IJWR

Optimizing Dental Anomaly Detection: A Region-Specific AI Framework with Hierarchical Attention Mechanisms

Mahdieh Dehghani[a], Reza Aghaeizadeh Zoroofi [b]*

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran; dehghani.mahdieh@ut.ac.ir

[b] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran; zoroofi@ut.ac.ir

ABSTRACT

Interpretation of dental panoramic radiographs which encompass all teeth as well as portions of the jaw and facial bones is critically important for preventive care and for devising appropriate treatment plans based on clinical findings. However, a high clinical workload or the absence of a specialist may compromise the accurate interpretation of even fundamental conditions, such as the detection of abnormalities. In such cases, artificial intelligence techniques can serve as valuable tools to enhance diagnostic accuracy. This research introduces a modified detection framework based on YOLOv11, incorporating two main architectural enhancements: the addition of a module designed to increase attention to specific regions, and improvements to the multi-scale blocks in the backbone of the network. The post-processing stage also employed methods capable of effectively distinguishing overlapping teeth. Experimental results demonstrate an improvement of over 7 percent in the F1-score compared to the baseline YOLOv11 architecture. The proposed model demonstrates competitive performance compared to models with similar architectures and exhibits satisfactory generalization on an independent dataset that was not utilized during training. Furthermore, relying on the real-time processing capability inherent to the YOLO framework, the proposed method can serve as an effective deep learning engine for integration into web software platforms and tools, enabling rapid and accurate dental radiograph analysis in clinical and telemedicine environments.

Keywords*— *Dental Abnormality Detection, YOLO, Panoramic Dental X-ray Images, Weighted Box Fusion.

1. Introduction

Medical image processing presents significant challenges due to ethical constraints and the sheer volume of imaging data. Artificial Intelligence (AI) has emerged as a transformative tool in this domain, enhancing diagnostic speed and accuracy, particularly in scenarios where direct clinician oversight is limited. Dental imaging encompasses two primary modalities: Intraoral and Extraoral X-ray images, with panoramic radiographs serving as a cornerstone for detecting caries, structural anomalies, and alveolar bone loss. These images provide a comprehensive view of the dentition and adjacent maxillofacial structures, making them indispensable for preventive care and treatment planning [1]. With the rapid advancement of technology, Human-Computer Interaction (HCI) has assumed a pivotal role, enabling the use of non-medical imaging for preliminary examinations. Nevertheless, despite the enhanced diagnostic capabilities provided by RGB imaging, it remains limited in accurately evaluating root morphology and detecting metallic artifacts, such as dental implants [2].

AI-driven techniques are not intended to replace clinical expertise but rather function as decision-support systems that augment diagnostic precision and workflow efficiency. Their integration has catalyzed advancements across multiple dental specialties, including endodontics, oral radiology, orthodontics, and prosthodontics. Panoramic radiographs are routinely employed for diverse applications, ranging from osseous evaluation to implant planning and disease screening [3]. Although conventional image processing methods

http://dx.doi.org/10.22133/ijwr.2025.526178.1290

Citation M. Dehghani, R. Aghaeizadeh Zoroofi, "Optimizing Dental Anomaly Detection: A Region-Specific AI Framework with Hierarchical Attention Mechanisms", *International Journal of Web Research*, vol.8, no.4,pp.1-13, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.526178.1290.

*Corressponding Author

Article History: Received: 25 May 2025; Revised: 31 August 2025; Accepted: 19 September 2025.

have historically addressed quality enhancement, persistent challenges—such as motion artifacts and anatomical variability—have necessitated the adoption of deep learning approaches to improve interpretability [4].

Automated tooth classification and numbering represent foundational tasks in AI-assisted dental image analysis. While contemporary algorithms achieve clinically acceptable accuracy in these operations, limitations persist due to dataset scarcity, edentulous regions, and intraoral artifacts [5]. Semantic segmentation via U-Net architectures has further enabled precise segmentation of caries, root structures, and other anatomical components [6]. However, the heterogeneous nature of dental pathologies precludes the development of a universal diagnostic tool capable of addressing all conditions autonomously. Deep learning architectures are often characterized by high computational complexity in medical image analysis. Given the critical importance of fine-grained details in dental panoramic radiographs, there is a pressing need for computationally efficient, lightweight models optimized for deployment in web-based platforms and software tools, which support real-time inference and seamless integration into clinical workflows.

This study introduces a deep learning framework optimized for detecting cavity, filling, implants, and impacted teeth with high spatial fidelity. Our objectives prioritize both accuracy enhancement and cross-dataset generalizability, achieved through an architecture engineered for small-object detection and region-specific precision. Section 2 reviews the related work, while Section 3 details the model architecture and the conditions of the employed dataset, along with the evaluation metrics used for model assessment. Section 4 presents the experimental results, and Section 5 offers a comprehensive comparison of the proposed approach with existing models from multiple perspectives. Finally, Section 6 concludes the paper by discussing key challenges and outlining potential directions for future research.

2. Related Work

The rapid advancement of artificial intelligence and its integration into medical applications have opened new avenues for the development of deep learning algorithms aimed at enhancing diagnostic accuracy and streamlining treatment planning. The increasing availability of publicly accessible datasets has significantly accelerated research efforts in applying deep learning techniques to medical image interpretation. Recently, artificial intelligence has been recognized as a transformative tool across various domains within dentistry, revolutionizing diagnostic and analytical processes [3]. Convolutional Neural Networks (CNNs) constitute a foundational approach for feature extraction and have been extensively employed in diagnosing pathologies from clinical dental radiographs. These networks facilitate the classification of disease type and severity; however, their accuracy in detecting early-stage or mild conditions remains limited [7].

Another prominent method is region-based convolutional neural networks, such as RCNNs, which operate through a two-stage process involving region proposal and classification. Mask R-CNN, in particular, has shown promising results in dental applications, achieving an average precision of 79.5% in identifying abnormal teeth, thereby enabling early detection of caries in dental X-ray images [8]. Beyond tooth detection, Mask R-CNN has also been utilized for identifying oral conditions such as cold sores [9]. Tooth segmentation has been further advanced through collaborative learning approaches integrated with Mask R-CNN, where main and edge images are processed simultaneously. This method extracts features from complementary images, with an attention mechanism determining the relative importance of each at different spatial locations, ultimately facilitating precise segmentation of teeth [10]. Mask R-CNN remains the predominant architecture for tooth and disease segmentation tasks [11].

Faster R-CNN is another widely adopted model for interpreting panoramic dental images. When combined with backbone networks such as GoogLeNet and AlexNet, Faster R-CNN has achieved an accuracy of 94.18% in detecting dental anomalies [12]. Tooth-type classification (incisors, molars, premolars, canines) serves as a critical preprocessing step in dental image analysis. Region-based CNN (R-CNN) variants leverage anchor-based detection and intersection-over-union (IoU) optimization to accelerate recognition, attaining more than 90% detection accuracy and 99% classification precision via bounding box regression [13]. Additionally, Faster RCNN has been employed to predict the number and location of carious lesions from periapical radiographs, attaining a precision of 73.49%. The integration of common feature extraction backbones like ResNet-50, Xception, and VGG16 has further enhanced network performance compared to single-stage detectors such as YOLO [14]. The automated classification of teeth in periapical radiographs has gained significance in forensic odontology, particularly for postmortem identification. The psychological trauma associated with manual examination of human remains has further emphasized the need for automated interpretation systems. Recent advancements demonstrate that R-FCN architectures, employing cascade feature aggregation, substantially outperform conventional methods—achieving 95.8% precision and 96.1% recall [15].

Among single-stage detectors, YOLO (You Only Look Once) has emerged as a dominant CNN-based architecture for real-time object detection. Its end-to-end design simultaneously predicts bounding box coordinates and class probabilities through a unified neural network, offering superior computational efficiency. This network demonstrates a high level of generalizability, and various versions have been introduced to date. Nevertheless, it exhibits limitations in accurately detecting objects that are either very small or in close proximity [16]. YOLOv3 has achieved approximately 80% accuracy in caries detection and tooth numbering when integrated with semantic segmentation techniques [17]. while YOLOv4 [18] demonstrates 99.31% accuracy in panoramic image analysis through feature pyramid networks (FPNs) and spatial attention modules. Subsequent versions exhibit progressive enhancements.

YOLOv7 [19] integrates channel attention blocks to detect caries in bitewing radiographs (precision: 0.833, recall: 0.866), and YOLOv8 [20] achieves dual-domain competency—simultaneously processing bitewing and panoramic imagery with 90% precision-recall balance. However, limited dataset availability remains a constraint for deep learning models, prompting the adoption of augmentation strategies (e.g., rotational transforms, multi-scale resampling) to enhance model robustness [18]. Recent innovations like the YEM-SAFN framework introduce multi-scale feature fusion for dental pathology detection, addressing size variance challenges in panoramic datasets. By incorporating hierarchical cross-spatial attention (HCSA) mechanisms, this architecture surpasses YOLOv8s in lesion detection accuracy, demonstrating the evolving potential of attention-guided detection in dental diagnostics [21]. Recent advancements in YOLO-based architectures have demonstrated significant potential for dental image analysis, particularly in the domain of lightweight segmentation frameworks. YOLO-DentSeg, an optimized variant of YOLOv8n-seg, integrates a Bidirectional Feature Pyramid Network (BiFPN) to improve multi-scale feature fusion, enabling precise localization of oral pathologies while maintaining computational efficiency suitable for clinical environments with limited resources. The model's ability to delineate diseased regions with competitive accuracy positions it as a viable solution for real-time diagnostic applications [22].

Concurrently, evaluations of YOLOv9, YOLOv10, and YOLOv11 on RGB dental images captured via mobile devices, reveal YOLOv11m as the top-performing variant for plaque cluster detection, owing to its advanced feature aggregation mechanisms. YOLOv9 employs Programmable Gradient Information (PGI) to counteract gradient dissipation in deep layers, preserving feature fidelity without sacrificing inference speed. YOLOv10 introduces a dual-label inference paradigm that combines one-to-one and one-to-many label assignments, reducing reliance on non-maximum suppression (NMS) and enhancing performance in occluded scenarios. The YOLOv11 framework builds upon YOLOv8 through the integration of C2PSA blocks for cross-scale context modeling and C3K2 blocks with optimized convolutional kernels, which collectively improve spatial resolution for detecting minute structures such as early-stage caries [23].

The spectrum of dental diseases is highly diverse, prompting the development of various models for their detection. Notably, the YOLOv3 model has demonstrated an accuracy exceeding 99% when evaluated on a dataset of 1,200 panoramic images [24]. The Spatial Pyramid Pooling Framework (SPPF) module has been incorporated into the newly developed versions of the YOLO architecture facilitating the detection of objects at varying scales through convolutional layers of different sizes. This module effectively extracts both global and local features, and the integration of the SPPF layer into the final processing stage has resulted in enhanced performance in the detection of abnormal teeth. Despite the increasing complexity of the architecture, the real-time processing capability inherent to YOLO is preserved, making this design particularly suitable for deployment in web-based tools aimed at automatic dental disease diagnosis [25].

VGG-16 [26] is a 16-layer convolutional network, comprising 13 convolutional layers and 3 fully connected layers, renowned for its robust feature extraction capabilities, albeit with a substantial number of parameters required for computations. In contrast, ResNet50 [27], which utilizes 50 layers, is designed for global feature extraction and is more lightweight than VGG-16. MobileNetV2 further exemplifies a lightweight architecture, employing spatial filtering and channel combination techniques for efficient convolution calculations, enabling rapid real-time feature extraction.

Feature extraction from panoramic images is crucial in dental imaging. Given the advantages of various feature storage models, combining these models allows for multifaceted feature extraction approaches. Traditional methods such as Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Random Forest have been employed for the detection of abnormal teeth. However, integrating feature extraction methods with the Swin Transformer architecture yields improved conditions and enhances model accuracy. Additionally, the application of bagging ensemble classifier methods for decision-making further boosts accuracy. The

optimal model identified is the combination of MobileNetV2 and Swin Transformer [28].

Various methods exist for integrating the bounding boxes generated by object detection networks. Non-Maximum Suppression (NMS) remains a foundational technique, eliminating redundant detections by retaining only the highest-confidence box when the IoU between overlapping predictions exceeds a predefined threshold. Soft-NMS mitigates these issues by decaying confidence scores of overlapping boxes proportionally to their IoU values, preserving occluded objects while penalizing low-confidence duplicates through a continuous suppression function. Weighted Box Fusion (WBF) offers a more sophisticated alternative by aggregating all candidate boxes through confidence-weighted averaging. This method computes fused box coordinates as the weighted mean of all overlapping predictions, with weights derived from their individual confidence scores. Figure 1 illustrates the operational differences between these methods, highlighting their unique handling of overlapping detections [29].

3. Material and Method

The proposed AI engine is based on the YOLOv11 [30] architecture. To effectively and quickly detect dental conditions in panoramic images, we have modified certain parts of the backbone and neck architecture. YOLOv11 exhibits robust adaptability across diverse datasets, with enhanced capability for detecting subtle object features. This version incorporates the Spatial Pyramid Pooling – Fast (SPPF) module, which enables multi-scale feature extraction—a design principle also employed in YOLOv8 [31]. The module utilizes successive fixed-size max-pooling operations, reducing computational complexity relative to earlier architectures while improving inference speed.

Key to its operation is a convolutional layer preceding max-pooling, which reduces input dimensionality, followed by channel-wise concatenation of convolutional and pooled outputs. The integration of the SiLU activation function further optimizes performance for fine-grained object detection [32]. The incorporation of the Spatial Pyramid Pooling Framework (SPPF) layer has significantly enhanced the performance of the YOLOv11 architecture in the detection of abnormal teeth [25]. Consequently, we have refined this model into a more comprehensive version by integrating additional layer connections and implementing a more precise mechanism for feature extraction.

Proper feature extraction from images is essential for improving the engine's performance. Figure 2 shows the proposed network architecture.

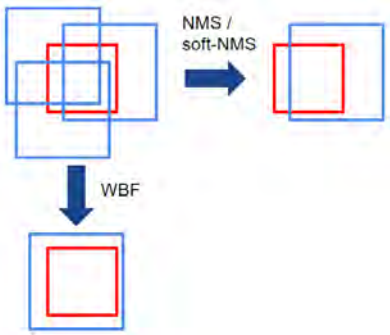

Figure. 1. NMS/soft-NMS vs. WBF [23]

The Spatial Pyramid Pooling Fast (SPPF) block aids in extracting detailed features. By compressing features after the pooling process in panoramic images, we enhance model performance. This compression happens in the backbone, before the last convolution block, which helps retain more details in the image. Feature extraction is advanced through the C3k2 module, which employs dual convolutional layers to capture intricate patterns with high flexibility.

Meanwhile, the Cross-Stage Partial with Spatial Attention (C2PSA) module enhances region-specific focus via a bifurcated processing pipeline: initial feature extraction through convolution is followed by division into dual branches, each processed by PSA blocks to weight spatially significant features. These branches are subsequently merged via convolutional fusion, preserving anatomical context in panoramic imagery through explicit spatial relationship modeling [30]. In our model, this module is repeated in the neck section, taking input from the SPPF and sending it to the Upsampler. This version of the module has less spatial attention than the original C2PSA, resulting in reduction in the model's time complexity. This block is called C2OSA, or two Convolutional Block with One Spatial Attention. Figure 3 depicts the architecture of this block.

In dental radiography, the consolidation of object detection outputs requires specialized methodologies to address challenges such as tooth occlusion and anatomical proximity. Predictions from detection networks consist of bounding box coordinates, categorical labels (e.g., molar, incisor), and confidence scores quantifying prediction certainty. NMS remains a foundational technique, eliminating redundant detections by retaining only the highest-confidence box when the IoU between overlapping predictions exceeds a predefined threshold [29]. However, NMS exhibits critical limitations in dental contexts: its performance is sensitive to IoU threshold selection, where overly stringent values may suppress valid detections of adjacent teeth, and it struggles to differentiate overlapping structures

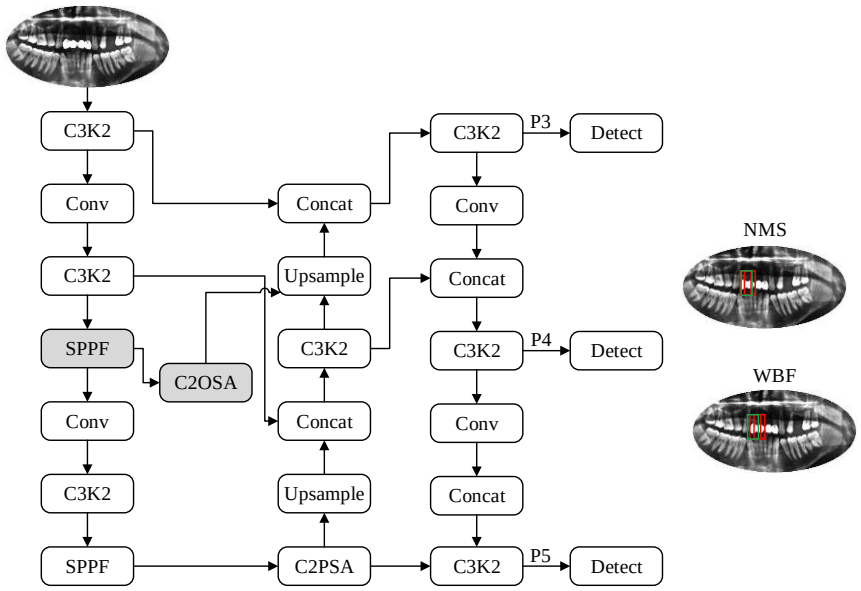

Figure. 2. Proposed Model Architecture

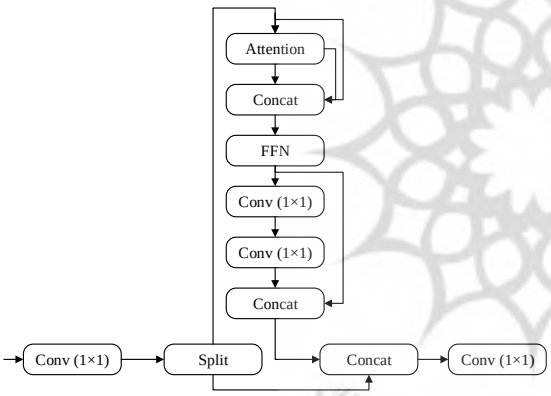

Figure. 3. C2OSA (Customized C2PSA)

due to its binary suppression logic. To resolve this issue, we applied the WBF [29] method for classification. This approach improves the detection accuracy of densely packed teeth by utilizing spatial consensus among multiple detections, rather than eliminating lower-confidence candidates. Furthermore, it enhances the reliability of dental abnormality classification.

During the training phase, images were processed at a resolution of 640×640 pixels over 100 epochs, using a batch size of 16. Notably, no preprocessing techniques were applied to the training or validation datasets. To ensure robustness and generalizability, a 5-fold cross-validation strategy was employed to split the data. All training and inference tasks were conducted on a computer with 128 GB of system RAM and dual RTX 3090 GPUs, each equipped with 48GB of dedicated memory.

3.1. Dataset

For the training and evaluation of the model, a publicly accessible dataset of dental panoramic X-ray images was employed [33]. This dataset was structured into three directories: validation, test, and training, encompassing a total of 1,269 panoramic images. In six instances, the same tooth was annotated with varying labels; these images were excluded from the dataset to improve the model's accuracy. Furthermore, all data were utilized in the 5-fold cross-validation method, with no fixed partitioning of the data.

The labels comprise four distinct categories: Implant, Fillings, Impacted Tooth, and Cavity. The remaining images within the dataset include 2,032, 6,039, 495, and 630 instances of each respective class, thereby illustrating the class imbalance inherent in the dataset. Labeling medical images is crucial, as the model's accuracy is highly dependent on the quality of the training labels. Consequently, it is essential to report on the labeling methodology and the expertise of the individuals involved. However, no information regarding this has been published, and this research relied solely on a public database.

Data augmentation was performed using copy-paste techniques with random 25%, 20%, and 15% for horizontal flips, rotation, and resizing. Given the significance of tooth position and type in these images, the augmented images were placed in the original location of the tooth, potentially overlapping with adjacent teeth. It is noteworthy that

the overall structure of the image was maintained throughout this process, and data augmentation was performed on two classes: Impacted Tooth and Cavity. Figure 4 presents examples of dataset images.

3.2. Evaluation metrics

The evaluation of the proposed model was conducted using standard criteria commonly employed for assessing deep learning methods. The calcification issue can be approached from two analytical frameworks: binary classification, which entails distinguishing between diseased and healthy teeth, and multi-class classification, which aims to differentiate among various dental pathologies in relation to healthy teeth. In the binary framework, a True Positive (TP) denotes that the model has accurately classified a tooth as diseased. Conversely, in the multi-class context, a TP is recorded when the model not only identifies the presence of a disease but also correctly classifies its specific type. An increased TP count is indicative of enhanced model performance in accurately diagnosing dental conditions.

Within the binary classification paradigm, a False Positive (FP) occurs when the model erroneously identifies a healthy tooth as diseased. Such misclassifications can lead to unnecessary clinical interventions, incurring both time and financial costs. In the multi-class setting, an FP is defined as the model identifying a tooth as diseased while incorrectly categorizing the specific disease type. Reducing the FP rate is crucial for minimizing the economic burden of misdiagnoses. A False Negative (FN) in the binary classification context is characterized by the model failing to recognize a diseased tooth, instead classifying it as healthy. This error is particularly critical as it conceals the disease from the clinician's attention. In the multi-class scenario, an FN arises when the model completely overlooks the disease, misclassifying it as healthy or neglecting to identify it altogether.

In binary classification, a True Negative (TN) indicates that the model has accurately predicted a tooth to be healthy. In the multi-class context, TN similarly pertains to the correct identification of a healthy state. A high TN value suggests that the model does not exhibit a significant bias towards diagnosing diseases and is proficient in detecting healthy teeth. The metrics reported in the confusion matrix are interpreted from a binary approach. The Precision metric, calculated using Equation (1), quantifies the proportion of teeth identified by the model as diseased that are indeed diseased, encompassing accurate diagnoses of disease types. The Recall metric evaluates the ratio of teeth classified as diseased by the model relative to the total number of teeth that are actually diseased or exhibit a specific disease type, as detailed in Equation (2).

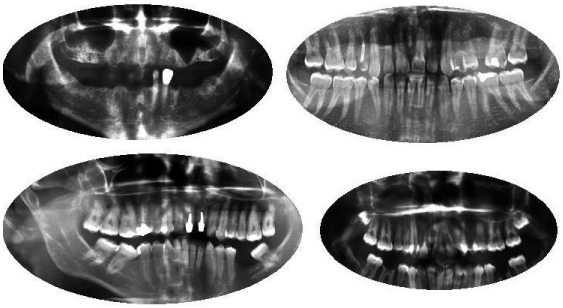

Figure. 4. Samples of dataset images (randomly selected)

It is essential to acknowledge the inherent trade-off between Precision and Recall.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

To assess the overall accuracy of the model, Equation (3) is employed especially when the data is imbalanced, the model might perform poorly on the smaller classes but still show high accuracy because it correctly classifies the majority class most of the time. To get a better understanding of the model's true performance, the F1-score is often used; it is calculated as shown in Equation (4) and is the harmonic mean of precision and recall. The F1-score balances these two metrics, making it especially useful in medical datasets where correctly identifying cases with abnormalities (recall) and avoiding false alarms (precision) are both very important. Since medical data often involves uneven class distributions, the F1-score provides a more accurate picture of how well the model can detect abnormalities across all categories, not just the most common ones.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

4. Results

The YOLOv11 architecture demonstrates enhanced detail detection through region-specific attention mechanisms, though its efficacy in panoramic X-ray imaging hinges on precise anatomical focus. Comparative analysis against the baseline YOLOv11 reveals superior performance in the proposed model, evidenced by balanced precision-recall metrics. As quantified in Table 1 (5-fold cross-validation averages), the model's outputs were validated via both Non-Maximum Suppression

Table 1. Evaluation Metrics of the Proposed Model

Model	*Precision*	*Recall*	*Accuracy*	*F1-score*
Proposed YOLO(NMS)	0.832	0.795	0.839	0.813
Proposed YOLO(WBF)	0.862	0.825	0.843	0.883

(NMS) and Weighted Boxes Fusion (WBF), with performance evaluated at an intersection-over-union (IoU) threshold more than 0.5. WBF proved particularly effective in dental applications, improving tooth localization accuracy while consolidating high-confidence predictions and retaining diagnostically relevant low-confidence boxes.

Class-specific improvements from WBF are detailed in Table 2. Medical imaging necessitates careful IoU calibration, as misclassification costs differ markedly between false positives (e.g., classifying a healthy tooth as abnormal) versus false negatives (e.g., missing a diseased tooth). WBF's weighted averaging reduced redundant detections compared to NMS, yielding more decisive predictions. However, performance disparities emerged due to dataset imbalances—notably, the Cavity class, underrepresented in training data, exhibited lower detection rates (F1-score: 0.815 vs. 0.872 for Filling). The influence of data augmentation on NMS and WBF is consistent when considering its effects on the training process. Overall, the model demonstrates superior performance in the Filling class compared to the other classes.

In this investigation, the F1-score metrics for cavity and impacted teeth are observed to be 8.3% and 3.6% lower, respectively, in the absence of data augmentation. While augmenting the data in panoramic radiographs enhances model performance, it does not entirely mitigate the class imbalance issue due to the intrinsic limitations associated with anatomical fidelity. As illustrated in Table 2, the model exhibits superior performance in classifying impacted teeth compared to cavities. Notably, the effect of data augmentation is more significant in the cavity class, as impacted teeth, primarily the third molars, are typically associated with higher detection accuracy owing to their distinct anatomical positioning. Cavity is the only class exhibiting a higher recall than precision. The model demonstrates a high sensitivity to identifying teeth within this class, even as the number of false positive alerts increases. Consequently, the model's accuracy in detecting cavities is lower compared to other classes. While data augmentation has enhanced performance in this regard, it has not entirely resolved the issue.

Table 2. Evaluation Metrics Across Classes

Model	*Class*	*Precision*	*Recall*	*Accuracy*	*F1-score*
Proposed YOLO(NMS)	Implant	0.886	0.797	0.889	0.839
	Fillings	0.867	0.843	0.872	0.855
	Impacted Tooth	0.812	0.763	0.821	0.786
	Cavity	0.765	0.777	0.775	0.771
Proposed YOLO(WBF)	Implant	0.906	0.816	0.912	0.859
	Fillings	0.884	0.861	0.894	0.872
	Impacted Tooth	0.847	0.801	0.882	0.823
	Cavity	0.807	0.823	0.832	0.815

Model performance was assessed using the normalized confusion matrix, which enables a detailed analysis of class-wise discrimination. Figure 5 illustrates the normalized confusion matrices corresponding to the NMS and WBF methods. As depicted, the WBF approach demonstrates improved detection performance across all classes. Notably, the model exhibits a tendency to misclassify instances of the Cavity class as the background or healthy. Conversely, in the Filling class, which has a larger sample size, the model occasionally misidentifies healthy teeth as fillings. Furthermore, due to visual similarities between dental fillings and implants, misclassification of implants as fillings is also observed (12%). To address class imbalance and enhance differentiation, class weighting adjustments were incorporated into the model. However, these modifications yielded limited performance gains, particularly due to the inherent challenge of distinguishing Cavity teeth from small implants or anatomical structures such as parts of the sinus visible in panoramic radiographs.

Figure 6 illustrates the ROC curves of the model for both the NMS and WBF methods, with standard deviations calculated from a 5-fold cross-validation procedure. Notably, the Impacted tooth class exhibits the highest standard deviation, which can be attributed to its limited number of training samples—an issue exacerbated by data partitioning in cross-validation. Despite this, the WBF method demonstrates superior performance, with reduced variability across folds. This is likely due to its ability to more effectively integrate detection results by down-weighting low-confidence predictions and emphasizing high-confidence boxes, particularly in scenarios where the model is biased toward classes with larger sample sizes.

Figure 7 compares the model outputs generated using NMS and WBF. To assess performance, a low IoU threshold of 0.2 was selected to highlight differences in the number and quality of predicted bounding boxes. As shown in part (a), WBF provides

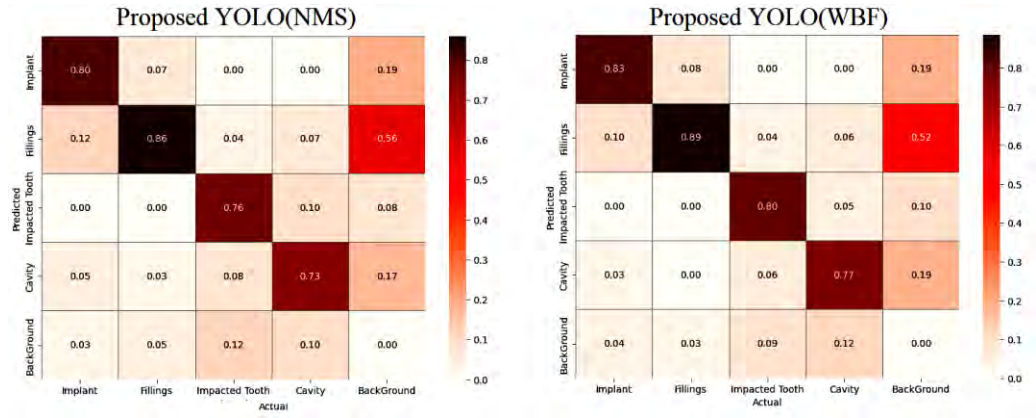

Figure. 5. Normalized Confusion Matrices

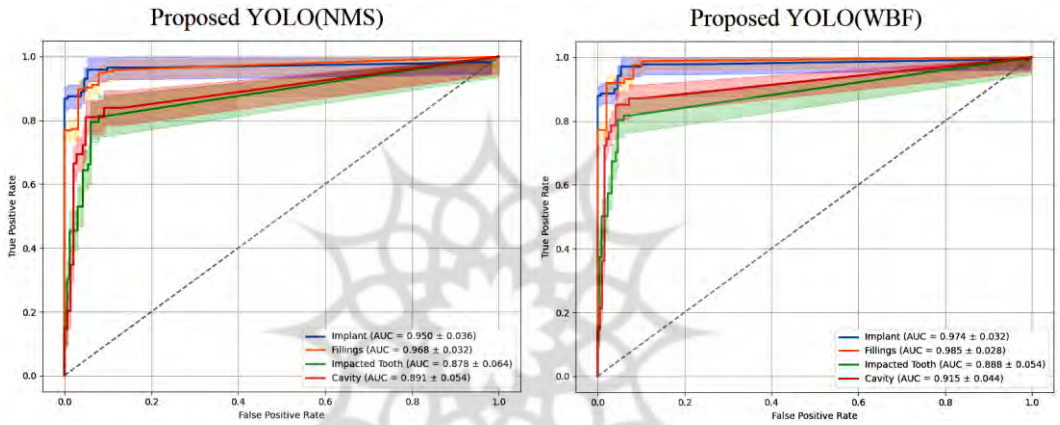

Figure. 6. ROC Curves

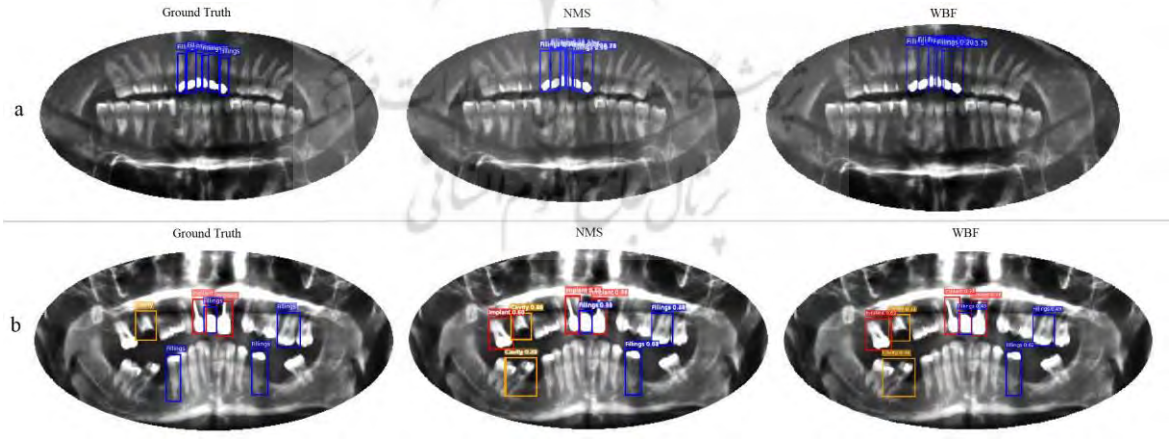

Figure. 7. Model Output

improved detection of overlapping teeth compared to NMS. In this sample, NMS produced 9 bounding boxes for the root canal structures, whereas WBF yielded only 4, reflecting a more concise and accurate output. Part (b) of Figure 7 presents another example, where NMS generated 32 boxes, in contrast to only 8 boxes detected by WBF at the same IoU threshold. In this case, although the filled tooth on the lower left was missed by both methods, WBF more clearly separated the bounding boxes for the three central teeth, indicating better spatial resolution in overlapping regions.

5. Discussion

YOLO models execute detection in a single step by extracting global features from the upper layers. These models are specifically designed for real-time processing, resulting in high-speed performance [16]. Additionally, the deeper layers extract local features, enabling the architecture to detect multi-scale objects effectively. Consequently, the performance of the model can be compared with similar architectures to provide a comprehensive analytical review of its efficacy.

5.1. Similar architectural frameworks

The enhanced YOLOv11-based [30] framework demonstrates superior performance over the original architecture through three key modifications: (1) a spatially adaptive sensitivity mechanism that prioritizes diagnostically critical regions, (2) hierarchical multi-scale feature integration for improved decision boundaries across anatomical structures, and (3) optimized utilization of primary convolutional features. Quantitative evaluation reveals consistent improvements of 5.5% in precision, 0.6% in recall, and 7.1% in F1-score (Table 3), with the balanced F1- score metric confirming the model's robustness against classification bias.

The overall efficacy of YOLOv11 [30] exceeds that of YOLOv9 [34] and YOLOv10 [34] when assessed using the F1-score metric. Conversely, YOLOv8 [31] exhibits superior precision relative to YOLOv11. The evaluation of the models presented in the first four rows of Table 3 employs the original architecture along with the pre-trained weights. Significantly, the proposed method demonstrates enhanced performance compared to versions 8 through 11 in terms of NMS. Regarding the WBF technique, precision enhancements of 5%, 10.8%, 8%, and 5.5% are recorded for versions 8 through 11, respectively. Additionally, for versions 8 to 10, akin to the proposed methodology, recall remains lower than precision, suggesting that these models are characterized by a reduced incidence of false positives.

Comparative analysis with the YEM-SAFN [21] model—an extension of YOLOv8 incorporating target-specific network structures—demonstrates the competitive advantage of our proposed approach. The integration of the Hybrid Cross-Scale Attention (HCSA) module effectively addresses challenges associated with anatomical overlap artifacts. When applied under identical experimental conditions, this model yielded the results presented in Table 3. Notably, our framework outperformed both YEM-SAFN and the baseline YOLOv11 in lesion detection tasks, while maintaining comparable computational complexity. The proposed methodology exhibits superior precision compared to YEM-SAFN; however, the recall rate of the YEM-SAFN model is greater than that of the proposed approach. To facilitate a comprehensive comparison, the harmonic mean F1-score was employed, indicating enhanced overall performance for the proposed model. Furthermore, the proposed approach effectively reduces the incidence of false positives.

Table 3. Comparison of Evaluation Criteria for Models with Similar Architecture

Model	*Precision*	*Recall*	*Accuracy*	*F1-score*
YOLOv8 [31]	0.812	0.751	0.795	0.780
YOLOv9 [34]	0.754	0.729	0.771	0.741
YOLOv10 [34]	0.782	0.764	0.795	0.771
YOLOv11 [30]	0.807	0.819	0.828	0.812
YEM-SAFN [21]	0.845	**0.873**	**0.862**	0.858
YOLO-DentSeg [22]	0.753	0.784	0.860	0.768
YOLO11+SPPF [25]	0.814	0.811	0.822	0.812
Proposed YOLO(NMS)	0.832	0.795	0.839	0.813
Proposed YOLO(WBF)	**0.862**	0.825	0.843	**0.883**

Further validation against YOLO-DentSeg [22]—a modified YOLOv5s variant that employs a triple-attention mechanism and a Bidirectional Feature Pyramid Network (BiFPN)—highlights the strengths of our architecture in analyzing cervical regions. YOLO-DentSeg leverages the CIoU loss function to enhance localization accuracy and is specifically tailored for the detection of caries, impacted teeth, periapical periodontitis, and bifurcated root lesions. On the same condition, our model achieved a 10.9% improvement in mean detection precision over YOLO-DentSeg (Table 3), primarily due to more effective feature fusion and a reduction in spatial redundancy during high-resolution image processing.

5.2. Architectural diversity

Two-stage methods, such as Faster R-CNN [35], require more time for detection compared to the YOLO family. While Faster R-CNN extracts significant local features using anchors, it demonstrates weaker performance than the proposed method, as indicated in Table 4. This is primarily due to the added block between the backbone and the neck in the proposed method, which mitigates the loss of critical features. DensNet-121 [36] excels in local feature extraction through dense connections between preceding and succeeding layers; however, it is slower and yields lower accuracy than the proposed method, which achieves 3.9% and 2.2% improvements in precision and recall, respectively. Feature extraction via VGG-16 [26] is more computationally intensive than the proposed model,

Table 4. Comparison of Evaluation Criteria for Models with Different Architecture

Model	*Precision*	*Recall*	*Accuracy*	*F1-score*
Faster R-CNN [35]	0.734	0.765	0.783	0.749
DensNet-121 [36]	0.823	0.803	0.827	0.812
VGG16 [26]	0.843	0.817	0.853	0.829
ResNet50 [27]	0.832	0.753	0.841	0.791
ResNet50 + Swin [28]	**0.921**	**0.911**	**0.934**	**0.915**
MobileNetV2 + Swin [28]	0.908	0.896	0.921	0.901
Proposed YOLO(NMS)	0.832	0.795	0.839	0.813
Proposed YOLO(WBF)	0.862	0.825	0.843	0.883

attributable to its higher parameter count. Nevertheless, using the perceptron, VGG-16 attains bounding box accuracy of 0.853, surpassing YOLOv11 with the original architecture, although it exhibits a 5.4% lower F1-score compared to the proposed architecture.

ResNet50 [27], which employs a head perceptron for decision-making, performs weaker than both the VGG-16 method and the proposed approach. However, the combined models demonstrate superior performance relative to the proposed model. The Swin Transformer [36] exhibits strong capabilities in global feature extraction, and its integration with local feature extraction methods enhances performance. The simple head MLP, which combines features from ResNet50 and the Swin Transformer, achieves the highest precision and recall values of 0.921 and 0.911, respectively.

Combining the Swin Transformer with MobileNetV2 results in a 1.8% improvement in F1-score over the proposed model; however, this configuration underperforms compared to the combination with ResNet50 due to the reduced parameter count of MobileNetV2. The combined models are more complex and demanding in terms of execution time than the proposed method, making them impractical for medical applications. Despite this, the combined features of ResNet50 and the Swin Transformer yield a 3.2% better F1-score. The proposed model, with its real-time capabilities, presents significant advantages for use as a medical tool.

5.3. Challenges and limitations

A critical challenge in the application of artificial intelligence within the medical domain is the assurance of reliability. Deep learning models are inherently dependent on labeled datasets; thus, any biases present in the labeling process can propagate through to the model, leading to skewed outcomes. Moreover, the process of labeling in the medical field is contingent upon the expertise of practitioners, which introduces variability and inconsistency. The lack of standardized benchmarks that provide accurate information regarding labeling conditions significantly undermines the reliability of models designed for the detection of abnormal teeth in panoramic radiographs. Furthermore, the hyperparameters utilized in these models may differ across various datasets, complicating the training process for models tasked with analyzing distorted images.

To assess model generalizability, we conducted experiments on the DENTEX [37] dataset, which contains 705 labeled images across four pathological classes: Lesion Caries, Deep Caries, Impacted, and Periapical. Although detailed metadata regarding image acquisition, annotation standards, and expert involvement is unavailable, the model demonstrated clinically acceptable performance. Specifically, for the Impacted class—the only category directly comparable between datasets—our model achieved a precision of 0.752, recall of 0.698, and F1-score of 0.723 without additional fine-tuning. These results underscore the model's robust transfer learning capability, particularly given the domain-specific training and the relatively limited number of corresponding samples in the dataset before augmentation.

A notable limitation in the detection of abnormal dental conditions in panoramic images is the disparate frequency of disease occurrence, which contributes to data imbalance. As a result, machine learning models tend to exhibit heightened sensitivity toward classes with greater representation, making it difficult to train effectively on less prevalent data. Employing weighting and data augmentation methods introduces further complications in panoramic images, as the overall anatomy of the mouth is critical, and the spatial arrangement of premolars and canines must remain consistent.

6. Conclusion

This study presents an optimized YOLOv11-based framework for the detection of four critical dental anomalies—implants, fillings, impacted teeth, and cavities—in panoramic radiographs. The proposed architecture enhances state-of-the-art performance through three principal innovations: (1) a location-sensitive attention mechanism, (2) hierarchical multi-scale feature extraction, and (3) a bounding-box score-weighted post-processing technique that refines detection confidence. One of the advantages of the proposed method is its ability to respond quickly based on the original YOLO architecture, making it well-suited for integration into web software platforms and tools. Comparative

evaluations demonstrate consistent improvements of over 0.6% in recall and over 5.5% in precision relative to the baseline YOLOv11, while also outperforming specialized models such as YEM-SAFN and YOLO-DentSeg in cross-dataset validation. Notably, the framework maintains strong generalization capabilities, achieving competitive accuracy on external datasets without retraining.

To effectively address the issue of data imbalance in dental imaging, future research should concentrate on the development of methodologies that are intricately aligned with the anatomical complexities of the oral cavity, ensuring comprehensive representation of each dental phenotype across a diverse range of imaging modalities. Moreover, subsequent to the data augmentation process, the integration of advanced image quality enhancement algorithms, specifically tailored to the unique transformations applied during augmentation, has the potential to significantly improve diagnostic accuracy. Additionally, exploring the incorporation of transformer fusion techniques within the model architecture, while maintaining critical real-time processing capabilities, presents a promising avenue for further investigation.

Declarations

Funding

This research did not receive any funding from public, commercial, or non-profit organizations.

Authors' Contributions

MD: Methodology, design and implementation, analysis of results, validation, and writing – editing.

RZ: Conceptualization, project management, supervision, validation, review, and editing.

Data Availability Statement

All datasets used in this study are publicly available, not owned by the authors, and were utilized exclusively for research purposes.

Ethics Approval

This study did not involve human or animal subjects; therefore, institutional review board approval was not required.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

[1] I. Shafi *et al.*, "A comprehensive review of recent advances in artificial intelligence for dentistry e-health," *Diagnostics*, vol. 13, no. 13, p. 2196, 2023, https://doi.org/10.3390/diagnostics13132196

[2] L. Lu, D. He, C. Liu and Z. Deng, "MASF-YOLO: An Improved YOLOv11 Network for Small Object Detection on Drone View", *arXiv preprint, arXiv:2504.18136*, 2025, https://doi.org/10.48550/arXiv.2504.18136

[3] Z. M. Semerci, S. Yardımcı, "Empowering modern dentistry: The impact of artificial intelligence on patient care and clinical decision making," *Diagnostics*,vol. 14, no. 12, p. 1260, 2024, https://doi.org/10.3390/diagnostics14121260

[4] R. Khan *et al.*, "Dental image enhancement network for early diagnosis of oral dental disease," *Scientific Reports*,vol. 13, no. 1, p. 5312, 2023, https://doi.org/10.1038/s41598-023-30548-5

[5] H. R. Choi *et al.*, "Automatic detection of teeth and dental treatment patterns on dental panoramic radiographs using deep neural networks," *Forensic Sciences Research*, vol. 7, no. 3, pp. 456-466, 2022, https://doi.org/10.1080/20961790.2022.2034714

[6] É da Silva Rocha, P. T. Endo, "A comparative study of deep learning models for dental segmentation in panoramic radiograph," *Applied Sciences*, vol. 12, no. 6, p. 3103, 2022, https://doi.org/10.3390/app12063103

[7] H. Chen, H. Li, Y. Zhao, J. Zhao, Y. Wang, "Dental disease detection on periapical radiographs based on deep convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, pp. 649-661, 2021, https://doi.org/10.1007/s11548-021-02319-y

[8] Y. Guo *et al.*, "Rapid detection of non-normal teeth on dental X-ray images using improved Mask R-CNN with attention mechanism." *International Journal of Computer Assisted Radiology and Surgery*, vol. 19, no. 4, pp. 779-790, 2024, https://doi.org/10.1007/s11548-023-03047-1

[9] R. Anantharaman, M. Velazquez and Y. Lee, "Utilizing mask R-CNN for detection and segmentation of oral diseases," *In 2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 2197-2204, https://doi.org/10.1109/BIBM.2018.8621112

[10] C. Wang, J. Yang, H. Liu, P. Yu, X. Jiang and R. Liu, "Co-Mask R-CNN: collaborative learning-based method for tooth instance segmentation," *Journal of Clinical Pediatric Dentistry*, vol. 48, no. 6, 2024, https://doi.org/10.22514/jocpd.2024.136

[11] M. Nandeesh, B. Naveen and C. N. Srividya, "Tooth Enamel segmentation from dental x-ray using MASK R-CNN Algorithm," In *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, B G Nagara, Mandya, India, 2024, pp. 1-5, https://doi.org/10.1109/ICRASET63057.2024.10895799

[12] S. L. Chen *et al.*, "Detection of various dental conditions on dental panoramic radiography using faster R-CNN," *IEEE Access*, vol. 11, pp. 127388-127401, 2023, https://doi.org/1109/ACCESS.2023.3332269

[13] A. Laishram and K. Thongam, "Detection and classification of dental pathologies using faster-RCNN in orthopantomogram radiography image," In *2020 7th international conference on signal processing and integrated networks (SPIN)*, Noida, India, 2020, pp. 423-428, https://doi.org/10.1109/SPIN48934.2020.9071242

[14] Y. Zhu *et al.*, "Faster-RCNN based intelligent detection and localization of dental caries," *Displays*, vol. 74, p. 102201, 2022, https://doi.org/10.1016/j.displa.2022.102201

[15] K. Zhang, J. Wu, H. Chen and P. Lyu, "An effective teeth recognition method using label tree with cascade network structure," *Computerized Medical Imaging and Graphics*, vol. 68, pp. 61-70, 2018, https://doi.org/10.1016/j.compmedimag.2018.07.001

[16] P. Jiang, D. Ergu, F. Liu, Y. Cai and B. Ma, "Review of Yolo algorithm developments," *Procedia computer*

science, vol. 199, pp. 1066-1073, 2022, https://doi.org/10.1016/j.procs.2022.01.135

[17] T. Sheryl Abraham, V. Jeyakumar, G. Marthi Krishna Kumar and P. Abraham Anandapandian, "Automated Analysis of Tooth Anatomy and Pathological Conditions from Orthopantomogram using Deep Neural Networks," *IETE Journal of Research*, vol. 70, no. 12, pp. 8702-8713, 2024, https://doi.org/10.1080/03772063.2024.2385044

[18] E. Kaya, H. G. Güneç, E. Ş. Ürkmez, K. C. Aydın and H. Fehmi, "Deep learning for diagnostic charting on pediatric panoramic radiographs," *International Journal of Computerized Dentistry*, vol. 27, no. 3, p. 225, 2024, https://doi.org/10.3290/j.ijcd.b4200863

[19] B. Ayhan, E. Ayan and Y. Bayraktar, "A novel deep learning-based perspective for tooth numbering and caries detection," *Clinical Oral Investigations*, vol. 28, no. 3, p. 178, 2024, https://doi.org/10.1007/s00784-024-05566-w

[20] M. Razaghi, H. E. Komleh, F. Dehghani and Z. Shahidi, "Innovative Diagnosis of Dental Diseases Using YOLO V8 Deep Learning Model," In *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*, Tehran, Islamic Republic of Iran, 2024, pp. 1-5, https://doi.org/10.1109/MVIP62238.2024.10491172

[21] Q. Wang, X. Zhu, Z. Sun, B. Zhang, J. Yu and S. Qian, "Optimized Yolov8 feature fusion algorithm for dental disease detection," *Computers in Biology and Medicine*, vol. 187, p. 109778, 2025, https://doi.org/10.1016/j.compbiomed.2025.109778

[22] Y. Hua, R. Chen and H. Qin, "YOLO-DentSeg: A Lightweight Real-Time Model for Accurate Detection and Segmentation of Oral Diseases in Panoramic Radiographs," *Electronics*, vol. 14, no. 4, p. 805, 2025, https://doi.org/10.3390/electronics14040805

[23] A. Ramírez-Pedraza *et al*., "Deep Learning in Oral Hygiene: Automated Dental Plaque Detection via YOLO Frameworks and Quantification Using the O'Leary Index. *Diagnostics*, vol. 15, no. 2, p. 231, 2025, https://doi.org/10.3390/diagnostics15020231

[24] H. Sadr *et al*., "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," *European Journal of Medical Research*, vol. 30, no. 1, p. 418, 2025, https://doi.org/10.1186/s40001-025-02680-7

[25] M. Dehghani and R. Aghaeizadeh Zoroofi, "AI-Driven Dental Disease Detection: A Web-Based Application Utilizing an AI Engine for Panoramic Image Analysis," In *2025 11th International Conference on Web Research (ICWR)*, Tehran, Islamic Republic of Iran, 2025, pp. 61-65, https://doi.org/10.1109/ICWR65219.2025.11006210

[26] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143-150, 2019, http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9420

[27] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In European conference on computer vision, pp. 630-645. Cham: Springer International Publishing, 2016.

[28] Y. M. Alsakar, N. Elazab, N. Nader, W. Mohamed, M. Ezzat and M. Elmogy, "Multi-label dental disorder diagnosis based on MobileNetV2 and swin transformer using bagging ensemble classifier," *Scientific Reports*, vol. 14, no. 1, p. 25193, 2024, https://doi.org/10.1038/s41598-024-73297-9

[29] R. Solovyev, W. Wang and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*,, vol. 107, p. 104117, 2021, https://doi.org/10.1016/j.imavis.2021.104117

[30] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024, https://doi.org/10.48550/arXiv.2410.17725

[31] X. Chi, Y. Sun, Y. Zhao, D. Lu, Y. Gao and Y. Zhang, "An Improved YOLOv8 Network for Detecting Electric Pylons Based on Optical Satellite Image," *Sensors*, vol. 24, no. 12, p. 4012, 2024, https://doi.org/10.3390/s24124012

[32] K. He, X. Zhang, S. Ren and . Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015, https://doi.org/10.1109/TPAMI.2015.2389824

[33] Dental Radiography Analysis and Diagnosis Dataset. Available online:. https://www.kaggle.com/datasets/imtkaggleteam/dental-radiography/data (Accessed on 28 Aug 2025).

[34] A. R. I. Davut and M. Burukanli, "Deep Learning for Dentistry: Yolo Variants in Tooth Abnormality Detection," *Computer Science Engineering*, 2025. https://www.gecekitapligi.com/Webkontrol/uploads/Fck/32-Bilgisayar_bilim_m%C3%BCh_ing_Haziran_2025_DK_V2.pdf#page=7

[35] R. Girshick, "Fast R-Cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448, https://doi.org/10.1109/ICCV.2015.169

[36] Y. Zhu and S. Newsam, "Densenet for dense flow," *2017 IEEE international conference on image processing (ICIP)*, Beijing, China, 2017, pp. 790-794. https://doi.org/10.1109/ICIP.2017.8296389

[37] I. E. Hamamci *et al*., "DENTEX: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays," *arXiv preprint, arXiv:2305.1911*, 2023, https://doi.org/10.48550/arXiv.2305.19112

Mahdieh Dehghani earned her B.Sc. in Computer Engineering from Shahid Bahonar University of Kerman in 2017, graduating as the top-ranked student in her cohort. She continued her academic excellence by obtaining an M.Sc. degree in Computer Engineering from the University of Tehran in 2019, where she again ranked first in her class. Her outstanding academic performance earned her a competitive scholarship from the National Elites Foundation of Iran in the same year. Currently, she pursuing the Ph.D. in Computer Engineering with a specialization in Software Engineering at the University of Tehran. Her research focuses on advanced applications of deep learning in computer vision, particularly in developing novel architectures for medical image analysis.

Reza Aghaeizadeh Zoroofi received the Ph.D. in Medical Image Analysis from Osaka University, Japan, in 1996. From 1996 to 1999, he conducted postdoctoral research at multiple prestigious Japanese institutions, including the Ministry of Industry of Japan, the Cardiovascular Research Center of Japan, and Osaka University's Faculty of Medicine. In March 2000, he joined the University of Tehran as an Assistant Professor and has since advanced to the rank of Full Professor at the School of Electrical and Computer Engineering. Since 1999, Professor Zoroofi has sustained active research collaborations with several Japanese universities as a visiting professor and researcher. His expertise spans medical image processing, image engineering, and the integrated management of medical imaging systems, with a focus on developing clinically translatable computational solutions.