | 0 (0%) | 8 (15.68%) | 7 (13.72%) |
| Not detected | 2 (3.92%) | 3 (5.88%) | 5 (9.80%) |
| Total | 51 (100%) | 51 (100%) | 51 (100%) |

Figure 7 presents a qualitative comparison between the reference annotation (ground truth) and the outputs of the three CNNs. In Figure 7(A), we see the tooth regions manually annotated; Figure 7(B–D) shows the detected/segmented areas from YOLOv11, U-Net, and DeepLabv3, respectively. It is observed that YOLOv11 identifies and segments the erupted tooth indicated by the black arrow in Figure 7(A), whereas U-Net and DeepLabv3 do not detect it. A difference in the type of segmentation is also noted: YOLOv11 performs instance segmentation (tooth by tooth), whereas U-Net and DeepLabv3 produce semantic segmentation, aggregating multiple teeth into a single "tooth" region.

## 6 Discussion

The use of CNNs in detecting patterns of normality and dental alterations using intraoral photographs is a growing topic (Moharrami *et al*. (2024)). However, research focused specifically on mixed dentition remains limited, with most previous investigations relying on radiographic imaging rather than photographic records (Asci *et al*. (2024); Mine *et al*. (2022). Recent studies by Beser *et al*. (2024) and Bumann *et al*. (2024) evaluated tooth detection and segmentation in mixed dentition using panoramic radiographs and demonstrated the promising performance of deep learning models. Only two studies using intraoral photographs aimed at tooth detection and segmentation (Ghorbani *et al*. (2025) and Liu *et al*. (2025)). Nevertheless, although panoramic radiographs provide comprehensive visualization of dental structures, they involve ionizing radiation and present limited feasibility for large-scale screening in pediatric populations. In contrast, intraoral photography represents a non-invasive, more accessible, and clinically applicable alternative, particularly in community and school-based contexts.

The mixed dentition period constitutes a stage of greater susceptibility to the development of carious lesions (Raja *et al*. (2025)). Currently, studies specifically evaluating CNN performance using intraoral photographs in children during the mixed dentition phase are still limited, although recent evidence has begun to emerge. In addition to the study by Ghorbani *et al*. (2025), Liu *et al*. (2025) also investigated automated tooth recognition and segmentation using intraoral photographs of children aged 7–9 years, a population predominantly in mixed dentition. However, the study by Ghorbani

**Table 2.** Performance of the CNNs on the evaluated metrics

| Model | Precision | Sensitivity | F1-score |
|---|---|---|---|
| YOLOv11 | 0.967 | 0.938 | 0.952 |
| U-Net | 0.953 | 0.951 | 0.951 |
| DeepLabv3 | 0.963 | 0.944 | 0.953 |

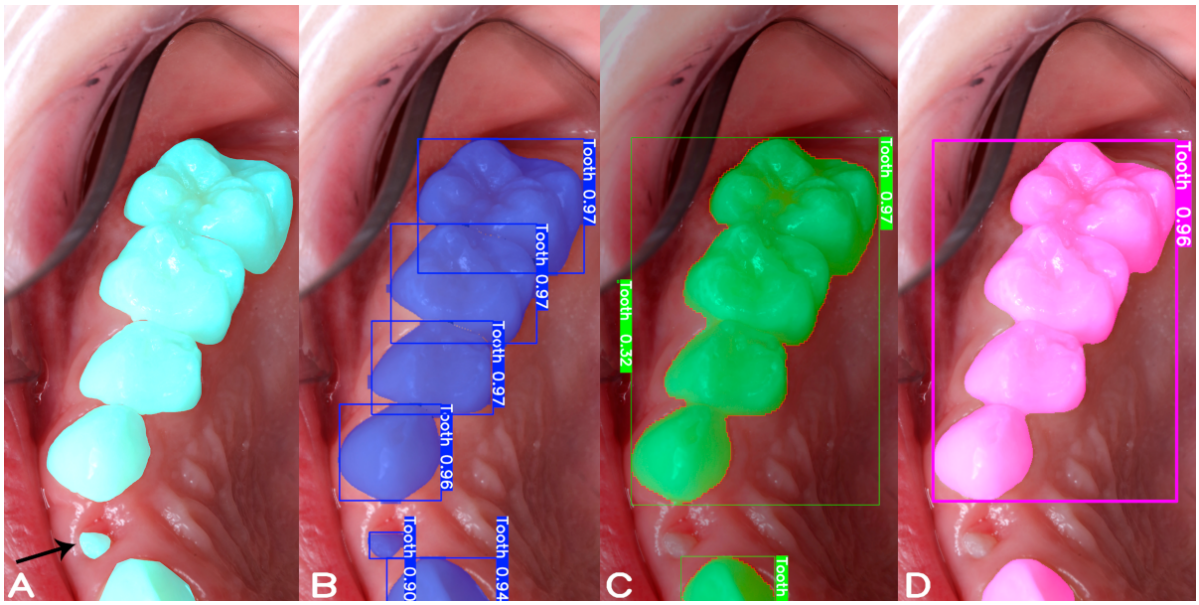


**Figure 7.** Visual comparison between the reference annotation (A) and the results of the networks: YOLOv11 (B), U-Net (C), and DeepLabv3 (D). Panel (A) shows the expected delineation of the teeth; panels (B–D) present the regions detected and segmented by each method, highlighting correct detections and local discrepancies at tooth boundaries and in the extent of the masks.

*et al*. (2025) included images from different stages of dental development—primary, mixed, and permanent—without stratifying the results according to each phase, while Liu *et al*. (2025) employed a multitask framework focused on tooth position recognition and segmentation rather than a direct comparative evaluation of multiple segmentation architectures.

This methodological heterogeneity highlights a relevant gap in the literature, considering that children in mixed dentition present distinct anatomical characteristics, with the coexistence of primary and permanent teeth at different stages of eruption, which may directly influence the performance and generalizability of artificial intelligence models.

The use of dental segmentation as an initial step in the clinical detection process of carious lesions is fundamental to promote methodological standardization and ensure the reproducibility of results obtained by artificial intelligence solutions applied as auxiliary tools for dental diagnosis (Park *et al*. (2022)). This prior step has been described in some recent studies (Park *et al*. (2022); Asci *et al*. (2024); Yoon *et al*. (2024)).

A systematic review conducted by Alharbi and Alhasson (2024) highlighted advances in CNNs for tooth detection and segmentation, emphasizing architectures such as Faster R-CNN, U-Net, and YOLO. Faster R-CNN, by integrating a region proposal network (RPN) with a detection network, has demonstrated high efficiency in identifying dental structures. U-Net is widely recognized for its accuracy and speed in segmentation, while YOLO models stand out for real-time detection with high reliability. In the present study, the YOLOv11 network showed slightly higher precision values and slightly lower sensitivity compared to U-Net.

Although the study by Ghorbani *et al*. (2025) achieved 95.72% precision with the YOLOv8 CNN, there was no discrimination of results considering only mixed dentition photographs. In the present study, YOLOv11 showed better performance in detecting partially erupted teeth, and it was also the only evaluated network capable of performing tooth-by-tooth segmentation. These findings reinforce the potential of CNNs, especially YOLO, in detecting and segmenting dental structures in dental practice, and it is important to conduct more studies on this topic.

In line with recent developments, Liu *et al*. (2025) demonstrated that deep learning models can accurately segment and recognize teeth in intraoral photographs of children in mixed dentition, achieving an IoU of 0.95 and an average F1-score of 0.906. Comparatively, the present study showed consistently high reliability across three different architectures (YOLOv11, U-Net, and DeepLabv3), with precision values exceeding 0.953 and F1-scores above 0.951. Unlike Liu *et al*. (2025), who adopted a multitask framework involving tooth-type recognition and image cropping into single-tooth datasets, the present investigation preserved full intraoral posterior images and focused on direct semantic segmentation in a clinically realistic context. This methodological approach may better reflect real-world diagnostic conditions, particularly in school-based and epidemiological scenarios.

A notable and clinically relevant difference of the present study lies in the specific evaluation of erupting and partially erupted permanent teeth, a subgroup that is rarely analyzed separately in previous artificial intelligence research. This subset, comprising 51 instances in the image database, repre-

sents a particularly challenging diagnostic scenario due to incomplete anatomical contours, partial occlusal exposure, and increased visual ambiguity. In this context, the evaluated architectures demonstrated distinct performance patterns, with YOLOv11 showing superior robustness by successfully detecting 96.08% of erupting teeth, whereas U-Net and DeepLabv3 achieved lower detection rates (78.43% and 76.47%, respectively). When specifically analyzing partially erupted teeth, considered the most ambiguous category, the limitations of segmentation-based architectures became more evident, with recognition rates of only 15.68% for U-Net and 13.72% for DeepLabv3. These findings suggest that object-detection-oriented models may be more adaptable to highly incomplete dental structures, which are intrinsically characteristic of the mixed dentition phase. From a clinical perspective, this is particularly relevant, as erupting permanent teeth, especially first molars, are more susceptible to biofilm accumulation and early caries development, reinforcing the importance of reliable automated detection in this developmental stage.

A recent study (Nguyen *et al*. (2025)) used SegmentAnyTooth for automated tooth segmentation and numbering in intraoral photographs at 5 angulations. However, a relevant limitation pointed out by the authors refers to the model's restricted applicability to permanent dentition, which reduces its usefulness in younger age groups. In this sense, the present study expands the use of CNNs by applying and comparing the performance of the YOLOv11, U-Net, and DeepLabv3 networks in a broader context, focusing on mixed dentition, which comprises both primary and permanent teeth at different stages of eruption, demonstrating the feasibility of segmentation and detection models with greater clinical applicability in Dentistry, especially in the field of Pediatric Dentistry.

# 7 Conclusion

This study successfully demonstrated the significant potential of Convolutional Neural Networks (CNNs) for the automated segmentation of posterior teeth in pediatric patients with mixed dentition. To this end, we conducted a comparative evaluation of three state-of-the-art network architectures, YOLOv11, U-Net, and DeepLabv3, on a dataset comprising 945 intraoral images from children aged 8 years. The experimental results indicated strong overall performance across all models. Specifically, for the core segmentation metrics of precision, recall, and the F1-score, all three networks consistently achieved values exceeding 0.93. Notably, the YOLOv11 model distinguished itself by achieving higher precision in identifying partially erupted teeth, a challenging subset of the mixed dentition phase.

Despite these promising results, we acknowledge a primary limitation of this work: the dataset's relatively small sample size. Future studies involving a larger and more diverse cohort of participants are likely to yield more accurate, generalizable, and statistically robust results. Furthermore, we emphasize that, to the best of our knowledge, this is the first work in the literature to perform semantic segmentation specifically on intraoral images focusing on mixed dentition. This focus on a clinically complex transitional stage underscores the novelty and contribution of our research.

Building upon this foundational segmentation work, our immediate next steps are twofold. First, we will isolate the individually segmented teeth. Subsequently, we will employ dedicated deep-learning classifiers to perform automated detection of carious lesions on each tooth, culminating in the assignment of a standardized severity score according to the International Caries Detection and Assessment System (ICDAS) index. This pipeline aims to progress toward a comprehensive, AI-assisted diagnostic tool for pediatric dentistry.

# Declarations

## Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 4000101606P8.

## Authors' Contributions

Bruna Cristine Dias collected the data, made the baseline markings in the CVAT software, performed the interpretation of the data, and wrote the manuscript. Mateus Felipe de Cássio Ferreira created the machine learning models, performed the statistical analysis, interpretration of data and wrote the manuscript. Luan Matheus Trindade Dalmazo performed the interpretation of data and performed the critical review of the maunscript. Luciana Reichert da Silva Assunção was one of the responsible for the study desing, performed the interpretation of data, and performed the critical review of the manuscript. Lucas Ferrari de Oliveira was the research adviser, was one of the responsible for the study design, interpreted the data and performed the critical review of the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors have no relevant financial or non-financial interests to disclose.

## Availability of data and materials

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Further relevant information

This study was conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments. The research protocol was approved by the Ethics Committee on Human Research of the Federal University of Paraná (CAAE: 25001219.5.0000.0102). Parental/guardian informed consent was obtained for all participants included in the study.

# References

Alharbi, S. S. and Alhasson, H. F. (2024). Exploring the applications of artificial intelligence in dental image detection: A systematic review. *Diagnostics*, 14(21):2442. DOI: 10.3390/diagnostics14212442.

Asci, E., Kilic, M., Celik, O., Cantekin, K., Bircan, H. B., Bayrakdar, İ. S., and Orhan, K. (2024). A deep learning approach to automatic tooth caries segmentation in panoramic radiographs of children in primary dentition, mixed dentition, and permanent dentition. *Children*, 11(6):690. DOI: 10.3390/children11060690.

Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. (2024). Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern*

*Analysis and Machine Intelligence*, 46(12):10076–10095. DOI: 10.1109/TPAMI.2024.3435571.

Beser, B., Reis, T., Berber, M. N., Topaloglu, E., Gungor, E., Kılıc, M. C., Duman, S., Çelik, , Kuran, A., and Bayrakdar, I. S. (2024). YOLO-V5 based deep learning approach for tooth detection and segmentation on pediatric panoramic radiographs in mixed dentition. *BMC Medical Imaging*, 24(1):172. PMID: 38992601; PMCID: PMC11238494. DOI: 10.1186/s12880-024-01338-w.

Bumann, E. E., Al-Qarni, S., Chandrashekar, G., Sabzian, R., Bohaty, B., and Lee, Y. (2024). A novel collaborative learning model for mixed dentition and fillings segmentation in panoramic radiographs. *Journal of Dentistry*, 140:104779. PMID: 38007173. DOI: 10.1016/j.jdent.2023.104779.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation.

Conrads, G. and About, I. (2018). Pathophysiology of dental caries. *Monographs in Oral Science*, 27:1–10. DOI: 10.1159/000487826.

Ghorbani, Z., Mirebeigi-Jamasbi, S. S., Hassannia Dargah, M., Nahvi, M., Hosseinikhah Manshadi, S. A., and Akbarzadeh Fathabadi, Z. (2025). A novel deep learning-based model for automated tooth detection and numbering in mixed and permanent dentition in occlusal photographs. *BMC Oral Health*, 25(1):455. DOI: 10.1186/s12903-025-05803-y.

Gomez, J. (2015). Detection and diagnosis of the early caries lesion. *BMC Oral Health*, 15(1):S3. DOI: 10.1186/1472-6831-15-S1-S3.

Hwang, J.-J., Jung, Y.-H., Cho, B.-H., and Heo, M.-S. (2019). An overview of deep learning in the field of dentistry. *Imaging Science in Dentistry*, 49(1):1–7. DOI: 10.5624/isd.2019.49.1.1.

Ismail, A. I. (2004). Visual and visuo-tactile detection of dental caries. *Journal of Dental Research*, 83(Spec No C):C56–C66. DOI: 10.1177/154405910408301s12.

Jocher, G. and Qiu, J. (2024). Ultralytics yolo11.

Kang, C. H. and Kim, S. Y. (2023). Real-time object detection and segmentation technology: an analysis of the yolo algorithm. *JMST Advances*, 5(2):69–76. DOI: 10.1007/s42791-023-00049-7.

Kang, S., Shon, B., Park, E. Y., Jeong, S., and Kim, E. (2024). Diagnostic accuracy of dental caries detection using ensemble techniques in deep learning with intraoral camera images. *Plos one*, 19(9):1–13. DOI: 10.1371/journal.pone.0310004.

Kühnisch, J., Meyer, O., Hesenius, M., Hickel, R., and Gruhn, V. (2022). Caries detection on intraoral images using artificial intelligence. *Journal of Dental Research*, 101(2):158–165. PMID: 34416824; PMCID: PMC8808002. DOI: 10.1177/00220345211032524.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. DOI: 10.1038/nature14539.

Lian, L., Zhu, T., Zhu, F., and Zhu, H. (2021). Deep learning for caries detection and classification. *Diagnostics*, 11(9):1672. DOI: 10.3390/diagnostics11091672.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88. DOI: 10.1016/j.media.2017.07.005.

Liu, Y., Cheng, Y., Song, Y., Cai, D., and Zhang, N. (2024). Oral screening of dental calculus, gingivitis and dental caries through segmentation on intraoral photographic images using deep learning. *BMC Oral Health*, 24(1):1287. DOI: 10.1186/s12903-024-05072-1.

Liu, Z., Zhao, S., Wang, Y., Cao, Y., Lin, H., and Pang, L. (2025). Automatic recognition of oral health status in mixed dentition via intraoral photography. *BMC Oral Health*, 25(1):1764. PMID: 41214645; PMCID: PMC12599049. DOI: 10.1186/s12903-025-06866-7.

Lynch, R. J. M. (2013). The primary and mixed dentition, post-eruptive enamel maturation and dental caries: a review. *International Dental Journal*, 63 Suppl 2(Suppl 2):3–13. DOI: 10.1111/idj.12076.

Marsh, P. D. and Zaura, E. (2017). Dental biofilm: ecological interactions in health and disease. *Journal of Clinical Periodontology*, 44(Suppl 18):S12–S22. DOI: 10.1111/jcpe.12679.

Mehdizadeh, M., Estai, M., Vignarajan, J., Patel, J., Granich, J., Zaniovich, M., Kruger, E., Winters, J., Tennant, M., and Saha, S. (2024). A deep learning-based system for the assessment of dental caries using colour dental photographs. *Studies in Health Technology and Informatics*, 310:911–915. PMID: 38269941. DOI: 10.3233/SHTI231097.

Mine, Y., Iwamoto, Y., Okazaki, S., Nakamura, K., Takeda, S., Peng, T.-Y., Mitsuhata, C., Kakimoto, N., Kozai, K., and Murayama, T. (2022). Detecting the presence of supernumerary teeth during the early mixed dentition stage using deep learning algorithms: A pilot study. *International Journal of Paediatric Dentistry*, 32(5):678–685. DOI: 10.1111/ipd.12946.

Moharrami, M., Farmer, J., Singhal, S., Watson, E., Glogauer, M., Johnson, A. E. W., Schwendicke, F., and Quinonez, C. (2024). Detecting dental caries on oral photographs using artificial intelligence: A systematic review. *Oral Diseases*, 30(4):1765–1783. DOI: 10.1111/odi.14659.

Moutselos, K., Berdouses, E., Oulis, C., and Maglogiannis, I. (2019). Recognizing occlusal caries in dental intraoral images using deep learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1617–1620, Berlin, Germany. IEEE. DOI: 10.1109/EMBC.2019.8856553.

Murat, A. A. and Kiran, M. S. (2025). A comprehensive review on yolo versions for object detection. *Engineering Science and Technology, an International Journal*, 70:102161. DOI: https://doi.org/10.1016/j.jestch.2025.102161.

Nguyen, K. D., Hoang, H. T., Doan, T.-P. H., Dao, K. Q., Wang, D.-H., and Hsu, M.-L. (2025). Segmentanytooth: An open-source deep learning framework for tooth enumeration and segmentation in intraoral photos. *Journal of Dental Sciences*, 20(2):1110–1117. DOI: 10.1016/j.jds.2025.01.003.

Nyvad, B., Fejerskov, O., and Baelum, V. (2011). Diagnóstico tátil-visual da cárie. *Kidd E, Fejerskov O. Cárie Dentária: A Doença e seu Tratamento Clínico. 2ª Ed. Tradução: Rossetti PHO. São Paulo: Santos*, pages 50–68.

Park, E. Y., Cho, H., Kang, S., Jeong, S., and Kim, E.-K. (2022). Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health*, 22(1):573. DOI: 10.1186/s12903-022-02589-1.

Pretty, I. and Ellwood, R. (2013). The caries continuum: Opportunities to detect, treat and monitor the re-mineralization of early caries lesions. *Journal of Dentistry*, 41:S12–S21. Establishing a new standard in cavity protection: Introducing Pro-Argin plus fluoride — a breakthrough technology. DOI: https://doi.org/10.1016/j.jdent.2010.04.003.

Raja, M., Nazzal, H., Cyprian, F. S., Matoug-Elwerfelli, M., and Duggal, M. (2025). Association of salivary proteins with dental caries in children with mixed dentition: a systematic review. *European Archives of Paediatric Dentistry*, 26(4):617–631. DOI: 10.1007/s40368-024-00994-4.

Rao, S. N. (2024). Yolov11 architecture explained: Next-level object detection with enhanced speed and accuracy. `https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speed-and-accuracy-2dbe2d376f71`. Accessed: 2025-12-17.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Schwendicke, F., Rossi, J. G., Göstemeyer, G., Elhennawy, K., Cantu, A. G., Gaudin, R., Chaurasia, A., Gehrung, S., and Krois, J. (2021). Cost-effectiveness of artificial intelligence for proximal caries detection. *Journal of Dental Research*, 100(4):369–376. DOI: 10.1177/0022034520972335.

Selwitz, R. H., Ismail, A. I., and Pitts, N. B. (2007). Dental caries. *Lancet*, 369(9555):51–59. DOI: 10.1016/S0140-6736(07)60031-2.

Shi, W., Qin, M., Chen, F., and Xia, B. (2016). Supragingival microbial profiles of permanent and deciduous teeth in children with mixed dentition. *PLOS ONE*, 11(1):e0146938. DOI: 10.1371/journal.pone.0146938.

Tuzoff, D. V., Tuzova, L. N., Bornstein, M. M., Krasnov, A. S., Kharchenko, M. A., Nikolenko, S. I., Sveshnikov, M. M., and Bednenko, G. B. (2019). Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, 48(4):20180051. DOI: 10.1259/dmfr.20180051.

Xiong, D., Marcus, M., Maida, C. A., Lyu, Y., Hays, R. D., Wang, Y., Shen, J., Spolsky, V. W., Lee, S. Y., Crall, J. J., and Liu, H. (2024). Development of short forms for screening children's dental caries and urgent treatment needs using item response theory and machine learning methods. *PLoS One*, 19(3):e0299947. PMID: 38517846; PMCID: PMC10959356. DOI: 10.1371/journal.pone.0299947.

Yoon, K., Jeong, H.-M., Kim, J.-W., Park, J.-H., and Choi, J. (2024). Ai-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation. *Journal of Dentistry*, 141:104821. DOI: 10.1016/j.jdent.2023.104821.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhang, K., Wu, J., Chen, H., and Lyu, P. (2018). An effective teeth recognition method using label tree with cascade network structure. *Computerized Medical Imaging and Graphics*, 68:61–70. DOI: 10.1016/j.compmedimag.2018.07.001.

Zhang, X., Liang, Y., Li, W., Liu, C., Gu, D., Sun, W., and Miao, L. (2022). Development and evaluation of deep learning for screening dental caries from oral photographs. *Oral Diseases*, 28(1):173–181. DOI: 10.1111/odi.13735.