Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2026 (44) 8** : 1–16.
ISSN-0037-8712
doi:10.5269/bspm.83900

YOLOv9c for Reliable Caries Detection on Intraoral Periapical Radiographs

D. Meghana and A. Manimaran

ABSTRACT: Automated detection of dental caries on intraoral periapical radiographs must couple high sensitivity to tiny, low-contrast radiolucencies with low latency for chairside use. While one-stage detectors have shown promise, YOLOv9 has not been systematically evaluated on periapical radiographs. We benchmark four YOLOv9 variants (e/m/s/c) under a unified, single-class protocol and identify YOLOv9c—which combines a GELAN backbone, training-time PGI, and decoupled multi-scale heads—as the most accurate and computationally efficient choice for this modality. The dataset comprises de-identified periapical radiographs collected at the Sibar Institute of Dental Sciences (Guntur, India), curated under institutional ethics approval with image-wise 80/10/10 splits; images were quality-screened and contrast-normalized prior to training. On a held-out test split, YOLOv9c attains Precision = 0.9830, Recall = 0.9974, F1 = 0.9902, mAP@0.5 = 0.9935, and mAP@0.5-0.95 = 0.8528, outperforming YOLOv9s/m/e and a two-stage Faster R-CNN (ResNet-18) baseline by large margins on both thresholded and strict-IoU criteria. Wilson 95% confidence intervals show high centers with tight bounds for precision and recall, and pairwise significance tests corroborate statistically reliable gains—especially in recall and tight-overlap localization—indicating that improvements are not sampling artefacts. Architecturally, GELAN's efficient, detail-preserving aggregation and PGI's gradient conditioning enhance shallow-stride features critical for subtle lesions, without adding inference cost. These results fill a key gap by providing the first modality-specific evaluation of YOLOv9 on periapicals and establish YOLOv9c as a strong, real-time baseline for clinical deployment and future research on periapical caries detection.

Keywords: Dental caries detection, intraoral periapical radiographs, one-stage detection, YOLOv9c, GELAN, PGI, Faster R-CNN (ResNet-18).

Contents

1 Introduction **2**

2 Related Works **3**

3 Materials and Methods **4**
3.1 Dataset Collection . . . 4
3.2 Data Preprocessing . . . 5
3.2.1 De-identification via manual cropping . . . 5
3.2.2 Image quality assessment and preprocessing . . . 5
3.2.3 Local Contrast Enhancement . . . 5
3.3 Annotation . . . 6
3.4 YOLO Overview . . . 6
3.5 Selecting the training model . . . 7
3.6 Proposed Model: YOLOv9c (GELAN + PGI + decoupled heads) . . . 8
3.7 Training Configuration . . . 8
3.8 Evaluation Metrics . . . 9

4 Experimental Results **10**
4.1 Overall Performance . . . 10
4.2 Statistical robustness and pairwise significance . . . 11

5 Discussion **13**

6 Conclusion and Future Work **15**

2020 *Mathematics Subject Classification:* 68T07, 68U10.
Submitted May 01, 2026. Published June 05, 2026.

1. Introduction

Dental caries remains one of the most prevalent chronic diseases worldwide and affects patients across the lifespan, with a particularly high burden among children and young adults. Untreated lesions may progress from initial enamel demineralization to deep dentin involvement, pulpal inflammation, pain, and ultimately tooth loss, with consequent effects on mastication, speech, quality of life, and health-care costs. Early identification of both cavitated and pre-cavitated lesions is therefore a central goal in preventive and restorative dentistry. Caries formation is driven by a dysbiotic biofilm that metabolizes fermentable carbohydrates and produces organic acids. These acids diffuse into enamel and dentin, dissolve hydroxyapatite, and create zones of subsurface demineralization. Clinically, early lesions may appear as subtle changes in opacity or contour; radiographically, the same process is seen as localized or diffuse radiolucencies at pits and fissures, proximal contact areas, and cervical regions. As lesions progress, mineral loss increases, radiographic contrast becomes more apparent, and the radiolucency may extend toward the pulp chamber and periapical region. Because the earliest radiographic signs are often faint and superimposed on complex anatomy, they may be overlooked in routine practice [1].

Conventional diagnosis relies on visual-tactile examination complemented by radiographic assessment. Bitewing radiographs are considered the standard for detecting approximal caries in posterior teeth because they provide high sensitivity for lesions obscured by contact areas. Periapical radiographs are routinely acquired for endodontic and restorative planning and can simultaneously depict occlusal, proximal, and root surfaces in fine detail. Panoramic radiographs provide a broader overview, although at lower resolution and with greater structural overlap, whereas high-quality intraoral photographs are increasingly used for documentation and teledentistry [2]. Across these modalities, subtle lesions, overlapping anatomy, and variable image quality contribute to missed or misclassified lesions, and several studies have reported substantial inter- and intra-observer variability in radiographic caries assessment [3].

Over the past decade, deep learning has emerged as a powerful tool for dental image analysis. Systematic reviews have consistently shown that convolutional neural networks (CNNs) can achieve high accuracy for caries detection across bitewing, periapical, panoramic, and photographic datasets, in some settings approaching or exceeding expert-level performance. Earlier studies relied on fully connected neural networks or patch-based CNNs, whereas more recent work has explored segmentation models such as U-Net, lesion-level detectors, and multitask architectures that jointly identify caries, restorations, and related anomalies. Among these approaches, one-stage object detectors such as the You Only Look Once (YOLO) family are especially attractive for clinical deployment because they localize and classify lesions in a single pass, enabling real-time inference on standard hardware while maintaining competitive accuracy relative to two-stage detectors [2].

Despite this progress, most published caries detectors still rely on earlier YOLO variants (e.g., YOLOv3-v8) or generic CNN backbones, and several limitations remain, including modest sensitivity for early or low-contrast lesions, limited calibration of predicted probabilities, and scarce evidence specific to periapical radiographs. To address information loss and gradient-flow issues that can hinder both large and lightweight detectors, Wang et al. recently introduced YOLOv9, which combines Programmable Gradient Information (PGI) with a Generalized Efficient Layer Aggregation Network (GELAN) backbone. PGI adds an auxiliary reversible branch that provides richer supervision during training without increasing inference cost, while GELAN improves feature reuse and efficiency across scales. Together, these innovations yield higher precision and better parameter-FLOP efficiency on standard benchmarks such as MS COCO compared with previous YOLO variants [4].

However, evidence for YOLOv9 in periapical radiographs and high-accuracy caries detection remains limited. Existing studies have focused primarily on bitewing radiographs or mixed image types, often report only moderate F1-scores and mAP values, and rarely examine probability calibration or small-lesion behavior in detail [5]. In particular, systematic benchmarking of multiple YOLOv9 variants on a single, well-curated periapical dataset with lesion-level annotations is still lacking. Given the central role of periapical radiographs in endodontic and restorative workflows, and the need for detectors that are both accurate and computationally efficient, this remains an important gap in the literature.

Motivated by these considerations, the present study investigates the use of YOLOv9 for automated cavity detection in intraoral periapical radiographs. We curate an annotated dataset of periapical images

with bounding-box labels for cavitated caries, compare multiple YOLOv9 variants under a unified training and evaluation protocol, and identify the configuration that offers the best trade-off between detection performance and efficiency. In addition, we analyze standard detection metrics Precision, Recall, F1-score, mAP@0.5, mAP@0.5-0.95 and, where relevant, evaluate probability calibration to assess the reliability of the model's confidence scores. By situating our results alongside recent YOLO-based caries detectors, we aim to clarify YOLOv9's role in dental radiograph analysis and to provide a strong baseline for future work on clinically deployable caries detection systems.

2. Related Works

Artificial intelligence has surfaced as a viable instrument for overcoming the constraints of conventional methodologies. A multitude of machine learning and deep learning models have been created to improve the precision of caries detection. Zhu et al. [6] presented CariesNet, an AI-based network for analyzing dental radiographs and accurately diagnosing carious lesions. The technology performed pixel-level classification, allowing dentists to identify deterioration with greater precision. Jusman et al. [7] investigated the application of texture features derived from the Gray Level Co-Occurrence Matrix (GLCM) technique, in conjunction with Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) classifiers. The study achieved notable accuracy rates of 95.7% with SVM and 94.9% with KNN; however, the limited dataset of 396 photos raised concerns about the generalizability of the findings.

Deep learning models have demonstrated significant potential in dental diagnosis. Juyal et al. [8] utilized YOLOv3 and Faster R-CNN models to identify dental cavities from real-time camera images, thereby obviating the necessity for X-rays. The research revealed accuracies of 75% with YOLOv3 and 80% with Faster R-CNN; however, the constrained sample of 300 photos underscored the need for larger, more varied datasets for training. Razaghi et al. [9] presented the YOLOv8 model for the automatic detection and classification of dental problems utilizing Bitewing and Orthopantomography (OPG) X-rays. Their model achieved a mean Average accuracy (mAP) of 71.6%, demonstrating the efficacy of deep learning in dental radiography analysis and suggesting potential for further accuracy improvement.

Saini et al. [10] evaluated these three CNN models and classifying early caries images. Kühnisch et al. [11] deployed the MobileNetV2 network model to detect individual cavities in oral photographs, attaining an accuracy over 90%. E.Y. Park [12] initially employed U-Net for segmenting dental characteristics and then utilized Faster R-CNN for detecting carious lesions, finding that darkened background regions enhanced accuracy metrics. Yoon et al. [13] employed the Cascade R-CNN model for tooth number recognition, attaining a mean Average Precision (mAP) score of 0.880, and an average mAP score of 0.769 in a three-stage caries detection task, thereby reinforcing the applicability of artificial intelligence in clinical environments.

Bayati et al. [14], 2025 used a modern, one-stage YOLOv8 pipeline on 552 bitewing radiographs to detect interproximal caries, reporting Precision 84.83%, Recall 79.77%, and F1 82.22% overall, with enamel-only lesions markedly more effortless ($\approx$ 96% precision) than dentin ($\approx$ 80% precision, $\approx$ 73% recall). The study mirrors a clinical workflow and underscores that deeper, lower-contrast defects remain the dominant failure mode even for strong one-stage detectors. As a single-center retrospective cohort without external validation, performance likely reflects device/projection idiosyncrasies; calibration and cross-site replication would strengthen clinical generalizability. Pérez de Frutos et al. [15], 2024 used YOLOv5—benchmarked against EfficientDet and RetinaNet on the large HUNT4 bitewing cohort for proximal caries—finding mAP = 0.647 and mean F1 = 0.548 at IoU 0.3, and surpassing both clinicians and the other one-stage baselines. The comparative design is valuable for situating detectors under a uniform protocol and for careful consensus labels, yet the sub-95% aggregate metrics highlight the intrinsic difficulty of bitewing imaging. Because the reported IoU differs from COCO-style summaries and the work is bitewing-only, portability to periapical radiographs has not been tested and should be interpreted cautiously.

Kaur et al., [16] 2024 used a three-way YOLO benchmark (YOLOv7/YOLOv8/YOLOv9) on $\sim$ 3,200 RVG periapical radiographs, reporting mAP@0.5 = 0.721 for YOLOv7, 0.832 for YOLOv9, and 0.982 for YOLOv8. The work shows one-stage detectors are feasible for deep-caries detection, with YOLOv9 competitive but not top on their data. However, heterogeneous image sources, missing per-class Precision/Recall/F1 metrics, and no external validation limit clinical transferability [5]. Elnady

et al. developed a YOLOv9 pipeline on 270 Kaggle images, aggressively augmented to $\sim$ 5,696 images, achieving mAP@0.5 $\approx$ 93.49%, with Precision $\approx$ 90.3%, Recall $\approx$ 90.5%, and overall F1 $\approx$ 89%; per-class mAP was $\sim$ 99.5% (caries) versus $\sim$ 88.1% (no-caries), but the tiny, photo-only dataset and lack of external testing constrain comparability to radiograph-based clinical settings.

Taken together, the above studies show steady progress in dental caries detection with one- stage detectors, yet the evidence is overwhelmingly bitewing-centric or photograph-based. To our knowledge, no peer-reviewed work has evaluated YOLOv9 on periapical radiographs for cavity detection. Accordingly, we introduce a YOLOv9-based detector for periapical images and conduct a controlled comparison across YOLOv9 variants under a unified training evaluation protocol. For context and fairness, we also benchmark a classical two-stage Faster R-CNN with a ResNet-18 backbone. Performance is reported with standard detection metrics (Precision, Recall, F1-score, mAP@0.5, mAP@0.5-0.95). Accordingly, the present contribution is primarily a modality-specific benchmarking and evaluation study rather than the introduction of a new detector architecture. The technical value lies in systematically assessing YOLOv9 variants under a unified protocol for periapical radiographs and in identifying YOLOv9c as the configuration that offers the most favorable accuracy-efficiency trade-off for this clinical setting.

3. Materials and Methods

Data preparation began with preprocessing to support high-quality labelling and stable training. After manual annotations were completed, we trained a YOLO-based detection model on the finalized dataset and measured its effectiveness using established evaluation criteria. The following sections provide a stage-by-stage description.

3.1. Dataset Collection

This study was conducted in partnership with the Sibar Institute of Dental Sciences (SIDS), Guntur, Andhra Pradesh, India. Formal clearance from the Institutional Ethics Committee as SIDS was granted on 13 December 2024 before any data access. Where required by hospital policy, informed consent was obtained. All records were de-identified before research use, and the project adhered to institutional and national regulations on patient privacy and data security. All procedures conformed to relevant institutional guidelines and the Declaration of Helsinki; owing to the retrospective design and full anonymization, the IEC granted a waiver of individual informed consent for secondary image use, except in instances where prospective consent was mandated by hospital policy.

Between November 2024 and February 2025, we assembled a cohort of 2,452 intraoral periapical radiographs with clinically confirmed dental caries. IOPA imaging is standard for detecting, grading, and planning treatment for cavities; it provides an appropriate foundation for developing and validating automated detection systems. Case selection and clinical tagging were performed under the supervision of experienced dental faculty at SIDS, with case identification and verification based on established diagnostic criteria so that labels reflected routine clinical practice. Images were then transferred over the hospital's secure Wi-Fi to research workstations, and data handling followed a standard operating procedure (access control, local encryption, and removal of identifiers) prior to analysis. As the present dataset was collected retrospectively from routine institutional clinical practice, image acquisition reflects real-world variation in positioning and exposure; however, device-wise stratified analysis was not separately performed in the present study.

Lesion localization was obtained through bounding-box annotation in the Computer Vision Annotation Tool (CVAT). Prior to annotation, clinical guidance on the identification and localization of carious lesions was provided by dental experts, and a predefined labeling protocol was established accordingly. Initial annotations were then created by the first author in accordance with this protocol and relevant clinical criteria. The resulting annotations were subsequently reviewed and verified by dental experts to ensure clinical accuracy, and any discrepancies were resolved before finalization. Since the annotations were created by a single primary annotator and subsequently verified by dental experts, formal inter-expert agreement statistics were not assessed in the present study.

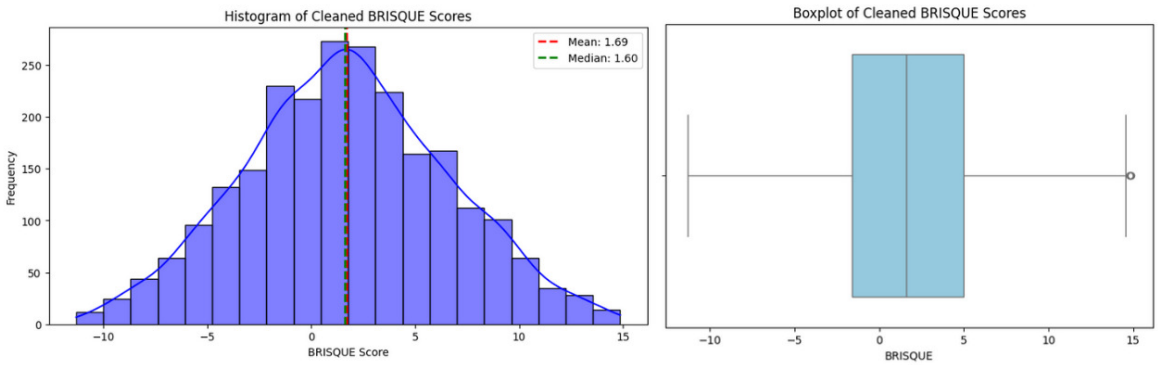


Figure 1: BRISQUE Score Distribution.

3.2. Data Preprocessing

High-quality preprocessing is essential for reliable training in medical imaging, where diagnostic appearance can shift with exposure settings, devices, and acquisition protocols. To promote consistency and preserve diagnostic detail, we adopted a standardized workflow designed to harmonize image presentation and safeguard patient privacy before any downstream use.

3.2.1. De-identification via manual cropping. To comply with ethical and regulatory requirements, each radiograph was manually cropped to remove any on-image identifiers (e.g., patient name, hospital ID, timestamp overlays) typically located along the borders. Only the de-identified crops were retained for analysis, and these were archived in a dedicated, access-controlled directory. This procedure ensured that no personally identifiable information was present at any stage of preprocessing, training, or evaluation.

3.2.2. Image quality assessment and preprocessing. Ensuring stable image quality is critical in dental radiography, so we adopted a standardized pre-training pipeline that screens quality, enhances diagnostically relevant contrast, and applies controlled augmentation. We evaluated each radiograph using the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) with the PIQ implementation in PyTorch. In this implementation BRISQUE is reported on a centered scale that can include negative values; lower values indicate better perceived quality, and scores are not directly comparable to the canonical 0-100 index, an appropriate no-reference choice for clinical datasets.

Quality control proceeded in two steps. First, images were retained if they lay within the central quality band of the cohort distribution on the centered BRISQUE scale. Second, we applied interquartile range (IQR) filtering to remove extreme outliers. From an initial 2,452 periapical radiographs 2,418 remained after quality screening. To characterize the cleaned set, we summarized BRISQUE values on the centered scale: the histogram was approximately normal with mean 1.69 and median 1.60, and the boxplot indicate a narrow IQR with few outliers; most scores lay roughly between -12 and +15. These summaries suggest a high and uniform baseline image quality after filtering. Figure 1 summarizes image quality after screening, showing a tight central BRISQUE scores, which supports stable training on uniformly presented periapicals.

3.2.3. Local Contrast Enhancement. To enhance diagnostically relevant detail without boosting noise, we applied Contrast-Limited Adaptive Histogram Equalization (CLAHE; tile grid 8×8, clip limit 2.0), which improved the visibility of subtle cues such as early carious radiolucencies and enamel boundaries. We then standardized geometry by resizing all images to 640×640 to match the YOLO input specification and, to improve generalization, used Albumentations to introduce controlled variability—horizontal and vertical flips (p=0.5), rotations by multiples of 90° plus small random rotations up to ±20°, and minor spatial transforms reflecting typical positioning differences—thereby broadening appearance diversity while preserving lesion morphology and reducing overfitting in downstream detection.

Figure 2: Original and Annotated Image

3.3. Annotation

Bounding-box labels were created in the Computer Vision Annotation Tool (CVAT), an open-source platform for vision datasets. A single-class project for caries was defined, and the rectangle tool was used to delineate visually apparent lesions across the full set of intraoral periapical radiographs. Each image was examined carefully to place boxes tightly around the suspected regions. Annotation was performed in accordance with the clinically guided labeling protocol described in Section 3.1, and all labels were stored under a single class with index 0.

After labeling, we exported annotations in YOLO detection format, which produces one text file per image containing the class index and normalized box coordinates in the order:

$$[\text{class\_id},\ x_{\text{center}},\ y_{\text{center}}, \text{width},\ \text{height}]$$

The export associates label files with their corresponding images automatically, enabling immediate ingestion by YOLO training scripts. During quality control, repeated images were removed. The curated set comprised 1887 images, partitioned 80/10/10 into training, validation, and test splits, respectively. Annotation consistency was maintained through a predefined labeling protocol, first-author annotation, and expert review before finalization.

In the present work, the detection task was formulated as a single-class problem, with all annotated lesions grouped under the category caries. Although lesion presentation reflected clinical diversity across cases, lesion-instance counts and class-wise lesion distributions were not separately analyzed in this study. Accordingly, conventional multi-class imbalance was not applicable under the adopted detection framework. This CVAT-based process yielded a consistent, model-ready dataset suitable for precise lesion localization and downstream benchmarking. Figure 2 illustrates original radiograph and its annotation.

3.4. YOLO Overview

"You Only Look Once" (YOLO) casts object detection as a single-pass, end-to-end prediction problem. Given an input radiograph $I \in \mathbb{R}^{H\times W}$, a backbone extracts hierarchical features, a neck fuses them across multiple strides (preserving both fine detail and global context), and a decoupled head simultaneously regresses bounding boxes and predicts confidence/classes at each spatial location. Concretely, each candidate yields geometry $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$, an objectness probability $p(\text{obj})$, and class posteriors $p(c \mid \text{obj})$; candidates are scored by $s = p(\text{obj}) \cdot p(c \mid \text{obj})$ and consolidated with Non-Maximum Suppression (NMS). Training minimizes a composite loss that couples an IoU-family box term (e.g., CIoU/DIoU) with BCE/focal-style classification and objectness terms, with dynamic (anchor-free) label assignment

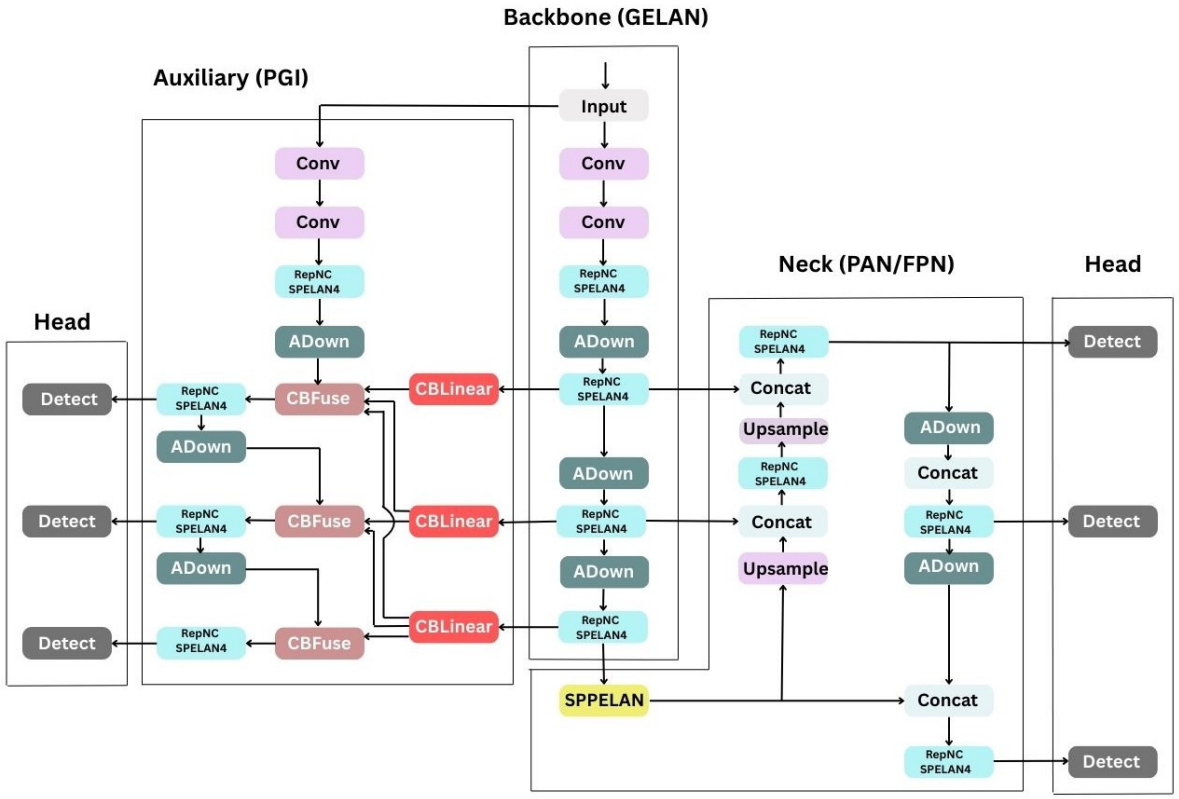


Figure 3: YOLOv9 Architecture

aligning positives by joint localization–classification quality. By eliminating proposal stages, YOLO delivers real-time throughput without sacrificing accuracy—an especially compelling fit for chairside dentistry, where small, low-contrast lesions demand multi-scale sensitivity and low latency is essential.

In periapical radiographs, these architectural choices are particularly consequential. Dental caries often occupies only a tiny fraction of the field and presents as low-contrast radiolucencies against complex anatomy. Modern YOLO variants therefore predict at multiple strides (e.g., 8/16/32), allowing the head to "see" very small structures at high spatial resolution while modelling larger context on coarser maps. The neck (FPN/PAN-style aggregation) routes shallow, high-frequency detail upward and propagates deep semantic context downward, which helps distinguish caries from overlapping structures. Recent implementations also employ decoupled heads and dynamic label assignment that selects positives samples using a joint measure of localization and classification quality, improving stability and sensitivity compared with fixed-IoU heuristics.

3.5. Selecting the training model

The dataset was partitioned into three splits—training (80%), validation (10%), and testing (10%). For each split, two parallel folders—`images/` and `labels/`—were created, and every image was paired with its corresponding YOLO label text file containing normalized $[class\_id, x\_center, y\_center, width, height]$ entries for the single class `caries`. The resulting directory tree was consolidated under a project root and described by a minimal dataset specification file (`dental_data.yaml`) that records the absolute root path, split-relative image directories, `nc:1`, and `names:['caries']`. This arrangement enables the trainer to resolve files unambiguously and guarantees that images and labels remain synchronized across splits.

Using this setup, we trained four YOLOv9 variants YOLOv9e, YOLOv9m, YOLOv9s, and YOLOv9c on the intraoral periapical dataset under an identical protocol (same splits, input resolution, epochs, and

single-class label schema). To contextualize the one-stage family with a classical proposal-based approach, we additionally trained a Faster R-CNN (ResNet-18) baseline on the same train/val/test partitions and evaluated it with the same metrics. Among all models, YOLOv9c delivered the strongest overall performance achieving the highest recall and F1 together with the leading strict-IoU mAP (mAP@0.5-0.95) and thus represents our proposed detector. To our knowledge, no prior study has trained a YOLOv9 model specifically on periapical radiographs for cavity detection; this study therefore presents a novel modality-specific evaluation of YOLOv9 on periapical radiographs and establishes a clear benchmark showing that YOLOv9c offers the most favorable accuracy-efficiency trade-off in our setting.

3.6. Proposed Model: YOLOv9c (GELAN + PGI + decoupled heads)

Our detector adopts the YOLO family's single-stage object detection architecture and tailors it to periapical radiographs. An input image is processed by a feature extractor (backbone), a multi-scale fusion module (neck), and a decoupled detection head that, at each spatial location on several feature maps, directly regresses a bounding box $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$ together with an objectness score and a class posterior. During training, a composite loss balances localization (IoU-based box loss), objectness, and classification terms so that the model learns both to place tight boxes and assign calibrated confidences. At inference, candidate boxes from all scales are ranked by the product of objectness and class probability; Non-Maximum Suppression removes duplicates, yielding a compact set of lesion hypotheses at real-time speed—an essential property for chairside triage and decision support. This single-pass, end-to-end design avoids proposal generation and per-region resampling, minimizing latency while preserving accuracy on small, low-contrast targets typical of dental caries. Dynamic, anchor-free label assignment aligns positives using joint localization–classification quality, improving stability over fixed-IoU heuristics.

Building on this paradigm, we use YOLOv9c as the base detector because it couples accuracy with efficient computation through two architectural ideas: GELAN and PGI. The Generalized Efficient Layer Aggregation Network (GELAN) forms the backbone and emphasizes multi- path feature reuse and cross-stage aggregation, improving gradient flow and parameter/FLOP efficiency. In Practice, GELAN preserves fine-grained detail at shallow depths, which is critical for subtle radiolucencies at enamel-dentin interfaces or near restorations. Above the backbone, a PAN/FPN-style neck fuses information top-down and bottom-up to produce feature maps at multiple spatial resolutions; the decoupled head then predicts boxes and class logits independently on each scale, allowing high-resolution maps to specialize on tiny lesions while coarser maps cover larger structures.

During training only, Programmable Gradient Information (PGI) attaches an auxiliary supervision branch that injects richer gradient signals into earlier layers; PGI strengthen optimization and stabilizes the decoupled heads without altering inference cost, because the auxiliary branch is discarded at test time. The final decoupled, multi-scale heads perform separate regression and classification/objectness prediction, reducing gradient interference between localization and recognition. Taken together—GELAN for efficient, detail-preserving representation; PGI for training-time guidance; and decoupled multi-scale heads for small- object sensitivity—YOLOv9c offers a compact, fast, and accurate detector tailored to periapical cavity localization, where small size, low contrast, and anatomical superposition are the dominant challenges. Figure 3 highlights the GELAN backbone and the training-time PGI branch within the YOLOv9c detector.

3.7. Training Configuration

All YOLOv9 variants were trained under a single, uniform protocol to ensure a fair comparison. Images were resized to 640 × 640 with letterbox padding; the task was framed as single-class detection (names: ['caries]); and we enforced image-wise 80/10/10 splits for training, validation, and testing. Each run consisted of 150 epochs with a batch size of 8, executed on GPU:0 when available, with automatic CPU fallback otherwise. Labels followed the canonical YOLO format with normalized coordinates [xcenter, ycenter, w, h], which keeps the loss terms consistently scaled across images and splits.

Optimization and scheduling were left at the Ultralytics trainer defaults so that observed performance differences reflect architectural changes rather than bespoke tuning. Concretely, optimizer = 'auto' selects a modern adaptive optimizer (typically AdamW with decoupled weight decay) and applies standard

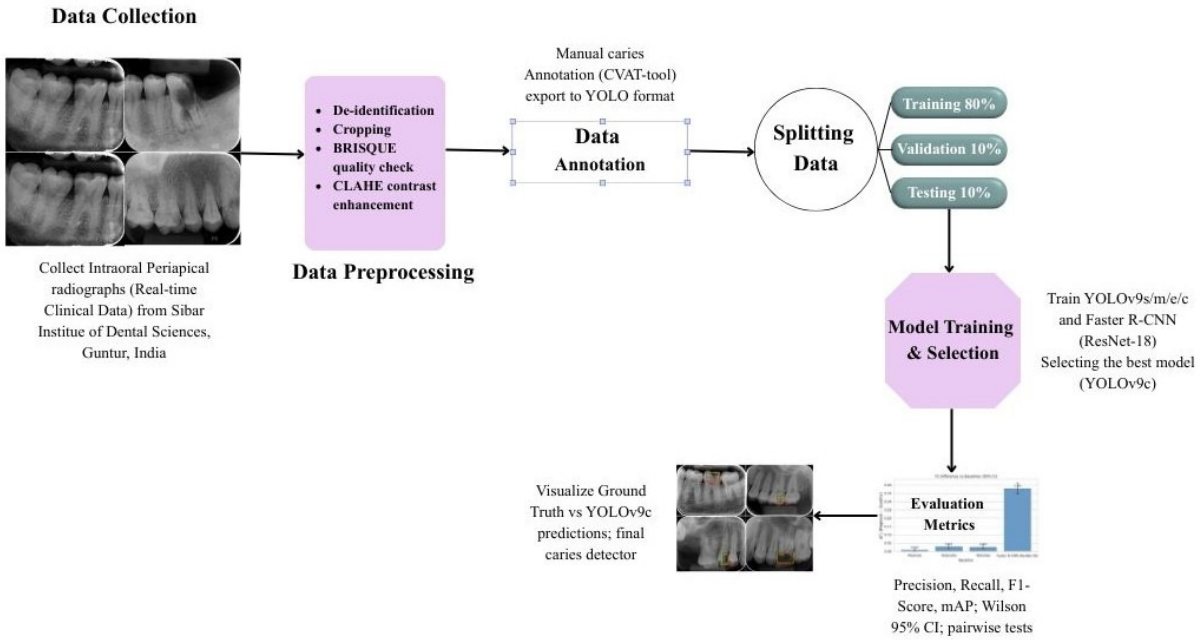


Figure 4: Overall workflow of the proposed YOLOv9c-based dental caries detection pipeline.

YOLO parameter-grouping so that biases, batch-norm parameters, and backbone layers receive appropriate learning-rate and decay settings. The learning rate undergoes a brief warm-up at the start of training and then follows the built-in smooth decay (one-cycle/cosine-style), which stabilizes early updates and mitigates late-epoch overfitting. The loss is the usual composite: an IoU-family box regression term (CIoU/DIoU) combined with BCE/focal-style objectness and classification terms; positives are assigned dynamically in an anchor-free manner using a task-aligned/OTA-style criterion that balances localization and classification quality.

Data handling likewise used the default detection augmentations: letterbox resize, random scale/translation, mild HSV jitter, and horizontal flip (probability ≈ 0.5). Mosaic/MixUp were used conservatively and ramped toward the end of training to avoid late-epoch instability. Validation and test sets were evaluated without augmentation. At inference, the engine's default confidence threshold (≈ 0.25) and NMS IoU (≈ 0.70) were applied with class-agnostic NMS, appropriate for a single-class task. All variants—YOLOv9e, YOLOv9m, YOLOv9s, and YOLOv9c—were trained and scored under this exact regimen; the Faster R-CNN (ResNet-18) baseline used the same splits and metric protocol, ensuring that any gains are attributable to the detector design, most notably GELAN + PGI in YOLOv9c. Figure 4 outlines the end-to-end pipeline from de-identification and CLAHE to training/validation, ensuring reproducibility of results and metrics.

3.8. Evaluation Metrics

We report standard detection metrics computed on the held-out test split.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

$$\text{mAP@0.5} = \text{mean Average Precision computed at IoU 0.5} \tag{3.4}$$

$$\text{mAP@0.5:0.95} = \text{mean Average Precision averaged over IoU thresholds 0.50–0.95} \tag{3.5}$$

4. Experimental Results

4.1. Overall Performance

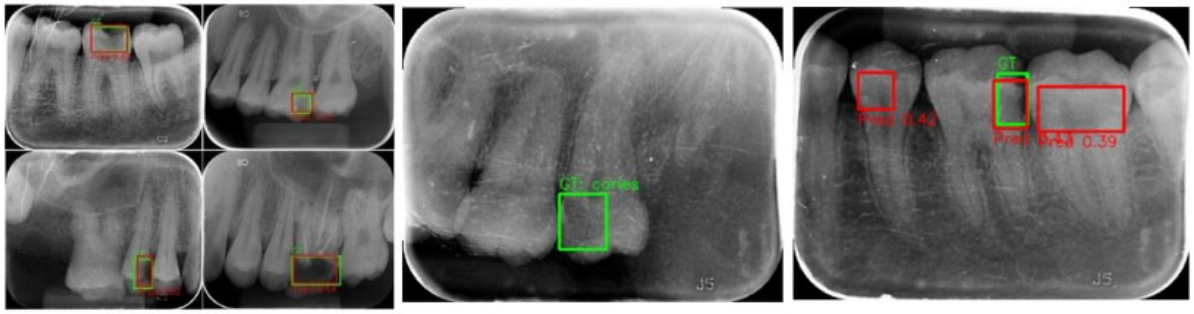
On the held-out test split, the four YOLOv9 variants exhibited a clear hierarchy. YOLOv9c achieved the strongest overall performance with Precision = 0.98, Recall = 0.99, F1-Score = 0.99, mAP@0.5 = 0.99, mAP@0.5-0.95 = 0.85. YOLOv9s followed Precision = 0.97, Recall = 0.98, F1-Score = 0.98, mAP@0.5 = 0.99, mAP@0.5-0.95 = 0.81, with YOLOv9m 0.97, 0.93, 0.95, 0.97, 0.78 and YOLOv9e 0.96, 0.95, 0.96, 0.97, 0.76 trailing. The two-stage Faster R- CNN (ResNet-18) baseline was substantially weaker Precision = 0.82, Recall = 0.48, F1-Score = 0.61, mAP@0.5 = 0.44, mAP@0.5-0.95 = 0.18. Taken together, the results indicate that YOLOv9c provides the best aggregate accuracy both thresholded metrics and strict IoU criteria. Representative predictions are shown in Figure 5 and table 1 summarizes the comparative performance of YOLOv9c and baseline models across Precision, Recall, F1-score, mAP@0.5, and mAP@0.5-0.95.

The margin between YOLOv9c and YOLOv9s is modest on thresholded scores ($\Delta F1 \approx +0.0102$) and mAP@0.5 ($\Delta \approx +0.0012$), but widens at strict IoU where YOLOv9c leads mAP@0.5:0.95 by +0.0380. Because mAP@0.5:0.95 averages AP over IoUs from 0.50 to 0.95, this advantage reflects tighter localization and greater robustness to small, low-contrast lesions—the dominant failure modes in periapical radiographs. Compared with YOLOv9s, YOLOv9c adds just enough capacity to improve feature expressiveness at shallow strides without inducing the optimization fragility or overfitting we observed in YOLOv9m/YOLOv9e. Relative to YOLOv9m and YOLOv9e, YOLOv9c improves F1 by $\sim$ 3.0–3.2 points and strict-IoU mAP by $\sim$ 7–9 points, suggesting that simply scaling depth/width did not translate into better generalization in this single-class regime. In practice, YOLOv9c achieves a well-balanced capacity–regularization trade-off that the other variants do not. Its GELAN backbone promotes multi-path feature reuse and cross-stage aggregation, preserving fine detail at small strides—precisely where subtle proximal radiolucencies reside. During training, PGI injects richer gradient signals that stabilize the decoupled heads; because this auxiliary branch is removed at inference, the model retains a lean runtime profile. Together, these design choices make YOLOv9c more expressive than YOLOv9s yet less overparameterized than YOLOv9m/YOLOv9e, improving mAP@0.5:0.95 without sacrificing sensitivity.

Faster R-CNN (ResNet-18) shows markedly lower recall and poorer localization (F1-Score 0.61, mAP@0.5 0.44, mAP@0.5-0.95 0.18). In periapical radiographs—where lesions are tiny and often overlap complex anatomy—the single-stage, multi-scale heads of YOLOv9 deliver higher small-object sensitivity and substantially better AP at tight overlaps. Operationally, YOLOv9 retains the efficiency advantages of a one-stage detector; however, explicit inference latency and FPS measurements were not benchmarked in the present study and should be established in future deployment-oriented evaluations.

From a deployment standpoint, Recall and F1 are safety-critical: missing disease is costlier than flagging a false positive. YOLOv9c achieves near-perfect recall (0.99) and the highest precision, yielding the top F1-Score (0.99). Its lead on mAP@0.5-0.95 demonstrates reliable box tightness across stricter overlaps, which supports trustworthy visualization and downstream decision support. Mechanistically, YOLOv9c's GELAN+PGI pairing appears to deliver the best accuracy-efficiency trade-off in our setting: GELAN preserves subtle signal at shallow strides; PGI improves optimization and calibration during training; and the decoupled multi-scale heads translate these gains into consistent localization without latency penalties.

Under a uniform training protocol, YOLOv9c consistently outperforms YOLOv9s/m/e and a Faster R-CNN (ResNet-18) baseline, with the highest F1 and the strongest mAP@0.5-0.95indicating tighter localization on small, low-contrast lesions. The gains are attributable to GELAN and PGI that together provide a superior accuracy-efficiency balance. These findings, together with convergent qualitative and quantitative evidence, indicate that the model is dependable and ready for real-time clinical use, where both accuracy and throughput are critical. Figure 6 compares core metrics across detectors; Figure 5



Predicted vs Ground Truth for YOLOv9c

False-negative for YOLOv9c.

False-positive for YOLOv9c.

Figure 5: Representative qualitative outcomes of YOLOv9c.

presents representative qualitative examples, including correctly localized lesions as well as illustrative false-negative and false-positive cases. These examples provide additional clinical insight into the model's residual errors, particularly in subtle low-contrast regions and anatomically confounding areas.

Table 1: Comparative performance of YOLOv9c and baseline models

Model	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5-0.95
YOLOv9s	0.97	0.98	0.98	0.99	0.81
YOLOv9m	0.97	0.93	0.95	0.97	0.78
YOLOv9e	0.96	0.95	0.96	0.97	0.76
Faster R-CNN (ResNet-18)	0.82	0.48	0.61	0.44	0.18
YOLOv9c	0.98	0.99	0.99	0.99	0.85

4.2. Statistical robustness and pairwise significance

table 2 summarizes precision and recall with Wilson 95% Confidence Intervals (CIs) to quantify uncertainty rather than relying on point estimates alone. YOLOv9c shows high centers with very tight bounds, indicating both superior accuracy and low variance, key for clinical reliability. In contrast, YOLOv9s, YOLOv9m, and YOLOv9e exhibit lower centers and/or wider intervals, and the Faster R-CNN baseline is markedly lower on both measures. These intervals demonstrate that YOLOv9c's advantage is not a sampling artefact and that its sensitivity is consistently maintained across test samples. Recall gains over all baselines are statistically significant, and the $\Delta F1$ confidence intervals lie entirely above zero for every contrast. Against YOLOv9s, the improvement is modest but meaningful, while gaps to YOLOv9m and YOLOv9e are larger and supported by highly significant recall p-values and bootstrap tests. The Faster R-CNN contrast shows the most pronounced effect, confirming that YOLOv9c delivers materially higher detection accuracy and localization fidelity. Collectively, these tests validate that YOLOv9c's superiority is systematic, not due to chance, and extends across both classification and localization metrics. table 3 presents pairwise significance tests, bootstrap p-values, and $\Delta F1$ with 95% CIs. In addition to point estimates, statistical significance and effect-size information were explicitly reported through pairwise $\Delta F1$ comparisons, corresponding p-values, and 95% confidence intervals, while Wilson 95% confidence intervals were used to quantify uncertainty in precision and recall.

Figure 7 summarizes precision and recall with Wilson 95% confidence intervals, showing YOLOv9c with both the highest central estimates and the tightest bounds—evidence of superior accuracy and low variance that supports clinical reliability. In contrast, the other YOLOv9 variants display either lower centers or wider intervals, and the two-stage baseline is markedly worse on both axes. Figure 8

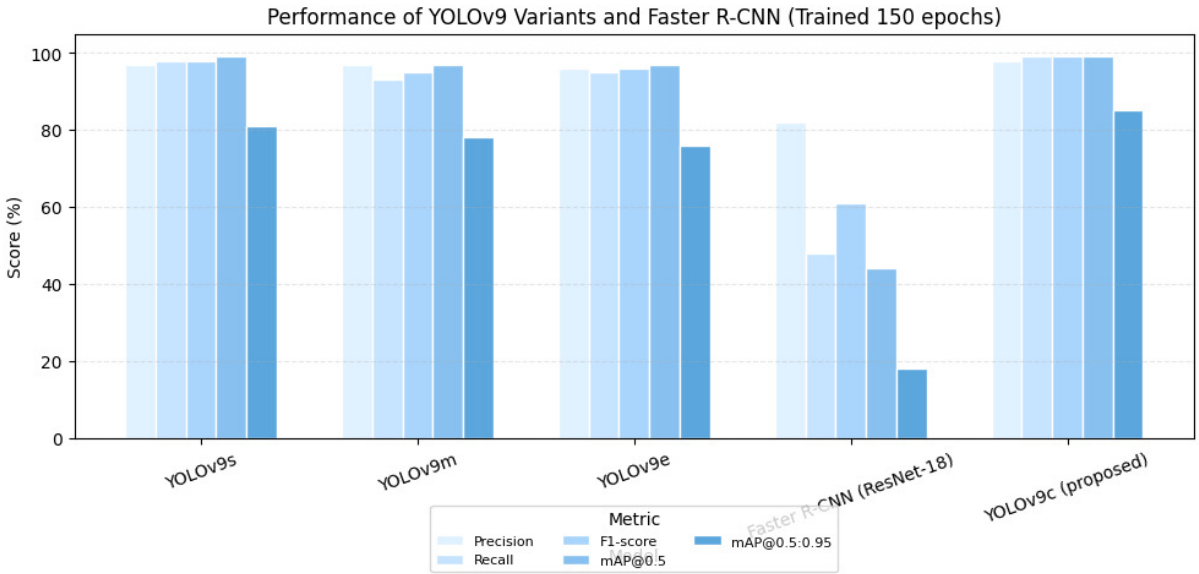


Figure 6: Performance comparison of YOLOv9 variants and Faster R-CNN

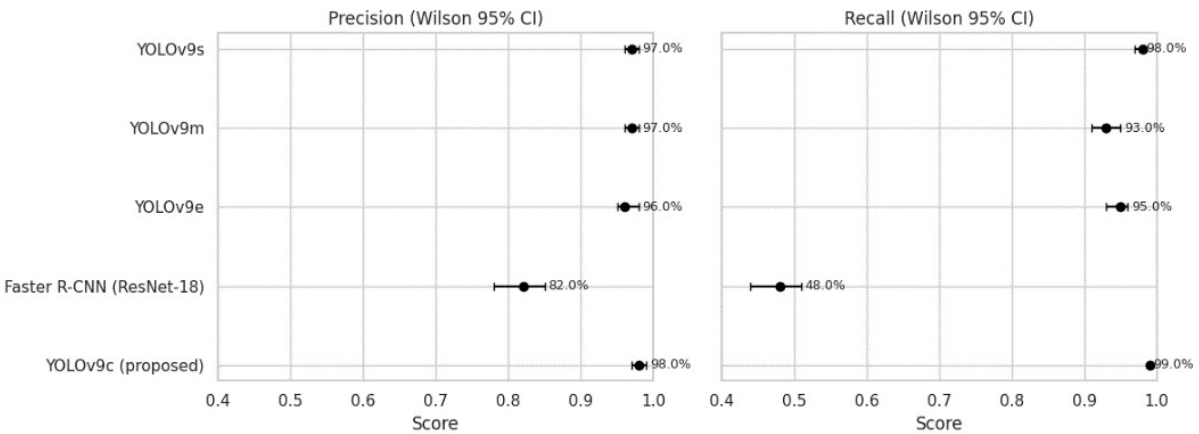


Figure 7: Wilson 95% Confidence Intervals (Precision/Recall).

Table 2: Performance with Wilson 95% Confidence Interval

Model	Precision	P_L	P_U	Recall	R_L	R_U
YOLOv9s	0.97	0.96	0.98	0.98	0.97	0.98
YOLOv9m	0.97	0.96	0.98	0.93	0.91	0.95
YOLOv9e	0.96	0.95	0.98	0.95	0.93	0.96
ResNet-18	0.82	0.78	0.85	0.48	0.44	0.51
YOLOv9c	0.98	0.97	0.99	0.99	0.99	0.99

Table 3: Pairwise significance vs. YOLOv9c

Baseline Model	Precision p-value	Recall p-value	ΔF1 Score	95% CI for ΔF1	Bootstrap p-value
YOLOv9s	4.405e-01	2.168e-03	0.010	(0.003, 0.018)	1.040e-02
YOLOv9m	6.448e-01	1.736e-10	0.031	(0.018, 0.045)	0.000e+00
YOLOv9e	9.413e-02	3.281e-08	0.029	(0.017, 0.042)	0.000e+00
Faster R-CNN (ResNet-18)	0.000e+00	0.000e+00	0.380	(0.348, 0.413)	0.000e+00

complements this by plotting the $\Delta F1$ against each baseline with bootstrap confidence intervals and p-values; for all contrasts the intervals lie entirely above zero, confirming that the gains of YOLOv9c are statistically robust rather than sampling artefacts. Together, these figures show that YOLOv9c improves not only point performance but also the certainty with which it is achieved.

Figure 9 provides a compact, model-wise comparison of the core metrics (Precision, Recall, F1, mAP@0.5, mAP@0.5:0.95), making clear that YOLOv9c attains the best balance of sensitivity and localization fidelity. The separation is most visible at strict IoU (mAP@0.5:0.95), where YOLOv9c maintains high scores while competing variants drop, indicating tighter boxes on small, low-contrast lesions. Read alongside Figure 7 and Figure 8, these point estimates contextualize the statistical evidence, showing that the observed advantages are both practically meaningful and statistically supported.

5. Discussion

This study demonstrates that, under a uniform training and evaluation protocol, YOLOv9c consistently delivers the best aggregate performance across precision, recall, F1, and strict-IoU mAP. The improvement over YOLOv9s is modest on thresholded scores but becomes pronounced at higher IoU thresholds, indicating tighter box localization—a crucial property for delineating small, low-contrast periapical lesions. The shortfall of YOLOv9m/e relative to v9c suggests that simply scaling capacity is not sufficient in this single-class, small-target regime; by contrast, v9c's GELAN backbone preserves shallow-stride detail while PGI supplies richer gradients during training, stabilizing the decoupled heads without adding inference overhead.

Compared with a representative two-stage detector (Faster R-CNN, ResNet-18), YOLOv9c achieves substantially higher recall and stricter-IoU accuracy while retaining real-time throughput. In clinical terms, the combination of near-perfect recall and high F1 reduces the risk of missed disease while keeping false positives manageable for chairside review. The tight Wilson 95% confidence intervals for precision and recall, together with significant pairwise tests, reinforce that the gains are systematic rather than sampling artefacts. From a workflow perspective, the proposed detector is intended to function as a decision-support tool rather than a replacement for clinical judgement. In practice, predicted bounding boxes may assist clinicians by highlighting suspicious radiolucent regions for focused review during radiographic interpretation, particularly in subtle cases. This box-level visualization also provides a basic level of interpretability, as the model output can be inspected directly against the underlying anatomy and either accepted or rejected by the clinician in context.

A brief qualitative error analysis was performed to contextualize the predictions shown in Figure 5. The false-negative case suggests that subtle low-contrast lesions may be missed, whereas the false-positive case indicates that overlapping anatomical structures or lesion-like radiolucencies may occasionally be misidentified as caries. Positionally, our results demonstrate that a carefully configured

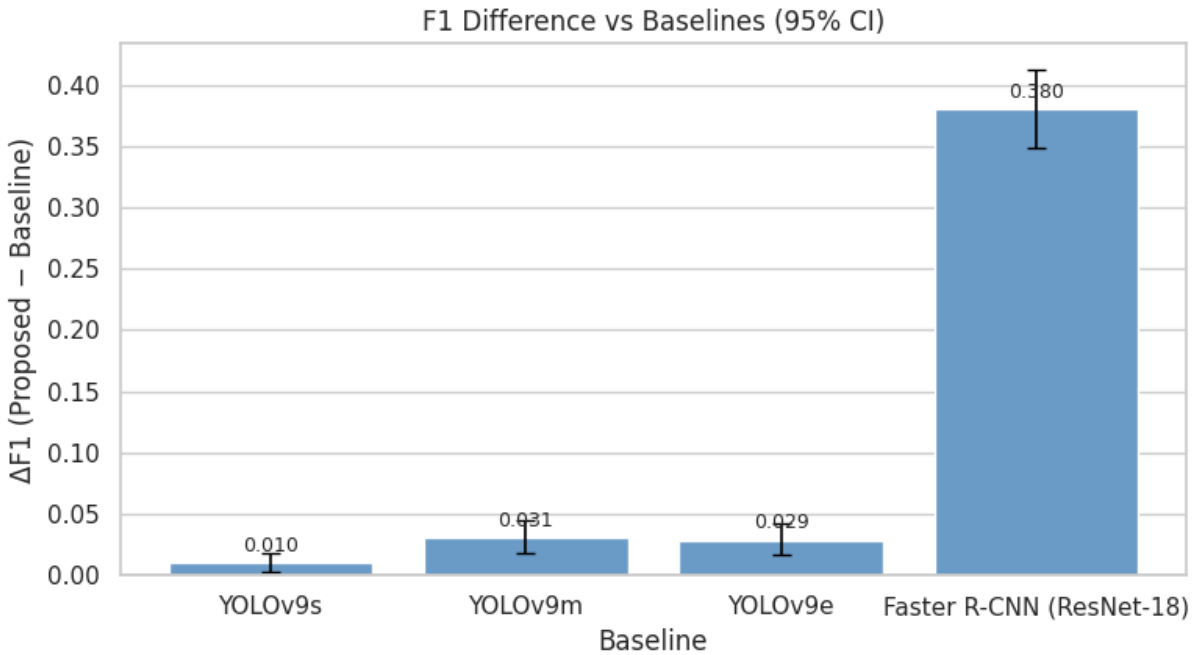


Figure 8: Pairwise ΔF1 with bootstrap CIs and p-values.

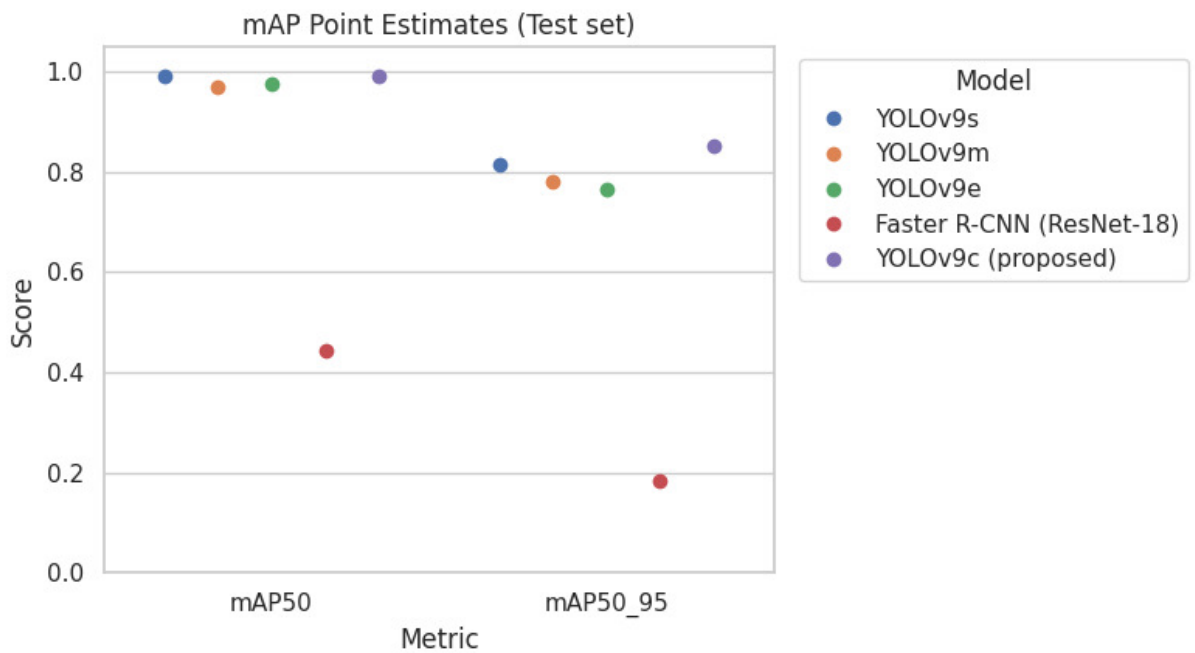


Figure 9: Point estimates for all models.

YOLOv9c—combining GELAN's efficient, detail-preserving aggregation with PGI's training-time gradient conditioning—achieves tight-overlap localization on periapical radiographs while retaining real-time throughput. The model delivers near-perfect recall alongside high strict-IoU mAP, indicating both reliable lesion sensitivity and precise box placement for subtle, low-contrast radiolucencies. Because PGI is discarded at inference, these gains incur no additional runtime cost, preserving a lean deployment profile. Overall, the evidence positions YOLOv9c as a balanced, clinically suitable detector purpose-built for the scale and contrast characteristics of periapical caries.

Beyond headline metrics, the study emphasizes rigor and reproducibility through standardized de-identification, quality gating, contrast normalization (CLAHE), fixed 640×640 inputs, and a transparent Ultralytics training/evaluation protocol. This consistency makes the results portable across labs and facilitates auditing, ablation, and retraining under governance requirements standard to clinical AI. Equally important, the combination of near-perfect recall, high strict-IoU mAP, and tight Wilson intervals argues for reliability at the box level— supporting confident visualization, scoring, and downstream review. Taken together, these elements strengthen the case for YOLOv9c as a credible periapical benchmark and a practical candidate for prospective validation, calibration studies, and multi-site replication en route to deployment. As the present study was based on a single-center retrospective dataset collected at SIDS, direct generalizability across institutions, imaging devices, acquisition protocols, and patient populations should be interpreted with caution until multi-center external validation is performed.

6. Conclusion and Future Work

This study introduces and rigorously evaluates a YOLOv9c-based detector for periapical caries localization. Leveraging GELAN's efficient, detail-preserving aggregation and PGI's training- time gradient conditioning, the model attains near-perfect recall with superior strict-IoU mAP, yielding tight boxes for subtle, low-contrast radiolucencies while maintaining real-time throughput. Under a uniform training/evaluation protocol, YOLOv9c outperforms alternative YOLOv9 variants and a two-stage baseline, establishing a strong modality-specific benchmark with promising potential for clinically usable chairside decision support, pending further operational validation.

Although YOLOv9c has not previously been reported for periapical caries detection, our findings indicate it is particularly well-suited to this setting and can anchor subsequent research. However, explicit calibration analyses such as reliability diagrams, Expected Calibration Error (ECE), and Brier score were not performed in the present study. These aspects will be addressed in future work to further assess the reliability of model confidence estimates for clinical triage. Building on this, we will: (i) investigate hybrid detectors that fuse v9c with transformer-style context modules or dental-biased attention blocks to enhance further small- lesion sensitivity; (ii) fine-tune v9c with calibration-aware training (e.g., temperature scaling, focal-tuning) to improve probability reliability for triage; (iii) assess domain generalization through cross-site validation and low-shift adaptation techniques (test-time adaptation, lightweight normalization alignment); and (iv) extend to multi-task formulations (e.g., joint caries/restoration detection and tooth-level mapping) to support more comprehensive chairside workflows.

Acknowledgments

The authors would like to express their sincere gratitude to Dr. Sagi Sai from SIBAR Institute of Dental Sciences, Guntur, for valuable support in the data collection process and for providing expert guidance on dental anatomy and caries classification.

References

1. Forouzeshfar, P., Safaei, A. A., Ghaderi, F., Hashemi Kamangar, S., Kaviani, H., Haghi, S., *Dental caries diagnosis using neural networks and deep learning: a systematic review*. Multimedia Tools Appl. 83, 30423-30466, (2024). 1

2. Bayraktar, Y., Ayan, E., *Diagnosis of interproximal caries lesions with deep convolutional neural network in digital bitewing radiographs*. Clin. Oral Investig. 26, 623-632, (2022). 2

3. Arzani, S., Karimi, A., Iranmanesh, P., Yazdi, M., Sabeti, M. A., Nekoofar, M. H., Dummer, P. M., *Examining the diagnostic accuracy of artificial intelligence for detecting dental caries across a range of imaging modalities: An umbrella review with meta-analysis*. PLoS One 20, e0329986, (2025). 3

4. Wang, C. Y., Yeh, I. H., Mark Liao, H. Y., *YOLOv9: Learning what you want to learn using programmable gradient information*. In: European Conference on Computer Vision, 1-21, (2024). 4

5. Kaur, A., Jyoti, D., Sharma, A., Yelam, D., Goyal, R., Nath, A., *Deep caries detection using deep learning: from dataset acquisition to detection*. Clin. Oral Investig. 28, 677, (2024). 5

6. Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J., *CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image*. Neural Comput. Appl. 35, 16051-16059, (2023). 6

7. Jusman, Y., Anam, M. K., Puspita, S., Saleh, E., Kanafiah, S. N. A. M., Tamarena, R. I., *Comparison of dental caries level images classification performance using knn and svm methods*. In: IEEE Int. Conf. Signal Image Process. Appl., 167-172, (2021). 7

8. Juyal, A., Tiwari, H., Singh, U. K., Kumar, N., Kumar, S., *Dental caries detection using faster R-CNN and YOLO V3*. ITM Web Conf. 53, 02005, (2023). 8

9. Razaghi, M., Komleh, H. E., Dehghani, F., Shahidi, Z., *Innovative diagnosis of dental diseases using YOLO V8 deep learning model*. In: Iranian/Int. Machine Vision Image Process. Conf., 1-5, (2024). 9

10. Saini, D., Jain, R., Thakur, A., *Dental caries early detection using convolutional neural network for tele dentistry*. In: Int. Conf. Advanced Computing Communication Systems, 958-963, (2021). 10

11. Kühnisch, J., Meyer, O., Hesenius, M., Hickel, R., Gruhn, V., *Caries detection on intraoral images using artificial intelligence*. J. Dent. Res. 101, 158-165, (2022). 11

12. Park, E. Y., Cho, H., Kang, S., Jeong, S., Kim, E. K., *Caries detection with tooth surface segmentation on intraoral photographic images using deep learning*. BMC Oral Health 22, 573, (2022). 12

13. Yoon, K., Jeong, H. M., Kim, J. W., Park, J. H., Choi, J., *AI-based dental caries and tooth number detection in intraoral photos: Model development and performance evaluation*. J. Dent. 141, 104821, (2024). 13

14. Bayati, M., Alizadeh Savareh, B., Ahmadinejad, H., Mosavat, F., *Advanced AI-driven detection of interproximal caries in bitewing radiographs using YOLOv8*. Sci. Rep. 15, 4641, (2025). 14

15. Pérez de Frutos, J., Holden Helland, R., Desai, S., Nymoen, L. C., Langø, T., Remman, T., Sen, A., *AI-Dentify: deep learning for proximal caries detection on bitewing x-ray-HUNT4 Oral Health Study*. BMC Oral Health 24, 344, (2024). 15

16. Elnady, N., Adel, A., Badawy, W., *Enhancing Dental Caries Detection with YOLOv9: A Comprehensive Analysis and Validation of an Automated Object Detection Model*. In: Int. Conf. Future Telecommunications Artificial Intelligence, 1-4, (2024). 16

D. Meghana and A. Manimaran,

Department of Mathematics,

School of Advanced Sciences,

VIT-AP University,

Beside AP Secretariat, Amaravati,

Andhra Pradesh 522241, India.

E-mail address: meghana.23phd7196@vitap.ac.in, manimaran.a@vitap.ac.in