Journal of Dentomaxillofacial Radiology, Pathology and Surgery

Spring 2025, Volume 14, Number 2

# Review article: Recent Advances in Deep Learning for Dental Imaging (2022–2025): A Narrative Review

**Soheil Vafaeian[1*], Pedram Hajibagheri[1]**

1. Dental Sciences Research Center, Department of Oral and Maxillofacial Medicine, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran

Use your device to scan and read the article online

**Citation** Vafaeian S, Hajibagheri P. Recent Advances in Deep Learning for Dental Imaging (2022–2025): A Narrative Review. Journal of Dentomaxillofacial Radiology, Pathology and Surgery. 2025; 14(2): 12-22

http://dx.doi.org/10.32592/3dj.14.2.--

## ABSTRACT

This narrative review synthesises 2022–2025 evidence on deep learning for dental imaging, focusing on diagnostic accuracy (precision, recall, F1), segmentation quality (Dice), and reporting of inference speed. Representative model families are also situated—YOLO one-stage detectors, attention-augmented U-Nets, 3-D CNNs, and Segment Anything Model–derived hybrids ("SAM-derived" = architectures adapted from Segment Anything for dental images)—within current clinical workflows. From database searches of PubMed, Scopus and IEEE Xplore (January 2022 – May 2025), 342 records were retrieved; after deduplication and screening, 17 primary studies were included. Detection/classification studies (n = 10) reported overall F1 values up to 0.97 (highest in peri-implantitis detection; internal test; n = 100 from an 800-image dataset); the lowest externally evaluated detector reported ~0.53 F1 on CBCT periapicals (external site; n = 195 scans). YOLOv8 achieved F1 ≈ 0.82 on bitewings in an internal test split (n = 150). Segmentation studies (n = 8) reported Dice ~0.49–0.98: Attention U-Net reached 0.963 for single-tooth CBCT on an internal test (n = 9 scans). A multi-structure Swin-UNETR reported Dice 0.936–0.965 for tooth/sinus/bone/canal on an external set (n = 55 scans). A SAM-derived model (Tooth-ASAM) achieved 0.909–0.975 Dice across mixed public/private datasets. While three studies included external-site validation, none were prospective or randomised. Key priorities for clinical translation were identified: consistent speed reporting (batch-1 latency/FPS on named hardware), metric harmonisation (Dice for segmentation; F1 at IoU = 0.5 for detection), dockerised inference pipelines, and multi-centre external testing followed by prospective trials to quantify clinical impact.

**Article info:**
**Received:** 20 May 2025
**Accepted:** 22 Jun 2025
**Available Online:** 28 Jun 2025

**Keywords:**
* Artificial intelligence
* Deep learning
* Machine learning
* Dental Radiography

## 1. Introduction

Manual interpretation of dental images—including bitewing, periapical, panoramic and cone-beam CT (CBCT) views—remains labour-intensive and suffers from inter-observer variability. For example, early inter-proximal caries on bitewings are subtle in contrast, periapicals and panoramics demand compensation for geometric distortion, and a single CBCT scan may contain > 400 axial slices, requiring 10 minutes or more of scrolling in a busy clinic. Two recent systematic reviews underscored both the promise and fragmentation of current evidence. Carvalho et al. pooled 25 studies of AI caries detection in bitewings and reported pooled sensitivity 0.87 and specificity 0.91 but noted small test sets and scarce external validation (1). Kot et al. synthesised 18 CBCT tooth-segmentation papers and found pooled Dice similarity coefficient (Dice) 0.93 while highlighting heterogeneous metrics and a lack of head-to-head model comparisons (2). These gaps motivate an updated, task-level synthesis. To orient the reader, we highlight three developments since January 2022 that shape current performance and reporting:

1. Mature one-stage detectors. The *You Only Look Once* (YOLO) family has evolved from v5 through v8. Bayati et al. reported YOLOv8 achieved F1 (F1-Score) 0.82 for inter-proximal caries on bitewings (3). Lee et al. reported YOLOv7 reached F1 0.97 for peri-implantitis detection (4). A benchmarking study across YOLO variants (v5–v9c) described "real-time" operation but did not disclose batch-1 latency or frame-rate figures; nevertheless, precision–recall trade-offs approached those of two-stage detectors such as Faster Region-Based Convolutional Neural Network (R-CNN) in several tasks (5).

**** Corresponding Authors:***
***Soheil Vafaeian***
***Address:*** *Dental Sciences Research Center, Department of Oral and Maxillofacial Medicine, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran*
***E-mail:*** *soheylvafa@gmail.com*

2. Task-specific segmentation networks. Attention-augmented U-Nets and 3-D CNN architectures consistently produce high-fidelity CBCT masks. Liu et al. used a Swin-U-Net Transformer (UNETR) backbone to segment tooth, sinus, bone and mandibular canal simultaneously, reporting Dice 0.94–0.97 across structures (6). Chen et al. integrated an Attention U-Net with V-Net to achieve Dice 0.963 for single-tooth CBCT segmentation and root-canal measurement (7). Palkovics et al. employed a multi-phase 3-D Segmentation Residual Network (SegResNet) for full-arch CBCT, reaching Dice 0.965 ± 0.010 on periodontal bone topography (8). Hsu et al. improved Dice to 0.96 by majority-voting a "3.5 D" U-Net ensemble (9).

3. Foundation-model adaptation. Transformer hybrids have entered dentistry via the Segment Anything Model (SAM). Here and throughout, we use "SAM-derived" to denote models adapted from SAM backbones (e.g., promptable/finetuned variants), distinct from classical U-Nets or pure ViT decoders. Wang et al. adapted SAM to multimodal tooth images (CBCT, panoramics, intra-oral photos), achieving Dice 0.909–0.975 while using ≈ 40 % of the manual labels required by traditional CNNs (10). Schneider et al. compared CNN, transformer and hybrid backbones across three dental segmentation tasks; hybrids retained CNN-level accuracy but demanded more GPU memory —still acceptable for back-office batch processing (11). These label-efficient, prompt-based approaches foreshadow task-agnostic AI workflows in dentistry.

Since January 2022, 17 primary studies have reported new dental AI models with transparent metrics and, in three cases, true external-site validation. Yet no narrative so far has compared YOLOv8 speed with Shifted-Window U-Net Transformer (Swin-UNETR) Dice or examined how SAM-derived models slot into existing clinical pathways. Moreover, federated learning for tooth segmentation on panoramics has recently outperformed local and central training without sharing raw data (12), and AI-based metal-artifact reduction is beginning to improve CBCT image quality upstream of segmentation (13). Incorporating these 2022-2025 advances provides a more realistic picture of what clinicians can expect today—and what gaps remain.

Consequently, the present narrative review:

- (i) collates 17 peer-reviewed primary studies (2022-2025) covering detection, segmentation and classification;
- (ii) quantifies pooled accuracy (precision, recall, F1, Dice) and runtime;
- (iii) discusses clinical readiness, remaining challenges and research priorities—including federated learning, artifact-aware networks and multi-centre prospective trials.

By triangulating these strands, the review aims to provide clinicians, researchers and developers with an up-to-date, task-level map of deep-learning performance in dental imaging and a clear agenda for bringing AI tools from bench to chair-side.

For clarity, all acronyms and technical terms used in this review (e.g., CBCT, PR, mIoU, Dice) are defined in the Abbreviations table (Table 1), and each is expanded at first use.

**Table 1.** List of abbreviations and definitions used in the review

| Abbreviation | Definition |
|---|---|
| AI | Artificial intelligence |
| AUC | Area under the ROC curve |
| BW | Bitewing radiograph |
| CBCT | Cone-beam computed tomography |
| CNN | Convolutional neural network |
| CV | Cross-validation (e.g., 5-fold) |
| Dice | Dice similarity coefficient (pixel-wise F1 for segmentation) |
| F1-score | Harmonic mean of precision and recall |
| FPS | Frames per second |
| IoU | Intersection-over-Union |
| mIoU | Mean Intersection-over-Union |
| MPA | Mean pixel accuracy |
| PA | Periapical radiograph |
| PR | Panoramic radiograph |
| RGB | Red–Green–Blue (photograph) |
| ROI | Region of interest |
| SAM | Segment Anything Model |
| SegResNet | Segmentation Residual Network |
| Swin-UNETR | Shifted-Window Transformer U-Net (UNETR) |
| TFLOPS | Tera floating-point operations per second |
| U-Net | Encoder–decoder segmentation network |
| ViT | Vision Transformer |
| YOLO | You Only Look Once (one-stage detector) |
| MFPT-Net | Multi-task Fine-Grained Progressive Training Network |
| MICCAI | Medical Image Computing and Computer-Assisted Intervention (challenge context) |

## 2. Materials and Methods

This is a narrative review with systematic elements (database searches, dual independent screening, agreement assessment). Screening counts are reported in text; no flow figure is provided.

We searched PubMed, Scopus and IEEE Xplore for English-language records published from 1 January 2022 to 31 May 2025. The Boolean string combined core terms for deep learning ("deep learning" OR "convolutional neural network" OR "transformer"), model names ("YOLO" OR "U-Net"), and imaging keywords ("dental imaging" OR "radiograph" OR "CBCT"). A hand search of reference lists completed the strategy.

We included peer-reviewed studies that (i) analysed human dental images using a deep-learning model and (ii) reported at least one quantitative performance metric from the set below.
To ensure a consistent comparison, only metrics that are widely reported and clinically meaningful were extracted.
(for TP = true positives, FP = false positives, FN = false negatives, TN = true negatives, P = prediction mask, G = ground-truth mask)

- Precision (positive-predictive value): TP / (TP + FP) – indicates how often a positive call is correct, limiting false-positive interruptions during chair-side work.
- Recall (sensitivity): TP / (TP + FN) – captures how many true lesions are detected; missed disease directly affects patient care.
- F1-score (harmonic mean of precision and recall): 2 × TP / (2 × TP + FP + FN) – the default summary number in object-detection studies because it balances FP and FN.
- Dice similarity coefficient (pixel-wise F1): 2 × |P ∩ G| / (|P| + |G|) – the dominant overlap metric for segmentation.
- Intersection-over-Union (IoU, Jaccard index): |P ∩ G| / |P ∪ G| – a stricter overlap measure than Dice; its class-averaged form is mean IoU (mIoU).
- Mean pixel accuracy (MPA): average over classes of $TP_k / (TP_k + FN_k)$ – occasionally reported in enamel-crack work as a complement to mIoU.
- Specificity: TN / (TN + FP) – relevant for fracture or bone-loss screening where over-referral must be minimised.
- Accuracy: (TP + TN) / (all) – retained for implant-brand classifiers, though less informative in imbalanced datasets.

Precision–recall metrics were extracted and reported for all detection studies. If a study did not report an F1-score, it was calculated from the provided confusion matrix or the reported precision and recall values:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The derived value was then included in the pooled statistics and marked in the tables as estimated.

For segmentation studies, we extracted whichever overlap metric was provided—Dice or mean Intersection-over-Union (mIoU) or pixel-wise F1—rather than pooling them together.
To enable pooled summaries, we harmonized segmentation metrics as follows. For pixel/voxel-wise segmentation, pixel-wise F1 and Dice are mathematically equivalent; therefore, when a study reported pixel-wise F1 we treated it as Dice with the same value. We did not convert object level detection F1-score to Dice.

Studies reporting Dice are summarized by their Dice values, and those reporting mIoU by their mIoU values. When only mIoU was available but a Dice approximation was useful for interpretation, we calculated

$$Dice_{\{est\}} = \frac{2 \times \{mIoU\}}{1 + \{mIoU\}}$$

which holds exactly for binary masks on the same region but serves only as an approximation when mIoU is averaged across multiple classes. The derived value was then included in the pooled statistics and marked in the tables as estimated.

Simple, unweighted means were reported to avoid over-representing large, single-site datasets within heterogeneous tasks. We also recorded split design (train/val/test counts; k-fold cross-validation) and whether evaluation was internal, external-site, or matched-control.
Claims of 'real-time' performance made without specific timings or hardware were recorded qualitatively; latency was not estimated in these cases.

We retrieved 342 records (PubMed 188, Scopus 122, IEEE Xplore 32). After removing 62 duplicates, 280 titles/abstracts were screened; 255 were excluded as off-topic or reviews. We assessed 25 full texts, excluding 8 (five lacked quantitative metrics; three were non-human/ex-vivo), leaving 17 primary studies for inclusion. Two reviewers screened independently; pilot κ = 0.89 and overall κ = 0.92 with consensus resolution of discrepancies (Table 2).

**Table 2.** A summary description of method steps

| | Description |
|---|---|
| Databases | PubMed, Scopus, IEEE Xplore (English; 1 Jan 2022 – 31 May 2025) |
| Search terms | "deep learning", "YOLO", "U-Net", "transformer", "dental imaging", "radiograph", "CBCT" |
| Inclusion | Human dental images analysed with deep learning; ≥ 1 metric (accuracy, precision, recall, F1, Dice, IoU) |
| Screening | De-duplication → title/abstract scan (n = 280) → full-text review (n = 25) |
| Data extraction | inter-reviewer κ = 0.92 |

## 3. Results

Seventeen primary studies met the inclusion criteria: nine investigate lesion or caries detection (or other image-level classifications) using one-stage convolutional detectors—predominantly YOLO variants alongside a few Faster R-CNN pipelines (3-5, 14-19). Seven focus on tooth or bone segmentation with attention-augmented U-Nets, Swin-UNETR or other 3-D CNN backbones, including a Segment-Anything adaptation (6-11, 20). The remaining one multimodal study pairs implant-brand classification with its own segmentation branch, bridging the two task categories (21).

Across nine datasets, reported precision ranges from 0.651 to 1.000 and recall from 0.727 to 1.000. The strongest single result is obtained on the 800-image peri-implantitis set from Taipei Medical University, where a YOLOv7 network achieves perfect precision with 0.94 recall (F1 = 0.97) (4). In caries detection, the Tehran University bite-wing series supports a YOLOv8 model that provides the most even trade-off (P 0.85, R 0.80, F1 0.82)(3), whereas a speed-optimised YOLOv9c trained at Kırıkkale University records the lowest YOLO accuracy (F1 0.69)(5). The YoCNET hybrid—YOLOv5 augmented with ConvNeXt features—raises precision to 0.99 on radiographs from Jinan Stomatological Hospital but at the cost of reduced recall (0.85; F1 0.92) (17). Two Faster R-CNN pipelines perform well on broader scenes: five-fold cross-validation on 4 083 panoramics from Pusan National University yields P = R = 0.90 with AUC 0.95 (18), and a Swin-Transformer variant reaches F1 0.95 on a 6 404-film mandibular-fracture set from Charité Berlin (19). The CBCT periapical-lesion study from Graz and Bern (Hadzić 2023) publishes a full confusion matrix but omits precision; using those counts we calculated precision 0.38 and F1 0.53, the lowest among the detectors (15).

Only two detection papers report genuine external-site evaluation: the 195-scan periapical-lesion CBCT study from Graz and Bern (15), the YoCNET periapical-radiograph detector evaluated on 200 images from Jiangsu Second Hospital (17). Several investigations use alternative split strategies: Pusan National University adopts five-fold cross-validation (18); Kırıkkale University merges training and validation counts and discloses only the 150-image test set (5); the fracture study balances its test cohort by matching control films rather than stating absolute numbers (19).

Six of the eight CBCT segmentation papers report Dice similarity coefficients ≥ 0.91; the exceptions are 0.911 (9) and 0.909 on the Tooth-CBCT subset in (10). When all modalities are considered (adding PR (Panoramic Radiograph)/BW (Bitewing) segmentation from (11) and optical-microscope cracks from (20), the additional Dice values are 0.89 (tooth), 0.85 (tooth-structure), 0.49 (caries) and ≈ 0.857 (cracks), which lowers the pooled all-modality mean. A multi-centre Chinese study trains Swin-UNETR on 451 CBCT volumes but withholds internal split counts, evaluating instead on a 55-volume external set and obtaining Dice values of 0.936 – 0.965 for tooth, sinus, bone and canal (6). Wang 2025 is the only investigation to combine public and private resources, blending the open NC-CBCT release, 45 Tooth-CBCT volumes from Hangzhou Dental Hospital, the MICCAI Tooth panoramic challenge dataset and 409 Vident-lab video frames; Wang 2025 compared four candidate backbones (baseline U-Net, Swin-UNETR, ViT-decoder and a Segment-Anything derivative) across its multi-source dataset but retained only the top performer (Tooth-ASAM**),** a SAM-adapted hybrid, for the final report; Tooth-ASAM reached its highest Dice (0.975) on the public Vident-lab video set and its lowest value (0.909) on the group's private 45-volume Tooth-CBCT collection from Hangzhou Dental Hospital, giving a cross-source mean Dice of 0.942. (10) . Classical CNNs remain competitive: an Attention U-Net on 39 Sichuan-University scans (27 / 3 / 9 split) reports Dice 0.963 (7); a multi-phase SegResNet trained on 57 / 13 volumes and tested on 10 external scans from Semmelweis University records Dice 0.965 ± 0.010 (8); and ensemble "3.5-D" U-Nets lift Dice from 0.93 to 0.96 using four-fold cross-validation on a 24-volume National Taiwan University cohort (9). The remaining paper, Xie 2024, evaluates 600 optical-microscope images of cracked enamel (Sun Yat-sen & Guangdong UT) and reports an improved FDB-DeepLabv3+ with mIoU 0.7507 and MPA 0.7552; following the procedure stated in the Methods section, this corresponds to Dice ≈ 0.857 for comparability (20). For comparability across studies, pixel-wise F1 was treated as Dice (11).

Only three segmentation papers provide a genuine external-site test set: the multi-structure Swin-UNETR study that added an independent 55-scan CBCT cohort from additional hospitals to its ten-centre training pool (6), the periodontal-bone SegResNet work evaluated on 10 CBCT volumes acquired at Bern after training solely on Semmelweis data (8), and the MFPT-Net multi-task model whose implant masks and classifications were challenged with 252 CBCT scans collected at two clinics outside the Peking University training cohort (21).

Split reporting is inconsistent: one large multi-centre CBCT project supplies total case numbers but omit detailed Train, Validation and Test allocation (6), while three others rely exclusively on k-fold cross-validation (9, 11, 18). Also, the Hadzic 2023 study trained its SCN + 3-D U-Net via four-fold cross-validation, and the authors subsequently evaluated the frozen model on an independent 195-scan set from Graz (15).

MFPT-Net simultaneously classifies implant brands and segments implant bodies on 3-D CBCT data drawn from Peking University and two collaborating clinics. Without disclosing its internal split, the network achieves F1 0.93 for multi-class implant recognition (5 categories) and Dice 0.98 for implant-body segmentation when tested on an independent set of 252 external implants (21; Table 3).

**Table 3.** Included studies characteristics

| # | Study year (Ref) | Clinical application | Image Type | Data Source (Size) | Train / Val / Test | Model family | Key overall metrics* |
|---|---|---|---|---|---|---|---|
| **Detection** | | | | | | | |
| 1 | Bayati 2025 (3) | Caries | BW | Tehran U; 1 506 images (552 BW) | 1 205 / 151 / 150 (80 / 10 / 10 %) | YOLOv8 | P 0.8483 · R 0.7977 · F1 0.8222 |
| 2 | Lee 2024 (4) | Peri-implantitis | PA | Taipei MU (800) | 600 / 100 / 100 | YOLOv7 | P 1.0000 · R 0.9444 · F1 0.9710 |
| 3 | Liu 2024 (17) | Periapical-lesion | PA | Nanjing Stom. Hosp. (1 305) + Jiangsu 2nd Hosp. (200 ext) | 1 044 / 261 / 200 ext | YoCNET (YOLOv5 + ConvNeXt) | P 0.9888 · R 0.8530 · F1 0.9159 |
| 4 | Hadzic 2023 (15) | Periapical-lesion | CBCT | Med U. Graz (195) | 4-fold CV 144 / NR / 36 195 ext | U-Net CNN | Sp 0.843 · R 0.867 Acc 0.895, $F1_{est}$≈0.53 $P_{est}$≈0.38 |
| 5 | Kunt 2023 (16) | Caries | BW | Prague multi-device (3 989) | 2 793 / 598 / 598 | CNN ensemble | P 0.83 · R 0.77 · F1 0.80 |
| 6 | Ayhan 2025 (5) | Caries | BW | Kırıkkale U. (2 150) | 2 000 train + val / 150 test | YOLOv9c | P 0.651 · R 0.727 · F1 0.687 |
| 7 | Ryu 2023 (18) | Bone-loss | PR | Pusan U. (4 083) | 5-fold CV (4 083) | Faster R-CNN | P 0.90 · R 0.90 · F1 0.90 (AUC 0.95) |
| 8 | Adnan 2024 (14) | Caries | RGB | Aga Khan U. (7 465) | 5 226 / 1 493 / 746 | YOLOv5s | P 0.907 · R 0.856 · F1 0.880 |
| 9 | Vinayahalingam 2022 (19) | Mandibular-fracture | PR | Charité Berlin – 6 404 PR (1 624 fx + 4 780 non-fx) | 1 310 / 165 / 149 (+ matched controls) | Faster R-CNN + Swin-T | P 0.977 · R 0.960 · F1 0.947 |
| **Segmentation** | | | | | | | |
| 10 | Liu 2024 (6) | Multi-structure | CBCT | 10 Chinese centres — 451 scans + 55 external scans | NR / NR / NR (+55 ext) | Swin-UNETR | Dice 0.965 (tooth) 0.936 (sinus) 0.954 (bone) 0.948 (canal) |
| 11 | Chen 2025 (7) | Tooth | CBCT | Sichuan U. (39) | 27 / 3 / 9 cases (800 / 71 / 275 ROIs) | Attention U-Net | Dice 0.9633 |
| 12 | Wang 2025 (10) | Tooth | CBCT + PR + RGB | NC-CBCT release (4 938) Tooth-CBCT — Hangzhou Dental Hosp. (45 vols) MICCAI-Tooth PR (2 000 + 500) Vident lab video (409) | NC 17 361 / 1 840 / 1 840 patches Tooth 5 900 / 720 / 720 patches MICCAI PR 1 600 / 200 / 200 imgs Video 300 / 29 / 80 vids | SAM derivative – ToothASAM | Dice range 0.909 – 0.975 (Avg 0.9418) |
| 13 | Palkovics 2025 (8) | Periodontal-bone | CBCT | Semmelweis U. (80) | 57 / 13 / 10 ext | Multi-phase SegResNet | Dice 0.9650 ± 0.0097 |
| 14 | Hsu 2022 (9) | Tooth | CBCT | NTU tri-centre (102) | 4-fold CV | 3.5-D U-Net | Dice 0.911 |
| 15 | Schneider 2025 (11) | Tooth / Tooth-structure / Caries | PR + BW | Charité (1881 PR 1625 BW 2689 BW) | 5-fold CV | U-Net / DeepLabV3+ | Tooth F1† 0.89 · Structure F1 0.85 · Caries F1 0.49 |
| 16 | Xie 2024 (20) | Crack | Optical microscope | Sun Yat sen & Guangdong UT 600 | 500 / — / 100 | FDB-DeepLabv3+ | mIoU 0.7507 MPA 0.7552 $Dice_{est}$≈0.857 |
| **Segmentation Classification** | | | | | | | |
| 17 | Zhao 2025 (21) | Implant | CBCT | Peking U + 2 clinics (437) + 252 ext | NR / NR /NR 252 ext | MFPT-Net (3-D CNN multitask) | Classification: P 93.15 % · R 93.31 % · F1 93.18 % \| Segmentation: Dice 0.9804 |

† Pixel-wise F1 reported; treated as Dice * P = precision, R = recall, F1 = F1-score, Sp = specificity, Acc = accuracy; Dice = Dice similarity coefficient; AUC = area under ROC curve; mIoU = mean Intersection-over-Union; MPA = mean pixel accuracy, MFPT-Net = Multi-task Fine-Grained Progressive Training Network. PR = Panoramic Radiograph, BW = Bitewing, CV = Cross Validation, est = Estimate, ext = Extrnal, NR = Not Reported

Table 4 pools the headline metrics for all 23 task–dataset pairs using simple, un-weighted averages, because four studies do not disclose the size of their independent test sets. We report unweighted means to avoid over-representing single-site large datasets; weighted estimates are not directly comparable across heterogeneous tasks.

**Table 4.** Pooled performance by task

| Task category | Metric (mean ± SD) | Range | Number of studies |
|---|---|---|---|
| Caries detection (bitewing / photo) | F1 = 0.80 ± 0.07 | 0.69 – 0.88 | 4 (3, 5, 14, 16) |
| Peri-implantitis detection | F1 = 0.97 | — | 1 (4) |
| Periapical-lesion detection (2-D radiograph) | F1 = 0.92 | — | 1 (17) |
| Periapical-lesion detection (CBCT) | F1 ≈ 0.53 † | — | 1 (15) |
| Periodontal bone-loss grading | F1 = 0.90 | — | 1 (18) |
| Mandibular-fracture detection | F1 = 0.95 | — | 1 (19) |
| Implant-brand classification | F1 = 0.93 | — | 1 (21) |
| Detection / classification* | F1 = 0.84 ± 0.14 | 0.53 – 0.97 | 10 |
| Tooth segmentation (CBCT) | Dice = 0.95 ± 0.02 | 0.91 – 0.97 | 4 (6, 7, 9, 10)‡ |
| Periodontal-bone segmentation | Dice = 0.97 | — | 1 (8) |
| Implant-body segmentation | Dice = 0.98 | — | 1 (21) |
| Sinus segmentation | Dice = 0.936 | — | 1 (6) |
| Bone segmentation | Dice = 0.954 | — | 1 (6) |
| Canal segmentation | Dice = 0.948 | — | 1 (6) |
| Tooth segmentation (PR) § | Dice = 0.89 | — | 1 (11) |
| Enamel-crack segmentation* | Dice ≈ 0.857 | — | 1 (20) |
| Tooth-structure segmentation § | Dice = 0.85 | — | 1 (11) |
| Caries segmentation § | Dice = 0.49 | — | 1 (11) |
| Segmentation* | Dice = 0.89 ± 0.13 | 0.49 – 0.98 | 13 |

* Dice for Xie 2024(20) estimated from mIoU = 0.7507 using the conversion described in the Methods section. ‡ Wang 2025 (10) reports multi-source Tooth-ASAM with per-source Dice; for the Tooth-CBCT subset the Dice is 0.909 (lowest of its sources). † Hadzić 2023 (15)did not publish precision or F1. Using the confusion-matrix counts provided, we calculated precision ≈ 0.38 and F1 ≈ 0.53, which is included here and in the pooled mean. * Pooled rows use simple, unweighted averages over task–dataset pairs. § Schneider 2025 (11) reports pixel-wise F1 (5-fold CV); for segmentation this is equivalent to Dice, so we list Dice = 0.89 (tooth), 0.85 (tooth-structure), 0.49 (caries).

Among the ten detection/classification papers the overall mean is F1 = 0.84 ± 0.14 (0.53 – 0.97). The peri-implantitis YOLOv7 detector evaluated on 800 periapicals from Taipei Medical University reports the highest F1 of 0.97 (4), whereas the CBCT periapical-lesion study from Graz/Bern yields the lowest performance; precision derived from its published confusion matrix gives F1 ≈ 0.53 (15). Caries detection is the most frequently explored sub-task, appearing in four datasets (3, 5, 14, 16) with a pooled F1 = 0.80 ± 0.07. Single-dataset categories reach F1 = 0.92 for 2-D periapical lesions (17), 0.90 for periodontal bone-loss grading (18), 0.95 for mandibular-fracture detection (19) and 0.93 for five-class implant-brand recognition (21).

Segmentation (all modalities) yields a pooled Dice = 0.89 ± 0.13 (0.49 – 0.98) across 13 task–dataset pairs (Table 4). For CBCT-only pairs, four independent tooth-only CBCT series (6, 7, 9, 10) cluster tightly at Dice = 0.95 ± 0.02 (0.91 – 0.97). Multi-structure Swin-UNETR reports Dice 0.936–0.965 for tooth, sinus, bone and canal across 451 CBCT volumes (6); periodontal-bone SegResNet achieves Dice = 0.97 (8), and implant-body MFPT-Net tops the group at Dice = 0.98 (21). The non-CBCT outlier is enamel-crack segmentation on 600 optical-microscope images, where Dice was estimated at ≈ 0.857 from the reported mIoU 0.7507 to permit comparison (20).

Inference speed reporting was largely absent. Only one primary study (Adnan 2024's smartphone caries prototype) published a numeric runtime, quoting ~14 frames per second (fps) (≈ 71 ms per frame) on a mobile system-on-chip (SoC) (14). Both papers invoke the YOLO family's reputation for real-time inference, citing the single-shot, one-stage design that characterises YOLOv7 and YOLOv8, but neither paper backs the "real-time" claim with concrete runtime data such as FPS or per-image latency (3, 4). Similarly, the benchmarking work by Ayhan 2025 (YOLOv9c) and Schneider 2025 (Vision-Transformer vs CNN) compared architectures qualitatively yet omitted hardware-specific timings (5, 11). None of the CBCT segmentation papers recorded inference throughput.

Across the 17 included studies, detector performance varied by model family and dataset. YOLO variants spanned F1 0.687–0.971 (lowest in a v9c caries model (5); highest in peri-implantitis with v7 (4)), with mid-range results on bite-wing caries (3) and intra-oral photos (14). The same study reports AUC ≈ 0.78 while achieving F1 ≈ 0.88; differences reflect metric choice and class balance (14). The hybrid YoCNET (YOLOv5+ConvNeXt) achieved the highest reported precision (P = 0.9888) with lower recall (R = 0.8530) (17). Faster R-CNN pipelines were strong on panoramics (F1=0.90 (18) and 0.947 with a Swin-T backbone (19)), though evaluated via cross-validation or matched controls (18, 19). Only two detection works used genuine external tests: a CBCT periapical-lesion study (195 scans) reporting Sp = 0.843, R = 0.867,

Acc = 0.895 with low F1≈0.53 and P≈0.38 (derived from the published confusion matrix) (15), and YoCNET on 200 external periapical images). Across all detectors, the lowest F1 was Hadzić 2023 (≈0.53) (15). For segmentation, U-Net derivatives reported high Dice (≈0.91–0.963) (7, 9) and SegResNet reached Dice = 0.965 ± 0.010 (8); Swin-UNETR achieved Dice 0.936–0.965 across tooth/sinus/bone/canal with a 55-scan external set (6), and a SAM-based model (Tooth-ASAM) reached Dice 0.909–0.975 (mean 0.9418) across mixed sources (10). DeepLabV3+ performed well for tooth/structure (Dice 0.89/0.85) but was weaker for caries (Dice 0.49) (11), while a multitask 3-D MFPT-Net showed strong external performance (classification F1≈93 % and segmentation Dice≈0.98 on 252 CBCT scans) (21). Split reporting was heterogeneous, NR counts and k-fold CV were common, limiting cross-study comparability (6, 9, 11, 18). Top performers**:** YOLOv7 delivered the highest single-site detector score (F1 = 0.971) (4); YoCNET offered the highest precision and is one of the only externally tested detectors (17); for segmentation, MFPT-Net led on external data (21), with SegResNet (8) and Swin-UNETR (6) the strongest non-implant CBCT results, and Tooth-ASAM the highest within-source Dice range (10; Table 5).

**Table 5.** Comparative performance of major model families

| Model family | Studies (Ref) | Task / Modality | Highest Reported key metrics (as given) | Strengths observed in this dataset | Weaknesses observed in this dataset |
|---|---|---|---|---|---|
| YOLO (v5/v7/v8/v9c) | Bayati 2025 (3); Lee 2024 (4); Ayhan 2025 (5); Adnan 2024 (14) | Detection (BW, PA, intra-oral RGB) | F1 0.687–0.971 (3) P 0.8483 · R 0.7977 · F1 0.8222 (4) P 1.000 · R 0.944 · F1 0.971 (14) P 0.907 · R 0.856 · F1 0.880 | Highest single-site detector (YOLOv7 F1 0.971) (4); solid mid-range on BW and RGB (3, 14) | Lowest YOLO F1 at 0.687 (v9c caries) (5); no external-site evaluation reported |
| Faster R-CNN | Ryu 2023 (18) | Detection (PR) | P 0.90 · R 0.90 · F1 0.90 · AUC 0.95 | Balanced precision/recall on panoramics | Reported via 5-fold CV only |
| Faster R-CNN + Swin-T | Vinayahalingam 2022 (19) | Detection (PR) | P 0.977 · R 0.960 · F1 0.947 | High F1 on mandibular fractures | Matched-control testing; no stated external site |
| Hybrid YoCNET (YOLOv5 + ConvNeXt) | Liu 2024 (17) | Detection (PA) | P 0.9888 · R 0.8530 · F1 0.9159 | Highest detector precision and external test of 200 images | Precision–recall gap (R 0.8530) |
| CNN ensemble | Kunt 2023 (16) | Detection (BW) | P 0.83 · R 0.77 · F1 0.80 | Reasonable multi-device BW performance | Lower F1 than top detectors; no external site |
| U-Net (detection, CBCT periapical) | Hadzić 2023 (15) | Detection (CBCT) | Sp 0.843 · R 0.867 · Acc 0.895; F1 ≈ 0.53; P ≈ 0.38 | Includes 195-scan external test | Lowest detector F1 and precision in set |
| U-Net variants (Attention / 3.5-D) | Chen 2025 (7); Hsu 2022 (9) | Segmentation (CBCT tooth) | Dice 0.9633 (7); 0.911 (9) | High Dice on internal cohorts | Small/CV splits; no external site |
| Swin-UNETR | Liu 2024 (6) | Segmentation (CBCT, multi-structure) | Dice 0.936–0.965 (tooth/sinus/bone/canal) | Strong multi-structure Dice; 55-scan external set | Train/val/test counts NR |
| SegResNet (multi-phase) | Palkovics 2025 (8) | Segmentation (CBCT periodontal bone) | Dice 0.9650 ± 0.0097 | High Dice; 10 external scans | External n is small |
| DeepLabV3+ / FDB-DeepLabV3+ | Schneider 2025(11); Xie 2024 (20) | Segmentation (PR/BW; optical microscope) | (11) Tooth Dice 0.89; Structure Dice 0.85; Caries Dice 0.49. (20) mIoU 0.7507 · MPA 0.7552 *(Dice ≈ 0.85)* | Effective for tooth/structure masks (11) | Caries segmentation weaker (Dice 0.49) (11) no external site stated (11, 20) |
| SAM-derivative (Tooth-ASAM) | Wang 2025 (10) | Segmentation (CBCT + PR + RGB video) | Dice 0.909–0.975 (mean 0.9418) | Highest within-source Dice range across mixed sources | Lower Dice on private Tooth-CBCT subset (0.909) |
| MFPT-Net (3-D multitask) | Zhao 2025 (21) | Classification + Segmentation (CBCT implants) | Cls F1 93.18 % · Seg Dice 0.9804 | Strongest external-set segmentation; multitask performance | Internal split NR |

## 4. Discussion

Deep-learning performance in the 17 primary studies is already strong enough to influence day-to-day dentistry, yet the way results are reported still lags behind their raw accuracy.

Our pooled analysis shows that detection only papers (excluding the implant-brand classifier; n = 9) reach a broad F1 band of 0.53 – 0.97 —the lower bound reflects the CBCT periapical-lesion detector in Hadzić 2023 with F1 ≈ 0.53 derived from its published confusion matrix (15); with YOLOv7 on peri-implantitis periapicals setting the high-water mark (F1 0.97; P 1.00; R 0.94) (4) and a speed-optimised YOLOv9c anchoring the lower end among YOLO variants (F1 0.69) (5). After metric harmonization, segmentation results across all modalities (n = 13 task–dataset pairs) yield Dice 0.89 ± 0.13 (range 0.49–0.98). CBCT-only tooth segmentation remains tighter at Dice 0.95 ± 0.02 (0.91–0.97), with the SAM-derived Tooth-ASAM topping out at Dice 0.975 (10); even the smallest CBCT dataset, a 3.5-D U-Net ensemble, clears Dice 0.91 (9). Harmonization steps included treating pixel-wise F1 as Dice (Schneider 2025 (11)) and converting mIoU to Dice for cracks (Xie 2024: mIoU 0.7507 → Dice ≈ 0.857) (20). One multitask network bridges the two worlds, pairing implant-brand classification (F1 0.93) with implant-body segmentation at Dice 0.98 (21). These figures confirm that transformer-enhanced U-Nets, SegResNets and SAM adaptations have converged on near-surgical CBCT overlap accuracy, whereas detector performance remains modality-sensitive—excellent on peri-implantitis hotspots but weaker on low-contrast enamel lesions.

Tooth segmentation on panoramic radiographs (PR) reaches Dice ≈ 0.89 (11), whereas CBCT tooth segmentation spans 0.909–0.965 across four independent series (6, 7, 9, 10). On PR, tooth-structure masks are lower (Dice ≈ 0.85) and caries segmentation is markedly weaker (Dice ≈ 0.49) (11). In multi-structure CBCT, component Dice are consistently high (tooth 0.965; sinus 0.936; bone 0.954; canal 0.948) (6). These modality-linked gaps likely reflect inherent contrast/resolution advantages of CBCT, 3-D context, and differences in annotation granularity.

On panoramic radiographs, reported sensitivity varies widely due to head-position variability, magnification and distortion; single-site reports can be high, whereas multi-site settings are lower. For intra-oral/smartphone photographs, the included study reports AUC ≈ 0.78 alongside overall F1 ≈ 0.88 (14), underscoring illumination/focus shifts and device heterogeneity. To mitigate these gaps, domain adaptation merits systematic evaluation—e.g., appearance/style transfer (CycleGAN-style unpaired translation), histogram or color normalization, and adversarial feature-level alignment; test-time adaptation and multi-source aggregation (as in Tooth-ASAM across NC-CBCT, Tooth-CBCT, PR and video (10)) are also promising. Future work should report cross-domain performance deltas (internal to external) and ablate adaptation components to quantify their contribution.

Preliminary reader studies suggest that, in specific cohorts and structures, high volumetric Dice (≈0.94–0.95) may be judged acceptable for planning; however, acceptability is task-, structure-, and surface-error–dependent and no field-wide thresholds exist (22, 23).We therefore refrain from adopting a numeric cut-off.

Across the corpus, performance closely tracked dataset composition. Higher scores were observed on homogeneous, single-site cohorts with consistent acquisition (e.g., CBCT tooth segmentation at Dice 0.91–0.97), whereas more heterogeneous or low-contrast tasks (e.g., PR caries masks at Dice ≈ 0.49) lagged (6, 7, 9-11). Annotation protocol (single rater vs. consensus), prevalence and case mix, and split design (external test vs. k-fold CV) all influenced headline figures (6, 8, 9, 11, 18, 21) . Multi-source aggregation improved robustness: Wang 2025 combined NC-CBCT, Tooth-CBCT, PR and video and achieved within-source Dice up to 0.975, albeit with a lower result (0.909) on the private Tooth-CBCT subset. Durable public endpoints are essential for replication; at the time of writing, (10) the NC-CBCT link cited in (10) was inaccessible during manuscript preparation; mirroring on durable repositories would aid reproducibility.

These advances in deep learning are beginning to reshape clinical workflows across diagnosis, treatment planning, and disease monitoring in dentistry. For diagnosis, second-reader systems based on YOLOv8 have demonstrated improved detection of enamel lesions on bitewing radiographs, achieving an F1 of 0.82, precision of 0.85, and recall of 0.80, thus reducing missed early caries cases (3). In treatment planning, high-precision segmentation of single-tooth CBCT volumes—with Dice scores ranging from 0.960 to 0.963—enables automated generation of implant surgical guides and clear aligner designs, reducing clinician workload and improving consistency (7, 9). For monitoring, convolutional neural networks trained on panoramic images can now grade periodontal bone loss with an AUC of 0.95 and an F1 of 0.90, supporting longitudinal surveillance of periodontal health with minimal manual input (19).

Despite these encouraging developments, several key challenges remain before deep-learning systems can be fully integrated into clinical practice. Generalisation across sites and imaging conditions continues to be a concern, with performance typically dropping by an appreciable amount in metrics like Dice or F1 when models are evaluated on external datasets. For instance, a CBCT periapical-lesion detector evaluated on an independent cohort reported Acc = 0.895 with F1 ≈ 0.53 derived from its confusion matrix (15); other studies report only internal or only external figures, limiting paired deltas (6, 8, 21). Explainability is another limiting factor; only a minority provide saliency/attention overlays or probability calibration (e.g., architecture

comparisons (11)). Embedding calibrated probabilities and case-level heatmaps directly in the viewer could improve interpretability and clinician acceptance.

Beyond saliency/attention overlays, clinician trust depends on usable evidence. Small reader studies in medical imaging suggest that heatmaps and calibrated probabilities can improve decision confidence, but standardized validation in dentistry remains limited. We therefore recommend (i) reporting probability calibration (e.g., temperature scaling with expected calibration error), (ii) embedding case-level overlays directly in the radiology viewer, and (iii) conducting user-centred reader studies that measure time-to-decision, confidence shifts, and error types with and without explainability. These endpoints would connect technical metrics to clinical usability.

Finally, the current body of work lacks prospective or randomized studies—none of the 17 reviewed papers meet these standards—leaving real-world impacts on diagnostic accuracy, clinical efficiency, or chair-time untested.

While deep-learning models have reached high performance on bitewings and standard CBCT scans, key modality-specific gaps still limit broader clinical deployment. On panoramic radiographs, diagnostic sensitivity can be unstable—owing to head-position variability and geometric distortion—despite strong single-site results (19). For intra-oral photographs, only one included detector study reported performance (F1 ≈ 0.88) (14), underscoring the need for broader external validation and domain-adaptation strategies for illumination and focus shifts. Micro-scale detection also remains under-developed: on optical-microscope enamel-crack images, the improved DeepLabv3+ reached mIoU = 0.7507 (≈ Dice 0.857) (20), and small-fissure sensitivity remains limited by dataset size. Beyond the included set, soft-tissue classification will require curated, diverse photo atlases and external-site testing before routine use. Finally, CBCT scans containing heavy-metal restorations continue to degrade image quality and segmentation performance, and none of the included segmentation papers explicitly evaluated metal-artifact-reduction within the AI pipeline (6-11, 20, 21). Targeted artifact-aware training remains an open need (13). To mitigate domain shift across modalities and scanners, multi-source aggregation (e.g., Wang 2025's Tooth-ASAM across NC-CBCT, Tooth-CBCT, PR and video (10)) and explicit domain-adaptation (e.g., style/appearance transfer or feature-level adaptation) merit systematic evaluation. Bridging these modality-specific gaps will require large, multi-centre datasets; artifact-aware or domain-adaptive architectures; and rigorous, externally validated evaluations to ensure reliability in real-world conditions. Our review could not compute a cross-study latency average because only one primary paper (14) published a numeric runtime (~14 fps on a mobile SoC). Flagship YOLO papers describe their detectors as "real-time" but the articles themselves omit frames-per-second or millisecond figures. By contrast, the MLCommons MLPerf Inference benchmark obliges every submission to disclose batch-1 latency, hardware and software stack, and "Queries per Second" —making vendor claims directly comparable (24). A one-page "speed sheet" following the MLPerf template would let journals verify that a model advertised for chair-side use really meets the <100 ms latency widely accepted in human-computer-interaction studies. Authors should also report effective Tera Floating point operations per second (TFLOPS) (images $s^{-1}$ × model-FLOPs) so throughput scales with hardware generations can be normalised across clinics that own different GPUs. Half of the included papers do not state framework versions, CUDA/cuDNN libraries, augmentation pipelines or even the random-seed policy. Such omissions block byte-for-byte replication and conceal latent implementation bugs that sometimes inflate headline metrics. At minimum, studies should publish inference scripts, weights and Docker files (or conda environments) listing exact package versions. The federated-learning work of Schneider 2023 (12) proves that privacy-preserving containers can be shared without moving raw patient data, so legal constraints are surmountable. Bringing the strands together, six inter-locking actions would move dental-imaging AI from promising prototypes to reproducible, clinically dependable tools:

1-Publish a speed-sheet in every paper. Authors should append a one-row table—mirroring the MLPerf template—listing batch-1 latency (ms), frames per second and effective TFLOPS on a named GPU, plus an edge-device figure whenever "chair-side" on a named GPU, including pre-/post-processing or "hand-held" use is claimed (24) (For clarity, we define "chair-side viable" as end-to-end batch-1 latency < 100 ms per image on named hardware, including pre/post-processing. Because, aside from one study on a mobile app (~14 fps), the included papers did not report model-specific latency or FPS, we treat "real-time" claims as unquantified and refrain from inferring < 100 ms). Transparent timing is essential; accuracy alone says nothing about workflow impact.

2-Standardise core metrics. Report Dice (or volumetric Dice) for every segmentation task and F1 at IoU = 0.5 for every detection task; treat pixel-wise F1 as Dice and convert mIoU→Dice as specified in Methods; confusion matrices or ROC curves should appear in the supplements to reveal class imbalance (25). Harmonising on these two anchors prevents today's scatter of pixel-F1, mIoU and specificity-only reports and lets investigators track real progress year-on-year.

3-Create an open benchmark with automatic scoring. A de-identified, multi-centre repository of bitewings, panoramics and CBCT volumes should host dockerised submissions. The server would grade each container on accuracy (Dice/F1) and latency, publishing a public leaderboard analogous to Common Objects in Context (COCO) for detection or Machine Learning Performance (MLPerf ) for inference (24-26).

4-Release full inference pipelines. Every study should provide Docker (or conda) images that fix framework versions, CUDA/cuDNN builds, augmentation parameters and random seeds. Such containers let independent groups reproduce results byte-for-byte and surface implementation bugs early; the federated-learning work of Schneider 2023 shows this is feasible without moving raw data(12).

5-Leverage federated learning and explainable AI. Multi-centre federated pipelines have already outperformed both local and centrally merged training while keeping data private(12). At inference time, saliency maps or attention overlays should be embedded directly into radiology viewers so clinicians can see why the network fired.

6-Validate prospectively with hybrid architectures. Hybrid CNN–Transformer designs promise to pair YOLO-level speed with ViT-level context. Their real-world impact on diagnostic yield, chair-time and cost now needs proof in multi-centre, prospective trials that use the standardised metrics and speed sheets outlined above.

If the field adopts these practices, the high accuracies already demonstrated by YOLO detectors(3-5, 16), transformer-enhanced U-Nets (7-9)and SAM derivatives(10) can translate into dependable, vendor-neutral systems that genuinely shorten chair-time and improve diagnostic care.

## 5. Conclusions

Between 2022 and 2025, dental-imaging AI moved from feasibility to credible pilot readiness. Detectors achieved high F1 on bitewings and periapicals, while CBCT segmenters consistently reached high Dice—topping out near 0.98 on select tasks. Across all modalities, pooled segmentation performance was strong (≈ Dice 0.89 ± 0.13), and CBCT-only tooth segmentation was tighter (≈ 0.95 ± 0.02). Detection remained modality-sensitive, varying with lesion salience and acquisition quality.

We judge three model families as pilot-ready in their target niches: (i) YOLOv8 for chair-side caries detection on bitewings, (ii) Tooth-ASAM for CBCT tooth segmentation across mixed sources, and (iii) SegResNet for CBCT periodontal-bone mapping.

Closing the evidence–deployment gap now depends less on raw accuracy and more on standards: 1. a one-row “speed sheet” (batch-1 latency, FPS, effective TFLOPS, named hardware), 2. metric harmonization (Dice for segmentation; F1@0.5 for detection; confusion matrices/ROC), 3. dockerized, version-locked inference pipelines, and 4. public multi-centre benchmarks that score both accuracy and latency. Adopting these practices will make progress measurable and reproducible; prospective, multi-centre trials can then determine the real-world impact on diagnostic yield, misses, and chair-time.

## Ethical Considerations

This study is a focused narrative review of previously published literature and does not involve any human participants, animal subjects, or identifiable personal data. Therefore, ethical approval was not required.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Authors’ Contributions

**Soheil Vafaeian:** Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing. **Pedram Hajibagheri:** Investigation, Writing - Review & Editing.

## Conflict of Interests

The authors declare no conflicts of interest.

## Availability of Data and Material

Not applicable.

## Acknowledgments

The authors acknowledge the use of ChatGPT (OpenAI, GPT-4) for language editing under close supervision. All generated content was reviewed and verified by the authors.

## References

1. Carvalho BKG, Nolden E-L, Wenning AS, Kiss-Dala S, Agócs G, Róth I, et al. Diagnostic accuracy of artificial intelligence for approximal caries on bitewing radiographs: A systematic review and meta-analysis. J Dent. 2024:151:105388. [DOI: 10.1016/j.jdent.2024.105388] [PMID]
2. Kot WY, Au Yeung SY, Leung YY, Leung PH, Yang W-f. Evolution of deep learning tooth segmentation from CT/CBCT images: a systematic review and meta-analysis. BMC Oral Health. 2025;25(1):800. [DOI: 10.1186/s12903-025-05984-6] [PMID] [PMCID]
3. Bayati M, Alizadeh Savareh B, Ahmadinejad H, Mosavat F. Advanced AI-driven detection of interproximal caries in bitewing radiographs using YOLOv8. Sci Rep. 2025;15(1):4641. [DOI: 10.1038/s41598-024-84737-x] [PMID] [PMCID]
4. Lee W-F, Day M-Y, Fang C-Y, Nataraj V, Wen S-C, Chang W-J, et al. Establishing a novel deep learning model for detecting peri-implantitis. J Dent Sci. 2024;19(2):1165–73. [DOI: 10.1016/j.jds.2023.11.017] [PMID] [PMCID]
5. Ayhan B, Ayan E, Karadağ G, Bayraktar Y. Evaluation of caries detection on bitewing radiographs: A comparative analysis of the improved deep learning model and dentist performance. J Esthet Restor Dent. 2025; 37(7):1949-61. [DOI: 10.1111/jerd.13470] [PMID] [PMCID]
6. Liu Y, Xie R, Wang L, Liu H, Liu C, Zhao Y, et al. Fully automatic AI segmentation of oral surgery-related tissues

based on cone beam computed tomography images. Int J Oral Sci. 2024;16(1):34. [DOI: 10.1038/s41368-024-00294-z] [PMID] [PMCID]
7. Chen Z, Liu Q, Wang J, Ji N, Gong Y, Gao B. Tooth image segmentation and root canal measurement based on deep learning. Front Bioeng Biotechnol. 2025;13:1565403. [DOI: 10.3389/fbioe.2025.1565403] [PMID] [PMCID]
8. Palkovics D, Molnar B, Pinter C, García-Mato D, Diaz-Pinto A, Windisch P, et al. Automatic deep learning segmentation of mandibular periodontal bone topography on cone-beam computed tomography images. J Dent. 2025; 159:105813. [DOI: 10.1016/j.jdent.2025.105813] [PMID]
9. Hsu K, Yuh D-Y, Lin S-C, Lyu P-S, Pan G-X, Zhuang Y-C, et al. Improving performance of deep learning models using 3.5 D U-Net via majority voting for tooth segmentation on cone beam computed tomography. Sci Rep. 2022;12(1):19809. [DOI: 10.1038/s41598-022-23901-7] [PMID] [PMCID]
10. Wang P, Gu H, Sun Y. Tooth segmentation on multimodal images using adapted segment anything model. Sci Rep. 2025;15(1):13874. [DOI: 10.1038/s41598-025-96301-2] [PMID] [PMCID]
11. Schneider L, Krasowski A, Pitchika V, Bombeck L, Schwendicke F, Buettner M. Assessment of CNNs, transformers, and hybrid architectures in dental image segmentation. J Dent. 2025;156:105668. [DOI: 10.1016/j.jdent.2025.105668] [PMID]
12. Schneider L, Rischke R, Krois J, Krasowski A, Büttner M, Mohammad-Rahimi H, et al. Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs. J Dent. 2023;135:104556. [DOI: 10.1016/j.jdent.2023.104556] [PMID]
13. Wajer R, Wajer A, Kazimierczak N, Wilamowska J, Serafin Z. The impact of AI on metal artifacts in CBCT oral cavity imaging. Diagnostics. 2024;14(12):1280. [DOI: 10.3390/diagnostics14121280] [PMID] [PMCID]
14. Adnan N, Faizan Ahmed SM, Das JK, Aijaz S, Sukhia RH, Hoodbhoy Z, et al. Developing an AI-based application for caries index detection on intraoral photographs. Sci Rep. 2024;14(1):26752. [DOI: 10.1038/s41598-024-78184-x] [PMID] [PMCID]
15. Hadzic A, Urschler M, Press J-NA, Riedl R, Rugani P, Štern D, et al. Evaluating a periapical lesion detection CNN on a clinically representative CBCT dataset-A validation study. J Clin Med. 2023;13(1):197. [DOI: 10.3390/jcm13010197] [PMID] [PMCID]
16. Kunt L, Kybic J, Nagyová V, Tichý A. Automatic caries detection in bitewing radiographs: part I-deep learning. Clin Oral Investig. 2023;27(12):7463–71. [DOI: 10.1007/s00784-023-05335-1] [PMID]
17. Liu J, Liu X, Shao Y, Gao Y, Pan K, Jin C, et al. Periapical lesion detection in periapical radiographs using the latest convolutional neural network ConvNeXt and its integrated models. Sci Rep. 2024;14(1):25429. [DOI: 10.1038/s41598-024-75748-9] [PMID] [PMCID]
18. Ryu J, Lee D-M, Jung Y-H, Kwon O, Park S, Hwang J, et al. Automated detection of periodontal bone loss using deep learning and panoramic radiographs: a convolutional neural network approach. Appl Sci. 2023;13(9):5261. [DOI: 10.3390/app13095261]
19. Vinayahalingam S, van Nistelrooij N, van Ginneken B, Bressem K, Tröltzsch D, Heiland M, et al. Detection of mandibular fractures on panoramic radiographs using deep learning. Sci Rep. 2022;12(1):19596. [DOI: 10.1038/s41598-022-23445-w] [PMID] [PMCID]
20. Xie Z, Lu Q, Guo J, Lin W, Ge G, Tang Y, et al. Semantic segmentation for tooth cracks using improved DeepLabv3+ model. Heliyon. 2024;10(4):e25892. [DOI: 10.1016/j.heliyon.2024.e25892] [PMID] [PMCID]
21. Zhao Y, Zhu L, Wang W, Lv L, Li Q, Liu Y, et al. Progressive multi-task learning for fine-grained dental implant classification and segmentation in CBCT image. Comput Biol Med. 2025;189:109896. [DOI: 10.1016/j.compbiomed.2025.109896] [PMID]
22. Dot G, Schouman T, Dubois G, Rouch P, Gajny L. Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework. Eur Radiol. 2022;32(6):3639–48. [DOI: 10.1007/s00330-021-08455-y] [PMID]
23. Xiang B, Lu J, Yu J. Evaluating tooth segmentation accuracy and time efficiency in CBCT images using artificial intelligence: A systematic review and Meta-analysis. J Dent. 2024;146:105064. [DOI: 10.1016/j.jdent.2024.105064] [PMID]
24. Reddi VJ, Cheng C, Kanter D, Mattson P, Schmuelling G, Wu C-J, et al. Mlperf inference benchmark. 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA); 2020: IEEE. [Link]
25. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:1–28. [DOI: 10.1186/s12880-015-0068-x] [PMID] [PMCID]
26. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13; 2014: Springer. [Link]