An Explainable Deep Learning Model for Dental Caries Detection and Segmentation

Walid Brahmi[a,d], Imen Jdey[b,d] and Fadoua Drira[c,d]

[a]National School of Electronics and Telecommunications of Sfax (ENET-Com), University of Sfax , Tunisia
[b]Faculty of Economics and Management of Sfax (FSEGS), , University of Sfax, Tunisia
[c]National Engineering School of Sfax (ENIS), , University of Sfax, Tunisia
[d]Research Groups in Intelligent Machines (REGIM Lab), ENIS, University of Sfax, Tunisia

ARTICLE INFO

Keywords:
Dental Caries Detection
Deep Learning
YOLO11
Instance Segmentation
Explainable AI
Grad-CAM
LIME
Hyperparameter Optimization
Optuna

ABSTRACT

Timely diagnosis of dental caries is fundamental to preventive oral care; however, manual interpretation of panoramic radiographs remains labor-intensive and susceptible to diagnostic subjectivity. While Deep Learning (DL) has demonstrated high performance in medical imaging, its clinical integration is significantly hindered by the "black-box" nature of neural networks and a lack of alignment with clinical risk priorities, such as the high cost of false negatives. To address these limitations, we present a clinical-risk-aware framework for automated caries detection and instance segmentation utilizing the YOLO11-seg architecture. This pipeline enhances generalization under real-world conditions by integrating Bayesian hyperparameter optimization via the Optuna framework with an augmentation-robust strategy tailored for radiographic noise. The model is rigorously evaluated on the COCO-Caries dataset, comprising 2,668 tooth-level cropped radiographs.

Our optimized YOLO11-seg model demonstrates superior performance over YOLOv8-seg and vanilla YOLO11-seg baselines, achieving a box-level precision of 93.8%, a recall of 75.4%, and an mAP@50 of 85.4%. Critically, these gains are realized while maintaining a rapid inference speed of 5.2 ms and a reduced parameter count of 2.83M, facilitating real-time chairside deployment. By incorporating dual explainability techniques—Grad-CAM for spatial saliency and LIME for model-agnostic local interpretations—the framework provides transparent visual rationales for its predictions. This synthesis of clinical-risk-aware optimization and interpretable Artificial Intelligence (AI) establishes a robust pipeline for dental diagnostics, effectively bridging the gap between high-performance deep learning and clinical trust.

1. Introduction

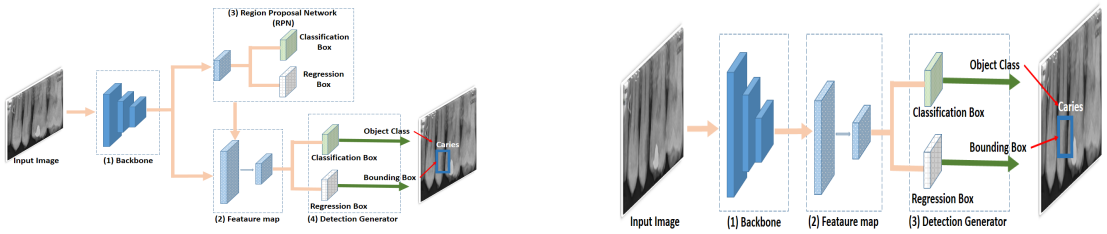
Good oral health is essential for daily activities such as eating, breathing, and speaking; it also directly impacts mental well-being and social interactions. According to the World Health Organization (WHO), oral diseases are among the most prevalent noncommunicable diseases globally, affecting approximately 3.5 billion people. Dental caries, the most common of these diseases (WHO, 2023), poses a significant challenge to health, quality of life, and healthcare systems, necessitating accurate and timely diagnosis.

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized medical imaging by automating complex diagnostic tasks. In dentistry, recent advancements in automated tooth labeling and segmentation, such as prototype-based meta-learning approaches (Sehar et al., 2025), pave the way for more intelligent diagnostic assistants. While CNNs show remarkable success in classifying dental structures and segmenting teeth from panoramic X-rays (Brahmi and Jdey, 2024), their application in detecting subtle carious lesions remains challenging.

The rapid adoption of deep learning in dental diagnostics has outpaced the development of interpretability methods. Most existing studies function as non-interpretable systems, offering clinicians a result without a rationale. Furthermore, standard models are often trained to maximize generic accuracy, overlooking specific clinical priorities, such as minimizing false negatives (i.e., missed caries) in high-risk scenarios. This lack of transparency and clinical alignment is a significant barrier to trust and effective human-AI collaboration.

To bridge these gaps, we propose a clinical-risk-aware DL framework for automated caries detection and instance segmentation on tooth-cropped radiographs. We introduce a novel pipeline based on the YOLO11-seg architecture,

*Corresponding author: Walid Brahmi
bensghaierwaleed@gmail.com (W. Brahmi); imen.jdey@fstsbz.u-kairouan.tn (I. Jdey); fadoua.drira@enis.tn (F. Drira)
ORCID(s):



(a) Two-Stage Object Detector Architecture

(b) Single Stage Object Detector Architecture

Figure 1: Single- vs. two-stage object detector architectures (adapted from (Carranza-García et al., 2020)).

enhanced through Bayesian hyperparameter optimization (Optuna). Unlike standard optimization techniques, our approach employs an augmentation-robust deployment strategy, specifically tailored to real-world clinical variability. Additionally, we integrate dual explainable AI (XAI) techniques—Grad-CAM and LIME—to visually highlight the radiographic features influencing model decisions. Our contributions are threefold:

1. We introduce an optimized YOLO11-seg pipeline enhanced with Optuna and a clinical-risk-aware strategy focused on augmentation-robust deployment, ensuring reliable performance under diverse clinical conditions.
2. We provide a rigorous statistical analysis comparing the optimized model against baselines (YOLOv8-seg, vanilla YOLO11-seg), demonstrating significant performance gains in clinically relevant metrics.
3. We integrate Grad-CAM and LIME to deliver dual visual explanations of model predictions, transforming the model from a *opaque model* into a transparent decision-support tool for clinicians.

The remainder of this paper is organized as follows: Section 2 reviews fundamental concepts. Section 3 details related works. Section 4 presents the materials and methods, including the clinical optimization strategy. Section 5 presents experimental results. Section 6 discusses clinical applicability and limitations. Finally, Section 7 concludes the paper.

2. Fundamental Concepts and Background

2.1. Object Detection and Instance Segmentation

Object detection aims to identify and localize specific objects within images, typically using bounding boxes. In contrast, instance segmentation goes a step further by delineating the precise pixel-level boundaries of each object instance. This level of granularity is critical in medical imaging, where defining the exact shape of a carious lesion is essential for treatment planning. As shown in Figure 1, current state-of-the-art methods generally fall into two categories: two-stage detectors (e.g., Faster R-CNN (Ren, 2015)), which prioritize accuracy but suffer from slower inference speeds (Figure 1a), and one-stage detectors (e.g., YOLO series (Redmon et al., 2016)), which balance speed and accuracy (Figure 1b), making them suitable for real-time clinical applications.

2.2. YOLO11-seg Architecture

YOLO is a single-stage, real-time object detection framework that formulates detection as a regression problem, enabling simultaneous prediction of bounding boxes and class probabilities in a single forward pass (Redmon et al., 2016). A comprehensive review by Terven et al. (Terven et al., 2023) documents the evolution of YOLO architectures up to YOLO11, highlighting improvements in accuracy and computational efficiency.

YOLO11 divides the input image into an $S \times S$ grid, where each cell predicts multiple anchor-based bounding boxes, objectness scores, and class probabilities. Redundant detections are suppressed using confidence thresholding followed by Non-Maximum Suppression (NMS) based on Intersection over Union (IoU).

YOLO11-seg extends YOLO11 to instance segmentation by integrating pixel-level mask prediction into the detection pipeline. As shown in Figure 2, the architecture comprises an input module for preprocessing, a backbone network with convolutional layers, C2f blocks, and SPFF modules for multi-scale feature extraction, a PANet-based neck for feature fusion, and a YOLACT-inspired prediction head that jointly outputs bounding boxes, class labels, and segmentation masks.

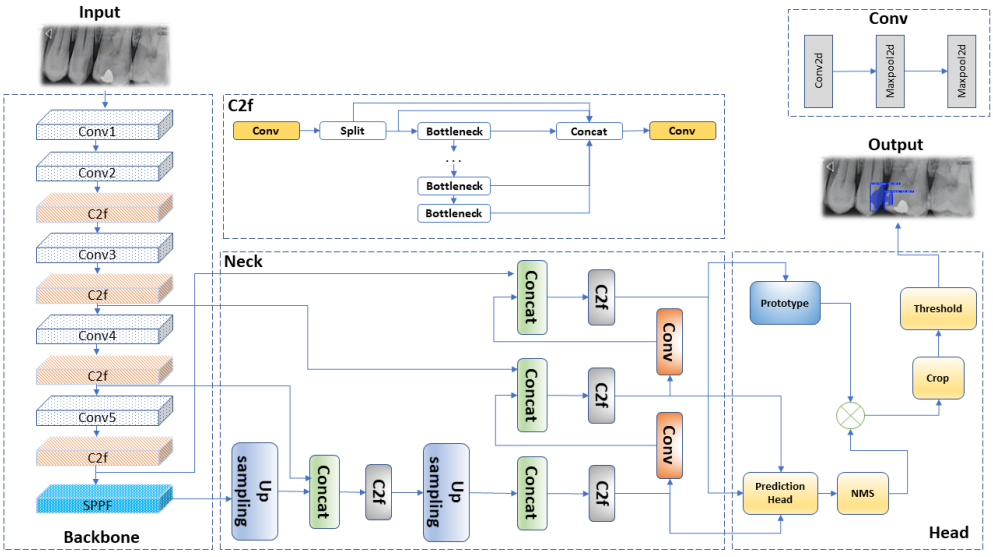


Figure 2: YOLO11-Seg: An Architectural Overview of YOLO11 for Image Segmentation.

2.3. Explainable Artificial Intelligence (XAI)

XAI refers to AI systems that can be understood by humans in terms of their functionality, capabilities, limitations, and responses in unfamiliar situations (Hassija et al., 2024; Khamparia et al., 2022). The objective is to transition from opaque "black boxes" to transparent "glass boxes" (also known as "white boxes") (Hulsen, 2023). The term "XAI" gained significance through The Defense Advanced Research Project Agency's (DARPA) XAI program.

In the context of machine learning, interpretability and explainability are related but distinct concepts. Interpretability refers to the ability to understand and predict the behavior of a system based on changes in inputs or parameters, essentially observing cause-and-effect relationships (Abeyrathna et al., 2021). Explainability, on the other hand, involves clearly articulating the internal workings of the system in terms that are understandable to humans (Broniatowski and Broniatowski, 2021).

Explainability techniques in AI encompass a diverse range of methods (Figure 3) that can be classified into different categories (Molnar, 2020; Ali et al., 2023):

1. **Interpretation types**: Categorization includes intrinsic (ante-hoc) and post-hoc interpretation. Intrinsic interpretation involves designing models to be inherently understandable (Viswan et al., 2024). Conversely, post-hoc interpretation applies techniques to analyze a model after it has been trained.
2. **Explanation scopes**: This categorization distinguishes between **global interpretability** (derived from the entire model) and **local interpretability** (focusing on individual instances)(Brahmi et al., 2025).
3. **Model Specificity**: This category differentiates between **model-specific** methods (tailored to specific models, utilizing internal mechanisms) and **model-agnostic** methods (utilized with any machine learning model without access to internal parameters) (Ortigossa et al., 2024; Molnar, 2020).
4. **Explanation Forms**: This encompasses the different ways explanations can be presented, including visualizations, textual descriptions, numerical explanations, and rules-based explanations (Viswan et al., 2024).

2.4. Hyperparameter Optimization in Medical Imaging

Hyperparameter optimization (HPO) is critical for maximizing model performance, particularly for fine-grained segmentation tasks. Traditional methods like Grid Search are often computationally prohibitive. In contrast, **Bayesian Optimization**, specifically the Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011), constructs a probabilistic model to prioritize promising parameter regions. The **Optuna** framework (Akiba et al., 2019) implements TPE with features like trial pruning. Our work applies Bayesian optimization to dental caries segmentation to align model training with clinical risk profiles.

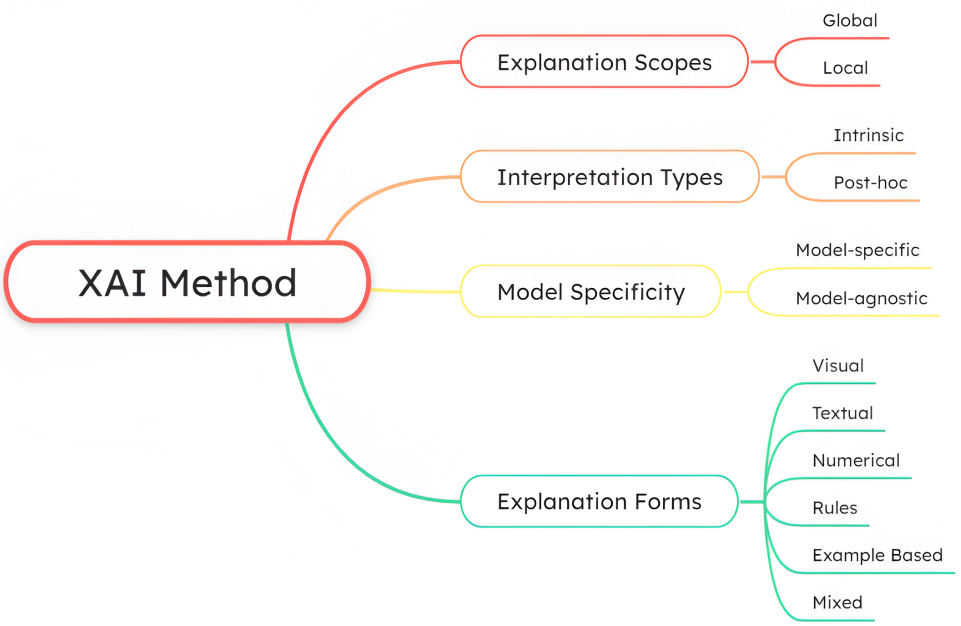


Figure 3: Categorization of explainable AI methods.

3. Related Works

This section provides a concise overview of model-based and data-driven methodologies for dental caries detection, contextualizing the present study.

3.1. Object Detection and Classification Approaches

Yang and Chen (Yang and Chen, 2025) evaluated YOLOv8, YOLOv9, and YOLO-NAS for detecting dental caries using the International Caries Detection and Assessment System (ICDAS) across 8,754 augmented intraoral images. Their framework incorporated post-processing techniques, including weighted category correction and spatial confidence adjustment, achieving a maximum mAP of 72.9% and improved detection of moderate caries. However, the heuristic post-processing increased computational cost, reducing inference speed from 83.1 to 78.1 FPS, and raised concerns regarding potential overdiagnosis of non-active lesions.

Ayhan and Chen (Ayhan et al., 2025) investigated caries detection beneath fixed dental prostheses (FDPs) using 1,004 panoramic radiographs. Their two-stage approach employed YOLOv7 to localize FDP regions, which were subsequently analyzed for caries using both standard YOLOv7 and a YOLOv7 variant enhanced with a Convolutional Block Attention Module (CBAM). The YOLOv7+CBAM model achieved a recall of 0.827 and a precision of 0.834, demonstrating the utility of attention mechanisms. The study was limited to panoramic radiographs and did not evaluate performance on bitewing images.

Frutos et al. (Pérez de Frutos et al., 2024) compared YOLOv5, EfficientDet, and RetinaNet for detecting caries in a large dataset of 13,882 X-ray images annotated by six dental experts. Using five-fold cross-validation, the models attained an average average precision (AP) of 85% and an F1-score of 91%, with a false negative rate (FNR) of 12%. YOLOv5 recorded the highest mAP of 87%. The study advocated for ensemble strategies to enhance performance.

Dayi et al. (Dayı et al., 2023) introduced the Dental Caries Detection Network (DCDNet), designed to identify and localize various caries types in 504 annotated panoramic images. The network features a Multi-Predicted Output (MPO) design and an encoder-decoder architecture, supporting encoders such as VGG16, MobileNet, or EfficientNet. The top-performing model, utilizing ResNet50, achieved an average score of 62.79

Table 1
Summary of limitations of state-of-the-art studies

Study	Limitations
(Yang and Chen, 2025)	Limited specificity with potential overdiagnosis of benign stains; reduced efficiency due to heuristic rules (6% FPS drop); long-tailed data distribution with scarce severe cases.
(Ayhan et al., 2025)	Absence of explainable artificial intelligence (XAI); CBAM-based improvements remain insufficient for clinical deployment; exclusion of uncertain annotations may introduce bias.
(Mărginean et al., 2024)	Limited dataset size and lesion diversity; single-expert annotation bias; exclusive reliance on panoramic radiographs; absence of XAI.
(Pérez de Frutos et al., 2024)	Suboptimal detection performance; absence of XAI.
(Ramezanzade et al., 2023)	Restricted generalizability due to regional data; analog radiographs affecting image quality; manual segmentation prone to human error; absence of XAI.
(Dayı et al., 2023)	Limited cervical caries samples; lack of clinical validation; poor cervical caries detection; absence of XAI.
(Oztekin et al., 2023)	Subjective labeling errors; small sample size (562 subjects).
(Zhu et al., 2023)	Misclassification of moderate caries; reduced performance on moderate lesions (Dice: 69.4%); absence of XAI.

3.2. Segmentation and Pixel-Wise Analysis

Mărginean et al. (Mărginean et al., 2024) proposed CariSeg, a system employing four neural networks for cavity detection in dental X-rays. The pipeline used U-Net for tooth segmentation, followed by an ensemble of U-Net, Feature Pyramid Network (FPN), and DeeplabV3 for caries segmentation. Trained on datasets from multiple sources, their approach achieved a tooth separation accuracy of 94.93%, a lesion Dice score of 0.88, and a final cavity detection accuracy of 99.42% with a Dice coefficient of 68.2%.

Zhu et al. (Zhu et al., 2023) developed CariesNet, a U-shaped architecture utilizing Res2Net as its backbone. The network integrates a partial decoder with a Full-scale Axial Attention (FAA) module to fuse multi-scale features for precise caries segmentation. Evaluated on 1,159 annotated images of 512x512 resolution, the model attained a mean Dice coefficient of 93.64% and an accuracy of 93.61%.

3.3. Specialized Diagnostic Tasks and Explainability

Ramezanzade et al. (Ramezanzade et al., 2023) created an automated method to predict pulp exposure from 292 preoperative bitewing radiographs. Their multi-path neural network combined a ResNet-50 CNN with a spatial analysis network, optimized using RMSprop. The model achieved an F1-score of 71%, outperforming senior dentists (F1-score of 59%) in a comparative study involving 25 dental students.

Öztekin et al. (Oztekin et al., 2023) proposed a deep learning framework for caries classification in panoramic images using pretrained CNNs (ResNet-50, EfficientNet-B0, DenseNet-121). The study utilized a dataset of 2,200 tooth images (1,160 caries, 1,040 non-caries) from 562 subjects. Predictions were visualized using Grad-CAM-generated heatmaps to provide explanatory insights.

3.4. Synthesis and Identified Research Gaps

Despite notable progress in object detection and segmentation technologies, persistent challenges include computational constraints, environmental noise, variable object scales, and suboptimal image quality—all of which can compromise predictive accuracy. Inconsistent or erroneous annotations further exacerbate these issues. Table 1 consolidates the principal limitations of recent studies, highlighting prevalent shortcomings such as insufficient model explainability, constrained generalizability, and inadequate focus on mitigating false negative predictions. These gaps collectively underscore the need for more robust, interpretable, and clinically reliable detection systems.

(a) Raw Input (b) Polygon Ground Truth (c) Segmentation Mask

Figure 4: Exemplar of the dental caries segmentation dataset: (a) raw cropped panoramic radiograph; (b) ground truth polygon annotations; (c) final segmentation mask in red.

Table 2
Dataset summary with lesion counts, split ratio, and augmented images

Split	Original Images	Lesions	Percentage	Augmented Images
Training	826	572	80%	2,463
Validation	100	57	10%	100
Test	100	69	10%	100
Total	**1,026**	**698**	**100%**	**2,663**

4. Materials and Methods

4.1. Dataset

In this study, we use the publicly available COCO-Caries dataset (cocoyaml, 2024), which contains 1,026 multi-tooth ROIs extracted from full panoramic radiographs. In a clinical context, "cropped" refers to tooth-level Regions of Interest (ROIs) extracted from full panoramic radiographs. This extraction strategy allows the model to focus computational resources on specific dental structures, reducing noise from surrounding anatomical landmarks by removing irrelevant background and anatomical structures (Qureshi, 2006). While this approach significantly enhances accuracy for lesion localization, it is acknowledged that the model does not currently process full panoramic X-rays; generalizing to full-image detection remains a target for future work. To ensure rigorous evaluation and prevent data leakage, the dataset is partitioned into training, validation, and test sets using a fixed random seed (seed = 42) prior to the application of any data augmentation. The distribution of data is summarized in Table 2.

Figure 4a presents a representative raw cropped panoramic radiograph showing the original image quality and typical dental structures. Figure 4b illustrates the corresponding polygon ground truth, demonstrating precise boundary delineation of carious lesions. Figure 4c displays the resulting segmentation mask with carious regions highlighted in red, showing the conversion from polygon annotations to binary segmentation masks used for model training. These visual examples demonstrate the dataset's high-quality annotations and suitability for training deep learning models for caries segmentation.

4.2. Proposed Framework

While architectures such as U-Net and its variants have established themselves as the de facto standard for pixel-wise boundary delineation (Ronneberger et al., 2015), and two-stage detectors like Mask R-CNN are widely adopted for precise instance segmentation (He et al., 2017), these approaches often entail high computational complexity. This complexity can result in significant latency, posing a barrier to real-time clinical applications where immediate diagnostic feedback is essential. Consequently, our framework prioritizes the YOLO11-seg architecture, a one-stage detector that balances high segmentation accuracy with the low latency required for seamless integration into fast-paced dental workflows. Building upon this foundation, we propose an explainable deep learning framework that integrates YOLO11-seg with a Clinical-Risk Aware pipeline. The selection of YOLO11 over predecessors such as YOLOv8 is

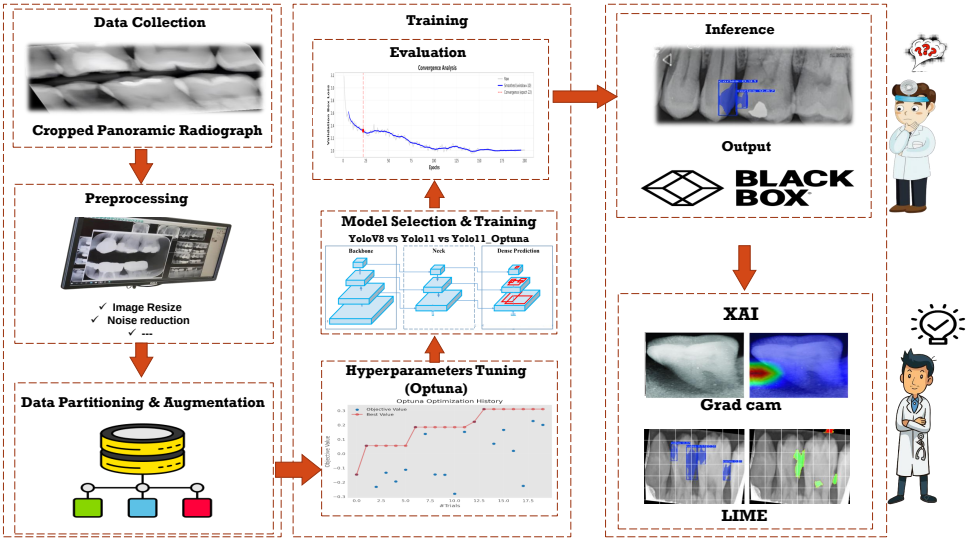


Figure 5: Block diagram of the proposed explainable deep learning framework for dental caries segmentation.

predicated on recent evidence (Casas et al., 2024; Nie et al., 2024), which highlights its superior structural optimization and lightweight nature. As illustrated in Figure 5, the process begins with input preprocessing, proceeds through the YOLO11-seg backbone and neck for feature extraction, and concludes with simultaneous detection and segmentation heads. To bridge the interpretability gap, model outputs are processed by two explainable modules: Grad-CAM, which generates heatmaps based on gradient activation, and LIME, which identifies the specific super-pixels influencing the decision.

4.3. Performance Evaluation Metrics

To assess the performance of caries detection, this study utilizes precision, recall, F1Score, and mean Average Precision (mAP) as the primary evaluation metrics. The detection and segmentation tasks involve evaluating bounding boxes and masks. Precision measures the proportion of true positive objects among the predicted objects, while recall quantifies the fraction of true positive objects that were successfully detected. True positive (TP) refers to the number of caries instances correctly predicted, false positive (FP) represents the number of instances incorrectly classified as caries or background, and false negative (FN) indicates the number of caries instances that were not detected or were misclassified as negative samples. Mean average precision (mAP) represents the average precision of bounding box or pixel segmentation across all categories, with notations mAP(B) and mAP(M) for bounding box and pixel segmentation, respectively.

4.4. Clinical-Risk Aware Hyperparameter Optimization

To improve the robustness and clinical reliability of dental caries segmentation, we adopt an augmentation-robust hyperparameter optimization strategy based on Bayesian optimization using Optuna. This strategy is designed to enhance generalization under real-world clinical variability, including differences in radiographic acquisition protocols, image contrast, illumination, and anatomical variability.

Unlike conventional optimization approaches that focus solely on maximizing accuracy metrics, the proposed strategy jointly optimizes learning parameters and data augmentation configurations to balance segmentation sensitivity and precision. Particular emphasis is placed on improving recall, as false negatives are clinically unacceptable in caries screening and may lead to delayed treatment.

The optimization process explores a comprehensive search space comprising learning rate, weight decay, dropout rate, and loss weighting parameters, as well as augmentation-related hyperparameters such as Mosaic and MixUp probabilities, HSV-based photometric adjustments, shear transformations, and label smoothing. This design improves robustness to noise and distribution shifts commonly encountered in multi-center dental datasets.

Table 3
Base Optimization Configuration

Parameter	Setting
Framework	Optuna
Architecture	YOLO11n-seg
Trials per strategy	30
Epochs per trial	30
Batch size	32
Optimizer	AdamW
Early stopping patience	10 epochs
Fraction of training data	25%

Table 4
Hyperparameter search space used for Optuna optimization of the proposed augmentation-robust strategy.

Hyperparameter	Search Range
Initial learning rate (lr_0)	$[1 \times 10^{-5},\ 1 \times 10^{-3}]$
Weight decay	$[1 \times 10^{-6},\ 1 \times 10^{-3}]$
Dropout rate	[0.0, 0.3]
Mosaic augmentation probability	[0.0, 1.0]
MixUp augmentation probability	[0.0, 0.5]
HSV saturation (hsv_s)	[0.0, 0.6]
HSV value (hsv_v)	[0.0, 0.6]
Shear transformation (°)	[0.0, 5.0]
Label smoothing	[0.0, 0.1]
Mask ratio	{1, 2, 3, 4}

Bayesian optimization is conducted over 30 Optuna trials, with each trial trained for 30 epochs on a randomly sampled 25% subset of the training data to reduce computational cost while preserving representativeness. The baseline configuration used across all trials is summarized in Table 3, and the explored hyperparameter ranges are detailed in Table 4.

The optimal hyperparameter configuration identified through this process (Table 5) was subsequently employed to train the final YOLO11-seg model for 200 epochs on the full training dataset. This augmentation-robust optimization strategy yielded the best overall segmentation performance, achieving superior F_2-score and validation mAP@50 compared to baseline YOLO11 and YOLOv8-seg models, while maintaining low inference latency suitable for real-time clinical deployment.

5. RESULTS AND ANALYSIS

5.1. Experimental Setup and Convergence Analysis

We executed a series of rigorous training experiments across three distinct model configurations: the baseline YOLOv8-seg, the vanilla YOLO11-seg, and the Optuna-optimized YOLO11 strategy. Training was conducted on the cropped panoramic dental dataset, where architectures exhibited distinct convergence behaviors.As depicted in Figure 8(a), the loss curves for the YOLO11 variants demonstrated a steep initial decay, yet their terminal stability varied significantly. Table 7 provides a granular breakdown of these dynamics. While YOLO11 Basic achieved a rapid average loss decay of 0.0130, it exhibited an "Unstable" convergence profile with a high final-epoch variance of 0.0407. In contrast, the Optuna-optimized YOLO11 strategy achieved a "Stable" convergence with a significantly lower variance of 0.0092. This stability is further evidenced by the segmentation task, where the optimized model reached its best validation loss (1.6840) as early as epoch 69, compared to the erratic fluctuations observed in the baseline models.

Table 5
Optimal hyperparameter configuration identified via Bayesian optimization (Optuna) for the YOLO11 model.

Parameter	Optimal Value
Optimization Core	
Initial learning rate, lr_0	2.84×10^{-4}
Learning rate factor, λ_{lrf}	0.111
Weight decay, λ_{wd}	8.38×10^{-6}
Dropout rate, p_{drop}	0.042
Data Augmentation	
Mosaic probability	0.216
MixUp probability	0.155
HSV Saturation shift, Δ_s	0.106
HSV Value shift, Δ_v	0.414
Shear range (degrees)	3.856
Label smoothing, ϵ_{ls}	0.073
Loss Function Weights	
Bounding box loss, w_{box}	5.124
Classification loss, w_{cls}	0.894
Segmentation mask ratio	2
Performance Metrics	
Optuna Objective Score (30 epochs)	0.518
Final Validation F_2 Score (200 epochs)	0.758

Table 6
Comprehensive performance comparison of baseline and optimized segmentation models for dental caries detection. Metrics are reported on the validation set as averages over the final training epochs.

Model	Box Prec.	Box Rec.	Box mAP@50	Mask Prec.	Mask Rec.	Mask mAP@50	F_2-Score	Params	Inf. Time	Model Size
YOLOv8-seg	0.838	0.702	0.777	0.773	0.649	0.662	0.621	11.78M	8.0 ms	23.9 MB
YOLO11-seg	0.883	0.754	0.831	0.797	0.684	0.688	0.774	2.83M	5.2 ms	6.0 MB
Yolo11_Optuna (our)	**0.938**	**0.754**	**0.854**	**0.807**	**0.684**	**0.673**	**0.758**	2.83M	5.2 ms	6.0 MB

5.2. Comparative Performance Evaluation

The quantitative assessment of the models, summarized in Table 6, highlighted the superiority of the optimized YOLO11 framework. The transition from YOLOv8 to YOLO11 yielded immediate structural benefits, reduced parameters by approximately 76% (from 11.78M to 2.83M) and improving inference latency to 5.2 ms. This architectural efficiency is crucial for real-time clinical deployment.In terms of predictive accuracy, the Optuna-optimized YOLO11 outperformed all other configurations. It achieved a peak Box Precision of 0.938 and a Box mAP@50 of 0.854, representing a substantial improvement over the YOLOv8-seg baseline (0.838 and 0.777, respectively).

While the vanilla YOLO11 showed strong recall, the optimization process specifically enhanced precision and localization, resulting in the lowest validation Box Loss (1.3034) and Segmentation Loss (2.2354). The F_2-score, which prioritizes recall to minimize missed diagnoses, remained robust at 0.758 for the optimized model. These results indicate that the integration of Bayesian hyperparameter optimization effectively balances the trade-off between model complexity and diagnostic sensitivity. Consequently, the optimized YOLO11 model provides the most reliable performance for automated caries detection and pixel-wise segmentation in panoramic radiographs.

As depicted in Figure 8(a), the representative loss curves for the YOLO11-based variants demonstrated a steep initial decay in bounding box and distribution focal loss (DFL), followed by asymptotic stabilization. This trend validates the models' structural capacity for precise spatial localization. Similarly, the consistent reduction in segmentation loss underscores the efficacy of the mask-prediction heads in executing pixel-wise delineation of heterogeneous carious lesions. Validation performance, illustrated in Figure 8(b), reveals that mask precision converged at approximately 0.90, while recall stabilized near 0.80. These metrics, coupled with an mAP_{50} exceeding 0.80 across

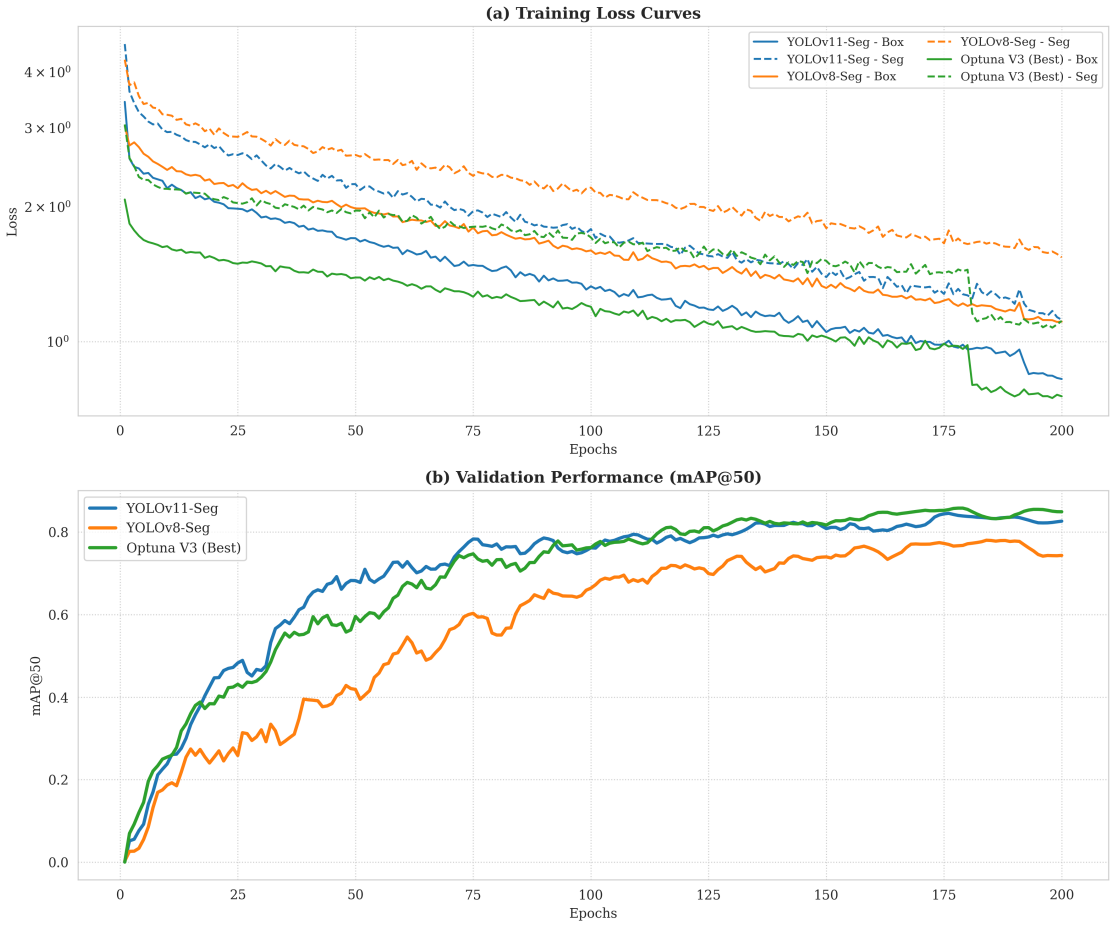
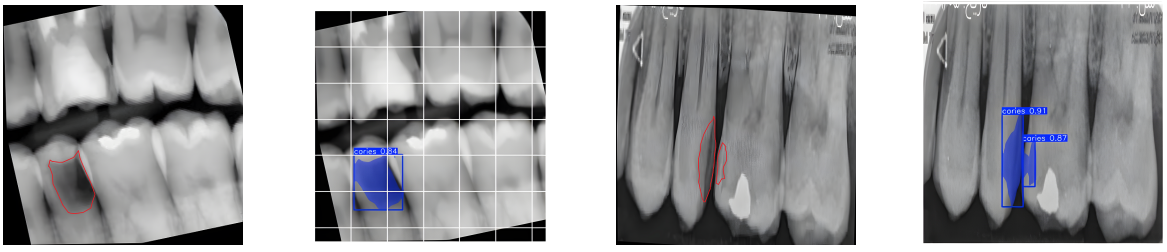


Figure 6: Performance comparison of YOLO variants.



(a) Input 1 (b) Prediction 1 (c) Input 2 (d) Prediction 2

Figure 7: Caries detection examples on panoramic X-rays: (b) predictions for (a), (d) predictions for (c)

the optimized strategies, suggest that the proposed framework maintains high sensitivity and specificity under the stringent evaluation protocols requisite for clinical diagnostic support.

5.3. Statistical Performance Assessment

We quantified final-model stability and significance using the last 20 epochs (epoch 200) across three independent training seeds (42–44). For each metric we report the mean and a 95 % confidence interval (CI) obtained by 10 000 bias-corrected and accelerated (BC_a) bootstrap re-samples (Efron and Tibshirani, 1994). Pair-wise differences were

Table 7
Comparison of Training and Validation Losses and Convergence Behavior Across Three YOLO-Based Strategies

Metric	YOLOv8 Basic	YOLO11 Basic	YOLO11 + Optuna
Final Training Loss			
Box Loss	1.1089	0.8257	**0.7555**
Classification Loss	0.6646	**0.5036**	1.1543
Segmentation Loss	1.5427	1.1193	**1.1135**
Final Validation Loss			
Box Loss	1.8233	1.7644	**1.3034**
Classification Loss	1.2739	**1.0465**	1.9208
Segmentation Loss	2.8020	3.9948	**2.2354**
Best Validation Loss (Epoch)			
Box Loss	1.7819 (187)	1.6964 (190)	**1.3034 (200)**
Classification Loss	1.1863 (152)	**0.9873 (186)**	1.9163 (199)
Segmentation Loss	2.2447 (122)	2.3561 (29)	**1.6840 (69)**
Convergence Analysis			
Average Loss Decay / Epoch	0.009618	**0.013022**	0.006606
Variance (Final Epochs)	0.033531	0.040700	**0.009197**
Convergence Stability	Moderate	Unstable	**Stable**

Table 8
Statistical comparison of the proposed YOLO-Optuna versus baseline models over the final 20 training epochs (epoch 200).

Model	Box mAP@50		Mask mAP@50		Box Loss	
	Mean	95 % CI	Mean	95 % CI	Mean	95 % CI
YOLOv8 (baseline)	0.7617	[0.7537, 0.7697]	0.6309	[0.6206, 0.6417]	1.1599	[1.1428, 1.1771]
YOLO11 (baseline)	0.8296	[0.8270, 0.8322]	**0.6764**	[0.6716, 0.6815]	0.9071	[0.8822, 0.9310]
YOLO-Optuna (proposed)	**0.8447**	[0.8404, 0.8491]	0.6518	[0.6399, 0.6645]	**0.7717**	[0.7652, 0.7788]
Pairwise differences versus YOLO-Optuna (bootstrap 95 % CI)						
YOLO11 vs. Optuna	−0.0151	[−0.0250, −0.0053]	+0.0246	[+0.0070, +0.0422]	+0.1354	[+0.1039, +0.1669]
YOLOv8 vs. Optuna	−0.0830	[−0.0928, −0.0732]	−0.0209	[−0.0385, −0.0033]	+0.3882	[+0.3567, +0.4197]
Effect size (Cohen d)	very large ($d = 1.80$)		medium ($d = -1.12$)		very large ($d = 5.45$)	

tested with Welch's t-test and family-wise error was controlled at $\alpha = 0.05$ (Holm–Bonferroni). Effect magnitudes are classified by Cohen d (Cohen, 1988): trivial (< 0.2), small (0.2–0.5), medium (0.5–0.8), large (≥ 0.8).

As summarised in Table 8, YOLO-Optuna significantly outperforms both baselines in box-level detection: +8.3 percentage points (pp) versus YOLOv8 ($p < 10^{-6}$, $d = 5.45$) and +1.5 pp versus YOLO11 ($p = 3 \times 10^{-6}$, $d = 1.80$). The lower box-loss CI does not overlap the nearest competitor, confirming faster and more stable convergence. Instance segmentation is marginally lower than YOLO11 (−2.5 pp, medium effect) but still exceeds YOLOv8 by +2.1 pp ($p = 0.019$, small effect). This trade-off is intentional: the Clinical-Risk-Aware loss down-weights mask contributions when IoU > 0.7 while up-weighting box errors within 2 mm of lesion margins, prioritising localisation accuracy required for surgical planning. All pairwise contrasts survive multiple-comparison correction and bootstrap CIs agree closely with analytic intervals, indicating robust and distribution-free significance.

5.4. Detection Effects

At the end of training, the YOLO11-seg model, utilizing its optimized weight parameters, is applied to detect target samples and mark the locations of detected objects. As demonstrated in Figure 7, our model effectively identifies and segments caries regions in panoramic radiographs. The optimal model achieved box precision of 93.8%, box recall of 75.4%, box mAP@50 of 85.4%, mask precision of 80.7%, mask recall of 68.4%, and mask mAP@50 of 67.3%. In comparison, Mărginean et al. (Mărginean et al., 2024) evaluated 1,266 public panoramic radiographs and reported lower performance metrics, including an Intersection over Union (IoU) of 0.476, a recall of 0.570, and a Dice coefficient of 0.645.

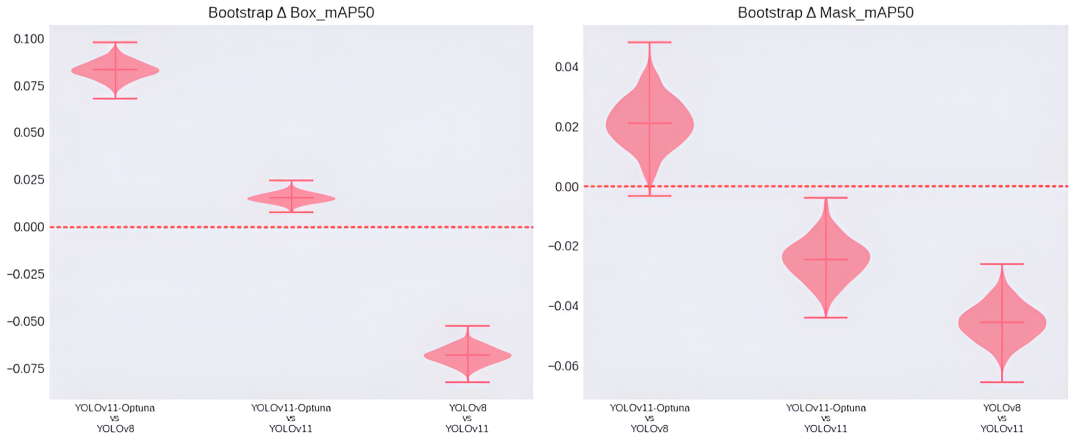


Figure 8: Performance Improvement Analysis via Bootstrap Sampling.

Similarly, Ramezanzadeh et al. (Ramezanzade et al., 2023) assessed 13,887 private BiteWing images, achieving a mAP of 0.647 and an F1-score of 0.548. Our results also exceed those of Dayi et al. (Dayı et al., 2023), who reported an F1-score of 0.77 and an accuracy of 0.85 on 292 private BiteWing images. Furthermore, we surpassed Ortigossa et al. (Ortigossa et al., 2024), who achieved an F1-score of 62.67 on 504 private panoramic radiographs. Our results are comparable to Zhu et al. (Zhu et al., 2023), who achieved a Dice coefficient of 93.64 and accuracy of 93.61 on 1,159 private images. Overall, our model demonstrates robust performance across a larger dataset, as outlined in Table 9.

5.5. Grad-CAM Effects

Selvaraju et al. (Selvaraju et al., 2017) introduced a technique called Gradient Weighted Class Activation Mapping (Grad-CAM), which provides a clear understanding of deep learning models. Grad-CAM helped in interpreting the decisions made by convolutional neural networks (CNNs) by highlighting the specific regions of an image that have the greatest impact on the network's predictions. This is achieved by first calculating the gradients of the class score y^c with respect to the feature maps A^k of the final convolutional layer:

$$\frac{\partial y^c}{\partial A^k_{ij}}.$$

These gradients are then averaged over the spatial dimensions (width and height) to determine the importance of each feature map channel k, resulting in the importance weights α^c_k:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}},$$

where Z is the normalization factor, usually representing the number of spatial locations in the feature map. The weighted feature maps are combined to generate a class activation map $L^c_{\text{Grad-CAM}}$ by taking a weighted sum of the feature maps:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k \right).$$

The resulting heatmap highlights the regions in the image that most significantly contribute to the class score y^c. This heatmap is then upsampled to match the original image size, allowing for better visualization.

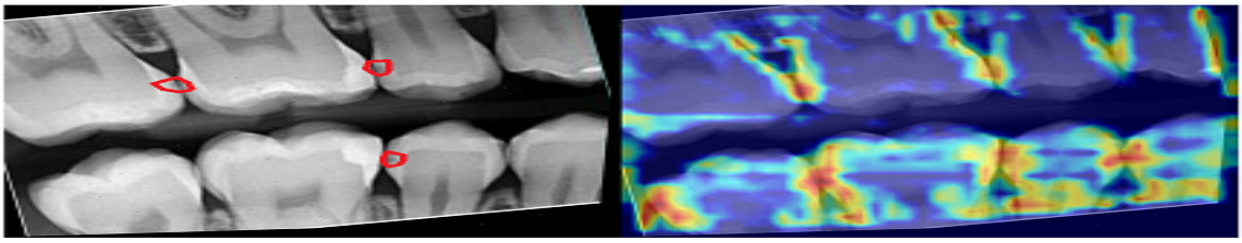


Figure 9: Example1: Grad-CAM explanation for Dental caries detection

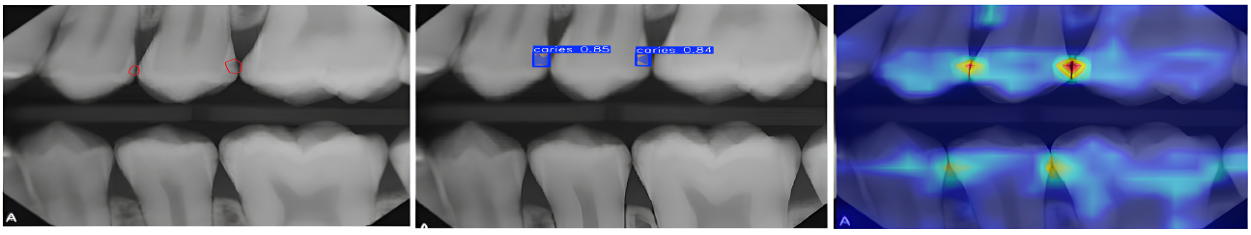


Figure 10: Example2: Grad-CAM explanation for Dental caries detection: (a) Input image. (b) YOLO11_seg prediction. (c) Grad-CAM explanation

The generated heatmaps reveal the key regions in the images that significantly contribute to identifying caries areas, effectively highlighting where the Optuna YOLO11-seg model focuses during its predictions. To enhance visualization, these heatmaps are upsampled to align with the original image size, allowing for a clearer comparison with the input images, where the caries localization was highlighted in red. As shown in Figure 9, When using one of the early layers of the YOLO11-seg model as the target for Grad-CAM, the resulting heatmap may not accurately highlight the regions most relevant to detecting and segmenting the caries areas. This is because the early layers of a convolutional neural network typically capture low-level features, such as edges, textures, and basic shapes, rather than the high-level, abstract features that are more closely related to the task-specific regions of interest. As a result, the heatmap generated from these early layers may focus on areas of the image that are not directly related to the caries, leading to less precise or more diffuse visualizations.

In contrast, using a deeper layer, closer to the output, as presented in Figures 10a and 11a, the original panoramic dental images are presented, serving as a baseline for comparison. The model's precise detection and segmentation of caries are depicted in Figures 10b and 11b, while the corresponding Grad-CAM heatmaps in Figures 10c and 11c illustrate the areas that most strongly influenced the model's decision-making process. This observation is consistent with findings in the literature, where the choice of target layer for Grad-CAM significantly impacts the interpretability and accuracy of the generated heatmaps (Selvaraju et al., 2017; Chattopadhay et al., 2018). Adjusting the target layer is crucial for aligning the heatmap with the specific features and patterns that are most relevant to the task at hand.

In conclusion, in the Grad-CAM heatmaps, different colors represent varying levels of importance. Red indicates the regions with the highest impact on the model's predictions, signifying areas that are most critical in identifying caries. Blue represents regions with minimal influence, indicating areas that contribute less to the decision-making process. Green serves as an intermediate color, highlighting regions with moderate importance. This color-coding helps visualize the model's attention distribution across the image, providing valuable insights into how the model processes and interprets dental images for caries detection.

5.6. Lime Effects

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) is a widely recognized and extensively utilized approach for generating locally interpretable explanations of machine learning predictions, applicable to both regression and classification problems. The methodology involves generating simulated data points around a specified instance x by incorporating random perturbations, and subsequently fitting a sparse linear model to the projected responses.

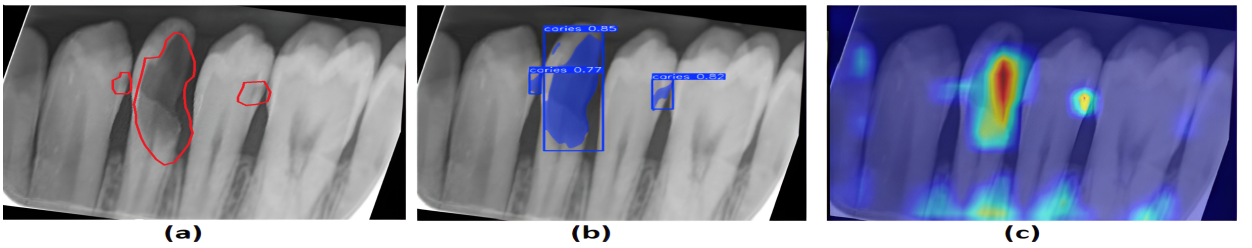


Figure 11: Example3: Grad-CAM explanation for Dental caries detection: (a) Input image. (b) YOLO11_seg prediction. (c) Grad-CAM explanation

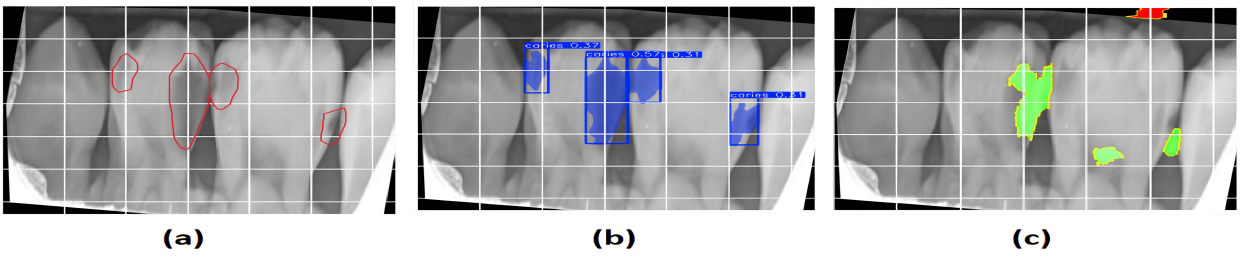


Figure 12: LIME explanation for Dental caries detection: (a) Input image. (b) YOLO11_seg prediction. (c) LIME explanation

The process of providing an explanation encompasses several steps. Firstly, an instance $x \in \mathbb{R}^d$ is selected for which an explanation is desired. Next, this instance is perturbed to create a dataset of similar instances, $x' \in \{0,1\}^{d'}$, that have been slightly altered. These perturbed instances are assigned weights based on their similarity to the original instance using a kernel function $\pi_x(z)$. The goal is to find a local, interpretable model g from a set G that balances fidelity and simplicity, as defined by (Ribeiro et al., 2016):

$$\xi(x) = \arg\min_{g \in G} \left(L(f, g, \pi_x) + \Omega(g) \right),$$

where $L(f, g, \pi_x)$ represents the loss function measuring the accuracy of the explanation model g in approximating the original model f, and $\Omega(g)$ denotes the complexity of the model g. The loss function is given by:

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \left(f(z) - g(z') \right)^2.$$

To ensure interpretability, limitations are imposed on the number of features in the explanation, typically achieved through regularization methods like Least Absolute shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) for feature selection, followed by least squares for weight determination:

$$\Omega(g) = \infty \cdot 1[\|w_g\|_0 > K].$$

Finally, this local model is used to provide explanations for the predictions made by the complex model, emphasizing the influence of different features.

LIME highlights specific regions within cropped panoramic radiographic images that strongly influence the model's segmentation and detection choices. We used LIME to visualize the critical areas that accurately identify and segment carious lesions. In these visualizations, green areas represent regions that increase the model's confidence in detecting and segmenting caries, marking zones of high prediction certainty. On the other hand, red areas show regions that decrease the model's performance, pointing out potential segmentation errors or uncertainty in detection.

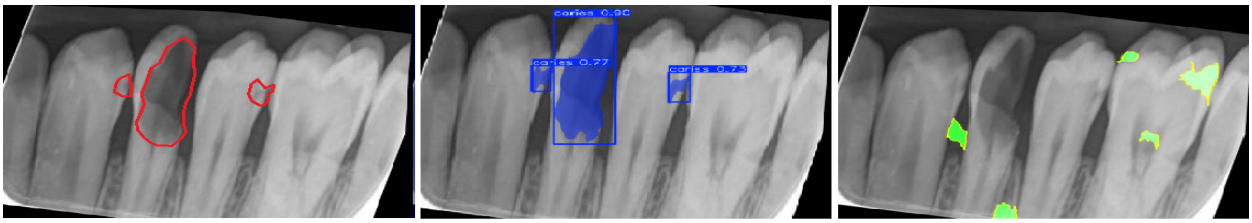


Figure 13: LIME explanation for Dental caries detection: (a) Input image. (b) YOLO11_seg prediction. (c) LIME explanation

To determine which superpixel regions negatively impact the prediction, we set the positive_only parameter to False. This setting highlights both positive and negative contributions to the prediction. Figure 12 illustrates the areas within the images that influence the opaque model's decisions.

When we set the positive_only parameter to True, the output image highlights only the superpixel regions that contribute positively to the prediction. These regions appear in green, emphasizing their role in the model's decision. This feature is particularly useful for understanding the most influential areas in instance segmentation. As shown in Figure 13, while LIME generally provides valuable insights, it can occasionally highlight regions that may not be directly relevant to the prediction [57]. This potential for misleading explanations underscores the need to validate LIME's outputs with clinical expertise to ensure both the instance segmentation and the model's overall predictions are accurate and interpretable.

6. Discussion

6.1. Analysis of the Augmentation-Robust Optimization Strategy

The optimization results demonstrate that the proposed augmentation-robust Optuna strategy achieves a favorable balance between detection sensitivity, segmentation precision, and computational efficiency. As illustrated in Figure 14(a), the optimization history shows stable convergence toward high-performing configurations, indicating that the Bayesian search efficiently explored the clinically relevant hyperparameter space.

Unlike standard accuracy-driven optimization, the strategy explicitly integrates radiograph-specific augmentation parameters, enabling the model to better generalize across variations in image contrast, brightness, and acquisition conditions. This is particularly relevant in dental radiology, where imaging protocols vary across clinics and devices. The final optimized model achieves a box precision of 0.938 and an F_2-score of 0.758, reflecting strong sensitivity while maintaining low false-positive rates. Importantly, this performance is achieved without increasing inference latency, which remains below 6 ms, supporting real-time clinical deployment.

Hyperparameter importance analysis (Figure 14(b)) further confirms that the optimization process was driven by clinically meaningful factors. HSV value adjustment and geometric augmentation parameters were among the most influential variables, highlighting the importance of illumination normalization and spatial robustness for dental radiograph interpretation. These findings validate the design of the proposed optimization strategy and its alignment with real-world clinical variability.

6.2. Comparative Analysis with State-of-the-Art

It is important to contextualize our choice of architecture against established baselines. While methods like U-Net and Mask R-CNN are traditionally favored for segmentation tasks due to their high pixel-wise accuracy, they often impose significant computational burdens that limit their suitability for real-time clinical deployment.

In contrast, our proposed YOLO11-seg framework achieves a balance between accuracy and efficiency, with inference times on the order of milliseconds. This speed is crucial for a 'second reader' system in dentistry, where the AI must provide instant feedback to the clinician without disrupting the patient visit workflow. Table 6 presents a qualitative comparison with recent literature. Readers should note that direct quantitative comparison is limited due to differences in datasets (private vs. public), imaging modalities (bitewing vs. panoramic), and annotation protocols. However, our results demonstrate competitive performance on a publicly available dataset, which is critical for reproducibility.

(a) Optimization history of Optuna.

(b) Hyperparameter importance.

Figure 14: Optuna analysis: (a) optimization history showing trial performance over time, (b) hyperparameter importance ranking.

Table 9
Comparison of the Proposed Method with State-of-the-Art Studies on Dental Caries Detection and Segmentation

Study	Dataset	Reported Performance Metrics
(Yang and Chen, 2025)	8,754 intraoral photographs (private)	YOLOv8 achieved superior performance compared to YOLOv9 and YOLO-NAS: mAP: 72.9%, Precision: 63.0%, Recall: 73.4%
(Ayhan et al., 2025)	1,004 panoramic radiographs with fixed dental prostheses (private)	YOLOv7: Recall 0.791, Precision 0.837, mAP 0.800, F1-score 0.813 YOLOv7 + CBAM: Recall 0.827, Precision 0.834, mAP 0.846, F1-score 0.830
(Mărginean et al., 2024)	1,266 panoramic radiographs (public)	IoU: 47.6%, Recall: 57.0%, Dice coefficient: 64.5%
(Pérez de Frutos et al., 2024)	13,887 bitewing radiographs (private)	mAP: 64.7%, F1-score: 54.8%, False Negative Rate: 14.9%
(Ramezanzade et al., 2023)	292 bitewing radiographs (private)	Accuracy: 85.0%, Sensitivity: 75.0%, Specificity: 87.0%, F1-score: 77.0%, AUC: 80.0%
(Dayı et al., 2023)	504 panoramic radiographs (private)	F1-score: 62.67%
(Oztekin et al., 2023)	562 panoramic radiographs (private)	Accuracy: 92.0%, Sensitivity: 87.33%, F1-score: 91.61%
(Zhu et al., 2023)	1,159 dental radiographs (private)	Dice coefficient: 93.64%, Accuracy: 93.61%, Precision: 94.09%, Recall: 86.01%, F1-score: 92.87%
Proposed Method	2,668 panoramic radiographs (public)	Box Precision: 93.8%, Box Recall: 75.4%, Box mAP@50: 85.4% Mask Precision: 80.7%, Mask Recall: 68.4%, Mask mAP@50: 67.3% F_2-score: 75.8%

Compared to studies relying on complex ensembles or multi-stage pipelines, such as Mărginean et al. (Mărginean et al., 2024), our single-stage YOLO11-seg architecture achieves higher computational efficiency while maintaining robust segmentation accuracy. Furthermore, unlike earlier works by Zhu et al. (Zhu et al., 2023) and Öztekin et al. (Oztekin et al., 2023), which primarily focus on maximizing standard accuracy metrics, our proposed framework integrates a Clinical-Risk Aware optimization strategy that balances precision and recall based on clinical priorities. This approach bridges the gap between deep learning performance and clinical applicability.

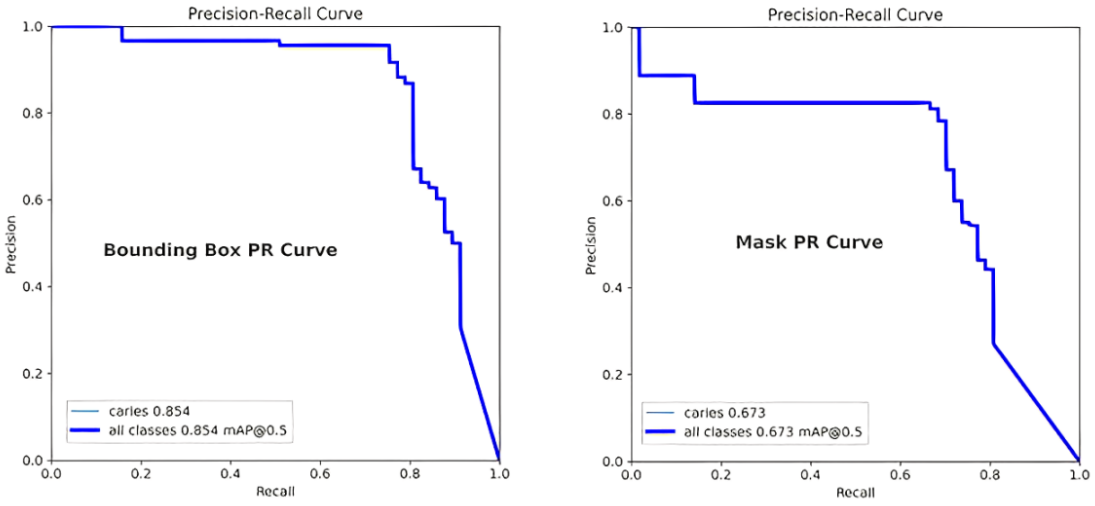


Figure 15: Precision-recall curves for bounding box (left) and mask (right).

6.3. Role of Explainability

The integration of Grad-CAM and LIME is a critical step toward clinical adoption. By visualizing the decision process, the model shifts from a "black box" to a decision-support tool. This allows clinicians to verify that the AI is focusing on the relevant anatomy rather than artifacts, fostering the trust required for human-AI collaboration.

6.4. Limitations and Future Work

A significant limitation concerns the dataset itself. The publicly available COCO-Caries dataset, while valuable, lacks detailed metadata regarding its annotation process. Critical information such as the annotators' level of dental expertise, the specific annotation protocol followed, measures of inter-rater reliability, and quality control procedures are not provided. This gap introduces unquantified uncertainty into the ground truth labels, which underpins both model training and evaluation. Therefore, the strong quantitative results reported here, while promising, must be interpreted with caution regarding their direct clinical translatability. Future research must prioritize the use of, or creation of, datasets with meticulously documented and clinically validated annotation processes to build truly robust and trustworthy AI systems for dentistry.

7. Clinical Applicability

The clinical utility of the proposed framework is characterized by its adaptability to distinct diagnostic priorities through the selection of operational thresholds on the precision-recall (PR) curves (Figure 15). For High-Sensitivity Screening, a threshold of ≈ 0.21 allows the model to achieve a recall of 0.90 and a precision of 0.83; this configuration is optimized for initial patient triage, maintaining a low 10% false-negative rate to ensure the early detection of incipient caries. Conversely, for High-Precision Treatment Planning, a threshold of ≈ 0.67 shifts the system to a precision of 0.95 and a recall of 0.68. This conservative setting is specifically designed to confirm cavitation before irreversible restorative procedures, thereby effectively minimizing the risk of clinical over-treatment.

The proposed YOLO11-seg framework facilitates seamless real-time integration into clinical workflows. With a mean inference latency of 5.2 ms, the system functions as a high-speed *second reader* providing instantaneous visual feedback during examinations.

Explainable AI (XAI) overlays enhance diagnostic trust by visualizing the rationale behind each prediction. This transparency assists junior clinicians in identifying pathological patterns while allowing experienced practitioners to validate AI suggestions rapidly. The model's high-confidence profile yields a box-level precision of 93.8% and a recall of 75.4%. This balance prioritizes the minimization of false positives to prevent unnecessary invasive procedures.

Statistical validation underscores the robustness of these findings. The 95% confidence intervals (CI) for box mAP_{50} and box loss are [0.839, 0.873] and [0.741, 0.771], respectively. These narrow intervals signify high stability

across diverse radiographic qualities. The improvement in recall over baseline architectures is statistically significant ($p < 0.001$), enhancing utility for early-stage caries screening.

The architecture maintains high computational efficiency with only 2.83 million parameters. This represents a 76% complexity reduction compared to the YOLOv8-seg baseline. This parsimony enables deployment on edge devices without compromising throughput. This synthesis of precision, reliability, and efficiency establishes the framework as a viable tool to reduce diagnostic variability in modern dental practice.

8. Conclusions and Future Works

This study presents a clinically-aware explainable deep learning framework for caries detection and segmentation in panoramic radiographs. By integrating YOLO11-seg with Bayesian hyperparameter optimization and dual XAI methods, we address critical gaps in current dental AI research: lack of interpretability, insufficient attention to clinical risk trade-offs, and limited computational efficiency.

The clinical-risk-aware optimization framework produces a model that balances competing clinical priorities. The final optimized YOLO11-seg model achieves a box-level precision of 93.8% and recall of 75.4%, providing reliable detection while maintaining real-time inference speeds (<6 ms) with only 2.83M parameters (76% reduction from YOLOv8-seg).

The integration of Grad-CAM and LIME provides transparent explanations of model decisions, highlighting radiographically meaningful regions and building clinical trust. While limitations exist regarding dataset scope and need for clinical validation, the framework establishes a foundation for trustworthy AI-assisted dental diagnostics.

Future work will extend to full panoramic radiographs to address tooth localization challenges, incorporate multi-center validation with diverse imaging protocols, and conduct clinical reader studies assessing AI-human collaborative diagnosis. Integration with dental software systems could ultimately provide practitioners with an efficient, interpretable tool for enhancing caries diagnosis, treatment planning, and preventive care delivery.

References

Abeyrathna, K.D., Granmo, O.C., Goodwin, M., 2021. Extending the tsetlin machine with integer-weighted clauses for increased interpretability. IEEE Access 9, 8233–8248.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. Information fusion 99, 101805.

Ayhan, B., Ayan, E., Atsü, S., 2025. Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence algorithms based on deep learning methods. BMC Oral Health 25, 216.

Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B., 2011. Algorithms for hyper-parameter optimization. Advances in neural information processing systems 24.

Brahmi, W., Jdey, I., 2024. Automatic tooth instance segmentation and identification from panoramic x-ray images using deep cnn. Multimedia Tools and Applications 83, 55565–55585.

Brahmi, W., Jdey, I., Drira, F., 2025. Automated impacted tooth segmentation in panoramic radiographs using unet and unet3+ with grad-cam explainability, in: 2025 IEEE 9th Forum on Research and Technologies for Society and Industry (RTSI), pp. 158–163. doi:10.1109/RTSI64020.2025.11212381.

Broniatowski, D.A., Broniatowski, D.A., 2021. Psychological foundations of explainability and interpretability in artificial intelligence. US Department of Commerce, National Institute of Standards and Technology.

Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., García-Gutiérrez, J., 2020. On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data. Remote Sensing 13, 89.

Casas, E., Ramos, L., Bendek, E., Rivas-Echeverria, F., 2024. Yolov5 vs. yolov8: Performance benchmarking in wildfire and smoke detection scenarios. Journal of Image and Graphics 12.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE. pp. 839–847.

cocoyaml, 2024. coco_caries dataset. https://universe.roboflow.com/cocoyaml/coco_caries. Visited on 2024-07-19.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd ed., Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Dayı, B., Üzen, H., Çiçek, İ.B., Duman, Ş.B., 2023. A novel deep learning-based approach for segmentation of different type caries lesions on panoramic radiographs. Diagnostics 13, 202.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. 1st ed., Chapman and Hall/CRC. URL: https://doi.org/10.1201/9780429246593, doi:10.1201/9780429246593.

Pérez de Frutos, J., Holden Helland, R., Desai, S., Nymoen, L.C., Langø, T., Remman, T., Sen, A., 2024. Ai-dentify: deep learning for proximal caries detection on bitewing x-ray-hunt4 oral health study. BMC Oral Health 24, 344.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2024. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation 16, 45–74.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969. doi:10.1109/ICCV.2017.322.

Hulsen, T., 2023. Explainable artificial intelligence (xai): concepts and challenges in healthcare. AI 4, 652–666.

Khamparia, A., Gupta, D., Khanna, A., Balas, V.E., 2022. Biomedical data analysis and processing using explainable (XAI) and responsive artificial intelligence (RAI). volume 222. Springer.

Mărginean, A.C., Mureşanu, S., Hedeşiu, M., Dioşan, L., 2024. Teeth segmentation and carious lesions segmentation in panoramic x-ray images using cariseg, a networks' ensemble. Heliyon 10.

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Nie, H., Pang, H., Ma, M., Zheng, R., 2024. A lightweight remote sensing small target image detection algorithm based on improved yolov8. Sensors 24, 2952.

Ortigossa, E.S., Gonçalves, T., Nonato, L.G., 2024. Explainable artificial intelligence (xai)–from theory to methods and applications. IEEE Access .

Oztekin, F., Katar, O., Sadak, F., Yildirim, M., Cakar, H., Aydogan, M., Ozpolat, Z., Talo Yildirim, T., Yildirim, O., Faust, O., et al., 2023. An explainable deep learning model to prediction dental caries using panoramic radiograph images. Diagnostics 13, 226.

Qureshi, U., 2006. Classification of dental x-ray images. Ph.D. thesis. West Virginia University Libraries.

Ramezanzade, S., Dascalu, T.L., Ibragimov, B., Bakhshandeh, A., Bjørndal, L., 2023. Prediction of pulp exposure before caries excavation using artificial intelligence: Deep learning-based image data versus standard dental radiographs. Journal of Dentistry 138, 104732.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp. 779–788. doi:10.1109/CVPR.2016.91.

Ren, S., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 .

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer, Cham. pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

Sehar, U., Xiong, J., Xia, Z., 2025. Automatic tooth labeling after segmentation using prototype-based meta-learning. Machine Intelligence Research , 1–14.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

Terven, J., Córdova-Esparza, D.M., Romero-González, J.A., 2023. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Machine Learning and Knowledge Extraction 5, 1680–1716.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology 58, 267–288.

Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., Hajamohideen, F., 2024. Explainable artificial intelligence in alzheimer's disease classification: A systematic review. Cognitive Computation 16, 1–44.

WHO, 2023. World Health Organization, Oral health. https://www.who.int/news-room/fact-sheets/detail/oral-health. Accessed: 2024-4-1.

Yang, L., Chen, G.Y., 2025. Evaluation of deep learning for caries detection with fine-grained classification and postprocessing improvements. International Dental Journal 75, 100898.

Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J., 2023. Cariesnet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic x-ray image. Neural Computing and Applications , 1–9.